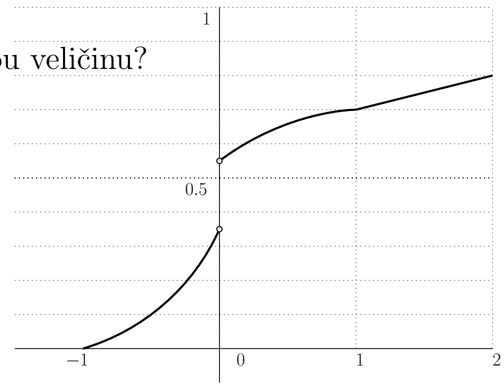
2A. zkoušková písemka NMAI059 Pravd. a Stat. 1 – řešení

1. (10 bodů) Na obrázku je zakreslená distribuční funkce náhodné veličiny X. Hodnota $F_X(0)$ není na obrázku vyznačena.

(a) Určete $F_X(0)$.
(b) Co můžete říct o hodnotě $F_X(4)$?
(c) Jedná se o diskrétní nebo o spojitou náhodnou veličinu?
(d) Spočtěte $P(X \leq 1)$.
(e) Spočtěte $P(X \geq 0)$.
(f) Spočtěte $P(0 < X < 1)$.
(g) Určete medián.



Řešení:

(a) Hodnota $F_X(0)$ není v grafu distribuční funkce zakreslená, ale díky spojitosti distribuční funkce zprava vidíme, že $F_X(0) = 0.55$.

(b) Hodnota $F_X(4)$ může být cokoli mezi 0.8 a 1. (Distribuční funkce je neklesající, omezená 1.)

(c) Distribuční funkce není spojitá, takže se nejedná o spojitou náhodnou veličinu. Není to ale ani po částech konstantní funkce, takže se nejedná ani o diskrétní n.v. (Jedná se o mix spojité n.v. a konstantní nuly, která má pravděpodobnost 0.2.)

(d) $P(X \leq 1) = F_X(1) = 0.7$

(e) $P(X \geq 0) = 1 - F_X(0_-) = 1 - 0.35 = 0.65$

(f) $P(0 < X < 1) = F_X(1_-) - F_X(0) = 0.7 - 0.55 = 0.15$ (F_X je spojitá v 1, ale kdyby nebyla, brali bychom limitu zleva, ne funkční hodnotu)

(g) Medián je nejmenší hodnota m, pro kterou platí $F_X(m) \geq 0.5$, tj. $m = 0$.

2. (10 bodů)

Házíme hrací kostkou, zapisujeme si jestli padlo sudé (S) nebo liché (L) číslo.

(a) Označme X počet výskytů „LL“, tj. takových i, že v i-tém i v $(i+1)$-ním hodu padlo liché číslo. Celkový počet hodů je 100. Určete $\mathbb{E}(X)$.

(b) Označme Y počet hodů, než se dočkáme výskytu dvojice „LL“, tj. kolikátým hodem padlo druhé liché číslo po sobě (házíme tak dlouho, dokud se to nestane). Určete $\mathbb{E}(Y)$.

Řešení: a) Označme X_i indikátorovou proměnnou, která je 1, pokud v i-tém i v $(i+1)$-ním hodu padlo liché číslo, a 0 jinak. Celkový počet výskytů „LL“ můžeme zapsat jako:

$$X = \sum_{i=1}^{99} X_i$$

.

Protože házíme normální hrací kostkou, pravděpodobnost, že padne liché číslo (1, 3, nebo 5), je $\frac{3}{6} = \frac{1}{2}$. Pravděpodobnost, že padne liché číslo dvakrát po sobě, je: $\frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$. Každá indikátorová proměnná X_i má tedy očekávanou hodnotu $\mathbb{E}(X_i) = \frac{1}{4}$.

Celkový očekávaný počet výskytů „LL" je podle linearity střední hodnoty

$$\mathbb{E}(X) = \sum_{i=1}^{99} \mathbb{E}(X_i) = 99 \cdot \frac{1}{4} = 24.75$$

b) Označme pro stručnost $\mathbb{E}(Y \mid LS)$ střední hodnotu Y, pokud první dva hody byly liché číslo, následované sudým číslem; analogicky ostatní zápisy. Jistě platí:

- $\mathbb{E}(Y) = \frac{1}{2} \cdot \mathbb{E}(Y \mid S) + \frac{1}{4} \cdot \mathbb{E}(Y \mid LL) + \frac{1}{4} \cdot \mathbb{E}(Y \mid LS)$ (věta o úplné střední hodnotě)
- $\mathbb{E}(Y \mid S) = \mathbb{E}(Y) + 1$ („začínáme znovu, ztratili jsme jeden hod")
- $\mathbb{E}(Y \mid LS) = \mathbb{E}(Y) + 2$ („začínáme znovu, ztratili jsme dva hody")
- $\mathbb{E}(Y \mid LL) = 2$ (dočkali jsme se)

Celkově máme tedy rovnici

$$\mathbb{E}(Y) = \frac{1}{2}\big(\mathbb{E}(Y) + 1\big) + \frac{1}{4} \cdot 2 + \frac{1}{4}\big(\mathbb{E}(Y) + 2\big)$$

jejímž řešením je $\mathbb{E}(Y) = 6$.

