

Malé statistické repetitorium

Verze s řešením

Příklad 1: Rozdělení náhodné veličiny, základní charakteristiky

Rozdělení diskrétní náhodné veličiny X je dáno následující tabulkou.

x	0	1	4	5
$P(X=x)$	0,05	0,5	0,25	0,2

- a) Nakreslete graf distribuční funkce veličiny X (pozn.: pamatujte, že $F(x) = P(X \leq x)$).
- b) Spočítejte EX .
- c) Spočítejte $\text{var } X$.

Řešení.

- a) Distribuční funkce diskrétní náhodné veličiny je neklesající „schodovitá“ funkce. V bodech blízkých $-\infty$ nabývá hodnoty 0, poblíž $+\infty$ hodnoty 1. Průběh mezi těmito krajními možnostmi lze popsat tak, že hodnota distribuční funkce se mění pouze v bodech, kterých může veličina X nabýt (zde v bodech x z množiny $\{0, 1, 4, 5\}$); v těchto bodech dochází ke skoku o $P(X=x)$. Obrázek pro konkrétní zadání **1a** ponecháme na čtenáři.

- b) Střední hodnotu vypočteme podle známého vzorce:

$$EX = \sum x \cdot P(X=x) = 0 \cdot 0,05 + 1 \cdot 0,5 + 4 \cdot 0,25 + 5 \cdot 0,2 = 2,5.$$

- c) Pro výpočet rozptylu bychom měli znát následující vztahy:

$$\text{var } X = E[(X - EX)^2] = E[X^2] - (EX)^2,$$

dopočítáme zvlášť

$$E[X^2] = \sum x^2 \cdot P(X=x) = 0^2 \cdot 0,05 + 1^2 \cdot 0,5 + 4^2 \cdot 0,25 + 5^2 \cdot 0,2 = 9,5,$$

celkem máme

$$\text{var } X = E[X^2] - (EX)^2 = 9,5 - 2,5^2 = 3,25.$$

Příklad 2: Lineární kombinace dvou nezávislých náhodných veličin

Jsou dány *nezávislé* náhodné veličiny X a Y , přičemž:

$$EX = 10, \quad \text{var } X = 1,$$

$$EY = 5, \quad \text{var } Y = 2.$$

Spočítejte:

- a) $E[4X]$.
- b) $E[X+5]$.
- c) $E[X+Y]$.
- d) $E[4X+3Y]$.
- e) $\text{var}[4X]$.
- f) $\text{var}[X+5]$.
- g) $\text{var}[X+Y]$.
- h) $\text{var}[4X+3Y]$.

Řešení.

- a) Měli bychom vědět, že konstanta lze „vytknout“ před střední hodnotu následujícím způsobem:

$$E[4X] = 4EX = 40.$$

Pro diskrétní náhodnou veličinu je to vidět přímo z výše uvedeného vzorce pro výpočet střední hodnoty – vytkneme konstantu před sumu:

$$E[4X] = \sum 4x \cdot P(X=x) = 4 \sum x \cdot P(X=x) = 4EX.$$

Chceme-li ověřit tento vztah pro spojité náhodné veličiny (zatím jsme s nimi zde nepracovali), musíme si nejprve uvědomit, jak je definována jejich střední hodnota. Potom plyne vztah $E[4X] = 4EX$ z vlastností integrálu (konkrétně z jeho linearity):

$$E[4X] = \int_{-\infty}^{\infty} 4x f(x) dx = 4 \int_{-\infty}^{\infty} x f(x) dx = 4EX.$$

- b) Podobnou úvahou jako v předchozím bodě (nebo prostou intuicí) lze snadno rozmyslet, že $E[X+5] = EX + 5 = 15$.
- c) Platí vztah, který lze slovně popsat větou „*střední hodnota součtu náhodných veličin je vždy rovna součtu jejich středních hodnot*“, neboli: $E[X+Y] = EX + EY = 15$. Tento vztah se ověří analogicky jako vztah z bodu **2a**.

- d) Nejprve použijeme postup z bodu 2c na náhodné veličiny $4X$ a $3Y$, dál postupujeme jako v 2a. Získáme postupně

$$E[4X + 3Y] = E[4X] + E[3Y] = 4EX + 3EY = 55.$$

- e) Pro rozptyl náhodných veličin platí podobné tvrzení, jako pro střední hodnotu (viz opět 2a) s tím, že tentokrát „vytýkáme“ konstantu s druhou mocninou:

$$\text{var}[4X] = 4^2 \text{var} X.$$

- f) Velmi zhruba řečeno, rozptyl veličiny nám říká, „jak daleko od sebe“ bývají hodnoty při různých realizacích uvažované náhodné veličiny. Pokud k náhodné veličině přičteme konstantu, posunou se (stejně) všechny hodnoty náhodné veličiny, jejich vzájemná vzdálenost se ovšem nezmění. Musí tedy platit

$$\text{var}[X + 5] = \text{var} X.$$

Tento vztah lze samozřejmě snadno formálně ověřit z definice rozptylu a z toho, co už víme o střední hodnotě z 2b:

$$\text{var}[X + 5] = E[(X + 5 - E[X + 5])^2] = E[(X + 5 - EX - 5)^2] = E[(X - EX)^2] = \text{var} X.$$

