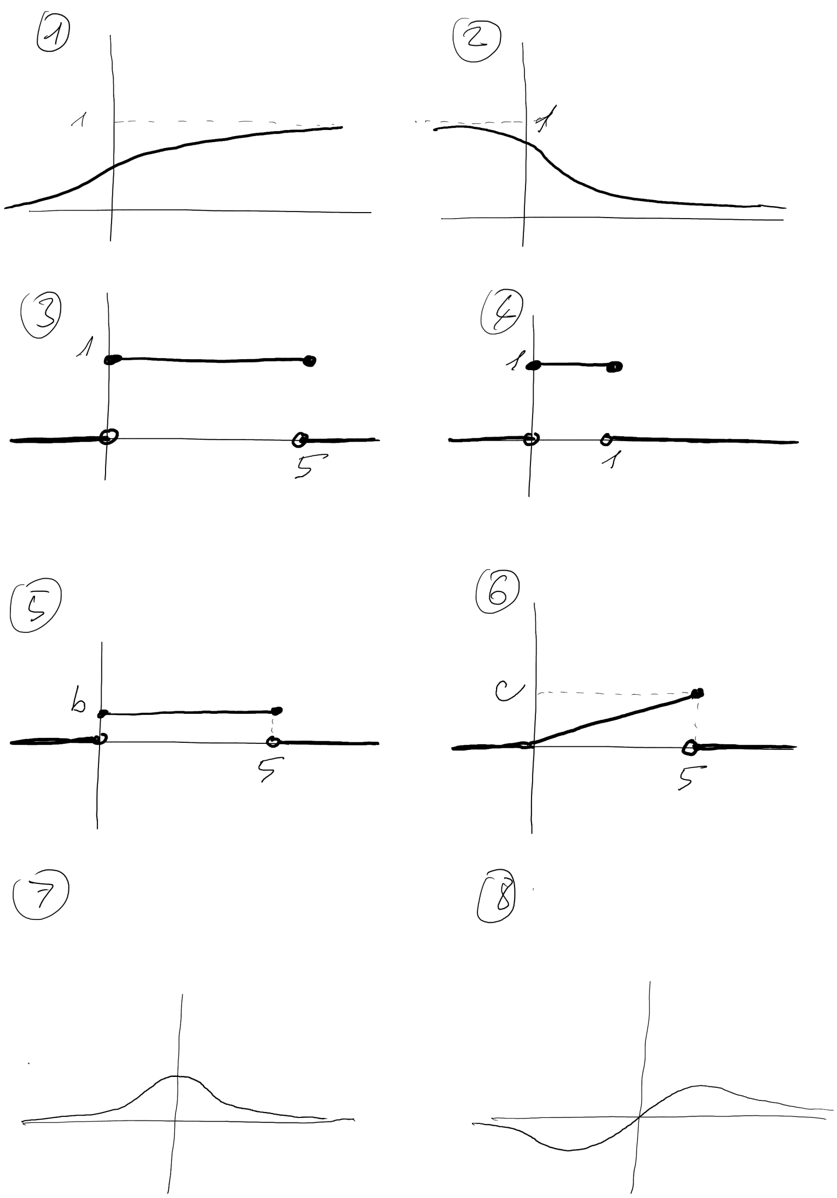
2. zkoušková písemka NMAI059 Pravd. a Stat. 1 – řešení – 22.6.2021

1. (10 bodů) (a) Rozhodněte, který z obrázků může popisovat hustotu nějaké náhodné veličiny. U obrázků 5 a 6 zvolte vhodnou hodnotu b, c, aby funkce byla hustota, je-li to možné. Další dvě části provádějte jen u těch obrázků, které hustotu zobrazují.

(b) Odhadněte střední hodnotu příslušného rozdělení.

(c) Seřaďte rozdělení podle hodnoty rozptylu. Tuto část dělejte jen pro obrázky 3–6, resp. ty z nich, které zobrazují hustotu.



Řešení:

(a) Víme, že hustota je nezáporná funkce, která má integrál přes celé $\mathbb{R}$ rovný jedné. Označme f_i funkci na i-tém obrázku a X_i příslušnou náhodnou veličinu.

