Stochastik - Lösungen

1. (10 Punkte)

a)	(i)	☒ wahr	☐ nicht wahr	**(0.5 Pkt)**
	(ii)	☒ wahr	☐ nicht wahr	**(0.5 Pkt)**
	(iii)	☐ wahr	☒ nicht wahr	**(0.5 Pkt)**
b)	(i)	☐ wahr	☒ nicht wahr	**(0.5 Pkt)**
	(ii)	☒ wahr	☐ nicht wahr	**(0.5 Pkt)**
	(iii)	☐ wahr	☒ nicht wahr	**(0.5 Pkt)**
c)	(i)	☐ wahr	☒ nicht wahr	**(0.5 Pkt)**
	(ii)	☐ wahr	☒ nicht wahr	**(0.5 Pkt)**
	(iii)	☒ wahr	☐ nicht wahr	**(0.5 Pkt)**
d)	(i)	☒ wahr	☐ nicht wahr	**(0.5 Pkt)**
	(ii)	☐ wahr	☒ nicht wahr	**(0.5 Pkt)**
	(iii)	☒ wahr	☐ nicht wahr	**(0.5 Pkt)**
e)	(i)	☒ wahr	☐ nicht wahr	**(0.5 Pkt)**
	(ii)	☐ wahr	☒ nicht wahr	**(0.5 Pkt)**
f)	(i)	☒ wahr	☐ nicht wahr	**(0.5 Pkt)**
	(ii)	☐ wahr	☒ nicht wahr	**(0.5 Pkt)**
	(iii)	☐ wahr	☒ nicht wahr	**(0.5 Pkt)**
g)	(i)	☒ wahr	☐ nicht wahr	**(0.5 Pkt)**
	(ii)	☐ wahr	☒ nicht wahr	**(0.5 Pkt)**
	(iii)	☒ wahr	☐ nicht wahr	**(0.5 Pkt)**

2. **(10 Punkte)** Nachdem die Coronakrise erfolgreich überstanden ist, möchte eine Gruppe Sozialforscher den Grad der vielbesagten Selbstverantwortung in der Bevölkerung analysieren. Genauer geht es darum den Anteil $0 < p < 1$ der Personen, die Quarantänevorschriften missachtet haben, zu schätzen.

Dafür wird anhand einer Gruppe von Personen, die eine Zeit lang in Quarantäne verbringen mussten, folgendes Experiment durchgeführt:

Sei $x \in \{0, \frac{1}{100}, \frac{2}{100}, \dots, 1\}$ bekannt. Eine Urne enthält hundert Umschläge. $100x$ von ihnen enthalten die Frage: "Haben Sie jemals Ihre Quarantänepflicht verletzt?" Die anderen $100(1-x)$ enthalten die Frage: "Haben Sie Ihre Quarantänepflicht stets eingehalten?"
N Personen werden nun zufällig aus der Gruppe ausgewählt. Nacheinander wird jede von ihnen gebeten, einen Umschlag aus der Urne zu nehmen, seinen Inhalt zu lesen und ihn dann wieder in die Urne zurückzulegen. Sie muss dann die Frage, die sie gelesen hat, laut mit "ja" oder "nein" beantworten. Also kennt nur sie allein die Frage, die in dem Umschlag enthalten war. Deshalb können wir, falls x nicht sehr nahe an 0 oder 1 ist, annehmen, dass jeder nur wahrheitsgemässe Antworten gibt. Es bezeichne Y die Anzahl derjenigen Personen, die mit "ja" geantwortet haben.

a) **(2 Punkte)** Zeige, dass Y binomialverteilt ist mit Parametern N und $p^* := p(2x-1) + 1 - x$.

b) **(1 Punkt)** Zeige, dass

$$\hat{p} = \frac{Y/N - 1 + x}{2x - 1}$$

ein Momentenschätzer für p ist.