- g) Obecně platí vztah

$$\text{var}[X + Y] = \text{var} X + \text{var} Y + 2\text{cov}(X, Y)$$

(o kovarianci dvou náhodných veličin $\text{cov}(X, Y)$ pojednává příklad 3d) Vzhledem k tomu, že jsou veličiny nezávislé, jsou také nekorelované. V takovém případě můžeme psát: $\text{var}[X + Y] = \text{var} X + \text{var} Y = 3$. Podrobněji viz příklad 3d.

- h) Kombinací dvou předchozích bodů máme $\text{var}[4X + 3Y] = 4^2 \text{var} X + 3^2 \text{var} Y = 34$. Poznamenejme, že jsou-li nezávislé veličiny X a Y , jsou nezávislé i veličiny $4X$ a $3Y$; díky tomu můžeme uvedený postup použít. Obecně platí, že z nezávislosti X a Y plyne nezávislost veličin $f(X)$ a $g(Y)$, kde f a g jsou libovolné nekonstantní reálné funkce.

Příklad 3: Sdružené dvojrozměrné rozdělení

Sdružené rozdělení náhodných veličin X a Y udává následující tabulka. Nejprve rozhodněte, zda jsou veličiny nezávislé, a pak spočítejte:

- $E[X|Y=1]$ a $E[Y|X=1]$.
- EX a EY .
- $\text{var} X$ a $\text{var} Y$.
- $\text{cov}(X, Y)$.
- $\text{var}[X + Y]$.

$P(X=x,Y=y)$		y		
		0	1	4
x	0	0	$\frac{1}{8}$	$\frac{1}{2}$
	1	$\frac{1}{8}$	$\frac{1}{4}$	0

Řešení.

Nezávislost X a Y : aby byly (diskrétní) veličiny X a Y nezávislé, muselo by pro každou kombinaci hodnot x a y platit

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

V našem zadání ovšem toto splněno není – například pro kombinaci hodnot $x=0$ a $y=0$, jak vzápětí ukážeme. Nejprve si rozmysleme, jak v tabulce sdruženého dvojrozměrného rozdělení vyčíslit „jednorozměrnou“ pravděpodobnost (tzv. *marginální pravděpodobnost*) $P(X=0)$. V tomto případě nám jakoby vůbec nezáleží na tom, jak dopadne Y a snadno si rozmyslíme, že nám stačí sečíst pravděpodobnosti všech kombinací x a y , kde má x hodnotu 0, tedy

$$P(X=0) = P(X=0, Y=0) + P(X=0, Y=1) + P(X=0, Y=4) = 0 + \frac{1}{8} + \frac{1}{2} = \frac{5}{8}.$$

Obdobně vypočteme, že $P(Y=0) = 0 + \frac{1}{8} = \frac{1}{8}$. Celkem tedy můžeme ověřit, že

$$0 = P(X=0, Y=0) \neq P(X=0) \cdot P(Y=0) = \frac{5}{8} \cdot \frac{1}{8} = \frac{5}{64},$$

veličiny X a Y tedy nejsou nezávislé.

O závislosti veličin lze rozhodnout i intuitivním způsobem. Obecně platí následující myšlenka: pokud nám informace o tom, jak dopadlo X , nějak pomůže při odhadu výsledné hodnoty Y , veličiny jsou závislé. Zde například: dozvíme-li se, že výsledná hodnota X je 0, můžeme vyloučit, že Y má rovněž hodnotu 0 (viz tabulka).

- a) Připomeňme nejprve pojem *podmíněné pravděpodobnosti*. Pokud víme, že veličina Y nabyla hodnoty 1, a zajímá nás pravděpodobnost, že X bude současně 0, mluvíme o podmíněné pravděpodobnosti $P(X=0|Y=1)$, kterou můžeme ze zadané tabulky snadno spočítat ze vztahu

$$P(X=0|Y=1) = \frac{P(X=0, Y=1)}{P(Y=1)} = \frac{\frac{1}{8}}{\frac{1}{8} + \frac{1}{4}} = \frac{1}{3},$$

Analogicky zjistíme, že $P(X=1|Y=1) = \frac{2}{3}$. Podmíněná střední hodnota $E[X|Y=1]$ se dopočítá podobně jako střední hodnota nepodmíněná (viz 1b) s tím, že namísto nepodmíněných pravděpodobností používáme podmíněné:

$$E[X|Y=1] = \sum x \cdot P(X=x|Y=1) = 0 \cdot \frac{1}{3} + 1 \cdot \frac{2}{3} = \frac{2}{3}.$$

Pro $E[Y|X=1]$ řešíme analogicky, vyjde opět $\frac{2}{3}$.

