Fachbereich Mathematik
Prof. Dr. Michael Kohler
Andreas Gärtner
Jens Mehnert

TECHNISCHE UNIVERSITÄT DARMSTADT

WS 09/10
03./04.02.2010

11. Übungsblatt zur „Statistik I für Human- und Sozialwissenschaft“

Aufgabe 41

Bei einer Erhebung über die Urlaubsdauer von Familien mit Kindern in der Bundesrepublik Deutschland wurde in einer Stichprobe von 1000 Familien (mit Kindern) eine mittlere Urlaubsdauer von $16,8$ bei einer empirischen Varianz von $8,2$ Tagen ermittelt. Wir nehmen an, dass die ermittelten Werte Realisierungen von unabhängigen und identisch verteilten Zufallsvariablen $X_1, \ldots, X_{1000}$ sind. Außerdem können wir auf Grund weiterer Untersuchungen davon ausgehen, dass die empirische Varianz mit der wirklichen Varianz übereinstimmt.

Wir möchten nun ein *zweiseitiges Konfidenzintervall* $K_n(x_1, \ldots, x_n)$ *zum Niveau* $\alpha = 0.95$ für den Erwartungswert der Zufallsvariablen X_1 bestimmen. D.h. es soll gelten:

$$\mathbf{P}\left[\mathbf{E}X_1 \in K_n(x_1, \ldots, x_n)\right] \;\geq\; \alpha = 0.95.$$

(a) Zeigen Sie mit Hilfe von bekannten Rechenregeln für Erwartungswert und Varianz, dass

$$\frac{\sum_{i=1}^n X_i - \mathbf{E}\left(\sum_{i=1}^n X_i\right)}{\sqrt{V(\sum_{i=1}^n X_i)}} = \frac{\sqrt{n}}{\sqrt{V(X_1)}}\left(\frac{1}{n}\sum_{i=1}^n X_i - \mathbf{E}X_1\right).$$

(b) Zeigen Sie nun, dass (für $u > 0$) aus $\left|\frac{\sqrt{n}}{\sqrt{V(X_1)}}\left(\frac{1}{n}\sum_{i=1}^n X_i - \mathbf{E}X_1\right)\right| \leq u$ folgt, dass

$$\mathbf{E}X_1 \in \left[\frac{1}{n}\sum_{i=1}^n X_i - \frac{\sqrt{V(X_1)}}{\sqrt{n}} \cdot u \;,\; \frac{1}{n}\sum_{i=1}^n X_i + \frac{\sqrt{V(X_1)}}{\sqrt{n}} \cdot u\right].$$

(c) Benutzen Sie nun den *zentralen Grenzwertsatz* und dass für eine standardnormalverteilte (d.h. $\mathcal{N}(0,1)$-verteilte) Zufallsvariable Z

$$P(|Z| \leq u_{0.025}) \;\geq\; 0.95 \quad \text{mit} \quad u_{0.025} \approx 1.96$$

ist, um das gesuchte Konfidenzintervall zu bestimmen.

Lösung:

(a) Nach der Vorlesung ist $\mathbf{E}(\sum_{i=1}^n X_i) = \sum_{i=1}^n \mathbf{E}X_i = n \cdot \mathbf{E}X_1$ und $V(\sum_{i=1}^n X_i) = \sum_{i=1}^n V(X_i) = n \cdot V(X_1)$, falls $X_1, \ldots, X_n$ unabhängig und identisch verteilt sind. Damit ergibt sich:

$$\begin{aligned}\frac{\sum_{i=1}^n X_i - \mathbf{E}\left(\sum_{i=1}^n X_i\right)}{\sqrt{V(\sum_{i=1}^n X_i)}} &= \frac{\sum_{i=1}^n X_i - n \cdot \mathbf{E}X_1}{\sqrt{n \cdot V(X_1)}} = \frac{n\left(\frac{1}{n}\sum_{i=1}^n X_i - \mathbf{E}X_1\right)}{\sqrt{n}\sqrt{V(X_1)}} \\ &= \frac{\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n X_i - \mathbf{E}X_1\right)}{\sqrt{V(X_1)}} = \frac{\sqrt{n}}{\sqrt{V(X_1)}}\left(\frac{1}{n}\sum_{i=1}^n X_i - \mathbf{E}X_1\right).\end{aligned}$$

