ETH Zürich HS 2019
RW, D-MATL, D-MAVT
Dr. Christoph Czichowsky

Koordinator
Florian Krach

Stochastik
Musterlösung 7

1. Die Randdichten eines zweidimensionalen Zufallsvektors $Z = (X, Y)$ sind wie folgt definiert:

$$f_X(x) = \begin{cases} \frac{1}{2} & -1 \le x \le 1 \\ 0 & \text{sonst} \end{cases} \qquad f_Y(y) = \begin{cases} \frac{3}{4}(2y - y^2) & 0 \le y \le 2 \\ 0 & \text{sonst} \end{cases}$$

Der Korrelationskoeffizient $\varrho_{X,Y}$ zwischen X und Y ist $\varrho_{X,Y} = \sqrt{\frac{1}{3}}$.

a) Berechne den Erwartungswert von $6X - 4Y + 2$.

b) Berechne die Kovarianz $\text{Cov}(6X, 4Y)$.

c) Berechne die Varianz von $6X - 4Y + 2$.

d) Berechne den Erwartungswert von $6X^2 - 4Y^2$.

Lösung:

e) Die Erwartungswerte von X und Y sind:

$$E(X) = \int_{-\infty}^{\infty} x f_X(x) dx = \int_{-1}^{1} x \frac{1}{2} dx = \Big[\frac{1}{4}x^2\Big]_{-1}^{1} = 0$$

$$E(Y) = \int_{-\infty}^{\infty} y f_Y(y) dy = \frac{3}{4} \int_{0}^{2} y(2y - y^2) dy = \frac{3}{4}\Big[\frac{2}{3}y^3 - \frac{y^4}{4}\Big]_0^2 = 1$$

Der Erwartungswert für $6X - 4Y + 2$ wird wie folgt berechnet:

$$E(6X - 4Y + 2) = 6E(X) - 4E(Y) + 2 = -2.$$

f) Die Varianzen für X und Y betragen:

$$\text{Var}(X) = E(X^2) - E(X)^2 \qquad \text{Var}(Y) = E(Y^2) - E(Y)^2$$

$$E(X^2) = \int_{-\infty}^{\infty} x^2 f_X(x) dx = \int_{-1}^{1} \frac{1}{2} x^2 dx = \Big[\frac{x^3}{6}\Big]_{-1}^{1} = \frac{1}{3}$$

Bitte wenden!

$$E(Y^2) = \int_{-\infty}^{\infty} y^2 f_Y(y)dy = \frac{3}{4}\int_0^2 y^2(2y - y^2)dy = \frac{3}{4}\Big[\frac{y^4}{2} - \frac{y^5}{5}\Big]_0^2 = \frac{6}{5}$$

$$\mathrm{Var}(X) = \frac{1}{3} \qquad \mathrm{Var}(Y) = \frac{1}{5}.$$

Daraus ergibt sich die folgende Berechnung der Kovarianz $\mathrm{Cov}(6X, 4Y)$:

$$\mathrm{Cov}(X, Y) = \varrho_{X,Y} \cdot \sqrt{\mathrm{Var}(X) \cdot \mathrm{Var}(Y)} = \sqrt{1/3} \cdot \sqrt{1/3 \cdot 1/5} = \sqrt{1/45},$$

$$\Rightarrow \mathrm{Cov}(6X, 4Y) = 6 \cdot 4 \cdot \mathrm{Cov}(X, Y) = 24 \cdot \sqrt{1/45}.$$

g)

$$\begin{aligned}\mathrm{Var}(6X - 4Y + 2) &= \mathrm{Var}(6X) + \mathrm{Var}(4Y) - 2\mathrm{Cov}(6X, 4Y)\\ &= 36 \cdot \mathrm{Var}(X) + 16 \cdot \mathrm{Var}(Y) - 2\mathrm{Cov}(6X, 4Y)\\ &= \frac{36}{3} + \frac{16}{5} - 2 \cdot 24 \cdot \sqrt{1/45} = 8.04.\end{aligned}$$

h)

$$E(6X^2 - 4Y^2) = 6E(X^2) - 4E(Y^2) = 6\frac{1}{3} - 4\frac{6}{5} = -\frac{14}{5}.$$

