Zulassungsprüfung Stochastik, 13.10.2017

Wir gehen stets von einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ aus.

Aufgabe 1 (15 Punkte)
Seien $a, b > 0$ und $x, y \in \mathbb{R}$ fest. Gegeben sei das Maß $\mu : \mathcal{B}^1 \longrightarrow [0, \infty)$, $\mu(A) := a1_A(x) + b1_A(y)$. Beweisen Sie, dass für alle Borel-messbaren Funktionen $f : \mathbb{R} \longrightarrow \mathbb{R}$

$$\int_{\mathbb{R}} f d\mu = af(x) + bf(y)$$

gilt.

Aufgabe 2 (15 Punkte)
Seien X, Y Zufallsvariablen mit $E(X|Y) = Y$, $E(X^2|Y) = Y^2$.

(a) Bestimmen Sie $E\left((X+Y)^2|Y\right)$ in Abhängigkeit von Y.

(b) Beweisen Sie, dass $X = Y$ fast sicher gilt.

Aufgabe 3 (20 Punkte)
Seien $Z \sim B\left(1, \frac{1}{3}\right)$, $D \sim B(1,p)$ mit $p \in (0,1)$ und $H \sim B\left(1, \frac{1}{2}\right)$ unabhängig. Sei

$$X := ZD + (1-Z)H. \tag{1}$$

(a) Begründen Sie, dass es ein $q \in (0,1)$ gibt mit $X \sim B(1,q)$

(b) Sei q wie in (a). Beweisen Sie $q = \frac{1}{3}(1+p)$ und bestimmen Sie Erwartungswert und Varianz von X.

(c) Bestimmen Sie für unabhängige Realisierungen $X_1, \ldots, X_n \overset{\text{iid}}{\sim} X$ aus (a) und (b) einen Maximum-Likelihood-Schätzer von p.

Aufgabe 4 (20 Punkte)
Seien $X_1, \ldots, X_n \overset{\text{iid}}{\sim} B\left(1, \frac{1+p}{3}\right)$ mit $p \in (0,1)$, $S := \sum_{i=1}^n X_i$ und

$$\hat{p} := \frac{3S}{n} - 1. \tag{2}$$

Für Berechnungen von Wahrscheinlichkeiten verwenden Sie stets die Approximation von S mit der Normalverteilung $N(E(S), \mathrm{Var}(S))$.

(a) Beweisen Sie, dass $\hat{p}$ ein erwartungstreuer Schätzer von p ist und dass $\mathrm{Var}(\hat{p}) = \frac{1}{n}(1+p)(2-p)$ gilt.

(b) Begründen Sie, dass $\hat{p}$ näherungsweise $N\left(p, \frac{1}{n}(1+p)(2-p)\right)$ verteilt ist.

(c) Sei $p = 0,3$. Bestimmen Sie mit der Normalverteilungsapproximation aus (b) für $n = 1203$ die Wahrscheinlichkeit $P(\hat{p} \geq 0,37)$.

(d) Für eine Stichprobe mit $n = 1203$ ist $s = \sum_{i=1}^{n} x_i = 550$.

(i) Bestimmen Sie einen Schätzwert $\hat{p}$ von p. Runden Sie auf zwei Nachkommastellen.

(ii) Wie beurteilen Sie die Hypothese $H_0 : p = 0,3$ gegen $H_1 : p > 0,3$ zu einem Niveau von 5 %? Verwenden Sie hierzu die Normalapproximation aus (b). *Hinweis*: (c).

