Prof. M. Maathuis | ETH Zürich | 29. Januar 2021

D–MATL D–MAVT RW

Stochastik – Lösungen

401-0603-00L

1. (9 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)**
	(iii)	☒ 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)**

2. **(7.5 Punkte)** Wir analysieren die Chancen eines Fussballclubs, das Viertelfinale der Champions League 2021 zu erreichen, falls er es in das Achtelfinale geschafft hat. Der Club schafft es in das Viertelfinale, falls er beide Spiele (d.h. Auswärts- und Heimspiel) des Achtelfinals gewinnt. Sollte jede der beiden Mannschaften genau eines der Spiele des Achtelfinals gewinnen, so wird das Weiterkommen per Elfmeterschiessen entschieden. Mit A und H bezeichnen wir das Ereignis, dass der Club das Auswärts- bzw Heimspiel gewinnt. Einem Online-Wettbüro entnehmen wir die Wahrscheinlichkeiten

$$\mathbb{P}(A) = \frac{1}{3}, \qquad \mathbb{P}(H) = \frac{2}{3}.$$

Kommt es zum Elfmeterschiessen, so gewinnt der Club es mit Wahrscheinlichkeit 1/2. Bezeichne mit E das Ereignis, dass der Club das Elfmeterschiessen gewinnt. Ausserdem bezeichnen wir mit V das Ereignis, dass der Club das Viertelfinale erreicht. Wir nehmen an, dass A, H und E unabhängig sind

a) **(1.5 Punkte)** Bestimme die Wahrscheinlichkeit, dass der Club das Viertelfinale *nach dem Elfmeterschiessen* erreicht.

Hinweis: Gesucht ist die Wahrscheinlichkeit, dass es zum Elfmeterschiessen kommt und der Club dieses gewinnt.

b) **(1.5 Punkte)** Bestimme die Wahrscheinlichkeit, dass der Club das Viertelfinale erreicht.

c) **(1.5 Punkte)** Das Auswärtsspiel resultiert in einem Sieg für den Club. Bestimme die Wahrscheinlichkeiten, dass der Club das Viertelfinale erreicht bzw. nicht erreicht, gegeben dieses Ergebnis der Hinrunde, d.h. bestimme $\mathbb{P}(V|A)$ und $\mathbb{P}(V^c|A)$.

Nehmen wir nun an, der Club hätte das Auswärtsspiel verloren. Die Wahrscheinlichkeit, dass er dennoch ins Viertelfinale kommt berechnet sich zu $p = \mathbb{P}(V|A^c) = 1/3$. Wir möchten überprüfen, ob $p = \mathbb{P}(V|A^c) = 1/3$ historisch belegt werden kann und entscheiden uns dazu, p eigenständig aus historischen Daten zu schätzen.

d) **(1.5 Punkte)** In den vergangenen 3 Jahren schafften es $x_1 = 37.72\%, x_2 = 30.79\%$ und $x_3 = 20.98\%$ der Teams, die das Auswärtsspiel im Achtelfinale verloren hatten, in das Viertelfinale. Wir nehmen an, dass dieser Prozentsatz X einer Betaverteilung mit Parametern $\alpha > 0$ und $\beta > 0$ folgt, d.h. $X \sim$Beta(α, β). Erwartungswert und Varianz dieser Verteilung sind gegeben als

$$\mathbb{E}[X] = \frac{\alpha}{\alpha + \beta} \qquad \mathrm{Var}[X] = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}.$$

Des weiteren nehmen wir an, dass $\beta = 2$, und möchten den verbleibenden, unbekannten Parameter α schätzen. Bestimme den Momentenschätzer $\hat{\alpha}^{\mathrm{MOM}}$ für α.

e) **(1.5 Punkte)** Berechne die Varianz der geschätzten Verteilung Beta $(\hat{\alpha}^{\mathrm{MOM}}, \beta)$. Wir entschliessen uns dazu, den konkreten Wert $p = 1/3$ anzunehmen, wenn p innerhalb zwei Standardabweichungen des Mittelwerts der geschätzten Verteilung Beta $(\hat{\alpha}^{\mathrm{MOM}}, \beta)$ liegt. Ist dies der Fall?