2. Es seien $\{X_i\}_{1\leq i\leq 50}$ unabhängige und normalverteilte Zufallsvariablen mit Erwartungswert $\mu = 1$ und Standardabweichung $\sigma = 2$. Darüber hinaus sind folgende Zufallsvariablen definiert:

$$S_n = X_1 + X_2 + \cdots + X_n$$

und

$$\bar{X}_n = \frac{1}{n}(X_1 + X_2 + \cdots + X_n) = \frac{S_n}{n}$$

Dabei ist $n = 50$.

a) Bestimme die Parameter der Normalverteilung von S_n sowie $\bar{X}_n$.

b) Berechne die Wahrscheinlichkeit $P(E[X_1] - 1 \leq X_1 \leq E[X_1] + 1)$.

c) Berechne $P(E[S_n] - 1 \leq S_n \leq E[S_n] + 1)$.

d) Berechne $P(E[\bar{X}_n] - 1 \leq \bar{X}_n \leq E[\bar{X}_n] + 1)$.

Lösung:

Siehe nächstes Blatt!

a) Für S_n gilt

$$E[S_n] = E\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} E[X_i] = 50 \cdot 1$$
$$Var(S_n) = Var\Big(\sum_{i=1}^{n} X_i\Big) = \sum_{i=1}^{n} Var(X_i) = 50 \cdot 4 = 200$$

S_{50} ist normalverteilt mit $\mathcal{N}(\mu_{S_{50}} = 50, \sigma^2_{S_{50}} = 200)$.
Für $\bar{X}_n$ gilt

$$E[\bar{X}_n] = \frac{1}{n}E\Big[\sum_{i=1}^{n} X_i\Big] = 1\ , \qquad Var(\bar{X}_n) = \frac{1}{n^2}Var\Big(\sum_{i=1}^{n} X_i\Big) = 0.08$$

$\bar{X}_{50}$ ist normalverteilt mit $\mathcal{N}(\mu_{\bar{X}_{50}} = 1, \sigma^2_{\bar{X}_{50}} = 0.08)$.

b) X_1 ist normalverteilt mit $\mathcal{N}(\mu_{X_1} = 1, \sigma^2_{X_1} = 4)$. Durch Standardisieren erhalten wir $Z \sim \mathcal{N}(0,1)$ wobei $Z = \frac{X_1 - 1}{2}$. Wir erhalten dann

$$\begin{aligned}
P(E[X_1] - 1 \leq X_1 \leq E[X_1] + 1) = P(0 \leq X_1 \leq 2) &= P\Big(-\frac{1}{2} \leq Z \leq \frac{1}{2}\Big) \\
&= \Phi\Big(\frac{1}{2}\Big) - \Phi\Big(-\frac{1}{2}\Big) = \Phi\Big(\frac{1}{2}\Big) - \Big(1 - \Phi\Big(\frac{1}{2}\Big)\Big) \\
&= \Phi\Big(\frac{1}{2}\Big) + \Phi\Big(\frac{1}{2}\Big) - 1 = 2 \cdot \Phi\Big(\frac{1}{2}\Big) - 1 \\
&= 2 \cdot 0.6915 - 1 = 0.3830
\end{aligned}$$

c) Durch Standardisieren: für $n = 50$ erhalten wir $Z = \frac{S_{50} - 50}{\sqrt{200}}$.
Es folgt:

$$P(49 \leq S_n \leq 51) = P(-0.07 \leq Z \leq 0.07) = 2 \cdot \Phi(0.07) - 1 = 2 \cdot 0.528 - 1 = 0.056$$

d) Durch Standardisieren: für $n = 50$ erhalten wir $Z = \frac{\bar{X}_{50} - 1}{\sqrt{0.08}}$.
Es folgt:

$$P(0 \leq \bar{X}_{50} \leq 2) = P(-3.5 \leq Z \leq 3.5) = 2 \cdot \Phi(3.5) - 1 = 2 \cdot 0.99977 - 1 = 0.99954$$

3. Die erwartete Lebensdauer μ eines Batterietyps ist unbekannt und soll durch das arithmetische Mittel der Lebensdauern von n unabhängigen Testbatterien dieses Typs geschätzt werden. Erfahrungsgemäss ist die Standardabweichung der Lebensdauer ungefähr 5 Stunden. Wie gross muss n mindestens sein, damit der Absolutbetrag der Differenz zwischen arithmetischem Mittel und μ mit einer Wahrscheinlichkeit von mindestens 99% höchstens 1 Stunde beträgt?