Aufgabe 5 (20 Punkte)

In einer Dopingstudie wurde Sportler in anonymisierter Form gefragt ob sie Dopingmittel verwenden. Neben der Antwort JA bzw. NEIN wurde die Antwortzeiten X bzw. Y in Sekunden erhoben mit der Annahme $X \sim N\left(\mu_X, \sigma_X^2\right)$ bzw. $Y \sim N\left(\mu_Y, \sigma_Y^2\right)$. Für die Antwortzeiten X bzw. Y ergibt sich die folgende Aufstellung für Mittelwert (MW), empirischer Varianz (emp. Var.), Minimum (Min.) und Maximum (Max.):

Antwort	Anzahl	MW	emp. Var.	Min.	Max.
JA (X)	653	22	16,8	1	153
NEIN (Y)	550	17	14,2	1	143
Summe	1203				

Beachten Sie auch die angegebenen Quantile am Ende der Aufgabe!

(a) Wie beurteilen Sie die Hypothese $H_0 : \sigma_X^2 = \sigma_Y^2$ zum Niveau von 95 %?

(b) Unter geeigneten Annahmen testen Sie die Hypothese, dass der Erwartungswert der Antwortzeit nicht von der Antwort abhängt, auf dem Niveau von $\alpha = 0,05$. Sind die Annahmen erfüllt?

(c) Es besteht der Verdacht, dass bei sehr kurzer Antwortzeit gedankenlos mit JA geantwortet wurde. Bereinigt man nun die Stichprobe um die Antworten mit einer Antwortzeit von weniger als 15 Sekunden, ergibt sich die folgenden Aufstellung:

Antwort	Anzahl	MW	emp. Var.
JA (X)	381	32,3	17,3
NEIN (Y)	221	32,5	14,3
Summe	602		

Wie beurteilen Sie nun die Hypothese in (b)?

Sie können folgende q-Quantile der F-Verteilung verwenden:

q	m	n	$F_{m,n}$
0,010	381	221	0,7602
0,010	380	220	0,7598
0,025	381	221	0,7936
0,025	380	220	0,7933
0,050	381	221	0,8236
0,050	380	220	0,8233
0,950	381	221	1,2214
0,950	380	220	1,2219
0,975	381	221	1,2695
0,975	380	220	1,2701
0,990	381	221	1,3279
0,990	380	220	1,3287

q	m	n	$F_{m,n}$
0,010	653	550	0,8270
0,010	652	549	0,8269
0,025	653	550	0,8521
0,025	652	549	0,8520
0,050	653	550	0,8743
0,050	652	549	0,8742
0,950	653	550	1,1448
0,950	652	549	1,1449
0,975	653	550	1,1748
0,975	652	549	1,1750
0,990	653	550	1,2109
0,990	652	549	1,2111

Lösungsvorschläge

Aufgabe 1 [15 Punkte]
Ist $f \geq 0$ eine Treppenfunktion, d.h. es gibt $\alpha_i \geq 0$, $A_i \in \mathcal{B}^1$, $i = 1, \ldots, n$ mit $f = \sum_{i=1}^{n} \alpha_i 1_{A_i}$, dann gilt

$$\begin{aligned} \int_{\mathbb{R}} f \, d\mu &= \sum_{i=1}^{n} \alpha_i \mu(A_i) = \sum_{i=1}^{n} \alpha_i (a 1_{A_i}(x) + b 1_{A_i}(y)) \\ &= a \sum_{i=1}^{n} \alpha_i 1_{A_i}(x) + b \sum_{i=1}^{n} \alpha_i 1_{A_i}(y) = a f(x) + b f(y). \end{aligned}$$

Ist nun $f \geq 0$, dann sei $(f_n)_{n \in \mathbb{N}}$ eine Folge von Treppenfunktionen mit $f_n \uparrow f$, $f_n \geq 0, n \in \mathbb{N}$. Mit dem Satz von der Monotonen Konvergenz folgt

$$\int_{\mathbb{R}} f \, d\mu = \lim_{n \to \infty} \underbrace{\int_{\mathbb{R}} f_n \, d\mu}_{= a f_n(x) + b f_n(y)} = \lim_{n \to \infty} (a f_n(x) + b f_n(y)) = a f(x) + b f(y).$$