(b) Auflösen des Betrages ergibt:

$$\begin{aligned}
& \left| \frac{\sqrt{n}}{\sqrt{V(X_1)}} \left(\frac{1}{n} \sum\nolimits_{i=1}^{n} X_i - \mathbf{E}X_1 \right) \right| \le u \\
\Leftrightarrow\quad & \frac{\sqrt{n}}{\sqrt{V(X_1)}} \left(\frac{1}{n} \sum\nolimits_{i=1}^{n} X_i - \mathbf{E}X_1 \right) \le u \quad \wedge \quad -\frac{\sqrt{n}}{\sqrt{V(X_1)}} \left(\frac{1}{n} \sum\nolimits_{i=1}^{n} X_i - \mathbf{E}X_1 \right) \le u \\
\Leftrightarrow\quad & \frac{1}{n} \sum\nolimits_{i=1}^{n} X_i - \mathbf{E}X_1 \le \frac{\sqrt{V(X_1)}}{\sqrt{n}} \cdot u \quad \wedge \quad \frac{1}{n} \sum\nolimits_{i=1}^{n} X_i - \mathbf{E}X_1 \ge -\frac{\sqrt{V(X_1)}}{\sqrt{n}} \cdot u \\
\Leftrightarrow\quad & \frac{1}{n} \sum\nolimits_{i=1}^{n} X_i - \frac{\sqrt{V(X_1)}}{\sqrt{n}} \cdot u \le \mathbf{E}X_1 \quad \wedge \quad \frac{1}{n} \sum\nolimits_{i=1}^{n} X_i + \frac{\sqrt{V(X_1)}}{\sqrt{n}} \cdot u \ge \mathbf{E}X_1
\end{aligned}$$

Was aber natürlich gleichbedeutend ist zu:

$$\mathbf{E}X_1 \in \left[\frac{1}{n} \sum\nolimits_{i=1}^{n} X_i - \frac{\sqrt{V(X_1)}}{\sqrt{n}} \cdot u \ , \ \frac{1}{n} \sum\nolimits_{i=1}^{n} X_i + \frac{\sqrt{V(X_1)}}{\sqrt{n}} \cdot u \right].$$

(c) Nach dem zentralen Grenzwertsatz und Teil (a) ist $Z := \frac{\sqrt{n}}{\sqrt{V(X_1)}} \left(\frac{1}{n} \sum_{i=1}^{n} X_i - \mathbf{E}X_1 \right)$ eine annähernd $\mathcal{N}(0,1)$-verteilte Zufallsvariable. Damit gilt mit Teil (b):

$$P(|Z| \le u_{0.025}) = P\left(\mathbf{E}X_1 \in \left[\tfrac{1}{n} \sum_{i=1}^{n} X_i - \tfrac{\sqrt{V(X_1)}}{\sqrt{n}} \cdot u_{0.025} \ , \ \tfrac{1}{n} \sum_{i=1}^{n} X_i + \tfrac{\sqrt{V(X_1)}}{\sqrt{n}} \cdot u_{0.025} \right] \right) \ \ge \ 0.95$$

Also ist das gesuchte Konfidenzintervall gegeben durch:

$$\left[16,8 - \frac{\sqrt{8,2}}{\sqrt{1000}} \cdot 1.96 \ , \ 16,8 + \frac{\sqrt{8,2}}{\sqrt{1000}} \cdot 1.96 \right] \approx [16.6225 \ , \ 16.9775] \,.$$

Aufgabe 42

Gegeben seien auf dem Intervall $[a, 3a]$ unabhängig identisch gleichverteilte Zufallsvariablen $X_1, \ldots, X_n$, $n \in \mathbb{N}$ und $a \in \mathbb{R}$. Wir möchten nun den Parameter

$$a = \frac{1}{2} \mathbf{E}X_1$$

mit Hilfe von Realisierungen von $X_1, \ldots, X_n$ schätzen. Geben Sie eine geeignete Schätzung dafür an und zeigen Sie, dass die Schätzung erwartungstreu und konsistent ist.
Hinweis: Modifizieren Sie den Schätzer aus der Vorlesung.

Lösung: Es ist bekannt, dass eine auf dem Intervall $[a, b]$ gleichverteilte Zufallsvariablen den Erwartungswert $\frac{a+b}{2}$ haben, wobei der Schätzer für den Erwartungswert aus der Vorlesung bekannt ist. Damit ist der gesuchte Schätzer

$$T(x_1, \ldots, x_n) = \frac{1}{2} \cdot \frac{x_1 + \ldots + x_n}{n}$$

mit

$$\mathbf{E}(T(X_1, \ldots, X_n)) = \mathbf{E}\left(\frac{1}{2n} \sum\nolimits_{i=1}^{n} X_i \right) = \frac{1}{2n} \sum_{i=1}^{n} \mathbf{E}(X_i) = \frac{1}{2} \mathbf{E}(X_1) = a$$

erwartungstreu. Er ist auch konsistent, weil $2 \cdot T(X_1, \ldots, X_n)$ nach Vorlesung ein konsistenter Schätzer ist.