Lösung:

a) Sei E das Ereignis, dass der Club das Elfmeterschiessen gewinnt. Gesucht ist die Wahrscheinlichkeit der disjunkten Ereignisse $A^c \cap H \cap E$ und $A \cap H^c \cap E$. Mit Hilfe der

Definition der bedingten Wahrscheinlichkeit ergibt sich

$$\begin{aligned}
&\mathbb{P}(\text{Der Club erreicht das VF nach dem Elfmeterschiessen}) = \\
&= \mathbb{P}\left((A^c \cap H \cap E) \cup (A \cap H^c \cap E)\right) = \\
&= \mathbb{P}(A^c \cap H \cap E) + \mathbb{P}(A \cap H^c \cap E) = \\
&= \mathbb{P}\left(A^c\right) \mathbb{P}\left(H\right) \mathbb{P}\left(E\right) + \mathbb{P}\left(A\right) \mathbb{P}\left(H^c\right) \mathbb{P}\left(E\right) \\
&= \left(\frac{2}{3}\right)^2 \frac{1}{2} + \left(\frac{1}{3}\right)^2 \frac{1}{2} = \frac{5}{18}.
\end{aligned}$$

b) Das Ereignis V kann disjunkt zerlegt werden als

$$V = (A \cap H) \cup E.$$

Somit ergibt sich

$$\mathbb{P}(V) = \mathbb{P}(A \cap H) + \mathbb{P}(E) = \frac{1}{3}\frac{2}{3} + \frac{5}{18} = \frac{1}{2}.$$

c)

$$\mathbb{P}(V|A) = \frac{\mathbb{P}(V \cap A)}{\mathbb{P}(A)} = \frac{\mathbb{P}((A \cap H) \cup (A \cap H^c \cap E))}{\mathbb{P}(A)} = \frac{\frac{2}{9} + \frac{1}{18}}{\frac{1}{3}} = \frac{5}{6},$$

$$\mathbb{P}(V^c|A) = 1 - \mathbb{P}(V|A) = \frac{1}{6}.$$

d) Laut Annahme ist X $\sim$Beta(α, β) =Beta$(\alpha, 2)$. Der entsprechende Erwartungswert ergibt sich aus der gegebenen Formel zu

$$\mathbb{E}[X] = \frac{\alpha}{\alpha + 2}.$$

Gleichsetzen dieses Erwartungswertes mit dem empirischen Mittel (aus drei Beobachtungen x_1, x_2, x_3) $\bar{x} = 0.2983$ ergibt

$$0.2983 = \frac{\alpha}{\alpha + 2} \quad \Longleftrightarrow \quad \alpha(1 - 0.2983) = 2 \cdot 0.2983 \quad \Longrightarrow \quad \hat{\alpha}^{\text{MOM}} = 0.8502.$$

e) Die gesuchte Varianz der Beta $(\hat{\alpha}^{\text{MOM}}, 2)$ Verteilung errechnet sich zu

$$\text{Var}[p] = \frac{2\hat{\alpha}^{\text{MOM}}}{(\hat{\alpha}^{\text{MOM}} + 2)^2(\hat{\alpha}^{\text{MOM}} + 3)} = 0.0544.$$

Da $\bar{x} + 2\sqrt{0.0544} = 0.7648 > 1/3$ (und $\bar{x} - 2\sqrt{0.0544} < \bar{x} < 1/3$) entscheiden wir uns für den konkreten Wert $p = 1/3$.

3. **(10 Punkte)** In einem Teamwettbewerb versuchen Bonnie und Clyde möglichst viele Sekunden auf je einer Slackline zu balancieren. In jeder Sekunde ist die Wahrscheinlichkeit von der Leine zu fallen gegeben als $p \in [0,1]$, sowohl für Bonnie als auch Clyde. Ausserdem sind die Zeiten, die Bonnie und Clyde auf der Slackline balancieren unabhängig voneinander. Bonnie und Clyde scheiden aus dem Bewerb, sobald einer der beiden von der jeweiligen Leine fällt. Sei X die Sekunde, in der Bonnie von der Leine fällt, sei Y die Sekunde, in der Clyde von der Leine fällt und sei Z die Sekunde, in der das Paar aus dem Bewerb ausscheidet.

   a) **(1 Punkte)** Gib die Verteilung von Y inklusive Parameter an.

   b) **(2 Punkte)** Gegeben, dass Clyde mindestens s volle Sekunden auf seiner Slackline balanciert, betrachte die Wahrscheinlichkeit, dass er noch mindestens t weitere volle Sekunden auf der Leine balancieren kann. Zeige, dass diese Wahrscheinlichkeit nicht von s abhängt. Wie lautet der konkrete Wert für $t = 3$ und $p = 0.5$?

   c) **(2 Punkte)** Zeige, dass Z geometrisch verteilt ist, d.h. $Z \sim$Geom(q), und bestimme den Parameter q.

   d) **(2 Punkte)** Nehmen wir nun an, Z ist geometrisch verteilt mit Parameter q. Bestimme die Loglikelihoodfunktion für q basierend auf n i.i.d. Beobachtungen $z_1, \ldots, z_n$ und leite damit den Maximum-Likelihood Schätzer $\hat{q}^{MLE}$ in allgemeiner Form her.