3. (10 bodů) O naměřených hodnotách 6, 4, 10, 8, 10 předpokládáme, že pocházejí z náhodného výběru z rozdělení $Pois(\lambda)$. (Jedná se o počet dotazů na webový server během jedné minuty.)

(a) Navrhněte bodový odhad parametru λ momentovou metodou.

(b) Navrhněte bodový odhad parametru λ metodou maximální věrohodnosti.

Řešení: a) První moment, tj. střední hodnota $Pois(\lambda)$ je λ, první výběrový moment je $\bar{x} = (6 + 4 + 10 + 8 + 10)/5 = 38/5 = 7.6$. Odhad pomocí momentové metody je tedy $\hat{\lambda} = \bar{x} = 7.6$.

b) Dle vzorce pro Poissonovo rozdělení a nezávislosti měření je pravděpodobnost, že dostaneme měření $x_1, \ldots, x_5$ rovna

$$L(x, \lambda) = \prod_{i=1}^{5} e^{-\lambda} \frac{\lambda^{x_i}}{x_i!} = e^{-5\lambda} \lambda^{5\bar{x}} \cdot C,$$

kde $C = 1/(x_1! \ldots x_5!)$ je konstanta. (Pro stručnost zápisu píšeme $5\bar{x}$ místo $x_1 + \cdots + x_5$.) Pro snazší výpočet zlogaritmujeme:

$$\ell(x, \lambda) = \log(L(x, \lambda) = -5\lambda + 5\bar{x} \log(\lambda) + \log(C).$$

Hledáme maximum, proto zderivujeme podle λ:

$$\ell'(x, \lambda) = -5 + \frac{5\bar{x}}{\lambda}.$$

Pro $\lambda \in (0, \infty)$ je tato derivace napřed kladná, pak nulová, pak záporná. Takže bod, kde je derivace nulová je kýžené maximum. Snadný výpočet dává i tady $\hat{\lambda} = \bar{x} = 7.6$.

4. (10 bodů) (a) Definujte pojem nezávislé náhodné veličiny (spojitý případ, dvě veličiny). Uveďte formulaci pomocí distribuční funkce i pomocí hustoty.

Rozhodněte, zda existují nezávislé X, Y takové, že $X \sim N(0,1)$ a $Y \sim N(0,2)$. Pokud ano, jakých hodnot může nabývat $\mathbb{E}(X \cdot Y)$ a $\mathbb{E}((2X+Y)^2)$.

(b) Definujte pojem kovariance náhodných veličin.

Označme X výsledek hodu šestistěnnou hrací kostkou a Y hodnotu $X/2$ zaokrouhlenou na celá čísla dolů. Určete kovarianci X a Y.

Řešení: a) Nechť X, Y jsou veličiny se sdruženou hustotou $f_{X,Y}$ a sdruženou distribuční funkcí $F_{X,Y}$. Pak X, Y jsou nezávislé právě když pro všechna reálná x, y platí

$$F_{X,Y}(x,y) = F_X(x)F_Y(y) \qquad \text{ekvivalentně} \qquad f_{X,Y}(x,y) = f_X(x)f_Y(y).$$

Ano, takové veličiny existují, stačí definovat hustotu výše uvedeným vzorcem, tj.

$$f_{X,Y}(x,y) = \varphi(x) \cdot \varphi(\sqrt{2}x).$$

Podle věty o střední hodnotě součinu nezávislých veličin je

$$\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y) = 0.$$

Podle linearity střední hodnoty platí

$$\mathbb{E}((2X+Y)^2) = \mathbb{E}(4X^2) + 4\mathbb{E}(XY) + \mathbb{E}(Y^2) = 4 + 2 = 6.$$

(Využíváme zde toho, že $\mathrm{var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = \mathbb{E}(X^2)$, neboť v našem případě je $\mathbb{E}(X) = 0$; obdobně pro Y.)

b) Kovariance je definována jako $\mathrm{cov}(X,Y) = \mathbb{E}((X-\mathbb{E}(X))(Y-\mathbb{E}(Y)))$, lze také počítat jako $\mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$.

V našem případu je $\mathbb{E}(X) = 3.5$, $\mathbb{E}(Y) = (0+1+1+2+2+3)/6 = 9/6 = 1.5$. Dále je $\mathbb{E}(XY) = (1\cdot 0 + 2\cdot 1 + 3\cdot 1 + 4\cdot 2 + 5\cdot 2 + 6\cdot 3)/6 = 41/6$. Je tedy $\mathrm{cov}(X,Y) = 41/6 - 21/4 = (82-63)/12 = 19/12$.

