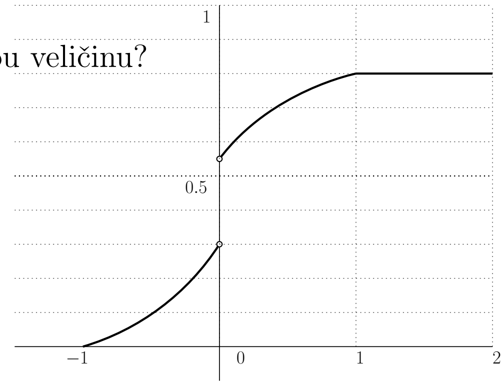
2B. zkoušková písemka NMAI059 Pravd. a Stat. 1 – řešení

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 $F_X(3)$?
(c) Jedná se o diskrétní nebo o spojitou náhodnou veličinu?
(d) Spočtěte $P(X < 1)$.
(e) Spočtěte $P(X > 0)$.
(f) Spočtěte $P(0 \leq X \leq 1)$.
(g) Určete osmdesátý percentil.



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 $F_X(3)$ 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 0.25.)

(d) $P(X < 1) = F_X(1_-) = 0.8$ (F_X je spojitá v 1, ale kdyby nebyla, brali bychom limitu zleva, ne funkční hodnotu)

(e) $P(X > 0) = 1 - F_X(0) = 1 - 0.55 = 0.45$

(f) $P(0 \leq X \leq 1) = F_X(1) - F_X(0_-) = 0.8 - 0.3 = 0.5$ (Odčítáme $P(X < 0)$, což není funkční hodnota distribuční funkce, ale její limita zleva.)

(g) Je to nejmenší hodnota q, pro kterou platí $F_X(q) \geq 0.8$, tj. $q = 1$.

2. (10 bodů)

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

(a) Označme X počet výskytů „LS“, tj. takových i, že v i-tém hodu padlo liché a v $(i+1)$-ním hodu sudé číslo. Celkový počet hodů je 400. Určete $\mathbb{E}(X)$.

(b) Označme Y počet hodů, než se dočkáme výskytu dvojice „LS“, tj. kolikátým hodem padlo sudé číslo předcházené lichým číslem (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 hodu padlo liché číslo a v $(i+1)$-ním hodu padlo sudé číslo, a 0 jinak. Celkový počet výskytů „LS“ můžeme zapsat jako:

$$X = \sum_{i=1}^{399} X_i$$

Protože házíme běžnou hrací kostkou, pravděpodobnost, že padne liché číslo, je $\frac{3}{6} = \frac{1}{2}$ a pravděpodobnost, že padne sudé číslo, je také $\frac{1}{2}$.

Každá indikátorová proměnná X_i má stejnou očekávanou hodnotu:

$$\mathbb{E}(X_i) = P(\text{Liché číslo v } i\text{-tém hodu a sudé číslo v } (i+1)\text{-ním hodu}) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$$

Celkový očekávaný počet výskytů „LS“ je:

$$\mathbb{E}(X) = \sum_{i=1}^{399} \mathbb{E}(X_i) = 399 \cdot \frac{1}{4} = 99.75$$

b) Nechť Y_1 je počet hodů, než se poprvé objeví liché číslo, Y_2 počet hodů, které po tomto prvním lichém čísle čekáme na první sudé číslo. Např., pokud padlo $SSSLLLS$, tak $Y_1 = 4$, $Y_2 = 3$.

Je zjevné, že Y_1 i Y_2 mají geometrické rozdělení s pravděpodobností úspěchu $1/2$, tj. se střední hodnotou 2. Dále si snadno rozmyslíme, že $Y = Y_1 + Y_2$, tj. z linearity střední hodnoty je $\mathbb{E}(Y) = \mathbb{E}(Y_1) + \mathbb{E}(Y_2) = 2 + 2 = 4$.

Alternativní postup: Označme pro stručnost $\mathbb{E}(Y \mid L)$ střední hodnotu Y, pokud první hod bylo liché číslo, atp. Jistě platí:

- $\mathbb{E}(Y) = \frac{1}{2} \cdot \mathbb{E}(Y \mid S) + \frac{1}{2} \cdot \mathbb{E}(Y \mid L)$ (věta o úplné střední hodnotě)
- $\mathbb{E}(Y \mid S) = \mathbb{E}(Y) + 1$ („začínáme znovu, ztratili jsme jeden hod“)
- $\mathbb{E}(Y \mid L) = 1 + \mathbb{E}(Geom(1/2))$ (jednou jsme hodili, teď „čekáme na úspěch“, tj. padnutí sudého čísla)

Celkově máme tedy rovnici

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

jejímž řešením je $\mathbb{E}(Y) = 4$.