Im allgemeinen Fall sei $f_{\pm} := \max(0, \pm f) \geq 0$. f_+ und f_- sind messbar, somit gilt

$$\int_{\mathbb{R}} f_+ \, d\mu = a f_+(x) + b f_+(x) \in \mathbb{R} \text{ und } \int_{\mathbb{R}} f_- \, d\mu = a f_-(x) + b f_-(y) \in \mathbb{R}.$$

Daraus folgt wegen $|f| = f_+ + f_-$, dass f integrierbar ist und aus $f = f_+ - f_-$ schließlich

$$\begin{aligned} \int_{\mathbb{R}} f \, d\mu &= \int_{\mathbb{R}} f_+ \, d\mu - \int_{\mathbb{R}} f_- \, d\mu = a f_+(x) + b f_+(y) - a f_-(x) - b f_-(y) \\ &= a f(x) + b f(y). \end{aligned}$$

Aufgabe 2 [(a) 6, (b) 9]
Zu (a)

$$\begin{aligned} E\left((X+Y)^2 \,|Y\right) &= E\left(X^2 + 2XY + Y^2 | Y\right) \\ &= E\left(X^2|Y\right) + E\left(2XY|Y\right) + E\left(Y^2|Y\right) \\ &= Y^2 + 2Y E\left(X|Y\right) + Y^2 = Y^2 + 2YY + Y^2 = 4Y^2 \end{aligned}$$

Für die zweite Zeile wird die Linearität der bedingten Erwartung verwendet, für die dritte die Voraussetzungen und die $\sigma(Y)$-Messbarkeit von Y und Y^2.
Zu (b)
Analog zu (a) ergibt sich

$$\begin{aligned} E\left((X-Y)^2 \,|Y\right) &= E\left(X^2 - 2XY + Y^2 | Y\right) \\ &= E\left(X^2|Y\right) - E\left(2XY|Y\right) + E\left(Y^2|Y\right) \\ &= Y^2 - 2Y E\left(X|Y\right) + Y^2 = Y^2 - 2YY + Y^2 = 0. \end{aligned}$$

Mit dem Satz von der iterierten Erwartung folgt damit

$$\int_\Omega (X-Y)^2\,dP = E((X-Y)^2) = E(E(X-Y)^2|Y)) = 0.$$

Wegen $(X-Y)^2 \geq 0$ folgt daraus $(X-Y)^2 = 0$ fast sicher und daraus $X = Y$ fast sicher.

In den folgenden Aufgaben wird auf die Statistik einer im Juli 2017 veröffentlichte Studie (vgl. [1]) über Doping im Sport zurückgegriffen. Die Sportler wurden indirekt mit der sogenannten randomisierten Antworttechnik befragt, die die Anonymität gewährleistet. Es werden gleichzeitig zwei Fragen D (Doping) und H (Harmlos) präsentiert mit den möglichen Antworten JA (=1) und NEIN (=0). Auf welche Frage der Sportler antwortet, wird mit der Zufallsvariablen Z zufällig bestimmt. Nur der Sportler weiß dabei, ob er auf D oder H antwortet. Dies führt zum Modell (1).

Aufgabe 3 [(a) 2 (b) 6, (c) 12]

Zu (a)

X nimmt die Werte 0 und 1 an, ist also Bernoulliverteilt.

Zu (b)

Die Zufallsvariable $X = ZD + (1-Z)H$ ist genau dann 1, wenn $Z = 1$, $D = 1$ oder $Z = 0$, $H = 1$. Damit gilt wegen der Unabhängigkeit von Z, D, H

$$\begin{aligned} P(X=1) &= P(ZD + (1-Z)H = 1) = P(Z=1, D=1) + P(Z=0, H=1) \\ &= P(Z=1)P(D=1) + P(Z=0)P(H=1) \\ &= \frac{1}{3}p + \frac{2}{3}\frac{1}{2} + = \frac{1}{3}(1+p). \end{aligned}$$