Hinweis: Betrachte hierzu die Zufallsvariable Y. Du kannst annehmen, dass $x \neq 1/2$.

c) **(1 Punkt)** Das Experiment wird mit $x = 0.25$ und $N = 100$ Personen durchgeführt und resultiert in 58% "ja"-Antworten. Wie lautet der realisierte Wert des Momentenschätzers?

d) **(1 Punkt)** Berechne den Erwartungswert des Schätzers $\hat{p}$ aus **b)**.

e) **(3 Punkte)** Bestimme die Varianz des Schätzers $\hat{p}$ aus **b)** für $N = 100, p = 0.3$ und sowohl $x = 0.55$ als auch $x = 0.95$. Nenne je einen Vorteil für die Wahl $x = 0.55$ und die Wahl $x = 0.95$.

f) **(2 Punkte)** Wie gross ist approximativ die Wahrscheinlichkeit, dass unter $N = 100$ befragten Personen mindestens 65 mit "ja" antworten, wenn $p = 0.3$ und $x = 0.25$? Benutze die Normalapproximation und die Tabelle der Standardnormalverteilung.

Lösung:

a) Sei X_i die Zufallsvariable, welche angibt, ob die i-te Person mit "ja" ($X_i = 1$) oder "nein" ($X_i = 0$) antwortet. Mit dem Satz von der totalen Wahrscheinlichkeit folgt

$$\begin{aligned} P[X_i = 1] &= P[\text{ja}|\text{Frage Quarantäne verletzt}]P[\text{Frage Quarantäne verletzt}] \\ &\quad + P[\text{ja}|\text{Frage Quarantäne eingehalten}]P[\text{Frage Quarantäne eingehalten}] \\ &= px + (1-p)(1-x) = p(2x-1) + 1 - x =: p^*, \text{ also } X_i \sim \text{ Ber}(p^*). \end{aligned}$$

Dieses Experiment wird N-mal unabhängig, identisch durchgeführt, und $Y = \sum_{i=1}^{N} X_i$. Also ist

$$Y \sim \text{Bin}(N, p^*).$$

b) Das erste theoretische Moment ist $E[Y] = Np^* = N(p(2x-1)+1-x)$. Gleichstellen mit dem empirischen Moment Y (es gibt nur eine Beobachtung von Y) und lösen nach p für $x \neq 1/2$ gibt

$$\hat{p} = \frac{Y/N - 1 + x}{2x-1}.$$

c) Für $x = 0.25$ und $Y/N = 0.58$ ergibt sich somit der realisierte Schätzwert von $\hat{p} = 0.34$.

d) Aus der Linearität des Erwartungswerts erhalten wir

$$E[\hat{p}] = \frac{E[Y]/N - 1 + x}{2x-1} = \frac{p(2x-1)+1-x-1+x}{2x-1} = p$$

(der Schätzer ist erwartungstreu).

e) Aus den Eigenschaften der Varianz ergibt sich (da x deterministisch ist)

$$\mathrm{Var}[\hat{p}] = \frac{\mathrm{Var}[Y]}{N^2(2x-1)^2} = \frac{Np^*(1-p^*)}{N^2(2x-1)^2}.$$

Für $N = 100, p = 0.3$ und $x = 0.55$ ergibt sich $p^* = 0.48$. Da $N^2(2x-1)^2 = 100$ folgt daraus

$$\mathrm{Var}[\hat{p}] = \frac{100 \cdot 0.48(1-0.48)}{100} = 0.48 \cdot 0.52 = 0.2496.$$

Analog erhält man für $N = 100, p = 0.3$ und $x = 0.95$, dass $p^* = 0.32$ und daraus folgt wegen $N^2(2x-1)^2 = 8100$, dass

$$\mathrm{Var}[\hat{p}] = \frac{100 \cdot 0.32(1-0.32)}{8100} = 0.2176\frac{1}{81} \approx 0.0027.$$

Der Vorteil von $x = 0.95$ ist also, dass der Schätzer eine kleinere Varianz aufweist und somit genauer ist. Ein Vorteil für $x = 0.55$ andererseits ist, dass die befragten Personen ehrlicher antworten, weil sie für beide Fragen eine etwa $50-50$-Chance haben.