   Bemerkung: Dieser Punkt lässt sich auch beantworten, wenn man Punkt c) nicht gelöst hat.

   e) **(1 Punkte)** Bevor sie an dem Wettbewerb teilnahmen haben sich Bonnie und Clyde 10 Mal getroffen und das Balancieren ausprobiert. Dabei haben sie die folgenden Zeiten aufgezeichnet

x	26	9	2	11	21	9	6	1	3	15
y	4	12	20	9	40	2	28	3	5	16

   Berechne den konkreten Wert des Schätzers aus d) für diese Beobachtungen.

   f) **(2 Punkte)** Clyde behauptet, er könne im Mittel länger auf einer Slackline balancieren als Bonnie. Führe an Hand der Stichproben aus Punkt e) einen Wilcoxon Test zum 5%-Niveau durch um seine Behauptung zu überprüfen. Gib geeignete Null- und Alternativhypothesen an. Welchen Wert nimmt die realisierte Teststatistik an? Wie lautet der Testentscheid?

   Hinweis: Wenn mehrere Werte den gleichen Rang bekommen, wird jedem dieser Werte der Mittelwert der entsprechenden Ränge zugeordnet.

Lösung:

a) Y ist geometrisch verteilt mit Parameter $p \in (0,1)$, d.h. $Y \sim$ Geom(p).

b) Gesucht ist die Wahrscheinlichkeit $P[Y > s + t | Y > s]$. Wegen $Y \sim$ Geom(p) gilt

$$\begin{aligned} P[Y > s+t | Y > s] &= \frac{P[Y > s+t, Y > s]}{P[Y > s]} = \frac{P[Y > s+t]}{P[Y > s]} \\ &= \frac{1 - F_Y(s+t)}{1 - F_Y(s)} = \frac{(1-p)^{s+t}}{(1-p)^s} = (1-p)^t. \end{aligned}$$

Alternativ folgt dies auch aus der Gedächtnislosigkeit der geometrischen Verteilung,

$$P[Y > s+t | Y > s] = P[Y > t] = 1 - F_Y(t) = (1-p)^t.$$

Für $t = 3$ und $p = 0.5$ folgt somit $P[Y > s+3 | Y > s] = (0.5)^3 = 0.125$.

c) Nach Voraussetzung sind X, Y i.i.d. und Geom(p)-verteilt. Z beschreibt die Sekunde, in der die beiden aus dem Bewerb scheiden; also ist $Z = \min\{X, Y\}$. Die Verteilungsfunktion von Z ist gegeben durch

$$\begin{aligned} F_Z(t) &= P[Z \le t] = P[\min\{X,Y\} \le t] \\ &= 1 - P[\min\{X,Y\} > t] = 1 - P[X > t, Y > t] \\ &= 1 - P[X > t]P[Y > t] \\ &= 1 - ((1-p)^t)^2 \\ &= 1 - (1 - p(2-p))^t. \end{aligned}$$

d.h. Z ist wieder geometrisch verteilt mit Parameter $\tilde{p} = p(2-p)$.

d) Wir nehmen an, dass $Z \sim$Geom(q) und bestimmen den Maximum-Likelihood Schätzer $\hat{q}^{MLE}$ in allgemeiner Form. Die Likelihoodfunktion für eine allgemeine Stichprobengrösse n und Realisierungen $z_1, \dots, z_n$ ist gegeben als

$$\ell(q|z_1, \dots, z_n) = \prod_{i=1}^{n} q(1-q)^{z_i - 1}.$$

Daraus ergibt sich die Log-Likelihoodfunktion

$$\log \ell(q|z_1, \dots, z_n) = \sum_{i=1}^{n} \log(q(1-q)^{z_i-1}) = n\log(q) + \log(1-q)\sum_{i=1}^{n}(z_i - 1).$$

