ETH Zürich HS 2021
Neur. Syst. Comp., D-MATL, D-MAVT
Patrick Cheridito

Koordinatorin
Hanna Wutte

Stochastik

Musterlösung Ferienserie

1. Bei den folgenden 10 Fragen ist jeweils genau eine Antwort richtig.

a) Eine Münze wird zweimal geworfen. Für $i = 1, 2$ sei K_i das Ereignis, dass beim i-ten Wurf Kopf erscheint. Dann gilt:

1. K_1 und K_2 sind unabhängig.
2. K_1 und K_2 sind disjunkt.
3. K_1 und K_2 sind unabhängig und diskunkt.

b) Seien A und B zwei Ereignisse mit $\mathbb{P}(A) > 0$ und $A \cap B = \emptyset$. Die Aussage $\mathbb{P}(B \mid A) = 0$ ist

1. wahr
2. nicht wahr.

c) Seien A und B zwei Ereignisse. Die Aussage $\mathbb{P}(A^c \cap B^c) = 1 - \mathbb{P}(A \cap B)$ ist

1. wahr
2. nicht wahr.

d) Seien A und B zwei Ereignisse. Das Additionsgesetz $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$ gilt falls:

1. A und B unabhängig sind.
2. A und B disjunkt sind.
3. A eine Teilmenge von B ist.

e) Seien A und B zwei Ereignisse mit $\mathbb{P}(A \cap B) = 1/4$ und $\mathbb{P}(B) = 1/2$. Dann kann gelten:

1. $\mathbb{P}(A) = 0.2$.
2. $\mathbb{P}(A) = 0.5$.
3. $\mathbb{P}(A) = 0.8$.

f) Seien A und B zwei Ereignisse mit $0 < \mathbb{P}(B) < 1$. Dann gilt:

1. $\mathbb{P}(A \mid B^c) = 1 - \mathbb{P}(A \mid B)$.
2. $\mathbb{P}(A \mid B^c) = 1 - \mathbb{P}(A^c \mid B^c)$.
3. $\mathbb{P}(A \mid B^c) = \mathbb{P}(A \cap B^c)/\mathbb{P}(B)$.

g) Wir untersuchen eine Bodenprobe auf zwei Schadstoffe X und Y. Wir definieren die Ereignisse A: "Bodenprobe enthält X" und B: "Bodenprobe enthält Y". Es sei $\mathbb{P}(A) = 0.1$ und $\mathbb{P}(B) = 0.3$. Zusätzlich wissen wir, dass die Wahrscheinlichkeit, dass die Bodenprobe keinen der beiden Schadstoffe enthält, 0.7 beträgt. Dann gilt:

1. $\mathbb{P}(A \cap B) = 0.03$.

2. $\mathbb{P}(A \cap B) = 0.1$.
3. Es gibt zu wenig Information, um $\mathbb{P}(A \cap B)$ zu berechnen.

h) Wir betrachten einen Labortest für eine Krankheit. Sei A das Ereignis, dass die getestete Person die Krankheit hat. Sei B das Ereignis, dass der Test positiv ist. Man weiss $\mathbb{P}(B \mid A) = 0.99$ und $\mathbb{P}(B^c \mid A^c) = 0.995$. Der Anteil der Bevölkerung, der die Krankheit hat ist 0.1%. Wie gross ist die Wahrscheinlichkeit, dass eine zufällig gewählte Person mit einem positiven Test tatsächlich die Krankheit hat?

1. $\mathbb{P}(A \mid B) \approx 0.065$.
2. $\mathbb{P}(A \mid B) \approx 0.165$.
3. $\mathbb{P}(A \mid B) \approx 0.560$.

i) Es sei X eine Zufallsvariable mit $\mathbb{E}[X] = 2$, $\mathrm{Var}(X) = 3$. Dann gilt:

1. $\mathbb{E}[X^2] = 4$.
2. $\mathbb{E}[X^2] = 7$.
3. Es gibt zu wenig Information um $\mathbb{E}[X^2]$ zu berechnen.

j) Wie oft muss man einen fairen Würfel im Schnitt werfen bis man zum ersten Mal eine 3 erhält?

1. 6 mal.
2. 6^6 mal.
3. $6!$ mal.