- b)** Střední hodnoty počítáme analogicky jako v **1b**, přičemž marginální pravděpodobnosti $P(X=x)$ a $P(Y=y)$ počítáme podle výše uvedeného postupu. Vyjde

$$EX = 0 \cdot \frac{5}{8} + 1 \cdot \frac{3}{8} = \frac{3}{8},$$

$$EY = 0 \cdot \frac{1}{8} + 1 \cdot \frac{3}{8} + 4 \cdot \frac{1}{2} = \frac{19}{8}.$$

- c)** Jde opět o běžný výpočet rozptylu jako v **1c**, práci navíc dá (podobně jako v minulém bodě) jedině výpočet marginálních pravděpodobností. Vyjde:

$$\text{var } X = \frac{15}{64},$$

$$\text{var } Y = \frac{237}{64}.$$

- d)** Kovariance náhodných veličin X a Y , značená $\text{cov}(X, Y)$, měří jistým způsobem lineární závislost mezi veličinami X a Y . Je-li tato závislost pozitivní, potom očekáváme, že veličiny mají tendenci „vyvíjet se stejným směrem,“ tj. při velkých hodnotách X očekáváme spíše velké hodnoty Y a opačně; v takovém případě má kovariance kladné znaménko. V naší tabulce je tomu spíše naopak: malá hodnota v proměnné X (tj. $X=0$) je spojena s velkými hodnotami Y . Závislost je tedy negativní a budeme očekávat záporné znaménko. Kovarianci definujeme pomocí střední hodnoty jako

$$\text{cov}(X, Y) = E[(X - EX)(Y - EY)] = E[XY] - EX \cdot EY.$$

Hodnoty EX a EY už známe z **3b**, stačí tedy dopočítat $E[XY]$. Platí:

$$E[XY] = \sum xy \cdot P(X=x, Y=y) = \frac{1}{4},$$

přičemž jednotlivé sčítance v uvedené sumě lze snadno spočítat ze zadané tabulky; snadno zjistíme, že jsou všechny kromě výrazu $1 \cdot 1 \cdot P(X=1|Y=1)$ nulové. Celkem tedy máme

$$\text{cov}(X, Y) = E[XY] - EX \cdot EY = \frac{1}{4} - \frac{3}{8} \cdot \frac{19}{8} = -\frac{41}{64}.$$

- e)** Vztah pro výpočet rozptylu náhodných veličin je uveden u řešení příkladu **2g**, po dosazení máme:

$$\text{var}[X+Y] = \text{var } X + \text{var } Y + 2\text{cov}(X, Y) = \frac{15}{64} + \frac{237}{64} + 2 \cdot \frac{-41}{64} = \frac{170}{64}.$$

Příklad 4: Podmíněná střední hodnota

Uvažujme opět náhodné veličiny X a Y z příkladu 3. V bodě **3a** jsme hledali střední hodnotu veličiny Y za podmínky, že víme, že veličina X nabyla hodnoty 0. Obecně bychom mohli hledat střední hodnotu Y při libovolné známé hodnotě x veličiny X . Zapište předpis funkce

$$f(x) = E[Y|X=x].$$

Funkci $f(x)$ se říká *podmíněná střední hodnota Y při daném X* a značí se často jako $E[Y|X]$.

Řešení.

Funkce $f(x)$ je definována pouze pro x rovno 0 nebo 1. Vzhledem k **3a** nám stačí dopočítat $E[Y|X=0] = \frac{17}{5}$, potom zapišeme funkci $f(x)$ ve tvaru

$$f(x) = \begin{cases} E[Y|X=0] & \text{pro } x=0, \\ E[Y|X=1] & \text{pro } x=1, \end{cases}$$

neboli po dosazení vypočítaných hodnot

$$f(x) = \begin{cases} \frac{17}{5} & \text{pro } x=0, \\ \frac{2}{3} & \text{pro } x=1. \end{cases}$$

Příklad 5: Sdružené dvojrozměrné rozdělení

Pomocí náhodných veličin X a Y z příkladu 3 definujeme *vektorovou náhodnou veličinu* (nebo též *náhodný vektor*) $\mathbf{Z}$ předpisem

$$\mathbf{Z} = \begin{bmatrix} X \\ Y \end{bmatrix} = (X, Y)^T.$$

Dále definujeme (konstantní) matici $\mathbf{A}$ jako

$$\mathbf{A} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

Pomocí výsledků z příkladu 3 určete

- a)** $E\mathbf{Z}$ a $\text{var}\mathbf{Z}$.
b) $E[\mathbf{AZ}]$ a $\text{var}[\mathbf{AZ}]$.

Řešení.