Ableiten und Nullsetzen liefert

$$\begin{aligned} \nabla_q \log \ell(q|z_1, \dots, z_n) &= \frac{n}{q} - \frac{\sum_{i=1}^{n}(z_i - 1)}{1-q} \overset{!}{=} 0 \\ \iff \quad & n(1-q) = q\sum_{i=1}^{n}(z_i - 1) \\ \iff \quad & n = q\left(n + \sum_{i=1}^{n}(z_i) - n\right) \\ \implies \quad & \hat{q}^{MLE} = \frac{n}{\sum_{i=1}^{n} z_i}. \end{aligned}$$

e) Die realisierten minimalen Zeiten lauten

z	4	9	2	9	21	2	6	1	3	15

Daraus ergibt sich der realisierte Schätzwert $\hat{q}^{MLE} \approx 0.14$.

f) Wir nehmen an, die Differenzen $D_1, \dots, D_{10}$ mit $D_i = X_i - Y_i$ sind i.i.d. mit Median μ. Wir testen die Hypothese

$$H_0 : \mu = 0 \quad \text{gegen} \quad H_A : \mu < 0.$$

Die realisierten Differenzen, sowie deren absolute Ränge ergeben sich als

d	22	-3	-18	2	-19	7	-22	-2	-2	-1
Rang$\|d\|$	9.5	5	7	3	8	6	-9.5	3	3	1

Daraus berechnen wir den realisierten Wert w der Teststatistik

$$W = \sum_{i=1}^{10} \text{Rang}|D_i| 1_{D_i > 0}$$

als $w = 3 + 6 + 9.5 = 18.5$. Dieser liegt nicht im einseitigen Verwerfungsbereich zum 5%-Niveau $K = \{W \leq 10\}$, weshalb wir die Nullhypothese nicht zu Gunsten der Alternativhypothese verwerfen können.

4. (9.5 Punkte) Der Bundesrat möchte den Anteil p der Corona-positiven Personen in der Bevölkerung schätzen. Dafür wird ein Corona-Massentest mit n zufällig ausgewählten Personen durchgeführt. Der Test, der dabei verwendet wird, liefert wahrheitsgetreue Ergebnisse.

a) **(1 Punkte)** Sei X_n die Anzahl derjenigen Personen, deren Testergebnis Corona-positiv ist. Bestimme die Verteilung von X_n inklusive Parameter.

b) **(3 Punkte)** Innerhalb der Personengruppe der Grösse $n = 500$ wurden 15 Personen positiv getestet. Betrachte $\frac{1}{n}X_n$ als Schätzer für p und bestimme den geschätzten Standardfehler. Bestimme ein 99%-Vertrauensintervall für p mittels Normalapproximation.

Neue ökonomische und sozialpolitische Massnahmen werden implementiert, von denen man sich eine Reduktion von p auf unter 0.02 erhofft. 14 Tage nach Implementierung dieser Massnahmen soll daher versucht werden, statistisch zu belegen, dass p auf unter 0.02 gefallen ist.

c) **(3 Punkte)** Betrachte die Nullhypothese $H_0 : p = 0.02$. Formuliere eine geeignete Alternativhypothese um diese Reduktion zu belegen. Bestimme mittels Normalapproximation und Transformation zur Standardnormalverteilung den zugehörigen approximativen Verwerfungsbereich für X_n auf dem Signifikanzniveau 0.05.

d) **(1 Punkte)** Nach Ablauf dieser 14 Tage nach Implementierung der Massnahmen werden nochmals 500 neu ausgewählte Testpersonen auf das Coronavirus getestet. Unter diesen Coronatests fallen 4 positiv aus. Wie lautet das Resultat des statistischen Tests aus Aufgabenteil c)?

e) **(1.5 Punkte)** Bestimme mittels Normalapproximation und Transformation zur Standardnormalverteilung die Macht des Tests für $p = 0.01$.

Lösung:

a) X_n ist binomialverteilt mit Parametern n und p.

b) Die Varianz des Schätzers $\hat{p} = \frac{1}{n}X_n$ errechnet sich zu

$$\mathbb{V}(\hat{p}) = \frac{1}{n^2}\mathbb{V}(X_n) = \frac{1}{n^2}np(1-p) = \frac{p(1-p)}{n}.$$

Der geschätzte Standardfehler ist somit $\sqrt{\frac{p(1-p)}{n}}$.

