Prof. H.R. Künsch Wahrscheinlichkeitsrechnung und Statistik Abt. IX, SS99

Musterlösung zur Übung 10

1. a) Seien $\underline{X} = (X_1, \ldots, X_n)$ und $\underline{x} = (x_1, \ldots, x_n)$. Da alle X_i unabhängig sind, ist die gemeinsame Dichte von $\underline{X}$ das Produkt der Dichten der X_i, und die Log-Likelihood-Funktion ist dann

$$l(\underline{x}) = \log\left(\prod_{i=1}^{n}(\theta-1)x_i^{-\theta}\right) = n\log(\theta-1) - \theta\sum_{i=1}^{n}\log x_i\,.$$

Ableiten und Nullsetzen liefert $n/(\theta-1) = \sum_{i=1}^n \log x_i$. Der Maximum-Likelihood-Schätzer ist

$$T_n(\underline{X}) = 1 + \frac{n}{\sum_{i=1}^n \log X_i} = 1 + \frac{1}{\frac{1}{n}\sum_{i=1}^n \log X_i}\,.$$

b) Zu zeigen ist für jedes $\varepsilon > 0$:

$$P[\,|T_n(\underline{X}) - \theta\,| > \varepsilon\,] = P\left[\,\left|\frac{1}{\frac{1}{n}\sum_{i=1}^n \log X_i} - (\theta-1)\right| > \varepsilon\,\right] \to 0 \quad \forall\theta \quad (n\to\infty)$$

Nun ist $\log X_i \sim \mathrm{Exp}(\theta-1)$ (nachrechnen!) mit Erwartungswert $\frac{1}{\theta-1}$. Bei der Exponentialverteilung existieren alle für das starke Gesetz der grossen Zahlen notwendigen Momente. Also konvergiert $\frac{1}{n}\sum_{i=1}^n \log X_i$ fast sicher gegen $\frac{1}{\theta-1}$ und damit der Betrag in der obigen Zeile fast sicher gegen 0. Daraus folgt unmittelbar die Behauptung.

Bemerkung: Man kann auch mit dem schwachen Gesetz der grossen Zahlen argumentieren. $\frac{1}{n}\sum_{i=1}^n \log X_i$ konvergiert dann stochastisch gegen $\frac{1}{\theta-1}$. Stochastische Konvergenz bleibt aber unter stetigen Transformationen f (hier: $f(x) = x^{-1}$) erhalten:

Da f stetig ist, existiert für jedes ε ein $\tilde\varepsilon$ mit $|f(X_i)-f(X)| \le \varepsilon$ für alle $|X_i - X| \le \tilde\varepsilon$. Also existiert für jedes ε ein $\tilde\varepsilon$ mit $P[|f(X_i)-f(X)| \le \varepsilon] \ge P[|X_i - X| \le \tilde\varepsilon]$. Wegen $X_i \to X$ stochastisch existiert aber für jedes $\tilde\varepsilon$ – insbesondere für obiges – und für jedes $\delta > 0$ ein i_0 mit $P[|X_i - X| > \tilde\varepsilon] \le \delta$ bzw. $P[|X_i - X| \le \tilde\varepsilon] \ge 1-\delta$ für alle $i \ge i_0$. Zusammengefasst: Zu jedem ε und jedem δ existieren ein $\tilde\varepsilon$ und ein i_0 mit $P[|f(X_i)-f(X)| \le \varepsilon] \ge P[|X_i - X| \le \tilde\varepsilon] \ge 1-\delta$ für alle $i \ge i_0$. Oder ohne das $\tilde\varepsilon$ ausgedrückt: Zu jedem ε und jedem δ existiert ein i_0 mit $P[|f(X_i)-f(X)| \le \varepsilon] \ge 1-\delta$ bzw. $P[|f(X_i)-f(X)| > \varepsilon] \le \delta$ für alle $i \ge i_0$. Das bedeutet stochastische Konvergenz für $f(X_i)$.

Wem beide Wege nicht zusagen, der kann den Beweis auch mit der Ungleichung von Čebyšev führen. . . .

2. a) Das interessierende Ereignis „vier Bauern“ kommt in den $n = 60$ (idealerweise) unabhängigen Austeilungen jeweils mit gleicher Wahrscheinlichkeit $p = \frac{1}{120}$ vor. Die Anzahl erfolgreicher Austeilungen ist daher binomialverteilt mit Parametern n und p bzw., da p klein ist, genähert poissonverteilt mit Parameter $\lambda = np = \frac{1}{2}$. „Exakt“ gilt $P[X=0] = \binom{60}{0}(\frac{1}{120})^0(\frac{119}{120})^{60} = 60.53\%$, und unter Verwendung der Poissonapproximation gilt $P[X=0] \approx e^{-1/2}(\frac{1}{2})^0/0! = e^{-1/2} = 60.65\%$.

b) $P[X=3] \approx e^{-1/2}(\frac{1}{2}^3)/3! = 1.26\%$ und
$P[X \ge 3] = 1 - P[X \le 2] \approx 1 - e^{-1/2}(\frac{1}{2})^0/0! - e^{-1/2}(\frac{1}{2})^1/1! - e^{-1/2}(\frac{1}{2})^2/2! = 1.44\%$.
Nur etwa jeden siebzigsten Jassabend ($100/1.44 \approx 70$), bei dem alles mit rechten Dingen zugeht, werden drei- oder mehrmal vier Bauern gewiesen. Das ist so selten, dass es wohl plausibler ist anzunehmen, jemand habe an diesem Abend mit gezinkten Karten gespielt.

