Stochastik - Lösung (BSc D-MAVT / BSc D-MATH / BSc D-MATL)

1. (6 Punkte)

a) (2 Punkte) Wir definieren die Ereignisse $K = \{\text{die Person ist krank}\}$ und $T = \{\text{der Test ist positiv}\}$. Laut Aufgabenstellung gilt $P(K) = 0.001$, $P(T|K) = 0.99$ und $P(T|K^c) = 0.002$. Der Satz von Bayes und der Satz der totalen Wahrscheinlichkeit liefern

$$\begin{aligned} P(K|T) &= \frac{P(K\cap T)}{P(T)} = \frac{P(T|K)P(K)}{P(T)} = \frac{P(T|K)P(K)}{P(T\cap K)+P(T\cap K^c)} \\ &= \frac{P(T|K)P(K)}{P(T|K)P(K)+P(T|K^c)P(K^c)} = \frac{P(T|K)P(K)}{P(T|K)P(K)+P(T|K^c)(1-P(K))} \\ &= \frac{0.99\times 0.001}{0.99\times 0.001 + 0.002\times 0.999} = \frac{0.00099}{0.002988} = 0.331\,. \end{aligned}$$

Die Wahrscheinlichkeit, dass die getestete Person wirklich krank ist, wenn der Test zu einem positiven Ergebnis führt, ist nur etwa 33.1%.

b) (1 Punkt) $P(K|T^c) = \dfrac{P(K\cap T^c)}{P(T^c)} = \dfrac{P(K) - P(K\cap T)}{1-P(T)} = \dfrac{0.001 - 0.00099}{1-0.002988} = 0.001\%$.

c) (1 Punkt) Sei die Zufallsvariable X die Gesamtzahl der Einwohner, die krank sind. Dann gilt $X \sim \mathrm{Bin}(n = 1000, p = 0.001)$. Die Wahrscheinlichkeit, dass kein Einwohner krank ist, ist also gegeben durch $P(X = 0) = (1-p)^n = 0.999^{1000} = 0.368$.

d) (1 Punkt) Sei die Zufallsvariable Z die Gesamtzahl der Einwohner, die mit dem billigeren Test positiv getestet wurden. Dann gilt $Z \sim \mathrm{Bin}(n = 1000, p = P(T) = 0.002988)$. Es beschreibe die Zufallsvariable Y die Gesamtkosten. Dann ist $Y = 1000 + 2Z$. Die durchschnittlichen Gesamtkosten sind gegeben durch den Erwartungswert der Zufallsvariablen Y, also $E(Y) = 1000 + 2E(Z) = 1000 + 2np = 1000 + 2 \times 1000 \times 0.002988 = 1005.98$. *Im Durchschnitt* spart der Arzt also $2000 - 1005.98 = 994.02$ Euro ein.

e) (1 Punkt) Die Wahrscheinlichkeit, dass das Budget überschritten wird, ist gegeben durch

$$\begin{aligned} P(Y > 1006) = P(1000 + 2Z > 1006) &= P(Z > 3) \\ &= 1 - P(Z \le 3) = 1 - \sum_{k=0}^{3} P(Z = k)\,. \end{aligned}$$

Wir können mit der Binomialverteilung rechnen:
$P(Z = 0) = (1-p)^n = (1 - 0.002988)^{1000} = 0.997012^{1000} = 0.050163$,

$$P(Z = k) = \binom{n}{k} p^k (1-p)^{n-k} = P(Z = k-1)\frac{p}{1-p}\frac{n-k+1}{k}\,,$$

$$P(Z \leq 3) = P(Z=0)\left(1 + \frac{p}{1-p}\frac{1000}{1}\left(1 + \frac{p}{1-p}\frac{999}{2}\left(1 + \frac{p}{1-p}\frac{998}{3}\right)\right)\right) = 0.650\,.$$

Die Wahrscheinlichkeit, dass das Budget überschritten wird, ist aslo 35%.