Lösung:

a) 1.

b) 1. $\mathbb{P}(B \mid A) = \mathbb{P}(A \cap B)/\mathbb{P}(A) = 0/\mathbb{P}(A) = 0$.

c) 2. $\mathbb{P}(A^c \cap B^c) = \mathbb{P}((A \cup B)^c) = 1 - \mathbb{P}(A \cup B)$.

d) 2.

e) 2. $\mathbb{P}(A) = 0.2$ stimmt nicht, weil $1/4 = \mathbb{P}(A \cap B) \leq \mathbb{P}(A)$. $\mathbb{P}(A) = 0.8$ stimmt auch nicht, weil das $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B) = 0.8 + 0.5 - 0.25 = 1.05 > 1$ ergibt. Also bleibt $\mathbb{P}(A) = 0.5$ übrig.

f) 2. Siehe die Formeln für bedingte Wahrscheinlichkeiten.

g) 2. Es ist gegeben, dass $\mathbb{P}(A^c \cap B^c) = 0.7$. Also $\mathbb{P}(A \cup B) = 1 - \mathbb{P}((A \cup B)^c) = 1 - \mathbb{P}(A^c \cap B^c) = 1 - 0.7 = 0.3$. Weiterhin gilt: $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$. Also $0.3 = 0.1 + 0.3 - \mathbb{P}(A \cap B)$. Also $\mathbb{P}(A \cap B) = 0.1$.

h) 2. Benutze den Satz von Bayes:

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(B \mid A)\mathbb{P}(A)}{\mathbb{P}(B \mid A)\mathbb{P}(A) + \mathbb{P}(B \mid A^c)\mathbb{P}(A^c)} = \frac{0.99 \cdot 0.001}{0.99 \cdot 0.001 + (1 - 0.995) \cdot 0.999} \approx 0.165$$

i) 2. $\mathrm{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$. Also $\mathbb{E}[X^2] = \mathrm{Var}(X) + (\mathbb{E}[X])^2 = 3 + 2^2 = 7$.

j) 1. Sei X die Anzahl Würfe bis zum ersten Mal eine 3 erscheint. Dann ist $X \sim \text{geometrisch}(1/6)$. Also $\mathbb{E}[X] = 1/(1/6) = 6$.

2. Bei den folgenden 10 Fragen ist jeweils genau eine Antwort richtig.

a) Wir erhalten eine grosse Lieferung Materialien. Aus Erfahrung wissen wir, dass im Schnitt 5% der Materialien mangelhaft sind. Nehmen wir an, dass die Materialien unabhängig voneinander sind. Dann gilt:

1. Die Wahrscheinlichkeit, dass genau zwei Materialien mangelhaft sind, ist:
$$\binom{10}{2} 0.05^8 (1-0.05)^2.$$
2. Die Wahrscheinlichkeit, dass genau zwei Materialien mangelhaft sind, ist:
$$\binom{10}{2} 0.05^2 (1-0.05)^8.$$
3. Die Wahrscheinlichkeit, dass höchstens zwei Materialien mangelhaft sind, ist:
$$\binom{10}{2} 0.05^2 (1-0.05)^8.$$
4. Wenn wir 20 Materialien anschauen, gibt es darunter sicher ein Mangelhaftes.

b) Seien X die Anzahl Einsen, die wir in 10 unabhängigen Würfen mit einem fairen Würfel werfen. Welche Verteilung kommt für X in Frage?

1. Bernoulli verteilt mit Parameter $p = 1/6$.
2. Geometrisch verteilt mit Parameter $p = 1/6$
3. Binomialverteilt mit Parameter $n = 10$ und $p = 1/6$.
4. Poisson verteilt mit Parameter $\lambda = 1/6$.

c) Sei $X \sim \text{Poisson}(\lambda)$ mit $\lambda > 0$. Dann gilt:

1. $\mathbb{P}(X > 5) = 1 - \mathbb{P}(X < 5)$.
2. $\mathbb{P}(X \geq 1 \mid X \leq 1) = \lambda/(\lambda + 1)$.
3. $2X \sim \text{Poisson}(2\lambda)$.

d) Betrachte eine normalverteilte Zufallsvariable $X \sim N(\mu, \sigma^2)$ mit $\mu = 1$ und $\sigma^2 = 3$. Dann gilt:

1. $\mathbb{P}(X \leq 0) < \mathbb{P}(X \geq 3)$.
2. Die Fläche unter der Dichte im Intervall $[1, 1 + \sqrt{3}]$ ist etwa 66%.
3. Die Fläche unter der Dichte im Intervall $[1 - 3, 1 + 3]$ ist etwa 66%.
4. Die Fläche unter der Dichte im Intervall $[1 - 2\sqrt{3}, 1 + 2\sqrt{3}]$ ist etwa 95%.

e) Sei X eine positive stetige Zufallsvariable. Das 95% Quantil von X beträgt 0.4. Dann gilt:

1. Der Median von X ist grösser als 0.4.
2. $\mathbb{P}(X \leq 0.95) = 0.4$.
3. Das 0.95% Quantil von $Y = 1 + 2X^2$ ist 1.32.
4. Die Fläche unter der Dichte im Intervall $[1 - 2\sqrt{3}, 1 + 2\sqrt{3}]$ ist etwa 95%.