3. (10 bodů) O naměřených hodnotách 6, 6, 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 $\bar{x} = (6 + 6 + 10 + 8 + 10)/5 = 40/5 = 8$. Odhad pomocí momentové metody je tedy $\hat{\lambda} = \bar{x} = 8$.

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 $C = 1/(x_1! \dots x_5!)$ je konstanta. 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 kyžené maximum. Snadný výpočet dává i tady $\hat{\lambda} = \bar{x} = 8$.

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 $\mathbb{E}((X - 2Y)^2)$.

(b) Definujte pojem kovariance náhodných veličin.

Označme X výsledek hodu šestistěnnou hrací kostkou a Y hodnotu $X/3$ 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.

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

Pak podle věty o střední hodnotě součinu nezávislých veličin je

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

Podle linearity střední hodnoty platí

$$\mathbb{E}((X - 2Y)^2) = \mathbb{E}(X^2) - 4\mathbb{E}(XY) + \mathbb{E}(4Y^2) = 1 + 4 \cdot 2 = 9.$$

(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 $\mathbb{E}(X) = 3.5$, $\mathbb{E}(Y) = (0+0+1+1+1+2)/6 = 5/6$. Dále je $\mathbb{E}(XY) = (1 \cdot 0 + 2 \cdot 0 + 3 \cdot 1 + 4 \cdot 1 + 5 \cdot 1 + 6 \cdot 2)/6 = 24/6 = 4$. Je tedy $\mathrm{cov}(X, Y) = 4 - 35/12 = 13/12$.

5. (10 bodů) Vyslovte Centrální limitní větu. Vysvětlete, k čemu se hodí. Zejména vysvětlete, jak se používá pro intervalové odhady.

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.

Jednou z hlavních aplikací CLV je konstrukce intervalových odhadů pro střední hodnotu populace. Intervalový odhad poskytuje rozmezí hodnot, ve kterém s určitou pravděpodobností leží skutečná střední hodnota populace. Postupujeme takto:

Mějme vzorek $X_1, X_2, \ldots, X_n$ z populace s neznámou střední hodnotou μ a rozptylem σ^2.

Spočteme výběrový průměr $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Podle CLV je výběrový průměr $\bar{X}$ pro velké n přibližně normálně rozdělen s očekávanou hodnotou μ a rozptylem $\frac{\sigma^2}{n}$.

Pro danou hladinu spolehlivosti $1-\alpha$ (např. 95 %) použijeme kritickou hodnotu $z_{\alpha/2}$ z normálního rozdělení, kde $P(Z > z_{\alpha/2}) = \frac{\alpha}{2}$.

Konfidenční interval pro μ je:

$$\bar{X} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

Tento interval znamená, že s pravděpodobností $1-\alpha$ leží skutečná střední hodnota μ v intervalu:

$$\left(\bar{X} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right)$$

6. (10 bodů) Vyslovte a dokažte větu o konvolučním vzorci pro diskrétní náhodné veličiny (tj. jaká je pravděpodobnostní funkce součtu).

Řešení:

Věta: Nechť X a Y jsou dvě diskrétní náhodné veličiny se sdruženou pravděpodobnostní funkcí $p_{X,Y}$. Pravděpodobnostní funkce součtu $S = X + Y$ je dána vzorcem

$$p_S(k) = \sum_{x\in Im(X)}^{\infty} p_{X,Y}(x, k-x).$$

Pokud X, Y jsou nezávislé, tak platí konvoluční vzorec

$$p_S(k) = \sum_{x\in Im(X)}^{\infty} p_X(x) p_Y(k-x).$$

Důkaz: Potřebujeme zjistit $P(S = k)$. Tuto pravděpodobnost zjistíme jako součet disjunktních jevů podle možných hodnot x náhodné veličiny X:

$$P(Z = k) = \sum_x P(X = x \,\&\, Y = k - x),$$

což nám dává první vzorec.

Pokud X a Y jsou nezávislé, tak platí:

$$P(X = x \,\&\, Y = k - x) = P(X = x) \cdot P(Y = k - x).$$

Označíme-li pravděpodobnostní funkce X a Y jako p_X a p_Y, pak můžeme přepsat:

$$p_S(k) = \sum_{x \in Im(X)} p_X(x) \cdot p_Y(k - x)$$

Tím je věta dokázána.