Wegen

$$P(X=0) = 1 - P(X=1) = 1 - \frac{1}{3}(1+p) = \frac{1}{3}(2-p)$$

ergibt sich

$$\begin{aligned} E(X) &= \frac{1}{3}(1+p) \\ \mathrm{Var}(X) &= \frac{1}{3}(1+p)\frac{1}{3}(2-p) = \frac{1}{9}(1+p)(2-p). \end{aligned}$$

Zu (c)

Die gemeinsame Dichte $f : \mathbb{R}^n \longrightarrow [0,\infty)$ von $(X_1, \ldots, X_n)$ ist auf ihrem Träger wegen der Unabhängigkeit gegeben durch

$$\begin{aligned} f(x_1, \ldots, x_n) &= \prod_{i=1}^n \left(\frac{1}{3}(1+p)\right)^{x_i} \left(\frac{1}{3}(2-p)\right)^{1-x_i} \\ &= \frac{1}{3^n}(1+p)^{\sum_{i=1}^n x_i}(2-p)^{n-\sum_{i=1}^n x_i},\ x_i \in \{0,1\}. \end{aligned}$$

Geht man zur Log-Likelihood-Funktion über folgt

$$
\begin{aligned}
\ell(p) &= -n \ln 3 + \ln(1+p) \sum_{i=1}^{n} x_i + \ln(2-p) \left(n - \sum_{i=1}^{n} x_i \right) \\
\frac{d}{dp}\ell(p) &= \frac{1}{1+p} \sum_{i=1}^{n} x_i - \frac{1}{2-p} \left(n - \sum_{i=1}^{n} x_i \right) \\
\frac{d^2}{dp^2}\ell(p) &= -\frac{1}{(1+p)^2} \sum_{i=1}^{n} x_i - \frac{1}{(2-p)^2} \left(n - \sum_{i=1}^{n} x_i \right) < 0.
\end{aligned}
$$

Damit ergibt sich mit $S := \sum_{i=1}^{n} S_i$ das Maximum $\hat{p}$

$$
\begin{aligned}
\frac{1}{1+\hat{p}} S &= \frac{1}{2-\hat{p}} (n-S) \\
(2-\hat{p})S &= (1+\hat{p})(n-S) \\
2S &= n + \hat{p}n - S \\
\hat{p} &= \frac{3S-n}{n}.
\end{aligned}
$$

Aufgabe 4[(a) 5 (b) 5 (c) 4 (d) 6]
Zu (a)
Wegen $S \sim B\left(n, \frac{1}{3}(1+p)\right)$ gilt

$$
\begin{aligned}
E(\hat{p}) &= E\left(\frac{3S}{n} - 1\right) = \frac{3}{n}E(S) - 1 = \left(\frac{3}{n}\right) n \left(\frac{1}{3}(1+p)\right) - 1 = p \\
\mathrm{Var}(\hat{p}) &= \mathrm{Var}\left(\frac{3S}{n} - 1\right) = \mathrm{Var}\left(\frac{3S}{n}\right) = \frac{9}{n^2}\mathrm{Var}(S) = \frac{9}{n^2} n \left(\frac{1}{3}(1+p)\right)\left(\frac{1}{3}(2-p)\right) \\
&= \frac{1}{n}(1+p)(2-p).
\end{aligned}
$$

Zu (b)
Allgemein gilt: $X \sim N(\mu, \sigma^2) \Longrightarrow aX + b \sim N(a\mu + b, a^2\sigma^2)$. Da S näherungsweise normalverteilt, folgt auch, dass $\frac{3}{n}S - 1$ näherungsweise normalverteilt mit den in (a) bestimmten Parameter für Erwartungswert und Varianz.
Zu (c)
Mit der Approximation in (b) ergibt sich hier, dass $\hat{p} \sim N\left(\frac{3}{10}, \frac{1{,}3 \cdot 1{,}7}{1203}\right)$ näherungsweise und somit

$$
P(\hat{p} \geq 0,37) \approx 1 - \Phi\left(\frac{0,37 - 0,3}{\sqrt{1,3 \cdot 1,7}}\sqrt{1203}\right) = 1 - \Phi(1,63) = 1 - 0,948 = 0,052.
$$