3. A: n Individuen, X = Anzahl Erfolge

a) $X \sim Bin(n,p)$, $\qquad P[X=x] = \binom{n}{x}p^x(1-p)^{n-x}$

b) $\log L_x(p) = \log\binom{n}{x} + x\log p + (n-x)\log(1-p)$.
Maximum-Likelihood-Schätzer für p :

$$T_A = \arg\max_{0<p<1} L_x(p) = \arg\max_{0<p<1}\ \log L_x(p)$$

$$\frac{d}{dp}\log L_x(p) = \frac{x}{p} - \frac{n-x}{1-p} = 0 \iff x - xp = np - xp \iff \widehat{p} = \frac{x}{n} \Longrightarrow T_A = \frac{X}{n}$$

c)

$$E[T_A] = E\left[\frac{X}{n}\right] = \frac{1}{n}E[X] = p, \quad \text{also erwartungstreu.}$$

d)

$$E\left[(T_A - p)^2\right] = E\left[(T_A - E[T_A])^2\right] = \mathrm{Var}[T_A] = \frac{1}{n^2}\mathrm{Var}[X] = \frac{p(1-p)}{n}$$

$p = 0.2,\ n = 5 \Longrightarrow E\left[(T_A-p)^2\right] = 0.032$

B: n Gruppen mit jeweils c Individuen, X = Anzahl Gruppen mit mindestens einem Erfolg.

a) $P[\text{ mindestens 1 Erfolg in der Gruppe }] = 1 - P[\text{ Kein Erfolg }] = 1 - (1-p)^c$.
Also $X \sim Bin(n, 1-(1-p)^c)$.

b) $\log L_x(p) = \log\binom{n}{x} + x\log(1-(1-p)^c) + c(n-x)\log(1-p)$.

$$\begin{aligned}
\frac{d}{dp}\log L_x(p) \quad &= \quad x\cdot\frac{c(1-p)^{c-1}}{1-(1-p)^c} - \frac{c(n-x)}{1-p} = 0\\
\iff \quad & cx(1-p)^c - c(n-x)\left(1-(1-p)^c\right) = 0\\
\iff \quad & (1-p)^c\cdot(x + (n-x)) = n - x \iff 1 - p = \sqrt[c]{\frac{n-x}{n}}\\
\iff \quad & \widehat{p} = 1 - \sqrt[c]{\frac{n-x}{n}} \quad \Longrightarrow T_B = 1 - \sqrt[c]{\frac{n-X}{n}}
\end{aligned}$$

c)

$$E[T_B] = E\left[1-\sqrt[c]{\frac{n-X}{n}}\right] = 1-E\left[\sqrt[c]{\frac{n-X}{n}}\right] \overset{\text{(Jensen)}}{>} 1-\sqrt[c]{E\left[\frac{n-X}{n}\right]}$$
$$= 1-\sqrt[c]{1-\frac{1}{n}E[X]} = 1-\sqrt[c]{1-\frac{1}{n}n(1-(1-p)^c)} = p$$

Also $E[T_B] > p$, d.h. T_B ist nicht erwartungstreu.

d)

$$E\left[(T_B-p)^2\right] = \sum_{x=0}^{n}(T_B-p)^2 P[X=x]$$
$$= \sum_{x=0}^{n}\left(1-\sqrt[c]{\frac{n-x}{n}}-p\right)^2\binom{n}{x}\left(1-(1-p)^c\right)^x\left(1-p\right)^{c(n-x)}$$
$$= 0.023$$

T_A ist erwartungstreu, T_B hingegen nicht. Aber T_B hat einen kleineren mittleren quadratischen Fehler $E[(T-p)^2]$ als T_A. Könnte man nicht den Fünfer und das Weggli haben? In der mathematischen Statistik (nächstes Semester!) wird der Frage nachgegangen werden, wie klein der mittlere quadratische Fehler eines erwartungstreuen Schätzers werden kann.

4. Für Bruchpunkte verwenden wir das Symbol ε^*. Ein einfacher Median hat Bruchpunkt $\varepsilon^* = \frac{1}{2}$ (siehe Vorbesprechung; auf Haarspaltereien, z. B. wenn n ungerade ist, wollen wir hier nicht eingehen!). Ausserdem ist folgendes nützlich und klar:

$$\text{median}((x_i - c)_{i=1,\ldots,n}) = \text{median}((x_i)_{i=1,\ldots,n}) - c$$

a) Sowohl der innere Median (über j) wie auch der äussere Median (über i) im MAD können bis zur Hälfte Ausreisser verkraften. Also hat die Median-Abweichung Bruchpunkt $\varepsilon^* = \frac{1}{2}$. Der Bruchpunkt der Quartilsdifferenz liegt bei $\varepsilon^* = \frac{1}{4}$.

b) Für den Hodges-Lehmann-Schätzer werden insgesamt n^2 Summen $X_i + X_j$ berechnet. Wenn k Beobachtungen beliebig gross werden, dann streben von diesen n^2 Summen $2kn - k^2$ gegen Unendlich:

	X_1	X_2	X_3	$\cdots$	X_n
X_1		∞	∞		
X_2	∞	∞	∞	∞	∞
X_3	∞	∞	∞	∞	∞
$\vdots$		$\vdots$	$\vdots$		
X_n		∞	∞		

(k Spalten und k Zeilen der Länge n enthalten je kn Einträge. Davon werden k^2 Einträge doppelt gezählt.)

Sobald die Hälfte aller Summen gegen Unendlich streben, bricht der Median zusammen. Es muss also gelten $2kn - k^2 < n^2/2$. Das ist der Fall, wenn $k < n(1-\sqrt{2}/2)$. Damit ist der Bruchpunkt des Hodges-Lehmann-Schätzers $\varepsilon^* = 1 - \frac{\sqrt{2}}{2}$.
Siegel's Repeated Median hat nach den einführenden Bemerkungen Bruchpunkt $\varepsilon^* = \frac{1}{2}$.