Bitte wenden!

Tipp: Zentraler Grenzwertsatz.

Lösung:

Sei L_i die Lebensdauer der i-ten Batterie mit erwarteter Lebensdauer $E(L_i) = \mu$ und Streuung $\sigma_{L_i} = 5$. Für n Batterien ist das arithmetische Mittel der Lebensdauern $L_1, \ldots, L_n$ definitionsgemäss die Zufallsvariable

$$\overline{L}_n := \frac{1}{n}\sum_{i=1}^{n} L_i \qquad \text{mit} \qquad E(\overline{L}_n) = \mu \qquad \text{und} \qquad \sigma_{\overline{L}_n} = \frac{5}{\sqrt{n}}.$$

Gesucht ist das kleinste n für welches

$$P(\left|\overline{L}_n - \mu\right| \leq 1) \geq 0.99 \tag{1}$$

gilt. Wir rechnen zunächst die linke Seite aus, d.h.

$$\begin{aligned} P(\left|\overline{L}_n - \mu\right| \leq 1) &= P(-1 \leq \overline{L}_n - \mu \leq 1) = P(\mu - 1 \leq \overline{L}_n \leq \mu + 1) \\ &= F_{\overline{L}_n}(\mu + 1) - F_{\overline{L}_n}(\mu - 1). \end{aligned}$$

Wegen des Zentralen Grenzwertsatzes nehmen wir nun an, dass $\overline{L}_n$ normalverteilt ist (mit Mittelwert μ und Streuung $\frac{5}{\sqrt{n}}$, wie oben ausgerechnet) und erhalten

$$\begin{aligned} &P(\left|\overline{L}_n - \mu\right| \leq 1) = F_{\overline{L}_n}(\mu + 1) - F_{\overline{L}_n}(\mu - 1) \\ &\approx \Phi\left(\frac{\mu + 1 - \mu}{\frac{5}{\sqrt{n}}}\right) - \Phi\left(\frac{\mu - 1 - \mu}{\frac{5}{\sqrt{n}}}\right) = \Phi\left(\frac{\sqrt{n}}{5}\right) - \Phi\left(-\frac{\sqrt{n}}{5}\right) \\ &= \Phi\left(\frac{\sqrt{n}}{5}\right) - \left(1 - \Phi\left(\frac{\sqrt{n}}{5}\right)\right) = 2\Phi\left(\frac{\sqrt{n}}{5}\right) - 1. \end{aligned}$$

Bedingung (1) lautet nun

$$\Phi\left(\frac{\sqrt{n}}{5}\right) \approx 0.995\ ,$$

aus der Tabelle der Standardnormalverteilung lesen wir das 0.995-Quantil ab und erhalten $\frac{\sqrt{n}}{5} \approx \Phi^{-1}(0.995) = 2.58$ und daraus $n \approx (2.58 \cdot 5)^2 = 166.41$. Somit muss $n \geq 167$ sein.

4. Die Krankenkasse easyHealth bietet einen sehr günstigen Studententarif für 160 Franken pro Jahr an. Sie hat pro Kunde einen jährlichen Administrationsaufwand von 100 Franken. Zusätzlich fallen die variierenden Gesundheitskosten (Arztrechnungen etc.) an. Zur Modellierung der *Gesamtkosten* X pro Kunde und Jahr benutzt die Krankenkasse die folgende Dichte

$$f(x) = \begin{cases} 0 & x \leq c \\ \frac{\alpha c^\alpha}{x^{\alpha+1}} & x > c \end{cases}$$

Siehe nächstes Blatt!

wobei $c > 0, \alpha > 1$ Parameter sind.

easyHealth hat zur Zeit 250 studentische Kunden. Was ist (approximativ) die Wahrscheinlichkeit, dass easyHealth mit diesem Tarif in einem Jahr Gewinn macht, wenn wir annehmen, dass $\alpha = 3$, $c = 100$?