f) Aus $p = 0.3$ und $x = 0.25$ folgt $p^* = 0.6$. Die Zufallsvariable Y ist approximativ normalverteilt mit Mittelwert

$$Np^* = 60$$

und Varianz

$$Np^*(1-p^*) = 24,$$

also $Y \overset{\text{approx.}}{\sim} \mathcal{N}(60, 24)$. Die gesuchte Wahrscheinlichkeit errechnet sich zu

$$\begin{aligned}
P[Y \geq 65] &= 1 - P[Y < 65] \\
&= 1 - P\left[\frac{Y-60}{\sqrt{24}} < \frac{65-60}{\sqrt{24}}\right] \\
&\approx 1 - \Phi\underbrace{\left(\frac{5}{\sqrt{24}}\right)}_{\approx 1.02} \\
&= 1 - 0.8461 = 0.1539.
\end{aligned}$$

3. (8 Punkte)

Die Anzahl X^{CH} der Bestellungen, die in der Schweiz pro Sekunde bei Amazon aufgegeben werden, ist Poisson-verteilt mit Ratenparameter $\lambda > 0$. Wir möchten diese Rate gerne schätzen.

a) **(2 Punkte)** Gib die Likelihoodfunktion $L(\lambda|x_1, \ldots, x_n)$ für n Beobachtungen $x_1, \ldots, x_n$ an. Bestimme den Maximum-Likelihood Schätzer $\hat{\lambda}^{\text{MLE}}$ als Funktion von x_i, $i = 1, \ldots, n$. Nenne den konkreten Schätzwert wenn innerhalb einer Minute (d.h. $n = 60$) 2100 Bestellungen aufgegeben werden.

Wir nehmen an, dass auch die Anzahl X^{AT} der Bestellungen, die im Nachbarland Österreich pro Sekunde bei Amazon aufgegeben werden Poisson-verteilt ist mit dem selben Ratenparameter $\lambda > 0$, d.h.

$$X^{AT} \sim \text{Poi}(\lambda).$$

Ausserdem nehmen wir an, dass X^{CH} und X^{AT} unabhängig sind.

b) **(1 Punkt)** Was ist die Verteilung der Gesamtzahl der in der Schweiz und Österreich pro Sekunde aufgegebenen Bestellungen? Begründe deine Antwort.

c) **(2 Punkte)** Berechne die Wahrscheinlichkeit, dass in Österreich und der Schweiz genau 70 Bestellungen pro Sekunde aufgegeben werden, gegeben, dass die Anzahl der Bestellungen pro Sekunde in Österreich 36 beträgt. Gib die Antwort als Funktion von λ an.

Wirtschaftsforscher sind skeptisch gegenüber der Unabhängigkeitsannahme. Sie vermuten, dass ein linearer Zusammenhang zwischen der Anzahl der Bestellungen in der Schweiz X^{CH} und jener der Bestellungen in Österreich X^{AT} besteht. Sie berechnen die empirische Korrelation r auf Grund von Paaren $(x_i^{\text{CH}}, x_i^{\text{AT}})$, $i = 1, \ldots, n = 3600$, die in einer Stunde beobachtet wurden, und erhalten den realisierten Wert 0.08 mit einem Standardfehler von 0.015.

d) **(2 Punkte)** Konstruiere ein 90%-Vertrauensintervall für die wahre Korrelation ϱ unter der Annahme, dass r normalverteilt (mit Erwartungswert ϱ) ist.

e) **(1 Punkt)** Benutze die Dualität zwischen Tests und Vertrauensintervallen um zu entscheiden, ob $H_0 : \varrho = 0$ zum Niveau $\alpha = 10\%$ verworfen wird versus $H_A : \varrho \neq 0$. Begründe deine Antwort!