- a) Střední hodnota vektorové náhodné veličiny (resp. náhodného vektoru) je definována jako vektor středních hodnot jednotlivých složek náhodného vektoru. Formálně zapsáno pro náš případ:

$$\mathbf{EZ} = \begin{bmatrix} EX \\ EY \end{bmatrix} = \begin{bmatrix} \frac{3}{8} \\ \frac{19}{8} \end{bmatrix} = \frac{1}{8} \begin{bmatrix} 3 \\ 19 \end{bmatrix},$$

numerické hodnoty obou složek jsme spočítali v příkladu 3b. Výraz „var $\mathbf{Z}$ “ se v případě vícerozměrných náhodných veličin nechte jako „rozptyl“, nýbrž představuje tzv. *varianční matici* náhodného vektoru $\mathbf{Z}$, která vypadá následovně (hodnoty jsou opět dosazeny z výpočtů v příkladech 3c a 3d):

$$\text{var}\mathbf{Z} = \begin{bmatrix} \text{var} X & \text{cov}(X,Y) \\ \text{cov}(X,Y) & \text{var} Y \end{bmatrix} = \frac{1}{64} \begin{bmatrix} 15 & -41 \\ -41 & 237 \end{bmatrix}.$$

- b) Násobíme-li náhodný vektor $\mathbf{Z}$ konstantní maticí $\mathbf{A}$, výsledkem je opět náhodný vektor, který má tentokrát 3 složky, jak se můžeme snadno přesvědčit:

$$\mathbf{AZ} = \begin{bmatrix} X+Y \\ Y \\ X \end{bmatrix}.$$

Charakteristiky veličiny $\mathbf{AZ}$ (tj. její střední hodnotu a varianční matici) můžeme počítat přímo z definice, nebo můžeme využít následujících vztahů, které jsou analogií ke vztahům pro násobení jednorozměrné náhodné veličiny reálným číslem:

$$E[\mathbf{AZ}] = \mathbf{A}(E\mathbf{Z}),$$

$$\text{var}[\mathbf{AZ}] = \mathbf{A}(\text{var}\mathbf{Z})\mathbf{A}^T.$$

Dosadíme-li do těchto vztahů výsledky z bodu a, dostaneme:

$$E[\mathbf{AZ}] = \frac{1}{8} \begin{bmatrix} 22 \\ 19 \\ 3 \end{bmatrix},$$

$$\text{var}[\mathbf{AZ}] = \frac{1}{64} \begin{bmatrix} 170 & 196 & -26 \\ 196 & 237 & -41 \\ -26 & -41 & 15 \end{bmatrix}.$$

Příklad 6: Normální rozdělení

Náhodné veličiny X a Y nechť jsou nezávislé a mají obě *normované normální rozdělení*. Připomeňme, že zápis $N(\mu; \sigma^2)$ představuje normální rozdělení se střední hodnotou μ a rozptylem σ^2 a že normované rozdělení je rozdělení $N(0;1)$. Definujme pomocí těchto veličin vektor $\mathbf{Z}$ opět předpisem $\mathbf{Z} = (X,Y)^T$. Určete pravděpodobnostní rozdělení následujících náhodných veličin:

- a) $3X+5$.
 b) $[1 \ 1]\mathbf{Z} = X+Y$.
 c) $(X+Y)/2$.
 d) $\mathbf{Z}$.
 e) $\mathbf{AZ}$, kde $\mathbf{A} = \begin{bmatrix} 1 & -1 \\ 2 & 0 \\ 0 & 1 \end{bmatrix}$.

Řešení.

- a) Měli bychom vědět, že má-li veličina X normální rozdělení, bude mít normální rozdělení i veličina $3X+5$ (tedy násobení konstantou ani její přičítání „nenaruší normalitu“). Konkrétní tvar normálního rozdělení je jednoznačně určen střední hodnotou a rozptylem, které v tomto případě snadno určíme – stačí si vzpomenout na příklad 2. Celkem můžeme psát $3X+5 \sim N(5;9)$.
 b) Obdobně platí, že součet dvou (nebo více) *nezávislých normálně rozdělených* náhodných veličin má opět normální rozdělení s příslušnými parametry. Zde po dosazení $X+Y \sim N(0;2)$.
 c) Spojením bodů a a b zjistíme, že jakákoli *lineární kombinace* nezávislých normálně rozdělených náhodných veličin má opět normální rozdělení. Pro výpočet v tomto příkladě můžeme použít výsledek z bodu b a dostaneme $(X+Y)/2 \sim N(0;1/2)$.
 d) Měli bychom ukázat, že náhodný vektor $\mathbf{Z}$ má *dvojrozměrné normální rozdělení*. Podle definice vícerozměrného normálního rozdělení má $\mathbf{Z}$ dvojrozměrné normální rozdělení, pokud každá lineární kombinace jeho složek je normálně rozdělená. Jak jsme ukázali v předchozích bodech, tato vlastnost je zaručena tím, že veličiny X a Y jsou nezávislé. Můžeme tedy psát

$$\mathbf{Z} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right) = N(\mathbf{0}; \mathbf{I}),$$

kde $\mathbf{0}$, resp. $\mathbf{I}$ představuje nulový vektor, resp. jednotkovou matici potřebných rozměrů.

