Lösung - Stochastik (401-0603-00L)

1. Jede richtige Antwort 1 Punkt.

a)	i) ☐	ii) ☒	iii) ☐
b)	i) ☐	ii) ☐	iii) ☒
c)	i) ☐	ii) ☐	iii) ☒
d)	i) ☐	ii) ☒	iii) ☐
e)	i) ☐	ii) ☒	iii) ☐
f)	i) ☐	ii) ☒	iii) ☐
g)	i) ☒	ii) ☐	iii) ☐
h1)	i) ☐	ii) ☐	iii) ☒
h2)	i) ☐	ii) ☐	iii) ☒
h3)	i) ☐	ii) ☒	iii) ☐

2. **a)** Wir treffen die Annahme, dass die einzelnen Tests unabhängig sind und immer die gleiche Falsch-Positiv-Rate von 2.5% haben. Da also i.i.d. Bernoulli-Versuche so oft wiederholt werden, bis es zum ersten positiven Ergebnis kommt, folgt X einer geometrischen Verteilung mit Parameter $p = 0.025$.

b) Nach den Formeln für die geometrische Verteilung gilt $\mathbb{E}[X] = 1/p = 40$.

c) Sei G die gesuchte Anzahl, dann soll gelten, dass

$$0.05 \geq \mathbb{P}(X > G) = 1 - F_X(G) = 1 - (1 - (1-p)^G).$$

Hier haben wir die Formel für die kumulative Verteilungsfunktion einer geometrischen ZV verwendet. (Alternativ kann auch direkt Unabhängigkeit der Bernoulli-Versuche verwendet werden.) Auflösen nach G liefert $G \geq \frac{\log(1-0.95)}{\log(1-p)} \approx 118.3$. Also ist die gesuchte Anzahl Tests 119 .

d) Ähnlich wie in der vorherigen Aufgabe bekommen wir hier

$$0.05 \geq \mathbb{P}(X > 10) = (1-p)^{10}.$$

Auflösen nach p liefert $p = 1 - (0.05)^{1/10} \approx 0.2589$, also in etwa 25.9%.

e) Sei T_1 das Ereignis, dass der Test positiv ist und K_1 das Ereignis, dass die Person die Krankheit bereits hatte. Wir wollen $\mathbb{P}(T_1)$ berechnen. Mit dem Satz der totalen Wahrscheinlichkeit gilt

$$\mathbb{P}(T_1) = \mathbb{P}(T_1|K_1)\mathbb{P}(K_1) + \mathbb{P}(T_1|K_1^c)\mathbb{P}(K_1^c) = 0.93q + 0.025(1-q) = 0.025 + 0.905q.$$

Für $q = 0.1$ ergibt sich also $\mathbb{P}(T_1) = 0.1155$ und für $q = 0.5$ erhalten wir $\mathbb{P}(T_1) = 0.4775$.

f) Mit denselben Ereignissen wie in der vorherigen Aufgabe, wollen wir $\mathbb{P}(K_1|T_1)$ berechnen. Wir verwenden den Satz von Bayes und erhalten

$$\mathbb{P}(K_1|T_1) = \frac{\mathbb{P}(T_1|K_1)\mathbb{P}(K_1)}{\mathbb{P}(T_1)} = \frac{0.93q}{0.025 + 0.905q}.$$

Einsetzten von $q = 0.1$ liefert $\mathbb{P}(K_1|T_1) = 0.8052$ und mit $q = 0.5$ bekommen wir $\mathbb{P}(K_1|T_1) = 0.9738$.

g) Sei X_i eine Zufallsvariable, die 1 ist, falls die i-te Testperson positiv getestet ist und sonst 0. Sei A_i das Ereignis, dass die i-te Testperson die Krankheit bereits hatte (also ist A_i^c das Ereignis, dass sie die Krankheit noch nicht hatte). Sei $p_1 := \mathbb{P}(X_i = 1|A_i) = 0.9$ und $p_2 := \mathbb{P}(X_i = 1|A_i^c) = 0.4$. Nach der Angabe gilt $\mathbb{P}(A_i) = q$ und $\mathbb{P}(A_i^c) = 1 - q$. Ausserdem wissen wir, dass alle X_i unabhängig sind. .
Ohne Informationen, welche Testperson die Krankheit bereits hatte, gilt (Satz der totalen Wahrscheinlichkeit)

$$\mathbb{E}[X_i] = 1 \cdot \mathbb{P}(X_i = 1) = \mathbb{P}(X_i = 1|A_i)\mathbb{P}(A_i) + \mathbb{P}(X_i = 1|A_i^c)\mathbb{P}(A_i^c) = p_1\, q + p_2\,(1-q).$$

