Wahrscheinlichkeit und Statistik

Musterlösung Serie 10

1. Seien $(X_i)_{i\in\mathbb{N}}$ i.i.d. Bernoulli-verteilte Zufallsvariablem mit unbekanntem Parameter $p \in [0,1]$. Wir betrachten zwei Schätzer für p. Der erste Schätzer ist $T_n^{(1)}(X_1,\dots,X_n) = X_1$. Der zweite Schätzer ist $T_n^{(2)}(X_1,\dots,X_n) = \overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$.

   a) Berechnen Sie den Erwartungswert und die Varianz beider Schätzer als Funktion von n. Kommentieren Sie dann das Resultat.
   b) Sind die Schätzer erwartungstreu?
   c) Sind die Schätzer konsistent?

   Lösung:

   a) Die Erwartungswerte lauten $\mathbb{E}[X_1] = p$ and $\mathbb{E}[\overline{X}_n] = p$, und die Varianzen sind gegeben durch $\mathrm{Var}(X_1) = p(1-p)$ and $\mathrm{Var}(\overline{X}_n) = \frac{1}{n^2}\sum_{i=1}^n p(1-p) = \frac{p(1-p)}{n}$. Offensichtlich ist der zweite Schätzer viel besser, da er eine viel kleinere Varianz und den gleichen Erwartungswert hat.
   b) Beide Schätzer sind erwartungstreue Schätzer für p, aufgrund von a).
   c) Der erste Schätzer X_1 ist nicht konsistent: Für $p \in (0,1)$, $\varepsilon = \min(\frac{p}{2}, \frac{1-p}{2})$ gilt $\mathbb{P}_p[|X_1 - p| > \varepsilon] = 1$ was offensichtlich nicht gegen 0 konvergiert. Der zweite Schätzer $\overline{X}_n$ ist jedoch ein konsistenter Schätzer für p, denn a) zeigte, dass $\lim_{n\to\infty} \mathrm{Var}[\overline{X}_n] = 0$ (und dies impliziert Konvergenz zum Erwartungswert mit Hilfe der Chebyshev-Gleichung).

2. Seien $(X_i)_{i\in\mathbb{N}}$ i.i.d. exponential verteilte Zufallsvariablem mit unbekanntem Parameter $\lambda \in (0,\infty)$. Der Momentenschätzer für λ ist gegeben durch $\Lambda_n(X_1,\dots,X_n) = \frac{1}{\overline{X}_n} = \frac{n}{\sum_{i=1}^n X_i}$.

   a) Ist dieser Schätzer erwartungstreu?
   b) Ist dieser Schätzer konsistent?

   Lösung: Zuerst erinnern wir uns, dass $\mathbb{E}[X_i] = \frac{1}{\lambda}$ und $\mathrm{Var}(X_i) = \frac{1}{\lambda^2}$. Dies impliziert, dass $\mathbb{E}[\overline{X}_n] = \frac{1}{\lambda}$ und durch Unabhängigkeit $\mathrm{Var}(\overline{X}_n) = \frac{1}{\lambda^2 n}$.

   a) Wir zeigen, dass $\Lambda_n(X_1,\dots,X_n)$ nicht erwartungstreu ist. Wenn wir den Beweis der Jensen-Ungleichung durchgehen, sehen wir ganz direkt, dass wir für eine streng konvexe Funktion g und eine nicht deterministische Zufallsvariable X eine strenge Jensen-Ungleichung $\mathbb{E}[g(X)] > g(\mathbb{E}[X])$ erhalten. Die Abbildung $x \mapsto \frac{1}{x}$ ist streng konvex auf $(0,\infty)$ und $\overline{X}_n$ ist nicht deterministisch. Dies führt zu der Ungleichung

   $$\lambda = \frac{1}{\mathbb{E}[\overline{X}_n]} < \mathbb{E}\left[\frac{1}{\overline{X}_n}\right] = \mathbb{E}[\Lambda_n(X_1,\dots,X_n)],$$

   sodass der Bias $b_{\Lambda_n(X_1,\dots,X_n)}(\lambda) = \mathbb{E}[\Lambda_n(X_1,\dots,X_n)] - \lambda > 0$ ist.

b) Wir zeigen, dass $\Lambda_n(X_1, \dots, X_n)$ ein konsistenter Schätzer für λ ist. Mit dem starken Gesetz der grossen Zahlen erhalten wir, dass $\overline{X}_n$ fast sicher gegen $\frac{1}{\lambda}$ konvergiert. Da die Abbildung $x \mapsto \frac{1}{x}$ auf $(0, \infty)$ stetig ist, erhalten wir, dass $\Lambda_n(X_1, \dots, X_n) = \frac{1}{\overline{X}_n}$ fast sicher gegen $\dfrac{1}{\frac{1}{\lambda}} = \lambda$ konvergiert, da

$$\left\{\omega \in \Omega : \lim_{n \to \infty} \overline{X}_n = \frac{1}{\lambda}\right\} \subseteq \left\{\omega \in \Omega : \lim_{n \to \infty} g(\overline{X}_n) = g\left(\frac{1}{\lambda}\right)\right\}$$

für jede stetige Funktion g. Fast sichere Konvergenz impliziert stochastische Konvergenz und somit ist $\Lambda_n(X_1, \dots, X_n)$ konsistent.