2. (6 Punkte)

a) (0.5 Punkte)

$$\begin{aligned} f_X(x) = \int_{\mathbb{R}} f_{X,Y}(x,y)dy &= \int_{\mathbb{R}} xe^{-x(1+y)} 1_{\{x \geq 0\}} 1_{\{y \geq 0\}} dy \\ &= xe^{-x} 1_{\{x \geq 0\}} \int_0^\infty e^{-xy} dy = e^{-x} 1_{\{x > 0\}}\,, \end{aligned}$$

denn $\int_0^\infty xe^{-xy}dy = 1$ für $x > 0$ (Integral der Dichte von $\text{Exp}(x)$).

b) (1 Punkt) Die Wahrscheinlichkeit, dass die Festplatte nicht mehr als a Jahre funktioniert, ist gegeben durch $P(X \leq a) = F(a) = \int_{-\infty}^a f_X(x)dx = \int_0^a e^{-x}dx = 1 - e^{-a}$. Also ist $P(X \leq 1.5) = 0.777$.

c) (1 Punkt) Wir suchen q_α, so dass gilt $P(X \leq q_\alpha) = \alpha$. Daraus folgt, dass $P(X \leq q_\alpha) = \alpha \Leftrightarrow F(q_\alpha) = 1 - e^{-q_\alpha} = \alpha \Leftrightarrow q_\alpha = -\log(1-\alpha)$.
Also ist $q_{97.5\%} = -\log(1 - 0.975) = 3.69$. Mit einer Wahrscheinlichkeit von 97.5% funktioniert die Festplatte nicht mehr als 3.69 Jahre.

d) (1 Punkt) Wir berechnen die bedingte Dichte von Y gegeben $X = x$, und zwar für $x > 0$. Das ist

$$f_{Y|X=x}(y) = \frac{f_{X,Y}(x,y)}{f_X(x)} = \frac{xe^{-x(1+y)} 1_{\{x \geq 0\}} 1_{\{y \geq 0\}}}{e^{-x} 1_{\{x > 0\}}} = xe^{-xy} 1_{\{y \geq 0\}}$$

für $x > 0$. Jetzt berechnen wir die bedingte Wahrscheinlichkeit, dass $Y > b$ gegeben $X = a$ als

$$P(Y > b | X = a) = \int_b^\infty f_{Y|X=a}(y)dy = \int_b^\infty ae^{-ay}dy = e^{-ab}\,.$$

Wenn man weiss, dass die Festplatte nach 2 Jahren kaputt war, so ist die Wahrscheinlichkeit, dass das RAM mehr als 1.5 Jahre funktioniert, gegeben durch
$P(Y > 1.5 | X = 2) = e^{-3} = 0.0498$.

e) (1 Punkt)
Wir berechnen den bedingten Erwartungswert von Y gegeben $X = a$; das ist

$$E(Y|X=a) = \int_{\mathbb{R}} y f_{Y|X=a}(y)dy = \int_0^\infty yae^{-ay}dy = \left[-ye^{-ay}\right]_0^\infty + \int_0^\infty e^{-ay}dy = \frac{1}{a}\,.$$

Alternativ: die bedingte Dichte von Y gegeben $X = a$ ist $ae^{-ay}1_{\{y \geq 0\}}$; also ist Y gegeben $X = a$ exponentialverteilt mit Parameter a und damit $E(Y|X = a) = \frac{1}{a}$. Wenn man weiss, dass die Festplatte nach 2 Jahren kaputt war, so ist die erwartete Lebensdauer des RAM gegeben durch $E(Y|X = 2) = 0.5$, also 1/2 Jahr.

f) (1.5 Punkte) Die Dichte von Y für $y \geq 0$ ist

$$f_Y(y) = \int\limits_{\mathbb{R}} f_{X,Y}(x,y)dx = 1_{\{y\geq 0\}} \int\limits_0^\infty xe^{-x(1+y)}dx = \frac{1}{(1+y)^2} 1_{\{y\geq 0\}} \,,$$

denn als Funktion von x ist der Integrand bis auf einen (fehlenden) Faktor $\frac{1}{1+y}$ die Dichte einer Exponentialverteilung mit Parameter $1+y$. Also ist

$$f_{X,Y}(x,y) = xe^{-x(1+y)} 1_{\{x\geq 0\}} 1_{\{y\geq 0\}} \neq e^{-x} 1_{\{x>0\}} \frac{1}{(1+y)^2} 1_{\{y\geq 0\}} = f_X(x) f_Y(y) \,,$$

d.h. X und Y sind nicht unabhängig.