Lösung:

a) Wir wissen, dass $X^{\text{CH}} \sim \text{Poi}(\lambda)$ und bestimmen den Maximum-Likelihood Schätzer $\hat{\lambda}^{MLE}$ in allgemeiner Form. Die Likelihoodfunktion für Stichprobengrösse n und Realisierungen $x_1, \ldots, x_n$ ist gegeben als

$$L(\lambda|x_1, \ldots, x_n) = \prod_i^n \frac{\lambda^{x_i} e^{-\lambda}}{x_i\,!}.$$

Daraus ergibt sich die Log-Likelihoodfunktion

$$\log L(\lambda|x_1, \ldots, x_n) = \sum_{i=1}^n (x_i \log(\lambda) - \lambda - \log(x_i\,!))\,.$$

Ableiten und Nullsetzen liefert

$$\nabla_\lambda \log L(\lambda|x_1, \ldots, x_n) = \frac{\sum_{i=1}^n x_i}{\lambda} - n \overset{!}{=} 0$$
$$\implies \quad \hat{\lambda}^{MLE} = \frac{\sum_{i=1}^n x_i}{n}.$$

Für $n = 60$ und $\sum_{i=1}^{60} x_i = 2100$ erhalten wir also den konkreten Schätzwert $\hat{\lambda}^{MLE} = 2100/60 = 35$.

b) Falls $X_1 \sim$Poi(λ_1), $X_2 \sim$Poi(λ_2) und X_1, X_2 unabhängig sind, dann gilt $X_1+X_2 \sim$Poi$(\lambda_1+\lambda_2)$. Daher ist $X^{CH} + X^{AT} \sim$Poi(2λ).

Alternative Lösung:
Wir bestimmen die Wahrscheinlichkeitsfunktion für die Zufallsvariable $X^{CH} + X^{AT}$. Für $k \in \mathbb{N}$ berechnen wir also

$$\begin{aligned}
P[X^{CH} + X^{AT} = k] &= \sum_{j=1}^{k} P[X^{CH} = j \wedge X^{AT} = k - j] \\
&= \sum_{j=1}^{k} P[X^{CH} = j | X^{AT} = k - j] P[X^{AT} = k - j] \\
&= \sum_{j=1}^{k} P[X^{CH} = j] P[X^{AT} = k - j] \\
&= \sum_{j=1}^{k} \frac{\lambda^j e^{-\lambda}}{j!} \frac{\lambda^{k-j} e^{-\lambda}}{(k-j)!} = e^{-2\lambda} \lambda^k \sum_{j=1}^{k} \binom{k}{j} \frac{1}{k!} \\
&= \frac{e^{-2\lambda} \lambda^k 2^k}{k!} \sum_{j=1}^{k} \binom{k}{j} \frac{1}{2^k} \\
&= \frac{e^{-2\lambda} (2\lambda)^k}{k!} \underbrace{\sum_{j=1}^{k} \binom{k}{j} \left(\frac{1}{2}\right)^{k-j} \left(\frac{1}{2}\right)^{j}}_{=1},
\end{aligned}$$

wobei der letzte Term als Summe über Wahrscheinlichkeiten einer Binomialverteilung mit Parametern k und $1/2$ auf eins summiert. Daher ist $X^{CH} + X^{AT} \sim$Poi(2λ).

c) Gesucht ist die Wahrscheinlichkeit

$$\begin{aligned}
P[X^{CH} + X^{AT} = 70 | X^{AT} = 36] &= P[X^{CH} = 34 | X^{AT} = 36] \\
&\overset{\text{u.a.}}{=} P[X^{CH} = 34] = \frac{\lambda^{34} e^{-\lambda}}{34!}
\end{aligned}$$

Alternative Lösung:

Gesucht ist die Wahrscheinlichkeit

$$
\begin{aligned}
P[X^{CH} + X^{AT} = 70 | X^{AT} = 36] &= \frac{P[X^{CH} + X^{AT} = 70 \wedge X^{AT} = 36]}{P[X^{AT} = 36]} \\
&= \frac{P[X^{CH} = 34 \wedge X^{AT} = 36]}{P[X^{AT} = 36]} \\
&\overset{\text{u.a.}}{=} \frac{P[X^{CH} = 34] P[X^{AT} = 36]}{P[X^{AT} = 36]} \\
&= \frac{\lambda^{34} e^{-\lambda}}{34!}
\end{aligned}
$$