- e) Budeme opět ukazovat, že vektor $\mathbf{Z}$ má vícerozměrné normální rozdělení, tentokrát však trojrozměrné, neboť

$$\mathbf{AZ} = \begin{bmatrix} X - Y \\ 2X \\ Y \end{bmatrix}$$

je trojrozměrný vektor. Opět vyjdeme z definice vícerozměrného normálního rozdělení. Libovolná lineární kombinace složek vektoru $\mathbf{AZ}$ vypadá následovně (označili jsme a, b a c koeficienty lineární kombinace):

$$a(X - Y) + b(2X) + cY = (a + 2b)X + (c - a)Y,$$

neboli libovolná lineární kombinace složek náhodného vektoru $\mathbf{AZ}$ lze zapsat jako lineární kombinace X a Y , a má tudíž (jak jsme ukázali výše) normální rozdělení. Abychom určili tvar rozdělení veličiny $\mathbf{AZ}$, je třeba dopočítat, jak vypadá $E[\mathbf{AZ}]$ a $\text{var}[\mathbf{AZ}]$. To provedeme zcela analogicky jako v příkladě 5b. Celkem můžeme zapsat:

$$\mathbf{AZ} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 2 & 1 & -1 \\ 1 & 2 & 0 \\ -1 & 0 & 1 \end{bmatrix}\right).$$

Pozn.: Obecně lze říci, že má-li náhodný vektor $\mathbf{X}$ n -rozměrné normální rozdělení a $\mathbf{B}$ je konstantní matice typu $(m \times n)$, má náhodný vektor $\mathbf{BX}$ m -rozměrné normální rozdělení.

Příklad 7: Některé transformace normálně rozdělených veličin

Náhodné veličiny $X_1, X_2, \dots, X_{10}$ jsou nezávislé a mají všechny normované normální rozdělení. Definujme náhodný vektor $\mathbf{X}$ předpisem $\mathbf{X} = (X_1, X_2, \dots, X_{10})^T$. Určete pravděpodobnostní rozdělení náhodných veličin

a) $Y = \frac{X_1 + X_2 + \dots + X_{10}}{10} = \frac{1}{10} \sum_{i=1}^{10} X_i = \frac{1}{10} \underbrace{\begin{bmatrix} 1 & 1 & \dots & 1 \end{bmatrix}}_{10 \text{ složek}} \mathbf{X}.$

b) $Z = X_1^2 + X_2^2 + \dots + X_{10}^2 = \sum_{i=1}^{10} X_i^2 = \mathbf{X}^T \mathbf{X}.$

c) $T = \frac{10Y}{\sqrt{Z}}.$

Řešení.

Náhodné veličiny $X_1, X_2, \dots, X_{10}$ jsou nezávislé a mají všechny normované normální rozdělení. Definujme náhodný vektor $\mathbf{X}$ předpisem $\mathbf{X} = (X_1, X_2, \dots, X_{10})$. Určete pravděpodobnostní rozdělení náhodných veličin

- a) Z příkladu 6 už víme, že veličina Y má normální rozdělení, stačí už jen spočítat jeho střední hodnotu a rozptyl. Vzhledem k nulové střední hodnotě $X_1, X_2, \dots, X_{10}$ je určitě $EY = 0$, rozptyl dopočteme snadno, když si uvědomíme, že díky nezávislosti veličin platí:

$$\text{var}\left[\frac{1}{10} \sum_{i=1}^{10} X_i\right] = \left(\frac{1}{10}\right)^2 \cdot \text{var}\left[\sum_{i=1}^{10} X_i\right] = \left(\frac{1}{10}\right)^2 \cdot \sum_{i=1}^{10} \text{var} X_i = \frac{1}{10}.$$

Celkem tedy zapíšeme $Y \sim N(0; \frac{1}{10})$.

- b) Zde není co počítat, pokud si pamatujete následující vztah: *mají-li nezávislé stejné rozdělené veličiny $X_1, X_2, \dots, X_n$ rozdělení $N(0;1)$, součet jejich druhých mocnin má chí-kvadrát rozdělení s n stupni volnosti*. Zapíšeme tedy $Z \sim \chi^2(10)$.
- c) Jde o podobný případ jako v b, správnou odpověď je opět jedno z „klasických“ spojitých rozdělení, tentokrát však jde o *studentovo t-rozdělení*. Obecně platí vztah: *jsou-li veličiny U a Z nezávislé s rozděleními $U \sim N(0;1)$ a $Z \sim \chi^2(n)$, potom veličina*