3. (6 Punkte)

a) (2 Punkte) Damit $f_{\alpha,\beta}$ eine Dichte ist, muss gelten $\int_{\mathbb{R}} f_{\alpha,\beta}(x)dx = 1$. Also berechnen wir $\int_{\mathbb{R}} f_{\alpha,\beta}(x)dx = \int_\beta^\infty cx^{-\alpha}dx = c\left[\frac{x^{-\alpha+1}}{-\alpha+1}\right]_\beta^\infty = c\frac{\beta^{-\alpha+1}}{\alpha-1}$. Somit erhalten wir $c = (\alpha-1)\beta^{\alpha-1}$. Daraus folgt

$$\begin{aligned} E(X^s) = \int\limits_{\mathbb{R}} x^s f_{\alpha,\beta}(x)dx &= \int\limits_\beta^\infty (\alpha-1)\beta^{\alpha-1}x^{s-\alpha}dx \\ &= (\alpha-1)\beta^{\alpha-1}\left[\frac{x^{s-\alpha+1}}{s-\alpha+1}\right]_\beta^\infty = \frac{(\alpha-1)\beta^s}{\alpha-1-s} \,, \end{aligned}$$

für $\alpha > s+1$. Insbesondere ist $E(X) = \frac{(\alpha-1)\beta}{\alpha-2}$, $E(X^2) = \frac{(\alpha-1)\beta^2}{\alpha-3}$,
$\mathrm{Var}(X) = E(X^2) - E(X)^2 = \frac{(\alpha-1)\beta^2}{(\alpha-3)(\alpha-2)^2}$, für $\alpha > 1$.

b) (0.5 Punkte) Wir wählen α so, dass das s-te theoretische Moment von X, definiert als $\mu_s = E(X^s)$, und das s-te empirische Moment von $x_1,\ldots,x_n$, definiert als $m_s = \frac{1}{n}\sum_{i=1}^n x_i^s$, übereinstimmen. Also haben wir für $s=1$

$$\mu_1(\alpha) = m_1 \Leftrightarrow \frac{(\alpha-1)\beta}{\alpha-2} = m_1 \Leftrightarrow \alpha = \frac{\beta-2m_1}{\beta-m_1} = 1 - \frac{m_1}{\beta-m_1} \Rightarrow \alpha = 1 - \frac{\frac{1}{n}\sum_{i=1}^n x_i}{\beta - \frac{1}{n}\sum_{i=1}^n x_i} \,.$$

Dies führt zum Momentenschätzer

$$\hat{\alpha}_{MoM} = 1 - \frac{\frac{1}{n}\sum_{i=1}^n X_i}{\beta - \frac{1}{n}\sum_{i=1}^n X_i} = 1 - \frac{\bar{X}_n}{\beta - \bar{X}_n}.$$

(1.5 Punkte) Da die Beobachtungen unabhängig sind, können wir die Likelihoodfunktion schreiben als $L(x_1,\ldots,x_n,\alpha,\beta) = \prod_{i=1}^n f_{\alpha,\beta}(x_i)$. Die log-Likelihoodfunktion ist

$$\begin{aligned} l(x_1,\ldots,x_n,\alpha,\beta) = \log\left(\prod_{i=1}^n f_{\alpha,\beta}(x_i)\right) &= \sum_{i=1}^n \log f_{\alpha,\beta}(x_i) \\ &= \sum_{i=1}^n \log\left((\alpha-1)\beta^{\alpha-1}x_i^{-\alpha}\right) \\ &= n\log(\alpha-1) + n(\alpha-1)\log\beta - \alpha\sum_{i=1}^n \log x_i \,. \end{aligned}$$