Funkce f_8 není nezáporná, integrály $\int_{-\infty}^{\infty} f_1$ i $\int_{-\infty}^{\infty} f_2$ jsou nekonečné a $\int_{-\infty}^{\infty} f_3 = 5$. Tyto čtyři funkce tedy hustotou být nemohou.

$\int_{-\infty}^{\infty} f_4 = 1$, takže se jedná o hustotu (a to uniformního rozdělení $U(0,1)$.

$\int_{-\infty}^{\infty} f_5 = 5b$ (můžeme počítat integrál, nebo obsah obdélníku) takže se jedná o hustotu, pokud položíme $b = 1/5$. Jedná se o uniformní rozdělení $U(0,5)$.

$\int_{-\infty}^{\infty} f_6 = 5c/2$ (můžeme počítat integrál, nebo obsah trojúhelníku) takže se jedná o hustotu, pokud položíme $c = 2/5$.

$\int_{-\infty}^{\infty} f_7$ se z obrázku poznat nedá, ale může to být jedna, takže se může jednat o hustotu. Dost možná se jedná o standardní normální rozdělení.

(b) Střední hodnota $U(a,b)$ je $(a+b)/2$. Je tedy $\mathbb{E}(X_4) = 1/2$ a $\mathbb{E}(X_5) = 5/2$. Zdá se, že funkce f_7 je sudá, takže $\mathbb{E}(X_7) = 0$ (nebo také $\mathbb{E}(X_7)$ neexistuje, pokud se jedná např. o Cauchyho rozdělení). Pro $\mathbb{E}(X_6)$ použijeme vzorec z definice

$$\mathbb{E}(X_6) = \int_{-\infty}^{\infty} x \cdot f_6(x) dx = \int_0^5 x \cdot \frac{2}{25} x dx = \Big[\frac{2}{25}\frac{x^3}{3}\Big]_0^5 = \frac{10}{3}.$$

Mohli bychom také použít geometrické znalosti o těžišti trojúhelníka (těžiště je ve třetině těžnice), se stejným výsledkem.

(c) Pokud si pamatujeme vzorec, tak víme rovnou, že $\operatorname{var}(X_4) = (1-0)^2/12 = 1/12$ a $\operatorname{var}(X_5) = (5-0)^2/12 = 25/12$. (I bez výpočtu by ale mělo být jasné, že X_4 má 25-krát menší rozptyl než X_5.) X_6 vypadá trochu více koncentrovaná než X_5, ověříme výpočtem:

$$\mathbb{E}(X_6^2) = \int_{-\infty}^{\infty} x^2 \cdot f_6(x) dx = \int_0^5 x^2 \cdot \frac{2}{25} x dx = \Big[\frac{2}{25}\frac{x^4}{4}\Big]_0^5 = \frac{25}{2}.$$

Rozptyl je tedy $\operatorname{var}(X_6) = \mathbb{E}(X_6^2) - \mathbb{E}(X_6)^2 = 25/2 - 100/9$. Což je od oka něco přes 1 (přesně $25/18 = 1.388\dots$). Žádané pořadí je tedy $\operatorname{var}(X_4) < \operatorname{var}(X_6) < \operatorname{var}(X_5)$.

2. (10 bodů) V osudí je sto míčků s čísly 1, 2, …, 100. Vytáhneme tři z nich (bez vracení).

(a) Jaká je pravděpodobnost, že budou všechny mít číslo nejvýše rovné 40?

(b) Jaká je střední hodnota součtu čísel na vytažených míčcích?

(c) Jaká je střední hodnota počtu vytažených míčků, jejichž číslo je nejvýše rovno 40?