f) Die stetige Zufallsvariable X hat die kumulative Verteilungsfunktion

$$F(x) = \frac{1}{1+\exp(-x)}, x \in \mathbb{R}.$$

Dann gilt:

1. Es ist $\mathbb{P}(X = 0) = 0.5$.
2. Für $x \geq 0$ gilt $\mathbb{P}(X > x) = \frac{\exp(-x)}{1+\exp(-x)}$.
3. Die Dichte von X ist $\frac{-\exp(-x)}{(1+\exp(-x))^2}$, $x \in \mathbb{R}$.

g) Es seien X und Y Zufallsvariablen mit $\mathrm{Var}(X) = 3$ und $\mathrm{Var}(Y) = 1$. Dann gilt:

1. $\mathrm{Var}(X - Y) = 2$.
2. $\mathrm{Var}(X - Y) = 4$.
3. Es gibt zu wenig Information um $\mathrm{Var}(X - Y)$ zu berechnen.

h) Es seien X und Y unabhängige und identisch verteilte Zufallsvariablen mit Varianz $\sigma^2 > 0$. Dann gilt:

1. $\mathrm{Var}(X + X - Y) = 3\sigma^2$.
2. $\mathrm{Var}(X + X - Y) = 5\sigma^2$.
3. $\mathrm{Var}(X + Y) = \mathrm{Var}(2X)$.

i) Seien X und Y zwei diskrete unkorrelierte Zufallsvariablen mit Wertebereichen W_X und W_Y. Dann gilt:

1. X und Y sind unabhängig.
2. $\mathbb{P}(X = x \mid Y = y) = \mathbb{P}(X = x)$ für alle $x \in W_X$ und $y \in W_Y$.
3. $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$.

j) Es seien $X \sim \mathrm{Poisson}(10)$ und $Y = X + 5$. Dann gilt:

1. $\mathrm{Corr}(X, Y) = 1$.
2. $Y \sim \mathrm{Poisson}(15)$.
3. $\mathrm{Cov}(X, Y) = 15$.

Lösung:

a) 2. Die zweite Binomialformel ist richtig. Die letzte Antwort ist falsch: $\mathbb{P}(\text{kein Material ist mangelhaft}) = \binom{n}{0}0.05^0 0.95^{20} = 0.95^{20} \approx 0.35$.

b) 3.

c) 2. Die erste Antwort ist falsch: $\mathbb{P}(X > 5) = 1 - \mathbb{P}(X \leq 5)$. Weil X diskret ist, ist das nicht gleich wie $1-\mathbb{P}(X < 5)$. Die zweite Antwort ist richtig: $\mathbb{P}(X \geq 1 \mid X \leq 1) = \mathbb{P}(X \geq 1 \,\text{und}\, X \leq 1)/\mathbb{P}(X \leq 1) = \mathbb{P}(X = 1)/\mathbb{P}(X \leq 1) = \frac{\lambda \exp(-\lambda)}{\exp(-\lambda)+\lambda \exp(-\lambda)} = \frac{\lambda}{\lambda+1}$. Die letzte Antwort ist falsch: Der Wertebereich von $2X$ ist $\{0, 2, 4, 6, \ldots\}$. Das zeigt, dass $2X$ nicht Poisson verteilt ist.

d) 4. Siehe Abbildung 2.9 im Skript. AAA

e) 3. Die erste Antwort ist falsch, weil der Median das 50% Quantil ist. Es kann also nicht grösser sein als das 95% Quantil. Die zweite Antwort ist falsch, weil $\mathbb{P}(X \leq 0.40) = 0.95$ gilt. Die Dritte Antwort ist richtig: weil $1+2X^2$ eine monotone Transformation ist, werden die Quantile genauso transformiert. Das 95% Quantil von Y ist damit $1 + 2 \cdot 0.40^2 = 1.32$.

f) 2. Die erste Antwort ist falsch: $F(0) = 0.5 = \mathbb{P}(X \leq 0)$ und $\mathbb{P}(X = 0) = 0$ weil X stetig ist. Die zweite Antwort ist richtig: Für $x \geq 0$ gilt $\mathbb{P}(X > x) = 1 - \mathbb{P}(X \leq x) = 1 - \frac{1}{1+\exp(-x)} = \frac{\exp -x}{1+\exp -x}$. Die dritte Antwort ist falsch: Die Dichte von X ist nicht negativ.

g) 3. $\text{Var}(X-Y) = \text{Var}(X+(-Y)) = \text{Var}(X)+\text{Var}(-Y)+2\text{Cov}(X,-Y) = \text{Var}(X)+(-1)^2\text{Var}(Y)-2\text{Cov}(X,Y) = 3+1-2\text{Cov}(X,Y)$. Aber $\text{Cov}(X,Y)$ ist nicht gegeben.