$$\frac{U}{\sqrt{Z}} \sqrt{n}$$

má studentovo t-rozdělení s n stupni volnosti (značené $t(n)$). Snadno ověříme, že pokud definujeme veličinu U předpisem $U = \sqrt{10} \cdot Y$, potom $U \sim N(0;1)$. Lze dále ukázat, že U a Z jsou nezávislé (to dá dost práce a nebudeme se tím zde zabývat). Celkem tedy máme U a Z nezávislé, $U \sim N(0;1)$, $Z \sim \chi^2(10)$. Odtud snadno ukážeme, že

$$\frac{U}{\sqrt{Z}} \sqrt{10} = \frac{10Y}{\sqrt{Z}} = T \sim t(10).$$

Příklad 8: Kvantilová funkce, interval spolehlivosti

- a) Najděte medián a 70% kvantil rozdělení náhodné veličiny X z příkladu 1.
- b) Najděte interval, do kterého spadne veličina Y z příkladu 7 na 95%. Přesně řečeno, najděte taková čísla l a u , aby platilo $P(Y \leq l) = P(Y \geq u) = 2,5\%$. Hledaný interval je potom (l, u) .
- c) Najděte obdobný interval pro veličinu T .

Řešení.

a) Připomeňme nejprve, co je to $(100p)$ -procentní kvantil náhodné veličiny. Jde-li o náhodnou veličinu, která má prostou a spojitou distribuční funkci (jako má například normální rozdělení), můžeme tento kvantil definovat jako číslo x_p , které splňuje vztah: $P(X \leq x_p) = p$. Pro ostatní veličiny definujeme $(100p)$ -procentní kvantil x_p jako nejmenší číslo, které splňuje vztah $P(X \leq x_p) \geq p$. Pro zadání **8a** můžeme tento vztah konkrétně přepsat jako $P(X \leq x_{0,7}) \geq 0,7$. Z tabulky, ve které je zadáno rozdělení náhodné veličiny X z příkladu 1, snadno zjistíme, že tento vztah splňují hodnoty x větší nebo rovny 4. Nejmenší z těchto hodnot je přirozeně číslo 4. Zapišme tedy $x_{0,7} = 4$. Obdobně spočítáme medián, čili 50% kvantil, jako $x_{0,5} = \tilde{x} = 1$.

b) Podle **7a** $Y \sim N(0;1)$, což je rozdělení se spojitou a prostou distribuční funkcí, číslo l je tedy přesně jeho 2,5% kvantil (viz bod **a**), který značíme $y_{0,025}$. Rozmysleme si nyní, jak bude vypadat číslo u . Platí, že $P(Y \geq u) = 1 - P(Y < u)$. Vzhledem k tomu, že u spojitých rozdělení je $P(Y = y) = 0$ pro libovolnou hodnotu y , platí také $P(Y < u) = P(Y \leq u)$, a vztah, kterým je v zadání definováno číslo u , můžeme přepsat jako

$$P(Y \leq u) = 1 - 0,025,$$

neboli u je vlastně 97,5% kvantil rozdělení veličiny Y , neboli $y_{0,975}$. Poznamenejme, že obecně se interval $(y_{0,025}; y_{0,975})$, tedy interval mezi 2,5% a 97,5% kvantilem, nazývá *oboustranným 95% intervalem spolehlivosti* pro veličinu Y .

Připomeňme si, jak najít kvantily normálního rozdělení pomocí statistických tabulek. V tabulkách najdeme pouze kvantily $N(0,1)$, které označujeme u_p , a to ještě pouze pro $p \geq 0,5$. Pomocí těchto tabulek najdeme snadno $u = y_{0,975}$, pokud si uvědomíme, že veličina Y lze vyjádřit pomocí veličiny $U \sim N(0;1)$ následovně

$$Y = \frac{U}{\sqrt{10}},$$

odkud vyjádříme

$$0,975 = P(Y \leq y_{0,975}) = P\left(\frac{U}{\sqrt{10}} \leq y_{0,975}\right) = P(U \leq \sqrt{10} \cdot y_{0,975}),$$

a tedy

$$\sqrt{10} \cdot y_{0,975} = u_{0,975}, \text{ neboli } y_{0,975} = \frac{1}{\sqrt{10}} \cdot u_{0,975}.$$