Wir berechnen die Ableitung der log-Likelihoodfunktion als

$$\frac{\partial l(x_1,\ldots,x_n,\alpha,\beta)}{\partial\alpha} = \frac{n}{\alpha-1} + n\log\beta - \sum_{i=1}^{n}\log x_i = \frac{n}{\alpha-1} + \sum_{i=1}^{n}\log\frac{\beta}{x_i}\,.$$

Aus $\frac{\partial l(x_1,\ldots,x_n,\alpha,\beta)}{\partial\alpha} = 0$ folgt also $\alpha = 1 - \frac{n}{\sum_{i=1}^{n}\log\frac{\beta}{x_i}}$. Da die zweite Abteilung

$$\frac{\partial l^2(x_1,\ldots,x_n,\alpha,\beta)}{\partial\alpha^2} = -\frac{n}{(\alpha-1)^2}$$

negativ ist, schliessen wir, dass α ein Maximum ist. Der gesuchte Maximum-Likelihood-Schätzer ist also

$$\hat{\alpha}_{MLE} = 1 - \frac{n}{\sum_{i=1}^{n}\log\frac{\beta}{X_i}} = 1 + \frac{1}{\overline{(\log X)}_n - \log\beta}\,,$$

wobei $\overline{(\log X)}_n = \frac{1}{n}\sum_{i=1}^{n}\log X_i$.

c) **(2 Punkt)** Wir nehmen an, dass die N strategischen Ölreserven unabhängig voneinander sind. Eine zweite Annahme ist, dass N gross genug ist, damit die Normalverteilung via den zentralen Grenzwertsatz eine gute Approximation liefert. Laut dem zentralen Grenzwertsatz ist S_N approximativ $\mathcal{N}(N\mu_{X_1}, N\sigma^2_{X_1})$-verteilt mit $\mu_{X_1} = E(X) = \frac{(\alpha-1)\beta}{\alpha-2}$ und $\sigma^2_{X_1} = \mathrm{Var}(X) = \frac{(\alpha-1)\beta^2}{(\alpha-3)(\alpha-2)^2}$.

Die Wahrscheinlichkeit, dass das Gesamtniveau der N strategischen Ölreserven die Mindestgrenze δ unterschreitet, ist approximativ gegeben durch

$$\begin{aligned} P(S_N \le \delta) &= P\left(\frac{S_N - \mu_{S_N}}{\sigma_{S_N}} \le \frac{\delta - \mu_{S_N}}{\sigma_{S_N}}\right) \approx P\left(Z \le \frac{\delta - \mu_{S_N}}{\sigma_{S_N}}\right) \\ &= \Phi\left(\frac{\delta - \mu_{S_N}}{\sigma_{S_N}}\right) = \Phi\left(\frac{\delta - N\mu_{X_1}}{\sqrt{N}\sigma_{X_1}}\right) = \Phi\left(\frac{\delta - N\frac{(\alpha-1)\beta}{\alpha-2}}{\sqrt{N\frac{(\alpha-1)\beta^2}{(\alpha-3)(\alpha-2)^2}}}\right). \end{aligned}$$

4. **(6 Punkte)** Seien $X_i \sim \mathcal{N}(\mu,\sigma^2)$ die gemessenen Nüchternblutzuckerwerte und wie immer die tatsächlich gemessenen Werte $x_1,\ldots,x_{10}$ Realisierungen davon. Der Parameter μ soll getestet werden. Da die Standardabweichung σ unbekannt ist, muss sie aus den Daten geschätzt werden, was uns zum t-Test führt.

a) **(1 Punkt)** Mit $t_{n-1;1-\frac{\alpha}{2}} = t_{9;0.995} = 3.25$ ergibt sich die Realisierung des 99%-Vertrauensintervalls zu $I_{99\%} = \left[\overline{x}_{10} - t_{9,0.995}\frac{s_{10}}{\sqrt{10}}\ ,\ \overline{x}_{10} + t_{9,0.995}\frac{s_{10}}{\sqrt{10}}\right] = [3.87, 5.17]$.