h) 2. $\text{Var}(X + X + Y) = \text{Var}(2X + Y) = \text{Var}(2X) + \text{Y}$, wobei die letzte Gleichheit wegen Unabhängigkeit gilt. Mit $\text{Var}(2X) = 2^2\text{Var}(X) = 4\sigma^2$ ergibt das $\text{Var}(2X + Y) = 5\sigma^2$. Die letzte Antwort ist falsch, weil $\text{Var}(X + Y) = 2\sigma^2$ und $\text{Var}(2X) = 4\sigma^2$.

i) 3. Unabhängigkeit impliziert, dass die Korrelation null ist, aber im Allgemeinen gilt die Umkehrung nicht. Darum ist die erste Antwort falsch. Die zweite Antwort ist die Definition von Unabhängigkeit und stimmt daher auch nicht. Die dritte Antwort ist richtig, weil $\text{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$. Aus $\text{Cov}(X,Y) = 0$ folgt also $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$.

j) 1. Die erste Antwort ist richtig, weil (X,Y) auf einer geraden Linie liegen. Die Zufallsvariable Y hat Wertebereich $\{5,6,7,\ldots\}$. Sie kann also nicht Poisson verteilt sein. (Alternativ: $\text{Var}(Y) = \text{Var}(X+5) = \text{Var}(X) = 10$. Also kann Y nicht Poisson(15) verteilt sein.) Die dritte Antwort ist falsch, weil $\text{Cov}(X,Y) = \text{Cov}(X,X+5) = \text{Cov}(X,X) = \text{Var}(X) = 10$.

3. Bei den folgenden 8 Fragen ist jeweils genau eine Antwort richtig.

a) Sei (X,Y) eine stetige Zufallsvariable mit einer uniformen Verteilung auf dem Viereck mit Ecken $(1,0)$, $(0,1)$, $(-1,0)$ und $(0,-1)$. Dann gilt:

1. Die Randverteilung von X ist Uniform$(-1,1)$
2. Für (x,y) in dem Viereck gilt $f(x,y) = 1/2$
3. X und Y sind unabhängig

b) Wir betrachten die gemeinsame Verteilung von zwei Zufallsvariablen X und Y. Dann gilt:

1. Aus der gemeinsamen Verteilung kann man immer die Randverteilungen berechnen.
2. Aus den Randverteilungen kann man immer die gemeinsame Verteilung berechnen.
3. Aus den Randverteilungen und $\text{Cov}(X,Y)$ kann man immer die gemeinsame Verteilung berechnen.
4. Es gilt immmer, dass $\text{Cov}(X,Y) \geq \text{Corr}(X,Y)$.

c) Wir betrachten die empirische Korrelation von (x_i, y_i), $i = 1, \ldots, 20$. Dann gilt:

1. Die empirische Korrelation liegt im Interval $[0,1]$.
2. Wenn wir eine empirische Korrelation von 1 beobachten, dann liegen die Daten auf einer Linie mit positiver Steigung.
3. Wenn wir eine empirische Korrelation von 0 beobachten, dann gibt es keinen Zusammenhang zwischen x_i und y_i in den Daten.

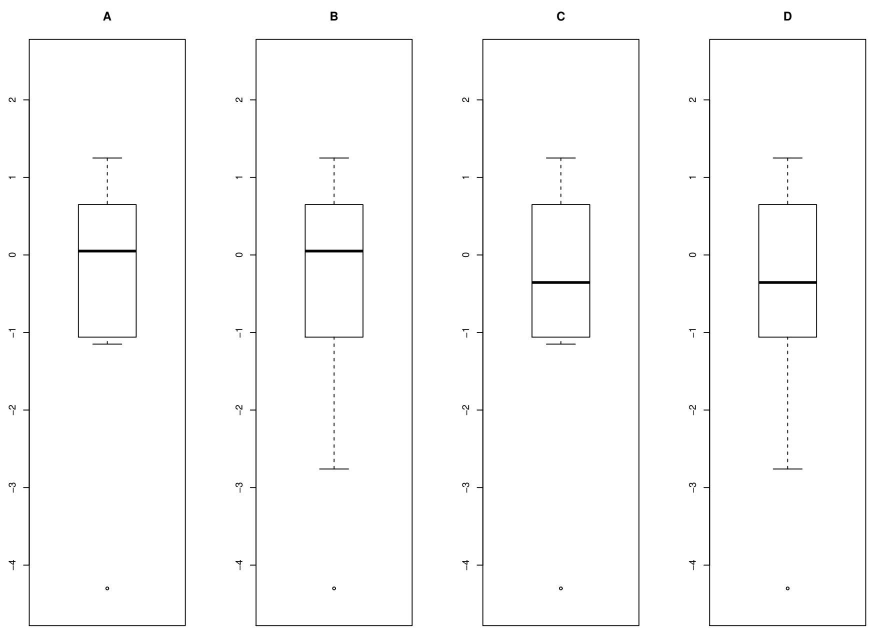
d) Wir haben eine Stichprobe $x_1, \ldots, x_n$ bestehend aus reellen Zahlen, wobei n eine ungerade Zahl ist. Welche Kennzahlen kann man direkt aus der empirischen Verteilungsfunktion dieser Stichprobe bestimmen?