Umformen nach q liefert $q = \frac{\mathbb{E}[X_i] - p_2}{p_1 - p_2}$.
Bei der Momentenmethode wird $\mathbb{E}[X_i]$ durch das empirische Moment $\bar{x}_n = \frac{1}{n}\sum_{i=1}^n x_i$ ersetzt. Somit erhalten wir den Momentenschätzer

$$\hat{q} = \frac{\bar{x}_n - p_2}{p_1 - p_2}.$$

Einsetzen der Werte $\bar{x}_n = \frac{240}{330}$, $p_1 = 0.9$, $p_2 = 0.4$ liefert das Ergebnis $\hat{q} \approx 0.6545$.

3. **a)** Es muss gelten, dass f über $\mathbb{R}$ zu 1 integriert. Somit folgt aus

$$1 = \int_0^\infty f(x|(k,\lambda))dx = c\left[-\frac{1}{k}\exp\left(-(x/\lambda)^k\right)\right]_0^\infty = \frac{c}{k},$$

dass $c = k$ gilt.

b) Die Likelihood-Funktion ist

$$L(\lambda) = \prod_{i=1}^n f(x_i|\lambda) = c^n(\lambda^{-k})^n \prod_{i=1}^n \left\{x_i^{k-1}\exp\left(-(x_i/\lambda)^k\right)\right\},$$

falls alle $x_i \geq 0$ und 0 sonst . Die Log-Likelihood-Funktion ist

$$l(\lambda) = \log(L(\lambda)) = n\log(c) - kn\log(\lambda) + (k-1)\sum_{i=1}^n \log(x_i) - \sum_{i=1}^n (x_i/\lambda)^k,$$

falls alle $x_i \geq 0$ und $-\infty$ sonst .

c) Wir leiten $l(\lambda)$ ab und setzen es gleich 0, um das Maximum zu finden. Es folgt also

$$0 = l'(\lambda) = -kn/\lambda - (-k)\lambda^{-k-1}\sum_{i=1}^n x_i^k,$$

und wenn wir nach λ auflösen:

$$\hat{\lambda}^k = \frac{1}{n}\sum_{i=1}^n x_i^k.$$

Einsetzen der Werte ergibt $\hat{\lambda} = 0.7088$.
Hinweis: Die zweite Ableitung ist

$$l''(\lambda) = kn/\lambda^2 - k(k+1)\lambda^{-k-2}\sum_{i=1}^n x_i^k,$$

und einsetzen von $\hat{\lambda} > 0$ ergibt

$$l''(\hat{\lambda}) = \tfrac{1}{\hat{\lambda}^2}\left[kn - k(k+1)n\right] = \tfrac{-k^2 n}{\hat{\lambda}^2} < 0,$$

daher handelt es sich bei dem Extremwert um ein Maximum.

d) Wir berechnen mit $c = k = 1$ und der Substitution $y = x/\lambda$ (i.e. $dy = dx/\lambda$) unter Verwendung des Tipps

$$\mathbb{E}[X] = \int_{-\infty}^\infty x\, f(x|k=1,\lambda)dx = \lambda^{-1}\int_0^\infty x\exp(-x/\lambda)dx = \int_0^\infty y\exp(-y)\lambda dy = \lambda\Gamma(2) = \lambda.$$

Analog berechnen wir

$$\mathbb{E}[X^2] = \int_{-\infty}^\infty x^2\, f(x|k=1,\lambda)dx = \lambda^{-1}\int_0^\infty x^2\exp(-x/\lambda)dx = \int_0^\infty \lambda y^2\exp(-y)\lambda dy = \lambda^2\Gamma(3) = 2\lambda^2.$$

und bekommen $\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \lambda^2$. .