Laut Normalapproximation ist $X_n \overset{appr.}{\sim} \mathcal{N}(np, np(1-p))$, also

$$\hat{p} \overset{appr.}{\sim} \mathcal{N}\left(p, \frac{1}{n}p(1-p)\right)$$

und damit ein approximatives Vertrauensintervall gegeben durch

$$\hat{p} \pm \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\sqrt{\hat{p}\left(1-\hat{p}\right)\frac{1}{n}},$$

wobei $\alpha = 0.01$. Weiter ergibt sich das realisierte Intervall mit $\hat{p} = \frac{1}{n}X_n, X_n = 15$ und $n = 500$ durch

$$0.03 \pm \Phi^{-1}(0.995)\sqrt{0.03 \times 0.97 \times \frac{1}{500}}.$$

Mit dem Wert $\Phi^{-1}(0.995) = 2.575$ erhält man

$$
\begin{aligned}
0.03 \pm 2.575\sqrt{0.03 \times 0.97 \times \frac{1}{500}} &= 0.03 \pm 2.575 \cdot 0.0076 \\
&= [0.0104, 0.0496]
\end{aligned}
$$

als realisiertes approximatives 99%-Vertrauensintervall für p.
Alternativ lässt sich ein approximatives Vertrauensintervall mit dem $(1-\alpha/2)$-Quantil $t_{n,1-\alpha/2}$ der t-Verteilung mit n Freiheitsgraden bestimmen als

$$\hat{p} \pm t_{n,1-\alpha/2}\sqrt{\hat{p}\left(1-\hat{p}\right)\frac{1}{n}}.$$

Analog erhält man mit dem Wert $t_{500,0.995} = 2.576$ das selbe realisierte approximative 99%-Vertrauensintervall $[0.0104, 0.0496]$ für p.

c) Zur Belegung der Aussage wählt man als Nullhypothese

$$H_0 : p = 0.02$$

und als Alternativhypothese die zu belegende Aussage, also

$$H_A : p < 0.02.$$

Der Verwerfungsbereich ist $\{x : x \leq c\}$, wobei $c \in \mathbb{N}_0$ so gross wie möglich mit $P_{0.02}[X_{500} \leq c] \leq 0.05$ ist. Zur Berechnung von $P_{0.02}[X_{500} \leq c]$ verwendet man die Normalapproximation.

$$
\begin{aligned}
P_{0.02}[X_{500} \leq c] &= P_{0.02}\left[\frac{X_{500} - 10}{\sqrt{0.02 \times 0.98 \times 500}} \leq \frac{c - 10}{\sqrt{0.02 \times 0.98 \times 500}}\right] \\
&\approx \Phi\left(\frac{c - 10}{\sqrt{0.02 \times 0.98 \times 500}}\right).
\end{aligned}
$$

Somit benötigen wir

$$\Phi\left(\frac{c - 10}{\sqrt{0.02 \times 0.98 \times 500}}\right) \leq 0.05,$$

beziehungsweise

$$1 - \Phi\left(\frac{c - 10}{\sqrt{0.02 \times 0.98 \times 500}}\right) = \Phi\left(\frac{10 - c}{\sqrt{0.02 \times 0.98 \times 500}}\right) \geq 0.95,$$

oder

$$c \leq 10 - \Phi^{-1}(0.95)\sqrt{0.02 \times 0.98 \times 500} \overset{\Phi^{-1}(0.95)\approx 1.645}{\approx} 10 - 1.645 \cdot 3.1305 \approx 4.85$$

Somit ist der approximative Verwerfungsbereich $\{x : x \leq 4\}$.

d) Da der realisierte Wert $x_{500} = 4$ im Verwerfungsbereich liegt, können wir belegen, dass die Aussage bzw. die gewählte Alternativhypothese am 5%-Niveau signifikant ist.

e) Die Macht ist gegeben als $P_{0.01}[X_{500} \leq 4]$. Durch Normalapproximation erhält man

$$
\begin{aligned}
P_{0.01}[X_{500} \leq 4] &= P_{0.01}\left[\frac{X_{500} - 5}{\sqrt{0.01 \times 0.99 \times 500}} \leq \frac{-1}{\sqrt{0.01 \times 0.99 \times 500}}\right] \\
&\approx \Phi\left(\frac{-1}{\sqrt{0.99 \times 0.01 \times 500}}\right)
\end{aligned}
$$

Mit $\frac{-1}{\sqrt{0.99 \times 0.01 \times 500}} \approx -0.4495$ und $\Phi(-0.4495) \approx 1 - \Phi(0.45) = 1 - 0.6736 = 0.3264$ ergibt sich die Macht des Tests daher als etwa 32.64%.