Hodnotu $u_{0,975}$ již v tabulkách snadno najdeme (je to přibližně 1,96) a hodnotu $y_{0,975}$ snadno dopočítáme. Zcela analogicky bychom ukázali, že

$$y_{0,025} = \frac{1}{\sqrt{10}} \cdot u_{0,025}.$$

Hodnotu $u_{0,025}$ sice v tabulkách nenajdeme, ale jelikož normované normální rozdělení je symetrické okolo nuly, musí platit, že

$$u_{0,025} = -u_{0,975}, \text{ tedy } y_{0,025} = -y_{0,975}$$

a nemusíme vlastně nic dalšího počítat. Hledaný interval je tedy přibližně

$$\left(-\frac{1}{\sqrt{10}} \cdot 1,96; \frac{1}{\sqrt{10}} \cdot 1,96\right).$$

c) Již víme, že veličina T má *studentovo t -rozdělení s 10 stupni volnosti*, a také víme, že budeme hledat jeho *oboustranný 95% interval spolehlivosti*, jehož hranice tvoří 2,5% a 97,5% kvantil příslušného rozdělení. Kvantily studentova rozdělení pro různé počty stupňů volnosti jsou opět tabelovány, $(100p)$ -procentní kvantil z rozdělení $t(n)$ značíme zpravidla $t_{p,n}$, můžeme tedy zapsat náš interval jako

$$(t_{0,025, 10}; t_{0,975, 10}) = (-t_{0,975, 10}; t_{0,975, 10}) = (-2,23; 2,23),$$

kde jsme využili toho, že i t -rozdělení je symetrické okolo nuly.

Příklad 9: Náhodný výběr z normálního rozdělení

Náhodné veličiny $X_1, X_2, \dots, X_{10}$ jsou nezávislé a pochází z rozdělení $N(\mu; \sigma^2)$. Označme symboly $\bar{X}$ a S^2 náhodné veličiny

$$\bar{X} = \frac{1}{10} \sum_{i=1}^{10} X_i,$$

$$S^2 = \frac{1}{9} \sum_{i=1}^{10} (X_i - \bar{X})^2.$$

Určete rozdělení následujících náhodných veličin:

a) $\bar{X}$.

b) $U = \frac{\bar{X} - \mu}{\sqrt{\frac{S^2}{10}}}.$

c) $Z = \frac{9S^2}{\sigma^2}.$

d) $T = \frac{\bar{X} - \mu}{\sqrt{\frac{S^2}{10}}}.$

Řešení.

- a) Veličina $\bar{X}$ je definována podobně jako veličina Y v příkladu 7a, rozdíl je pouze v tom, že veličiny X_i tentokrát pocházejí z obecného, nikoli normovaného normálního rozdělení. I zde bude zřejmě výsledkem normální rozdělení, jeho parametry budou $E\bar{X}$ a $\text{var } \bar{X}$, přičemž

$$E\bar{X} = \frac{1}{10} \sum_{i=1}^n E X_i = \frac{1}{10} \sum_{i=1}^n \mu = \mu,$$

$$\text{var } \bar{X} = \left(\frac{1}{10}\right)^2 \cdot \sum_{i=1}^n \text{var } X_i = \left(\frac{1}{10}\right)^2 \cdot \sum_{i=1}^n \sigma^2 = \frac{1}{10} \sigma^2.$$

Můžeme celkem shrnout $\bar{X} \sim N(\mu; \frac{1}{10} \sigma^2)$.

Pozn.: Představme si, že sledujeme statistický znak X , který je v populaci přibližně normálně rozdělen (např. výška jedince nebo IQ), a chceme znát parametry tohoto rozdělení. Z toho důvodu můžeme provést náhodný výběr o rozsahu n jedinců (získáme tak vlastně n nezávislých realizací náhodné veličiny X). Zobecníme-li výsledek tohoto příkladu, můžeme říci, že výběrový průměr má potom normální rozdělení $N(\mu; \sigma^2/n)$.

- b) Vzhledem k výsledku bodu a a známým pravidlům pro práci s normálně rozdělenými veličinami a konstantami (viz 6a) se snadno přesvědčíme, že veličina U má normované normální rozdělení $N(0;1)$.
- c) Odvození tohoto vztahu dá trochu práce, proto je lepší si jej pamatovat: platí, že veličina Z má rozdělení $\chi^2(9)$. Pokud bychom podobným způsobem pracovali namísto deseti veličin $X_1, X_2, \dots, X_{10}$ obecně s n (totožnými) veličinami $X_1, X_2, \dots, X_n$, platilo by, že

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1).$$

- d) Nejprve si uvědomíme, že platí následující rovnost:

$$T = \frac{\bar{X} - \mu}{\sqrt{\frac{S^2}{10}}} = \frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{10}}} \cdot \frac{1}{\sqrt{\frac{9S^2}{9\sigma^2}}} = \frac{U}{\sqrt{\frac{Z}{9}}}.$$

Už víme, že $U \sim N(0;1)$ a $Z \sim \chi^2(9)$, a pokud si připomeneme vztah z řešení příkladu 7c, snadno odhalíme, že veličina T má studentovo t -rozdělení s 9 stupni volnosti (opět zde pro jednoduchost nebudeme ověřovat, že U a Z jsou nezávislé).