Alternative Lösung: Für $k = 1$ handelt es sich bei der Verteilung gerade um eine Exponentialverteilung mit Parameter $1/\lambda$. Daher folgt direkt von den Formeln aus der Vorlesung, dass $\mathbb{E}[X] = \frac{1}{1/\lambda} = \lambda$ und $\mathrm{Var}(X) = \frac{1}{1/\lambda^2} = \lambda^2$.

e) Seien X_i, $1 \leq i \leq n$ Zufallsvariablen, die die Anzahl (in Millionen) gesichteter Meteoroiden am Tag i in der Zukunft modellieren. Wir nehmen an, dass alle X_i unabhängig und gleichverteilt mit Dichte $f(x|k = 1, \lambda)$ sind. Insbesondere gilt also $\mathbb{E}[X_i] = \lambda$ und $\mathrm{Var}(X_i) = \lambda^2$.
Die Anzahl (in Millionen) insgesamt gesichteter Meteoroiden in den nächsten 50 Tagen ist also $S = \sum_{i=1}^{n} X_i$, mit $n = 50$. Wir wenden den ZGWS an, um die approximative Verteilung von S zu bestimmen und erhalten $S \approx \mathcal{N}(n\lambda, n\lambda^2)$.

f) Es gilt $Z := \frac{S - n\lambda}{\lambda\sqrt{n}} \approx \mathcal{N}(0, 1)$ und somit

$$1 - \alpha = \mathbb{P}(-z_{1-\alpha/2} \leq Z \leq z_{1-\alpha/2}).$$

Einsetzen von Z und Umformen führt zu dem Vertrauensinterval $[n\lambda - z_{1-\alpha/2}\lambda\sqrt{n}, n\lambda + z_{1-\alpha/2}\lambda\sqrt{n}]$ für S .
Einsetzen der Werte $n = 50$, $\lambda = 0.7$ und $\alpha = 0.01$ ergibt $[22.25, 47.75]$.

4. **a)** Da wir wissen, dass es sich um normalverteilte Daten handelt, wir aber nicht die wahre Standardabweichung kennen, verwenden wir einen t-Test . Da die Daten nicht gepaart sind, müssen wir einen 2 Stichproben t-Test verwenden .

b) Unter den Annahmen gilt $X_i \sim \mathcal{N}(\mu_X, \sigma^2)$ i.i.d. und $Y_i \sim \mathcal{N}(\mu_Y, \sigma^2)$ i.i.d., alle unabhängig, mit einheitlicher aber unbekannter Varianz σ^2. X_i beschreiben dabei die Zeiten vom Segelflieger vom Typ 1 und Y_i die Zeiten vom Segelflieger vom Typ 2. . Da wir nur testen wollen, ob ein signifikanter Unterschied besteht (ohne Vorwissen über die Richtung der Abweichung) verwenden wir die Null- und Alternativhypothesen $H_0 : \mu_X = \mu_Y$ und $H_A : \mu_X \neq \mu_Y$. Die Teststatistik unter der Nullhypothese ist

$$T = \frac{\bar{X}_n - \bar{Y}_n}{\hat{\sigma}_{\text{pool}} \sqrt{2/n}},$$

welche einer t-Verteilung mit $2n - 2 = 18$ Freiheitsgraden folgt .

c) Der realisierte Wert der Teststatistik ist $t = 1.842$, wobei verwendet wurde, dass $n = 10$, $\hat{\sigma}_{\text{pool}} \approx 3.684$.
Daher ist der p-Wert gegeben durch $\mathbb{P}(|T| > |t| \mid H_0) = 2\mathbb{P}(T > t|H_0) = 2(1 - \mathbb{P}(T \leq t|H_0)) \approx 0.082$. Wir können die Nullhypothese also nicht auf Niveau $\alpha = 0.05$ verwerfen .