Řešení:

(a) Použijeme vzorec pro postupné podmiňování. Pokud A_i znamená jev „i-tý míček má číslo nejvýše 40“, tak

$$P(A_1 \cap A_2 \cap A_3) = P(A_1) P(A_2 \mid A_1) P(A_3 \mid A_1 \cap A_2) = \frac{40}{100} \cdot \frac{39}{99} \cdot \frac{38}{98}.$$

(b) Buď X_i číslo na i-tém míčku. Když odhlédneme od ostatních míčků, tak je X_i uniformně náhodné číslo od 1 do 100, tedy $\mathbb{E}(X_i) = (1+100)/2$. Veličiny X_1, X_2, X_3 *nejsou* nezávislé (když vytáhneme míček s číslem 1, tak ho nemůžeme vytáhnout podruhé, takže podmíněná střední hodnota dalších tažených míčků je trochu vyšší). Nicméně to linearita střední hodnoty nevyžaduje, je tedy

$$\mathbb{E}(X_1 + X_2 + X_3) = \mathbb{E}(X_1) + \mathbb{E}(X_2) + \mathbb{E}(X_3) = 303/2.$$

Mohli bychom počítat i se závislostí X_2 na X_1, tj. určit střední hodnotu $\mathbb{E}(X_2 \mid X_1 = 1)$, atd. V důsledku to znamená zjistit průměr všech součtů tříprvkových podmnožin $\{1, \dots, 100\}$ pomocí trojité sumy. Je to ale zbytečné a zdlouhavé.

(c) Pro pohodlí obarvíme míčky s číslem 1, …, 40, červeně, Zajímá nás tedy R, počet vytažených červených míčků. Jedná se o hypergeometrické rozdělení, podle vzorce pro jeho střední hodnotu je $\mathbb{E}(R) = 3 \cdot \frac{40}{100} = 1.2$.

Pokud bychom na hypergeometrické zobrazení zapomněli, můžeme počítat i přimočaře podle definice:

$$\mathbb{E}(R) = 3 \cdot P(R = 3) + 2 \cdot P(R = 2) + 1 \cdot P(R = 1) + 0 \cdot P(R = 0).$$

Podle části (a) je $P(R = 3) = \frac{40}{100} \cdot \frac{39}{99} \cdot \frac{38}{98}$. Další možnosti jsou trochu pracnější. Pokud $R = 2$, máme tři možnosti, který míček nebyl červený, pro každou z nich je stejná pravděpodobnost. Tj. $P(R = 2) = 3 \cdot \frac{40}{100} \cdot \frac{39}{99} \cdot \frac{60}{98}$. Analogicky $P(R = 1) = 3 \cdot \frac{40}{100} \cdot \frac{60}{99} \cdot \frac{59}{98}$. Dosazení do vzorce pro $\mathbb{E}(R)$ dává výsledek. Vyčíslovat při zkoušce netřeba, ale vyjde stejných 1.2.

Pak jsou další, trikovější (ale výpočetně snazší) způsoby, jak spočítat $\mathbb{E}(R)$, využívající postupu, kterým jsme na přednášce odvodili obecný vzorec pro střední hodnotu hypergeometrického rozdělení. Klíčem je vyjádřit R jakou součet tří indikátorů, popisujících, zda první, druhý, třetí míček je červený, a pak využít linearitu střední hodnoty.

3. (10 bodů) Doba trvání ústní zkoušky má exponenciální rozdělení se střední hodnotou 20 minut. Objednáni jsou dva studenti, jeden na 10:00, druhý na 10:20. Pokud se zkoušení prvního protáhne, zkoušení druhého začne až bude první hotový, jinak začne přesně v 10:20. Jaká je střední hodnota času, kdy bude druhý student dozkoušený?

Řešení: Označme X dobu zkoušení prvního studenta, Y dobu pro druhého studenta. Máme zadáno, že $X, Y \sim Exp(\lambda)$ pro $\lambda = 1/20$ (střední hodnota je $1/\lambda$), Považujme X, Y za nezávislé.

Výpočet rozdělíme podle toho, jestli druhý student musí čekat, nebo ne. Pokud ne, což nastane s pravděpodobností $P(X \le 20) = 1 - e^{-\lambda \cdot 20} = 1 - e^{-1}$, tak celkový čas je $S = 20 + Y$ a tedy $\mathbb{E}(S \mid X \le 20) = 20 + \mathbb{E}(Y) = 40$.