1. Sowohl den empirischen Mittelwert als auch den Median.
2. Den empirischen Mittelwert , aber nicht den Median.
3. Den Median , aber nicht den empirischen Mittelwert.

e) Wir betrachten folgende sortierten Daten:

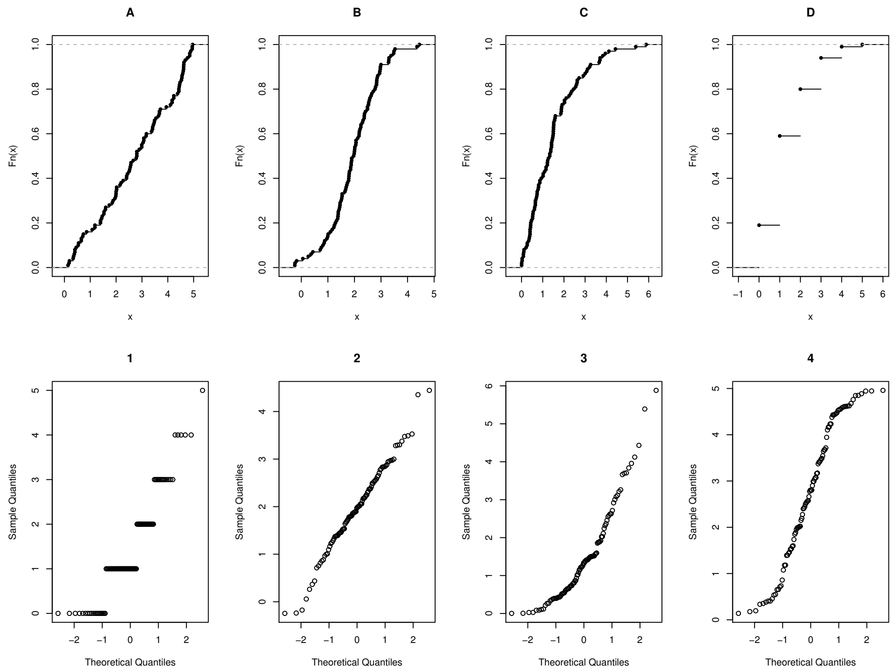
$$-4.30, -1.15, -1.06, -0.48, -0.41, -0.30, 0.50, 0.65, 0.95, 1.25$$

Welcher Boxplot gehört dazu?



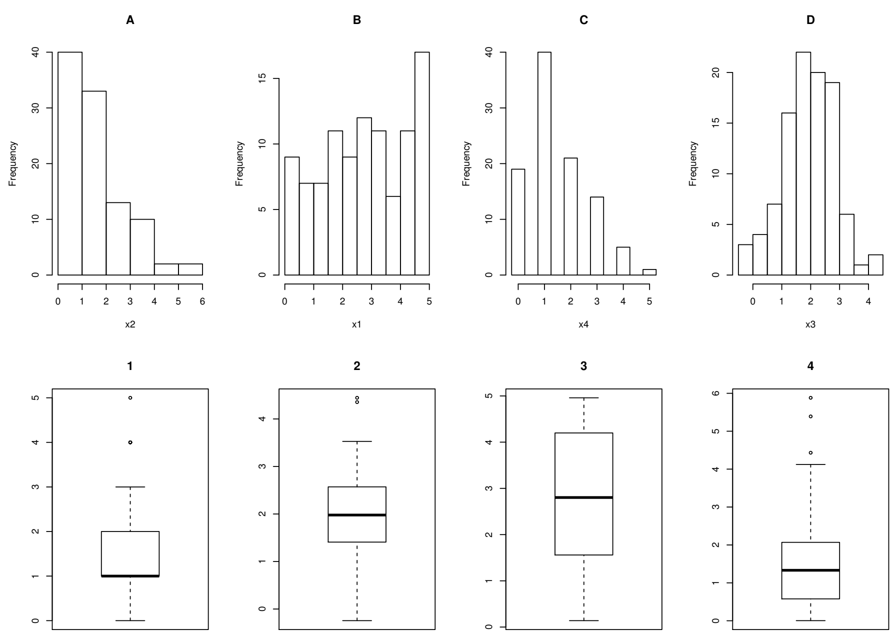
1. A
2. B
3. C
4. D

f) Was ist die richtige Zuordnung der empirischen kumulativen Verteilungsfunktionen und der QQ Plots (die Standardnormalverteilung wurde als Referenzverteilung genommen)?