b) **(1 Punkt)** Wir wollen wissen, ob das μ des Nüchternblutzuckers grösser als 4.1 mmol/l ist. Dazu führen wir einen *nach oben einseitigen* Test zum Niveau 5% mit Null- und Alternativhypothese $H_0 : \mu = \mu_0 := 4.1$ und $H_A : \mu > \mu_0$ durch. Der Verwerfungsbereich ist somit gegeben durch $K = [t_{n-1,1-\alpha}, \infty) = [t_{9,0.95}, \infty) = [1.833, \infty)$. Die Realisierung der Teststatistik ist

$$t = \sqrt{10}\frac{\overline{x}_{10} - \mu_0}{s_{10}} = \sqrt{10}\frac{4.52 - 4.1}{0.63} = 2.11 \in K\,,$$

weshalb die Nullhypothese verworfen wird.

c) (1 Punkt) Wir suchen das Niveau α, das gerade den Verwerfungsbereich $[2.12, \infty)$ hat; das heisst, wir suchen das Niveau α, für das $t_{n-1,1-\alpha} = 2.12$ gilt. Durch Rückwärtsauslesen aus der Tabelle der t-Quantile sieht man lediglich, dass der Wert zwischen 2.5% und 5% liegen muss (linear interpoliert bekommt man 3.3%). Dieses Niveau heisst P-Wert.

Wenn $\sigma = 0.5$ bekannt ist, verwenden wir jetzt den z-Test. Der einzige Unterschied ist, dass die Quantile der t-Verteilung durch jene der Standardnormalverteilung und die geschätzte Streuung s_{10} durch die tatsächliche Streuung σ ersetzt wird.

d) (1 Punkt) $K = [z_{1-\alpha}, \infty) = [z_{0.95}, \infty) = [1.645, \infty)$.

Die Teststatistik lautet nun $Z = \sqrt{10}\frac{\overline{X}_{10}-\mu_0}{\sigma}$, und ihre Realisierung ist gegeben durch

$$z = \sqrt{10}\frac{\overline{x}_{10} - \mu_0}{\sigma} = \sqrt{10}\frac{4.52 - 4.1}{0.5} = 2.66 \in K\,.$$

Also wird H_0 auch hier verworfen.

e) (1 Punkt) Wegen $Z \sim \mathcal{N}\left(\sqrt{n}\frac{\mu-\mu_0}{\sigma}, 1\right)$ ist $Z \overset{\mu=4.2}{\sim} \mathcal{N}(0.63, 1)$, und somit ergibt sich die Wahrscheinlichkeit für einen Fehler 2.Art zu

$$\begin{aligned}\mathbb{P}_{\mu=4.2}(Z \notin K) &= \mathbb{P}_{\mu=4.2}(Z < 1.645) = F_{\mathcal{N}(0.632,1)}(1.645) = \Phi(1.645 - 0.632)\\ &= \Phi(1.013) = 0.8445 \approx 84\%,\end{aligned}$$

was ziemlich hoch ist.

f) (1 Punkt) Die Breite des Vertrauensintervalls ist $2\frac{\sigma}{\sqrt{n}}z_{0.975}$. Somit ergibt sich die Bedingung $\frac{\sigma}{\sqrt{n}}z_{0.975} < 0.1$ und weiter $\sqrt{n} > \frac{\sigma z_{0.975}}{0.1} = \frac{0.5 \cdot 1.96}{0.1} = 9.8$ bzw. $n > 9.8^2 = 96.04$, d.h. n sollte mindestens 97 sein.

Die Frage ist nun, wie gross $\overline{x}_{97}$ sein muss, damit $z = \sqrt{97}\frac{\overline{x}_{97}-4.1}{0.5} \geq 1.645$ gilt. Diese Bedingung ist äquivalent zu $\overline{x}_{97} \geq 1.645\frac{0.5}{\sqrt{97}} + 4.1 = 4.18$. Bei einem Stichprobenumfang von $n = 97$ würde also bereits ein empirisch gemessener Mittelwert von 4.18 zum einseitigen Verwerfen der Nullhypothese führen.