Alternative Lösung: Aus der Chebyshev-Ungleichung folgt direkt, dass $\overline{X}_n$ in Wahrscheinlichkeit gegen $\frac{1}{\lambda}$ konvergiert. Da die Abbildung $x \mapsto \frac{1}{x}$ auf $(0, \infty)$ stetig ist, erhalten wir, dass $\Lambda_n(X_1, \dots, X_n) = \frac{1}{\overline{X}_n}$ in Wahrscheinlichkeit gegen $\dfrac{1}{\frac{1}{\lambda}} = \lambda$ konvergiert, und somit ist $\Lambda_n(X_1, \dots, X_n)$ konsistent.

3. Um die Anzahl N der Fische in einem Teich zu schätzen, werden 5 verschiedene Fische eingefangen, markiert und wieder in den Teich eingesetzt. Man nimmt an, dass sich nach einer gewissen Zeit die markierten Fische mit den übrigen Fischen gut vermischt haben und fängt dann 11 Fische. Von diesen 11 Fischen sind 3 markiert und 8 nicht markiert. Um die gesamte Anzahl Fische N zu schätzen berechnen Sie die Wahrscheinlichkeit dieser Beobachtung als Funktion von N und nehmen Sie als Schätzer das N^* für welches diese Wahrscheinlichkeit maximiert wird. (So ein Schätzer wird Maximum-Likelihood-Schätzer genannt)

   Welches N^* erhalten Sie?

   Lösung: Sei X die Anzahl markierter Fische im zweiten Fang. Beobachtet wurde $X = 3$. Die Wahrscheinlichkeit des Ereignisses, das wir beobachtet haben, hängt in folgender Weise von N ab: Es ist

$$\mathbb{P}_N[X = 3] = \frac{\binom{N-5}{8}\binom{5}{3}}{\binom{N}{11}} = \frac{9900\,(N-5)!\,(N-11)!}{(N-13)!\,N!} =: g(N).$$

   Wir müssen $N_{\max}$ finden, sodass $g(N_{\max}) = \sup_{N \in \mathbb{N}} g(N)$. Wir müssen also die Maximumsstelle der Funktion g bestimmen. Wegen

$$\frac{g(N+1)}{g(N)} = \frac{(N-4)(N-10)}{(N-12)(N+1)} \begin{cases} > 1 & \text{falls } N \leq 17 \\ < 1 & \text{falls } N \geq 18 \end{cases}$$

   ist $g(18)$ maximal. Also ist $N^* = 18$ der Maximum-Likelihood-Schätzer für N.

4. *Vergleich von arithmetischem Mittel und Stichprobenmedian*

   Seien $(X_i)_{i \in \mathbb{N}}$ i.i.d. $\mathcal{N}(\mu, \sigma^2)$-verteilt, wobei μ und σ^2 unbekannt sind. Wir betrachten zwei Schätzer von μ,

$$T^{(1)}_{2n+1}(X_1, \dots, X_{2n+1}) = \frac{1}{2n+1} \sum_{i=1}^{2n+1} X_i$$
$$T^{(2)}_{2n+1}(X_1, \dots, X_{2n+1}) = X_{(n+1)},$$

   wobei $X_{(1)} < X_{(2)} < \ldots < X_{(2n+1)}$ die der Grösse nach geordneten ersten $(2n+1)$-Werte bezeichnet.

a) Berechnen Sie mit Hilfe des Zentralen Grenzwertsatzes die Werte $c_n^{(1)}$ und $c_n^{(2)}$ so dass

$$\mathbb{P}\left[\left|T^{(i)}_{2n+1}(X_1,\ldots,X_{2n+1})-\mu\right|\leq c_n^{(i)}\right]\xrightarrow{n\to\infty}95\%.$$

b) Berechnen Sie ein $q\in\mathbb{R}_+$, so dass

$$\frac{c^{(2)}_{qn}}{c^{(1)}_n}\xrightarrow{n\to\infty}1.$$

Wie kann man q in Worten interpretieren?