Zu (d)
(i) $\hat{p} = \dfrac{3 \cdot 550}{1203} - 1 \approx 0,37.$
(ii) Die Nullhypothese wird nicht abgelehnt, wenn $\hat{p}$ unter H_0 (also $p = 0,3$) nicht zu stark von $0,3$ abweicht, also wenn

$$
P(\hat{p} \leq 0,37) \leq 0,95
$$

gilt. Mit (c) gilt

$$P(\hat{p} \leq 0,37) \approx 0,948 \leq 0,95.$$

Damit wird H_0 nicht abgelehnt.

Aufgabe 5 [(a) 4 (b) 9 (c) 7]
Zu (a)
Es wird der F-Test durchgeführt:

$$t = \frac{16,8}{14,2} = 1,183 \qquad F_{652,549,1-\frac{975}{1000}} = 1,175$$

also wird H_0 wegen $|t| = 1,183 > 1,175$ verworfen
Zu (b)
Man könnte den Zweistichproben t-Test durchführen:

$$t = \frac{17-22}{s}\sqrt{\frac{653 \cdot 550}{1203}}, \qquad \text{mit } s^2 = \frac{652 \cdot 16,8 + 549 \cdot 14,2}{1201} = 15,61$$
$$= -21,86.$$

Damit würde H_0 abgelehnt, wegen $t_{1201,1-\frac{975}{1000}} \approx 1,96$ und $|t| = 21,86 > 1,96$. Jedoch ist $\sigma_X^2 = \sigma_Y^2$ eine der Voraussetzungen für den zweieitigen t-Test. Dies wurde aber in (a) verworfen, die Voraussetzung ist also nicht erfüllt.
Alternativ kann man die Konfidenzintervalle für μ_X und μ_Y bestimmen, es ergibt sich

$$\left[\overline{x} - t_{652,1-\frac{975}{1000}} \frac{s_X}{\sqrt{653}}, \overline{x} + t_{652,1-\frac{975}{1000}} \frac{s_X}{\sqrt{653}}\right] = \left[\frac{217}{10}, \frac{223}{10}\right]$$
$$\left[\overline{y} - t_{549,1-\frac{975}{1000}} \frac{s_Y}{\sqrt{550}}, \overline{y} + t_{549,1-\frac{975}{1000}} \frac{s_Y}{\sqrt{550}}\right] = \left[\frac{167}{10}, \frac{173}{10}\right].$$

Da sich die Intervalle nicht schneiden, wird H_0 abgelehnt.
Zu (c)
Mit F-Test und t-Test erhält man nun mit der Vorgehensweise von (a) und (b)

$$t = \frac{17,3}{14,3} = 1,21, \qquad F_{380,220,1-\frac{975}{1000}} = 1,2701 \qquad (F\text{-Test})$$
$$t = \frac{32,5-32,3}{s}\sqrt{\frac{381 \cdot 221}{602}}, \qquad s = \frac{380 \cdot 17,3 + 220 \cdot 14,3}{600} = 16,2 \qquad (t\text{-Test})$$
$$= 0,588 \qquad t_{600,1-\frac{975}{1000}} \approx 1,96$$

somit verwirft weder der F-Test noch der Zweistichproben t-Test. Die Antwortzeiten der beiden Gruppen unterscheiden sich hier nicht.

Literatur

[1] Ulrich, Rolf et al.: Doping in Two Elite Athletics Competitions Assessed by Randomized-Response Surveys, *Sports Med*, DOI 10.1007/s40279-017-0765-4, Springer International Publishing AG 2017.