Aufgabe 43

Im Rahmen einer Untersuchung zum Thema Studienabbruch, muss Pädagoge P. schätzen, wieviele Studenten schon im Verlauf des ersten Semesters ihr Studium abbrechen. Nehmen wir einmal an, dass P. eine Liste mit den Adressen aller Studienanfänger eines Jahrgangs zur Verfügung hat. P. wählt zufällig von dieser Liste Personen aus, die er befragt. Er überlegt sich folgende Modellannahmen: Jeder Studienanfänger entscheidet sich unabhängig von allen anderen Studenten mit der selben Wahrscheinlichkeit p für einen Abbruch des Studiums im ersten Semester. Die Zufallsvariable X_i verwendet er zur mathematischen Modellierung der Antwort des i-ten Befragten. X_i erhält den Wert 1, falls der i-te Befragte sein Studium abgebrochen hat und 0 sonst. Die Gesamtanzahl aller Befragten sei n. Geben Sie einen erwartungstreuen und konsistenten Schätzer für den Parameter p an und begründen Sie, dass der von ihnen angegebene Schätzer diese Eigenschaften hat.

Lösung: $X_1, \ldots, X_n$ sind jeweils binomialverteilt mit Parametern 1 und p. Außerdem sind sie wegen der gemachten Modellannahmen unabhängig. Für die Binomialverteilung mit Parametern 1 und p gilt

$$p = \mathbf{E}X.$$

Nach der Vorlesung ist

$$T(x_1, \ldots, x_n) = \frac{x_1 + \cdots + x_n}{n}$$

ein konsistenter und erwartungstreuer Schätzer für den Erwartungswert, deshalb haben wir mit $T(X_1, \ldots, X_n)$ einen erwartungstreuen und konsistenten Schätzer für p.

Aufgabe 44

Wir wollen bei einem Multiple-Choice-Test überprüfen, ob ein Teilnehmer die Antworten durch einfaches Raten ermittelt hat. Dazu beschreiben wir die (richtige) Beantwortung jeder Frage mit einer $b(1, p)$-verteilten Zufallsvariable und nehmen an, dass die Fragen unabhängig von einander beantwortet wurden. Bei jeder der $n = 8$ Aufgaben war genau eine der 4 angegebenen Antworten richtig. Um nun zwischen den beiden Hypothesen

$$H_0: \ p = 0.25 \qquad H_1: \ p > 0.25$$

zu entscheiden, soll der Test

$$\varphi(x_1, \ldots, x_n) = \begin{cases} 1 & \text{falls } \sum_{i=1}^n x_i \geq c \\ 0 & \text{falls } \sum_{i=1}^n x_i < c \end{cases}$$

verwendet werden. Bestimmen Sie den kleinsten Parameter $c \in \{1, \ldots, n\}$ so, dass dies zu einem Test zum Niveau $\alpha = 0.05$ wird.

Lösung: Für unabhängige $b(1, p)$-verteilte Zufallsvariablen $X_1, \ldots, X_n$ ist $\sum_{i=1}^n X_i$ eine $b(n, p)$-verteilte Zufallsvariable. Daher ist die Fehlerwahrscheinlichkeit erster Art dieses Tests gegeben durch:

$$\begin{aligned} \mathbf{P}_{p=0.25}[\varphi(X_1, \ldots, X_n) = 1] &= \mathbf{P}_{p=0.25}[\sum_{i=1}^n X_i \geq c] \\ &= \mathbf{P}_{p=0.25}[\sum_{i=1}^n X_i = n] + \mathbf{P}_{p=0.25}[\sum_{i=1}^n X_i = n-1] + \ldots + \mathbf{P}_{p=0.25}[\sum_{i=1}^n X_i = c] \\ &= \binom{n}{n} \cdot p^n \cdot (1-p)^0 + \binom{n}{n-1} \cdot p^{(n-1)} \cdot (1-p)^1 + \ldots + \binom{n}{c} \cdot p^{(n-c)} \cdot (1-p)^c \end{aligned}$$

Für $n = 8$ und $p = 0.25$ erhalten wir:

$$\binom{8}{8} \cdot 0.25^8 \cdot 0.75^0 = 0.000015 \qquad \binom{8}{7} \cdot 0.25^7 \cdot 0.75^1 = 0.000366$$

$$\binom{8}{6} \cdot 0.25^6 \cdot 0.75^2 = 0.003845 \qquad \binom{8}{5} \cdot 0.25^5 \cdot 0.75^3 = 0.023071$$

$$\binom{8}{4} \cdot 0.25^4 \cdot 0.75^4 = 0.086517$$

Also ist $c = 5$ der kleinste Parameter, so dass $\mathbf{P}_{p=0.25}[\varphi(X_1, \ldots, X_n) = 1] \leq \alpha = 0.05$.