Lösung:

a)
- $i=1$:
  Da die X_i unabhängig und alle identisch $\mathcal{N}(\mu,\sigma^2)$-verteilt sind, aus dem Zentralen Grenzwertsatz gilt für das arithmetische Mittel:

$$\lim_{n\to\infty}\mathbb{P}\left(\frac{\sum_{i=1}^{2n+1}X_i-(2n+1)\mu}{\sigma\sqrt{2n+1}}\leq x\right)=\Phi(x),$$

  wobei die Wahrscheinlichkeit auf der linken Seite umgeschrieben werden kann wie folgt

$$\begin{aligned}\mathbb{P}\left(\frac{\sum_{i=1}^{2n+1}X_i-(2n+1)\mu}{\sigma\sqrt{2n+1}}\leq x\right)&=\mathbb{P}\left(\frac{\frac{1}{2n+1}\sum_{i=1}^{2n+1}X_i-\mu}{\frac{\sigma}{\sqrt{2n+1}}}\leq x\right)\\&=\mathbb{P}\left(\frac{1}{2n+1}\sum_{i=1}^{2n+1}X_i-\mu\leq x\frac{\sigma}{\sqrt{2n+1}}\right)\\&=\mathbb{P}\left(T^{(1)}_{2n+1}(X_1,\ldots,X_{2n+1})-\mu\leq x\frac{\sigma}{\sqrt{2n+1}}\right).\end{aligned}$$

  Da wir an $\mathbb{P}\left(\left|T^{(1)}_{2n+1}(X_1,\ldots,X_{2n+1})-\mu\right|\leq c_n^{(1)}\right)$ interessiert sind, erhalten wir

$$\begin{aligned}\mathbb{P}\left(\left|T^{(1)}_{2n+1}(X_1,\ldots,X_{2n+1})-\mu\right|\leq c_n^{(1)}\right)&=\mathbb{P}\left(\left|\frac{\sum_{i=1}^{2n+1}X_i-(2n+1)\mu}{\sigma\sqrt{2n+1}}\right|\leq x\right)\\&\xrightarrow{n\to\infty}\Phi(x)-\Phi(-x),\end{aligned}$$

  daher bekommt man unter Benutzung der Symmetrie von Φ, dass $c_n^{(1)}=x\frac{\sigma}{\sqrt{2n+1}}$, wobei man den Wert von x aus $95\%=\Phi(x)-\Phi(-x)=2\Phi(x)-1$ erhält:

$$x=\Phi^{-1}(97.5\%)=1.96.$$

- $i=2$:
  Andererseits gilt für das Stichprobenmedian $T^{(2)}_{2n+1}(X_1,\ldots,X_{2n+1})=X_{(n+1)}$, mit $X_i\sim\mathcal{N}(\mu,\sigma^2)$, dass der Median von $\widetilde{X}_i:=X_i-\mu$ null ist, i.e. $F^{-1}\left(\frac{1}{2}\right)=0$, wobei F die Verteilungsfunktion von $\widetilde{X}_i$ bezeichnet. Dann ist das Beispiel *Asymptotik des Medians* der Folie 273 für $\widetilde{X}_i$ anwendbar, und liefert

$$\mathbb{P}\left(\sqrt{(2n+1)}\widetilde{X}_{(n+1)}\leq x\right)\xrightarrow{n\to\infty}\Phi(2F'(0)x),$$

  wobei $F'(0)=\frac{1}{\sqrt{2\pi}\sigma}$. Die Wahrscheinlichkeit auf der linken Seite kann folgendermassen umgeschrieben werden

$$\mathbb{P}\left(\sqrt{2n+1}\widetilde{X}_{(n+1)}\leq x\right)=\mathbb{P}\left(\widetilde{X}_{(n+1)}\leq\frac{x}{\sqrt{2n+1}}\right).$$

Da wir an $\mathbb{P}\left(\left|T^{(2)}_{2n+1}(X_1,\dots,X_{2n+1})-\mu\right|\leq c_n^{(2)}\right)$ interessiert sind, muss x von der Form

$$95\% = \Phi(2F'(0)x) - \Phi(-2F'(0)x) = \Phi\left(\frac{\sqrt{2}}{\sqrt{\pi}\sigma}x\right) - \Phi\left(-\frac{\sqrt{2}}{\sqrt{\pi}\sigma}x\right)$$

sein und somit analog wie oben

$$c_n^{(2)} = \frac{x}{\sqrt{(2n+1)}} = \frac{1.96\ \sigma\ \sqrt{\pi}}{\sqrt{2(2n+1)}}.$$

Man sagt, dass das Stichprobenmedian asymptotisch normalverteilt ist mit asymptotischer Varianz $\sigma^2\frac{\pi}{2}$.

b) Es gilt, dass

$$\frac{c_{qn}^{(2)}}{c_n^{(1)}} = \frac{\frac{1.96\sigma\sqrt{\pi}}{\sqrt{2(2nq+1)}}}{\frac{1.96\sigma}{\sqrt{2n+1}}} = \frac{\sqrt{\pi}\sqrt{2n+1}}{\sqrt{2}\sqrt{2nq+1}} \xrightarrow{n\to\infty} 1,$$

wenn $q=\frac{\pi}{2}$. Der Stichprobenmedian ist also bei der Normalverteilung weniger genau als das arithmetische Mittel. Man braucht für den Stichprobenmedian $q=\frac{\pi}{2}$-mal so viele Beobachtungen als für das arithmetische Mittel. Man sagt, dass er $\sqrt{\frac{\pi}{2}}$-mal so viel streut und nennt q die *relative effizienz.*