d) Aus der Normalverteilungsannahme folgt

$$\frac{r - \varrho}{\mathrm{sd}(r)} \sim \mathcal{N}(0, 1).$$

Ein 90%-Konfidenzintervall für ϱ ist daher gegeben als

$$r \pm 1.645 \ \mathrm{sd}(r).$$

Dieses Konfidenzintervall errechnet sich wiefolgt. Es gilt

$$
\begin{aligned}
0.9 &= \mathbb{P}[q_{0.05} < \frac{r - \varrho}{\mathrm{sd}(r)} < q_{0.95}] \\
&= \mathbb{P}[r - q_{0.95} \ \mathrm{sd}(r) < \varrho < r - q_{0.05} \ \mathrm{sd}(r)] \\
&= \mathbb{P}[r - q_{0.95} \ \mathrm{sd}(r) < \varrho < r + q_{0.95} \ \mathrm{sd}(r)] \\
&= \mathbb{P}[r - 1.645 \ \mathrm{sd}(r) < \varrho < r + 1.645 \ \mathrm{sd}(r)],
\end{aligned}
$$

und daher erhalten wir mit $r = 0.08$ und Standardfehler 0.015 das Konfidenzintervall

$$0.08 \pm 1.645 \cdot 0.015 \approx (0.055, 0.105).$$

e) Die Nullhypothese wird verworfen, da $\varrho = 0$ nicht im Vertrauensintervall liegt.

4. (8 Punkte) A und B programmieren jeweils einen Generator, der standardnormalverteilte (pseudo) Zufallszahlen liefert. Sie erhalten die folgenden Realisierungen:

	1	2	3	4	5	6	7	8	9	10	11
a_i	-1.27	-0.65	-0.76	1.59	2.05	0.01	0.68	0.18	-0.69	-0.13	0.09
b_i	0.29	0.23	-0.17	0.73	0.86	-0.94	0.28	0.33	-0.14	-1.44	-0.33
$d_i = a_i - b_i$	-1.56	-0.88	-0.59	0.86	1.19	0.95	0.4	-0.15	-0.55	1.31	0.42

mit

$$\bar{a} = 0.1, \quad \bar{b} = -0.027, \quad \bar{d} = 0.127,$$

und

$$s_a = 0.964, \quad s_b = 0.654, \quad s_d = 0.895.$$

B misstraut den Realisierungen des Generators von A. Er vermutet, dass der Mittelwert der von Generator A erzeugten Zahlen vom Mittelwert seiner Zufallszahlen abweicht und möchte dies mit einem Zweistichprobentest belegen.

a) **(1 Punkt)** Wir bezeichnen die Zufallsvariablen generiert von A mit A_i, die Zufallsvariablen generiert von B mit B_i und setzen $D_i := A_i - B_i$. Wir nehmen an, dass

$$A_i, \ldots, A_n \overset{\text{iid}}{\sim} F_A$$
$$B_i, \ldots, B_n \overset{\text{iid}}{\sim} F_B$$
$$D_i, \ldots, D_n \overset{\text{iid}}{\sim} F_D.$$

Wie lauten die Annahmen eines Zweistichproben Z-Tests mit bekannter Varianz $\sigma^2 = 1$ bezüglich F_A und F_B?

b) **(4 Punkte)** Führe einen Zweistichprobentest mit bekannter Varianz $\sigma^2 = 1$ zum Niveau $\alpha = 0.05$ durch:

(i) Formuliere Null- und Alternativhypothese sowie die Teststatistik.

(ii) Berechne den realisierten Wert der Teststatistik.

(iii) Berechne den Verwerfungsbereich für die Teststatistik.

(iv) Formuliere das Testergebnis im Sinne der Zufallsgeneratoren.

c) **(1 Punkt)** Was ist die Verteilung der Teststatistik, wenn der mittlere Wert, den Generator A ausgibt den mittleren Wert des Generators B um 0.1 übersteigt?

d) **(2 Punkte)** Zur Vermeidung von Folgefehlern nehmen wir nun folgendes über den Test in **b)** an: die Verteilung der Teststatistik unter $H_0 : \mu_A - \mu_B = 0$ ist $\mathcal{N}(0, 1)$, die Verteilung der Teststatistik unter $H_A : \mu_A - \mu_B = 0.1$ ist $\mathcal{N}(0.1, 1)$, der Verwerfungsbereich ist $(-\infty, -2.12) \cup (2.12, +\infty)$, der realisierte Wert der Teststatistik ist 0.45 und der p-Wert ist 0.6528. Berechne die Macht des Tests.