Falls nur mit Tabelle berechnet: Von der Tabelle der t-Verteilung lesen wir ab, dass $t_{18,0.95} < t < t_{18,0.975}$ gilt. Also gilt $\mathbb{P}(T > t_{18,0.95}|H_0) > \mathbb{P}(T > t|H_0) > \mathbb{P}(T > t_{18,0.975}|H_0)$. Daher folgt, dass der p-Wert kleiner ist als $2(1 - 0.95) = 0.10$, aber grösser als $2(1-0.975) = 0.05$. Insbesondere können wir die Nullhypothese also nicht verwerfen .

d) Da wir wissen, dass es sich um normalverteilte Daten handelt, wir aber nicht die wahre Standardabweichung kennen, verwenden wir einen t-Test. Da die Daten nun aber gepaart sind, können wir einen 1-Stichproben t-Test mit den Differenzen durchführen . Da wir nun mehr Information über die Testbedingungen haben (das unterschiedliche "Niveau" aufgrund der Versuchsbedingungen kürzt sich nun weg), sollte es leichter möglich sein, ein signifikantes Testergebnis zu erhalten. Insbesondere erwarten wir, dass der p-Wert kleiner wird .

e) Seien $D_i = X_i - Y_i$. Unter den Annahmen gilt $D_i \sim \mathcal{N}(\mu = \mu_X - \mu_Y, \tilde{\sigma}^2)$, mit einheitlicher und unbekannter Varianz $\tilde{\sigma}^2$. Da wir nur testen wollen, ob ein signifikanter Unterschied besteht (ohne Vorwissen über die Richtung der Abweichung), verwenden wir die Null- und Alternativhypothesen $H_0 : \mu = 0$ und $H_A : \mu \neq 0$. Die Teststatistik unter der Nullhypothese ist

$$T = \frac{\bar{D}_n}{\hat{\sigma}\sqrt{1/n}},$$

welche einer t-Verteilung mit $n - 1 = 9$ Freiheitsgraden folgt .

f) Der realisierte Wert der Teststatistik ist $t = 2.608$, wobei verwendet wurde, dass $n = 10$, $\hat{\sigma} \approx 3.679$. (Der p-Wert ist $\mathbb{P}(|T| > |t| \mid H_0) \approx 0.028$. *Nicht Notwendig!*) Für das Niveau $\alpha = 0.05$ ergibt sich der Verwerfungsbereich $K = (-\infty, -t_{n-1,1-\alpha/2}] \cup [t_{n-1,1-\alpha/2}, \infty)$. Durch Ablesen aus der Tabelle bekommen wir $t_{n-1,1-\alpha/2} = 2.262$. Insbesondere gilt also $t \in K$, also können wir die Nullhypothese auf dem gegebenen Niveau verwerfen .

g) Unter der Alternativhypothese gilt, dass $\mu = 2$. Insbesondere folgt also unter der Alternativhypothese

$$T_A = \frac{\bar{D}_n - 2}{\hat{\sigma}\sqrt{1/n}},$$

einer t-Verteilung mit $n-1=9$ Freiheitsgraden. Es gilt $T_A = T - \frac{2}{\hat{\sigma}\sqrt{1/n}}$, wobei weiterhin T als Teststatistik verwendet wird . Damit können wir die Macht berechnen als

$$
\begin{aligned}
\beta &= \mathbb{P}(T > t_{n-1,1-\alpha/2}|H_A) + \mathbb{P}(T < -t_{n-1,1-\alpha/2}|H_A) \\
&= \mathbb{P}(T_A > t_{n-1,1-\alpha/2} - \frac{2}{\hat{\sigma}\sqrt{1/n}}|H_A) + \mathbb{P}(T_A < -t_{n-1,1-\alpha/2} - \frac{2}{\hat{\sigma}\sqrt{1/n}}|H_A) \\
&= \mathbb{P}(T_A > 0.543|H_A) + \mathbb{P}(T_A < -3.981|H_A) \\
&\approx 0.3 + 0 = 0.3,
\end{aligned}
$$

wobei wir die Werte aus der Tabelle verwendet haben (mit der Tabelle haben wir $\mathbb{P}(T_A < -3.981|H_A) < 0.005$, was wir durch 0 approximieren) .