1. A2, B4, C1, D3
2. A2, B4, C3, D1
3. A4, B2, C1, D3
4. A4, B2, C3, D1

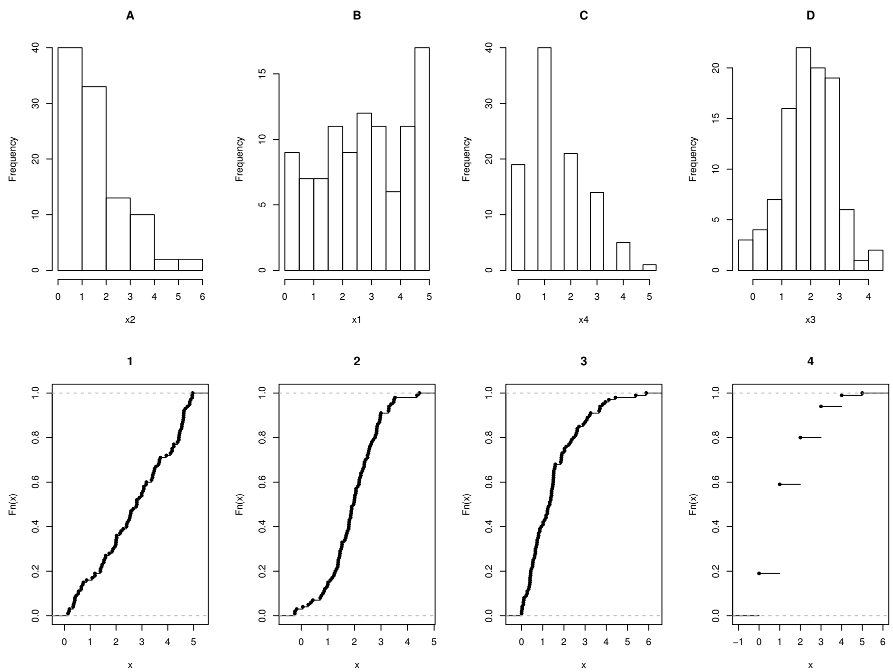
g) Was ist die richtige Zuordnung der Histogramme und Boxplots?



1. A1, B2, C4, D3
2. A1, B3, C4, D2
3. A4, B2, C1, D3

4. A4, B3, C1, D2

h) Was ist die richtige Zuordnung der Histogramme und empirische kumulatieve Verteilungsfunktionen?



1. A3, B1, C4, D2
2. A3, B2, C4, D1
3. A4, B1, C3, D2
4. A4, B2, C3, D1

Lösung:

a) 2. Die erste Antwort ist falsch: ohne zu rechnen sieht man, dass die Randverteilung von X eine höhere Dichte hat nahe bei null als weiter weg von null. Die zweite Antwort ist richtig. Die Fläche des Vierecks ist $\sqrt{2} \cdot \sqrt{2} = 2$. Also ist die Höhe der Dichte dort 1/2. Die letzte Antwort ist falsch: wenn man weiss, dass zum Beispiel X nahe bei 1 ist, dann muss Y nahe bei null sein.

b) 1. Die erste Antwort ist richtig: die gemeinsame Verteilung enthält alle Information. Die zweite und dritte Antworten sind falsch: man hat zu wenig Information über die Zusammenhang zwischen X und Y. Die vierte Antwort ist falsch: $\mathrm{Corr}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}$. Wenn $\sigma_X < 1$ und $\sigma_Y < 1$, dann ist also $\mathrm{Corr}(X,Y) > \mathrm{Cov}(X,Y)$.

c) 2. Die erste Antwort ist falsch: die empirische Korrelation liegt im Intervall $[-1, 1]$. Die dritte Antwort ist falsch: es kann einen nicht-lineare Zusammenhang geben.

d) 3. Den Median kann man direkt ablesen. Den Mittelwert nicht; den könnte man nur berechnen wenn man zuerst alle x_i abliest.

e) 3. Der Median ist $(-0.41 + -0.30)/2 = -0.355$. Also kommen nur noch C und D in Frage. Bei Plot D geht der lower whiskerëtwa bis -2.8. Es gibt aber keinen Datenpunkt mit diesem Wert. Also kann D nicht stimmen und nur C bleibt übrig.

f) 4. A ist etwa Uniform verteilt. Diese Verteilung ist mehr “heavy-tailedäls die Normalverteilung. A passt also zu 4. B ist symmetrisch und sieht etwa aus wie die kumulative Normalverteilung. B passt also zu 2. C ist rechttschief und stetig. C passt also zu 3. D ist rechtsschief und diskret. D passt also zu 1.

g) 4. B ist symmetrisch verteilt mit einem grossen Median und einer grossen Streuung. Darum passt B zu 3. D ist auch symmetrisch mit einem Median von etwa 2. D passt also zu 2. C ist rechtsschief mit diskreten Werten ≤ 5. C passt also zu 1. A ist rechtsschief mit werten ≤ 6 und passt zu 4.

h) 1. A ist rechtsschief mit Werten bis zu 6 und passt zu 3. B ist etwa uniform verteilt und passt zu 1. C ist rechtsschief mit diskreten Werten ≤ 5 und passt zu 4. D ist symmetrisch und etwa normalverteilt und passt zu 2.