Tipp: Die folgenden Formel können ohne Herleitung benutzt werden:

$$E[X] = c\frac{\alpha}{\alpha - 1}; \qquad Var(X) = \frac{c^2\alpha}{(\alpha-1)^2(\alpha-2)}.$$

Lösung:

Wir definieren die Zufallsvariable G als Differenz aus Einnahmen und Kosten: $G = 250 \cdot 160 - \sum_{i=1}^{250} X_i$, wobei X_i iid Kopien von X sind. Dann gilt mit dem zentralen Grenzwertsatz:

$$\begin{aligned} P(G > 0) &= P\left(\sum_{i=1}^{250} X_i < 250 \cdot 160\right) = P\left(\frac{\sum_{i=1}^{250} X_i - 250\mu}{\sqrt{250\sigma^2}} < \frac{250(160-\mu)}{\sqrt{250\sigma^2}}\right) \\ &\approx \Phi\left(\sqrt{\frac{250}{\sigma^2}}(160-\mu)\right) \end{aligned}$$

wobei $\mu = E[X] = 150$ und $\sigma^2 = \text{Var}(X) = 7500$. Somit erhalten wir

$$P(G > 0) \approx \Phi(1.8257) = 0.96$$

5. Auf `http://onlinestatbook.com/stat_sim/sampling_dist/index.html` findest du ein Tool, das sehr hilfreich ist beim Verständnis des Zentralen Grenzwertsatzes (der Link ist auch auf der Homepage zur Vorlesung).

Gegeben seien i.i.d. Zufallsvariablen $X_1, \cdots X_N$. Wie üblich sei $\bar{X}_N = \frac{1}{N}\sum_{i=1}^{N} X_i$ das arithmetische Mittel.

Eine kurze Anleitung: Wähle 'Beginn' im Menü auf der linken Seite (um über weitere Funktionalitäten zu erfahren kannst du auch zuerst die 'Instructions' lesen).

- Der oberste Plot zeigt die Verteilung eines einzelnen X_i (bzw. ein Histogramm davon). Diese Verteilung ist zu Beginn auf 'Normal' eingestellt, kann aber in einem Dropdown-Menü gewählt oder per Mausklick auf den Plot manuell geändert werden.

- Per Default ist die Anzahl an Beobachtungen (die 'Sample Size') auf $N = 5$ gestellt, im Dropdown-Menü weiter unten kann diese erhöht werden. Ein Klick auf 'Animated' generiert nun eine Beobachtung von $X_1, \cdots, X_N$ und zeichnet alle in den Plot 'Sample Data'. Dann wird der Mittelwert $\bar{X}_N$ dieses Samples im Plot 'Distribution of Means' eingezeichnet.
- Ein weitere Klick auf 'Animated' wiederholt diesen Vorgang. Wiederhole dies einige Male um sicherzugehen, dass du den vorherigen Punkt verstanden hast. Ein Klick auf 5, $10'000$ oder $100'000$ wiederholt den Vorgang dann 5, $10'000$ bzw. $100'000$ Mal. Dabei werden aber die Beobachtungen nicht mehr im Plot 'Sample Data' eingezeichnet, sondern für jede Ziehung wird direkt das arithmetische Mittel der N Beobachtungen im Plot "Distribution of Means' eingezeichnet.
- Dieser Plot zeigt nun die empirische Verteilung von $\bar{X}_N$. Wenn die Sample Size N gross genug ist, sollte diese laut dem Zentralen Grenzwertsatz ungefähr wie eine Normalverteilung aussehen. Indem du ein Häkchen bei 'Fit normal' setzt, kannst du überprüfen wie gut dies stimmt.

Experimentiere nun mit diesem Tool um zu bestimmen wie gross die Sample Size N in etwa sein muss, damit die Verteilung von $\bar{X}_N$ ungefähr wie eine Normalverteilung aussieht, wenn X_i folgende Verteilung hat:

a) Normal

b) Uniform

c) 'Skewed' (siehe Dropdown Menü).

Lösung:

a) Für beliebiges N. Für X_i normalverteilt ist nämlich auch $\bar{X}_N$ normalverteilt (siehe 5.2 im Skript).

b) Bereits für $N = 5$ sieht die Verteilung von $\bar{X}_N$ einer Normalverteilung sehr ähnlich.

c) Wie zu erwarten sieht die Verteilung von $\bar{X}_N$ für $N = 25$ schon viel eher aus wie eine Normalverteilung als für $N = 2$, aber selbst bei $N = 25$ gibt es noch eine leichte Verschiebung. N müsste also noch grösser sein.