5. (6 Punkte)

a) (0.5 Punkte) Die Stichprobe ist gepaart. Es geht um die gleichen Läufer, mit verschiedenen Schuhen.

b) (1.5 Punkte) Sei $D_i = X_i - Y_i$, wobei X_i die Zeit des i-ten Langstreckenläufers mit den alten Laufschuhen sei, und Y_i die Zeit des i-ten Langstreckenläufers mit den neuen Laufschuhen. Wir wollen testen, ob sich die mittlere Zeit der Langstreckenläufer mit beiden Laufschuhen unterscheiden, und zwar hoffen wir, dass die neuen Laufschuhe besser und damit die Laufzeiten kürzer sind. Es gibt keine Annahme der Normalverteilung. Die Voraussetzungen für den entsprechenden Wilcoxon-Test sind daher, dass die D_i unabhängig und identisch verteilt, mit einer um den Mittelwert μ symmetrischen stetigen Verteilung, sind. Für den t-Test kommt hinzu, dass die D_i normalverteilt sein müssen. Für den Vorzeichentest würde man die Symmetrie nicht brauchen, sondern nur, dass der Median der D_i Null ist. Daher ist der Wilcoxon-Test am besten geeignet um die gegebenen Informationen bestmöglich auszunutzen.

c) **(2 Punkte)** Wir wollen herausfinden, ob der Mittelwert μ der Differenzen $D_i = X_i - Y_i$ grösser als 0 ist. Dazu führen wir einen einseitigen Test zum Niveau 5% mit Null- und Alternativhypothese $H_0 : \mu = 0$ und $H_A : \mu > 0$ durch. Der Verwerfungsbereich ist somit gegeben durch $K = \{W \geq u\}$. Nach der Tabelle haben wir $K = \{W \geq 45\}$. Wir berechnen die entsprechenden Ränge Rang($|d_i|$) und die V_i-Indikatoren.

Langstreckenläufer (i)	1	2	3	4	5	6	7	8	9	10
alte Laufschuhe (x_i)	27.42	27.47	28.20	28.55	28.29	28.92	28.85	29.51	29.63	30.58
neue Laufschuhe (y_i)	27.34	27.40	28.14	28.70	28.53	29.07	29.00	29.57	29.69	30.56
$d_i = x_i - y_i$	0.08	0.07	0.06	-0.15	-0.24	-0.15	-0.15	-0.06	-0.06	0.02
Rang($\|d_i\|$)	6	5	3	8	10	8	8	3	3	1
v_i	1	1	1	0	0	0	0	0	0	1

Die Realisierung der Teststatistik ist

$$w = \sum_{i=1}^{10} \text{Rang}(|d_i|)v_i = 6 + 5 + 3 + 1 = 15 \notin K = \{W \geq 45\},$$

weshalb die Nullhypothese **nicht** verworfen wird. Es kann also nicht nachgewiesen werden, dass die neuen Laufschuhe eine Verbesserung der Leistung bewirken.

d) **(2 Punkte)** Wenn die Daten wirklich (eventuell verschoben) Exp(λ)-verteilt sind, erwarten wir ungefähr eine Gerade mit Steigung $1/\lambda$, und wenn die Daten wirklich $\mathcal{N}(\mu, \sigma^2)$-verteilt sind, erwarten wir ungefähr eine Gerade mit Steigung σ und Achsenabschnitt μ. Mithilfe dieser zwei QQ-Plots können wir schliessen, dass die Zeitdifferenzen eher $\mathcal{N}(\mu, \sigma^2)$-verteilt sind.

Unter Annahme der Normalverteilung sollte der Club den t-Test wählen, da er eine bessere Macht hat (unter dieser Annahme) als der Wilcoxon-Test oder der Vorzeichen-Test. Der z-Test ist keine Option, denn die Varianz ist nicht bekannt.