4. Der Patissier Giorgio hat ein altes Rezept für Butterkekse im Schrank seiner Grossmutter gefunden und hat es sofort ausprobiert. Für die neuen Kekse hat er so viele Komplimente von seinen Kunden bekommen, dass er sich entschlossen hat, diese im grossen Stil zu verkaufen. Dazu braucht er aber eine neue Maschine. Ihm stehen zwei Maschinen A und B zur Auswahl. Nach langem Abwägen von Vor- und Nachteilen der einzelnen Maschinen, kommt er zum Schluss, dass er die Maschine kauft, welche in einer Stunde am meisten produzieren kann. Man lässt die beiden Maschinen 10 mal jeweils eine Stunde Kekse produzieren und man erhält folgende Werte:

Stunde Nr.	1	2	3	4	5	6	7	8	9	10
Maschine A: x_i	995	970	955	1005	980	975	980	985	1000	965
Maschine B: y_i	985	1020	975	1005	1000	990	1000	1010	985	1015

$\bar{x}_{10} = 981,\ s_x^2 = 248.89,\ \bar{y}_{10} = 998.5,\ s_y^2 = 211.39,\ s_{x-y}^2 = 479.17,\ s_{pool}^2 = 230.14.$

Man kann davon ausgehen, dass die Anzahl Kekse pro Stunde gut durch eine Normalverteilung approximiert wird. Führe einen geeigneten t-Test durch, um zu sehen, ob sich die Produktionsmengen pro Stunde der beiden Maschinen signifikant unterscheiden.

a) Wie lautet die Nullhypothese?

(i) Maschine A produziert mehr Kekse in einer Stunde als Maschine B.
(ii) Maschine A produziert gleichviele Kekse in einer Stunde wie Maschine B.
(iii) Maschine A produziert weniger Kekse in einer Stunde als Maschine B.
(iv) Die Maschinen A und B produzieren unterschiedlich viele Kekse pro Stunde.

b) Wie lautet die korrekte Alternativ-Hypothese?

(i) Maschine A produziert mehr Kekse in einer Stunde als Maschine B.
(ii) Maschine A produziert weder mehr noch weniger Kekse in einer Stunde als Maschine B.
(iii) Maschine B produziert mehr Kekse in einer Stunde als Maschine A.
(iv) Entweder Maschine A oder B produziert im Schnitt mehr Kekse pro Stunde als die andere Maschine.

c) Wie muss der zugehörige Test durchgeführt werden?

(i) Einseitig und gepaart.

(ii) Einseitig und ungepaart.

(iii) Zweiseitig und gepaart.

(iv) Zweiseitig und ungepaart.

Führe nun gemäss deine Antworte bei den Aufgaben a), b) und c) den geeigneten t-Test zum Niveau 5% durch.

d) Gib den Verwerfungsbereich an! Kann Giorgio nach diesem Test eine eindeutige Kaufentscheidung fällen? Falls ja, wie entscheidet er sich? Begründe die Antworten!

Lösung:

a) (ii)

b) (iv)

c) (iv)

d) Die zu betrachtende Teststatistik lautet:

$$T := \sqrt{5}\frac{\bar{X}_{10} - \bar{Y}_{10}}{s_{pool}}.$$

Unter der Nullhypothese ist die Zufallsvariable T t-verteilt mit 18 Freiheitsgraden. Da der Test zweiseitig ist, wird die Nullhypothese verworfen, wenn

$$|T| \geq t(18, 97.5\%),$$

wobei $t(18, 97.5\%)$ das 97.5%-Quantile der t-Verteilung mit 18 Freihaitsgraden darstellt. Von der Tabelle der Quantile der t-Verteilung liest man ab, dass $t(18, 97.5\%) = 2.101$. Also lautet der Verwerfungsgbereich:

$$VB = \{T \in (-\infty, -2.101] \cup [2.101, \infty)\}.$$

Da die Teststatistik den Wert

$$T = \sqrt{5}\frac{\bar{x}_{10} - \bar{y}_{10}}{s_{pool}} = -2.579$$

annimmt, wird die Nullhypothese verworfen und Giorgio kann eine eindeutige Kaufentscheidung fällen. Er entscheidet sich für die Maschine B, da T negativ ist.