Lösung:

a) Wir führen einen Zweistichproben-Z-Test für die unabhängigen Stichproben $A_1, \ldots, A_{11}$ und $B_1, \ldots, B_{11}$ durch. Wir nehmen an, dass

$$A_1, \ldots, A_{11} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_A, 1), \quad B_1, \ldots, B_{11} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_B, 1).$$

Das heisst konkret, dass $F_A = \mathcal{N}(\mu_A, 1)$ und $F_B = \mathcal{N}(\mu_B, 1)$.

b) (i) Um zu testen, ob der Mittelwert μ_B vom Mittelwert μ_A abweicht, betrachten wir die Hypothesen

$$H_0 : \mu_A = \mu_B \qquad H_A : \mu_A \neq \mu_B.$$

Die Teststatistik Z ist gegeben als

$$Z := \frac{\frac{1}{11}\sum_{i=1}^{11} A_i - \frac{1}{11}\sum_{j=1}^{11} B_j}{\sqrt{\frac{2}{11}}}.$$

(ii) Der realisierte Wert der Teststatistik errechnet sich zu

$$z = \frac{0.127}{\sqrt{\frac{2}{11}}} = 0.2978 \approx 0.3.$$

(iii) Aus (i) bestimmen wir den Verwerfungsbereich zum Niveau $\alpha = 0.05$ via

$$\begin{aligned}
0.05 &= P[|Z| > c_\alpha \,|H_0] \\
&= 2P[Z > c_\alpha|H_0] \\
&= 2(1 - P[Z \leq c_\alpha|H_0]) \\
&= 2(1 - \Phi(c_\alpha)) \\
\iff c_\alpha &= \Phi^{-1}(0.975) = 1.96,
\end{aligned}$$

d.h. wir erhalten den Verwerfungsbereich $(-\infty, -1.96) \cup (1.96, +\infty)$.

(iv) Der realisierte Wert der Teststatistik 0.3 liegt nicht im Verwerfungsbereich aus (iii). Daher können wir die Nullhypothese nicht zugunsten der H_A verwerfen. Das bedeutet, dass die Daten nicht genug Evidenz dafür geben um zu sagen, dass der mittlere Wert, den Zufallsgenerator B ausgibt, vom mittleren Wert, den Generator A ausgibt, auf dem Fehlerniveau 1. Art von 0.05 abweicht.

c) Unter der $H_A : \mu_A - \mu_B = 0.1$ gilt

$$\frac{\frac{1}{11}\sum_{i=1}^{11} A_i - \frac{1}{11}\sum_{j=1}^{11} B_j - 0.1}{\sqrt{\frac{2}{11}}} \sim \mathcal{N}(0,1),$$

also

$$Z = \frac{\frac{1}{11}\sum_{i=1}^{11} A_i - \frac{1}{11}\sum_{j=1}^{11} B_j}{\sqrt{\frac{2}{11}}} \sim \mathcal{N}\left(0.1\sqrt{\frac{11}{2}}, 1\right).$$

d) Die Macht des Tests zu dieser $H_A : \mu_A - \mu_B = 0.1$ berechnen wir als

$$\begin{aligned}
P[|Z| > 2.12 \,|H_A] &= 1 - P[-2.12 < Z < 2.12 \,|H_A] \\
&= 1 - P\left[-2.22 < Z - 0.1 < 2.02 \,|H_A\right] \\
&= 1 - \Phi\left(2.02\right) + \Phi\left(-2.22\right) \\
&= 2 - \Phi\left(2.02\right) - \Phi\left(2.22\right) \\
&= 2 - 0.9783 - 0.9868 = 0.0349,
\end{aligned}$$

wobei wir verwendet haben, dass

$$Z - 0.1 \quad \overset{H_A}{\sim} \quad \mathcal{N}(0,1).$$

Die Macht ist also 0.0349.