Pokud druhý student musí čekat, tak $X > 20$, což nastane s pravděpodobností e^{-1}. Využijeme toho, že exponenciální rozdělení nemá paměť, a proto $X - 20 | X > 20$ má zase stejné exponenciální rozdělení $Exp(1/20)$. V takovém případě je tedy

$$\mathbb{E}(S \mid X > 20) = 20 + \mathbb{E}(X \mid X > 20) + \mathbb{E}(Y) = 20 + 20 + 20 = 60.$$

Podle věty o rozboru případů (o úplné střední hodnotě) je

$$\mathbb{E}(S) = P(X \le 20) \cdot \mathbb{E}(S \mid X \le 20) + P(X > 20) \cdot \mathbb{E}(S \mid X > 20) = (1 - e^{-1}) \cdot 40 + e^{-1} \cdot 60 = 40 + 20/e.$$

To je, pro zajímavost, rovno přibližně 47.3.

4. (10 bodů) (a) Definujte pojem nezávislost náhodných jevů.
(b) Definujte pojem podmíněná střední hodnota diskrétní náhodné veličiny.

Řešení:

(a) Říkáme, že jevy A_i, $i \in I$ jsou nezávislé, pokud pro každou konečnou množinu $S \subseteq I$ platí

$$P(\bigcap_{i \in S} A_i) = \prod_{i \in S} P(A_i).$$

(b) Pokud $X : \Omega \to \mathbb{R}$ je náhodná veličina na pravděpodobnostním prostoru $(\Omega, \mathcal{F}, P)$ a $A \in \mathcal{F}$, tak definujeme

$$\mathbb{E}(X \mid A) = \sum_{x \in ImX} x \cdot \Pr(X = x|A).$$

5. (10 bodů) Vysvětlete princip testování hypotéz. (Objasněte mimo jiné, co je to chyba 1. a 2. druhu.)

Řešení: Chceme testovat platnost nějakého tvrzení, např. zda je mince spravedlivá. Označíme jako H_0 to tvrzení, které považujeme za očekávané, tzv. nulovou hypotézu. Alternativní hypotézu značíme H_1 (v přednásce jsme uvažovali jen případ dvou hypotéz, kde H_1 je negace H_0).

Vytvoříme statistický model situace (pokud házíme n-krát mincí, tak typicky předpokládáme, že jednotlivé hody jsou nezávislé, a příslušný model je binomické rozdělení $Bin(n, \vartheta)$. Zde ϑ je neznámý parametr a nulová hypotéza říká, že $\vartheta = 1/2$.

Před naměřením dat se rozhodneme, jaký bude tzv. kritický obor, tj. množina W možných měření, při kterých H_0 zamítneme. Tuto množinu můžeme volit více způsoby, ale chceme zajistit, aby byla předem daná pravděpodobnost chyby 1. druhu: chybného zamítnutí. (Nechceme moc často hlásat zásadní objevy kvůli náhodnému kolísání dat.) Omezení pravděpodobnosti chyby 1. druhu značíme α.

Zároveň se snažíme o co nejmenší pravděpodobnost chyby 2. druhu, neboli chybného přijetí (značíme β). Pravděpodobnost, že chybnou hypotézu zamítneme nazýváme síla testu.

6. (10 bodů) Vyslovte a dokažte větu o konvolučním vzorci (součet nezávislých náhodných veličin), případ diskrétních náhodných veličin.

Řešení:

Věta:

Pokud X, Y jsou nezávislé diskrétní náhodné veličiny, tak pro $Z = X + Y$ platí

$$P(Z = z) = \sum_{x \in ImX} P(X = x) P(Y = z - x).$$

Důkaz: Podle věty o úplné pravděpodobnosti je

$$P(Z = z) = \sum_{x \in Im(X)} P(X = x) \cdot P(Z = z \mid X = x).$$

Podle definice Z je $P(Z = z \mid X = x) = P(Y = z - x \mid X = x)$. Z nezávislosti X a Y dostáváme, že $P(Y = z - x \mid X = x) = P(Y = z - x)$. Po dosazení máme důkaz hotov.