5. Die Wahrscheinlichkeit, dass an einem beliebigen Tag in einer Grossstadt erhöhte (d.h. einen kritischen Wert überschreitende) Schadstoffemissionen gemessen werden sei p. Zusätzlich nehmen wir an, dass die Emissionshöhen an unterschiedlichen Tagen unabhängig sind.

a) Sei in einem Zeitraum von n Tagen X_n die Anzahl derjenigen Tage, an denen erhöhte Emissionswerte gemessen werden. Welche Verteilung hat X_n?

b) Innerhalb der letzten 365 Tage wurden an 219 Tagen erhöhte Emissionen gemessen. Finde eine vernünftige Schätzung für p und bestimme unter Verwendung der Normalapproximation ein 99%-Vertrauensintervall für p.

Neue verkehrspolitische Massnahmen werden implementiert, von denen man sich eine Reduktion von p auf unter 0.4 erhofft. 100 Tage nach Implementierung dieser Massnahmen soll daher versucht werden, statistisch zu belegen, dass p auf unter 0.4 gefallen ist.

c) Formuliere die für die Belegung letzterer Aussage geeigneten Null- und Alternativhypothesen und bestimme mittels Normalapproximation den Verwerfungsbereich für das Signifikanzniveau 0.05.

d) Während dieses Zeitraums von 100 Tagen nach Implementierung der Massnahmen werden an 28 Tagen erhöhte Emissionen gemessen. Erkläre kurz das Resultat des Tests aus Aufgabenteil c) und bestimme mittels Normalapproximation den relevanten p-Wert.

Lösung:

a) X_n is binomialverteilt mit Parametern n und p.

b) Die Schätzung für p ist

$$\hat{p} = \frac{219}{365} = 0.6.$$

Laut Normalapproximation ist das Vertrauensintervall gegeben durch

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

wobei $\alpha = 0.01, x = 219, n = 365$, also durch

$$0.6 \pm \Phi^{-1}(0.995)\sqrt{0.6 \times 0.4 \times \frac{1}{365}}.$$

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

$$[0.5340, 0.6660]$$

(auf 4. Stelle gerundet) als 99%-Vertrauensintervall für p.

c) Zur Belegung der Aussage wählt man als Nullhypothese

$$H_0 : p = 0.4$$

und als Alternativhypothese die zu belegende Aussage, also

$$H_A : p < 0.4.$$

Der Verwerfungsbereich ist $\{x : x \leq c\}$, wobei $c \in \mathbb{N}_0$ so gross wie möglich mit $P_{0.4}[X_{100} \leq c] \leq 0.05$ ist. Zur Berechnung von $P_{0.4}[X_{100} \leq c]$ verwendet man die Normalapproximation:

$$\begin{aligned} P_{0.4}[X_{100} \leq c] &= P_{0.4}\left[\frac{X_{100} - 40}{\sqrt{0.4 \times 0.6 \times 100}} \leq \frac{c - 40}{\sqrt{0.4 \times 0.6 \times 100}}\right] \\ &\approx \Phi\left(\frac{c - 40}{\sqrt{0.4 \times 0.6 \times 100}}\right). \end{aligned}$$

Somit benötigen wir

$$\Phi\left(\frac{c - 40}{\sqrt{0.4 \times 0.6 \times 100}}\right) \leq 0.05,$$

beziehungsweise

$$1-\Phi\left(\frac{c-40}{\sqrt{0.4\times 0.6\times 100}}\right)=\Phi\left(\frac{40-c}{\sqrt{0.4\times 0.6\times 100}}\right)\geq 0.95,$$

oder

$$c \leq 40-\Phi^{-1}(0.95)\sqrt{0.4\times 0.6\times 100} \overset{\Phi^{-1}(0.95)\approx 1.645}{\approx} 31.94.$$

Somit ist der Verwerfungsbereich $\{x : x \leq 31\}$.

d) Aus der Beobachtung folgt, dass die Aussage bzw. die gewählte Alternativhypothese, signifikant ist. Der P-Wert ist gegeben durch $P_{0.4}[X_{100} \leq 28]$. Durch Normalapproximation erhält man

$$\begin{aligned} P_{0.4}[X_{100} \leq 28] &= P_{0.4}\left[\frac{X_{100}-40}{\sqrt{0.4\times 0.6\times 100}} \leq \frac{-12}{\sqrt{0.4\times 0.6\times 100}}\right] \\ &\approx \Phi\left(\frac{-12}{\sqrt{0.6\times 0.4\times 100}}\right) \\ &= 1-\Phi\left(\frac{12}{\sqrt{0.6\times 0.4\times 100}}\right). \end{aligned}$$

Mit $\frac{12}{\sqrt{0.6\times 0.4\times 100}} \approx 2.45$ und $\Phi(2.45) \approx 0.99286$ erhält man also den P-Wert 0.00714.