5. (10 bodů) Vyslovte Centrální limitní větu. Vysvětlete, k čemu se hodí.

Odhadněte pomocí CLV pravděpodobnost, že z 3600 hodů kostkou nám padne 683 šestek.

Řešení: Centrální limitní věta říká, že součet velkého počtu nezávislých a identicky rozdělených náhodných veličin s konečnou střední hodnotou a rozptylem má přibližně normální rozdělení. Formálně:

Nechť $X_1, X_2, \ldots$ jsou nezávislé a identicky rozdělené náhodné veličiny s očekávanou hodnotou μ a rozptylem σ^2. Pak pro součet $S_n = X_1 + X_2 + \cdots + X_n$ platí, že:

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{d} N(0,1)$$

kde $\xrightarrow{d}$ značí konvergenci v distribuci. Jinými slovy, pokud F_n znamená distribuční funkci náhodné veličiny $\frac{S_n - n\mu}{\sigma\sqrt{n}}$, tak platí, že $\lim_{n\to\infty} F_n(x) = \Phi(x)$ pro všechna x.

CLV se hodí pro aproximaci distribuce součtu nebo průměru velkého počtu náhodných veličin normálním rozdělením, což usnadňuje výpočty pravděpodobností a statistické odhady.

Jako příklad uvažujme zadaný příklad s kostkami. Označme X_i indikátorovou proměnnou, která je 1, pokud i-tý hod kostkou padne šestka, a 0 jinak. Pak $X_i \sim Ber(1/6)$.

Součet těchto indikátorových proměnných, $S_n = \sum_{i=1}^{3600} X_i$, představuje počet šestek v 3600 hodech. Očekávaná hodnota a rozptyl S_n jsou (podle vzorců pro Bernoulliho rozdělení)

$$\begin{aligned}\mathbb{E}(S_n) &= 3600 \cdot \frac{1}{6} = 600\\ \operatorname{var}(S_n) &= 3600 \cdot \frac{1}{6}\left(1 - \frac{1}{6}\right) = 3600 \cdot \frac{1}{6} \cdot \frac{5}{6} = 500.\end{aligned}$$

Podle CLV můžeme aproximovat S_n normálním rozdělením: $S_n \approx N(600, 500)$, neboli $Z = \frac{S_n - \mu}{\sigma}$ má přibližně standardní normální rozdělení. Chceme odhadnout pravděpodobnost, že $S_n = 683$. Vzhledem k tomu, že normální rozdělení je spojité, pravděpodobnost přesné hodnoty je 0. Můžeme však odhadnout pravděpodobnost, že S_n bude v blízkosti 683, například v intervalu $[682.5, 683.5]$.

$$\begin{aligned}\mathbb{P}(682.5 \leq S_n \leq 683.5) &= \mathbb{P}\left(\frac{682.5 - 600}{\sqrt{500}} \leq Z \leq \frac{683.5 - 600}{\sqrt{500}}\right)\\ &= \mathbb{P}\left(\frac{82.5}{\sqrt{500}} \leq Z \leq \frac{83.5}{\sqrt{500}}\right)\end{aligned}$$

Neboli pomocí distribuční funkce normálního rozdělení:

$$P(S_n = 683) \approx \Phi(\frac{83.5}{\sqrt{500}}) - \Phi(\frac{82.5}{\sqrt{500}}).$$

(Po dosazení bychom zjistili, že nám vzorec dává hodnotu $\approx 1.8 \cdot 10^{-5}$, oproti přesné hodnotě z binomického rozdělení $\approx 2.2 \cdot 10^{-5}$.)

6. (10 bodů) Vyslovte a dokažte větu o střední hodnotě součinu nezávislých diskrétních náhodných veličin.

Řešení: Věta: Nezávislé diskrétní náhodné veličiny X a Y mají střední hodnotu součinu rovnou součinu středních hodnot. Neboli, pokud $\mathbb{E}(X)$ i $\mathbb{E}(Y)$ existují, tak

$$\mathbb{E}(XY) = \mathbb{E}(X) \cdot \mathbb{E}(Y).$$

Důkaz: Protože X a Y jsou nezávislé, sdružená pravděpodobnostní funkce je:

$$P(X = x, Y = y) = P(X = x) \cdot P(Y = y).$$

Střední hodnota součinu je („pravidlo naivního statistika“)

$$
\begin{aligned}
\mathbb{E}(XY) &= \sum_x \sum_y xy \cdot P(X = x, Y = y) \\
&= \sum_x \sum_y xy \cdot P(X = x) \cdot P(Y = y) \\
&= \left(\sum_x x \cdot P(X = x) \right) \cdot \left(\sum_y y \cdot P(Y = y) \right) \\
&= \mathbb{E}(X) \cdot \mathbb{E}(Y)
\end{aligned}
$$

V sumách se sčítá přes všechna $x \in Im(X)$ a $y \in Im(Y)$.

Tím je věta dokázána.