Příklad 10: Test hypotézy o střední hodnotě

Náhodné veličiny $X_1, X_2, \dots, X_{10}$ jsou nezávislé a mají všechny stejné rozdělení (označme jej L). Známe hodnoty jedné realizace těchto náhodných veličin (tj. výběrový soubor) $x_1, x_2, \dots, x_{10}$, jejichž hodnoty (zapsané do sloupcového vektoru $\mathbf{x}$) jsou následující:

$$\mathbf{x} = [2 \ 3 \ 0 \ 1 \ 0 \ 3 \ 2 \ 3 \ 1 \ 0]^T.$$

Testujte na 95% hladině významnosti hypotézu, že střední hodnota rozdělení L je rovna 2, proti oboustranné alternativě, tj. testujte:

$$H_0: \mu = 2,$$

$$H_1: \mu \neq 2,$$

kde μ je střední hodnota rozdělení L . Předpokládejte přitom, že víme...

- a) ... že rozptyl L je 1.
- b) ... že L je normální rozdělení (s neznámým rozptylem).

Řešení.

- a) Pro test hypotézy o střední hodnotě budeme používat aritmetický průměr hodnot $x_1, x_2, \dots, x_{10}$. Vyjdeme z toho, že se jedná o realizace nezávislých stejně rozdělených náhodných veličin $X_1, X_2, \dots, X_{10}$. Pro účely testování hypotéz potřebujeme především znát pravděpodobnostní rozdělení aritmetického průměru těchto veličin

$$\bar{X} = \frac{1}{10} \sum_{i=1}^{10} X_i.$$

Označíme-li střední hodnotu rozdělení L jako μ a jeho rozptyl jako σ^2 (podle zadání tedy $\sigma^2 = 1$), můžeme podobně jako v příkladu 9a můžeme určit, že

$$E\bar{X} = \mu,$$

$$\text{var } \bar{X} = \frac{1}{10} \sigma^2 = \frac{1}{10}.$$

Zatím jsme ovšem nic neřekli o tvaru rozdělení veličiny $\bar{X}$, který budeme pro testování potřebovat. Tady nám přijde na pomoc *centrální limitní věta*, která zhruba říká, že s rostoucím počtem nezávislých náhodných veličin se rozdělení jejich aritmetického průměru blíží k rozdělení normálnímu. Budeme předpokládat, že v našem případě je počet veličin (10) dostatečný, a tedy že

$$\bar{X} \sim N(\mu; \frac{1}{10}).$$

Za platnosti nulové hypotézy známe dokonce rozdělení pro $\bar{X}$ zcela přesně, neboť potom známe $\mu = 2$, a tedy

$$\bar{X} \sim N(2; \frac{1}{10}).$$

Přístupme nyní k samotnému testování hypotézy o střední hodnotě. Snadno se přesvědčíme, že

$$\bar{x} = \frac{1}{10} \sum_{i=1}^{10} x_i = 1,5.$$

Test hypotézy probíhá následovně. Vyjdeme z nulové hypotézy ($\mu = 2$) a zjistíme, zda se nám za její platnosti náhodou nezda naměřený výsledek $\bar{x} = 1,5$ „příliš podezřelý“. Pokud ano, nulovou hypotézu zamítneme ve prospěch H_1 . Co přesně znamená formulace „příliš podezřelý“? Je to docela prosté: podezřelá hodnota je taková, která padne moc daleko od střední hodnoty $\mu = 2$, přesněji řečeno taková hodnota, která opustí *oboustranný 95% interval spolehlivosti pro $\bar{X}$* (procento intervalu spolehlivosti odpovídá hladině významnosti testu, *oboustranný interval* uvažujeme proto, že formulovaná alternativní hypotéza je *oboustranná*). Vzhledem k tomu, že rozdělení pro veličinu $\bar{X}$ jsme již určili, není problém interval spolehlivosti stanovit – provedeme to podobným způsobem, jako v příkladu 8b. Vyjde

$$\left(\mu - u_{0,975} \cdot \frac{\sigma}{\sqrt{n}}; \mu + u_{0,975} \cdot \frac{\sigma}{\sqrt{n}} \right) = \left(2 - u_{0,975} \sqrt{\frac{1}{10}}; 2 + u_{0,975} \sqrt{\frac{1}{10}} \right) = (1,38; 2,62).$$

Hodnota $\bar{x} = 1,5$ do tohoto intervalu spadá, tedy H_0 nezamítáme.

- b)** Podobně jako v bodě **a** zde platí, že $\bar{X} \sim N(\mu; \sigma^2/10)$, tentokrát však σ^2 neznáme, nemůžeme proto na tomto rozdělení založit naše „testovací úvahy“. Nezbyvá nám, než rozptýl σ^2 odhadnout. Učiníme tak pomocí výběrového rozptylu

$$s^2 = \frac{1}{9} \sum_{i=1}^{10} (x_i - \bar{x})^2,$$

který můžeme chápat jako realizaci náhodné veličiny S^2 definované jako v příkladu 9. Jak už víme z příkladu 9d, má veličina

$$T = \frac{\bar{X} - \mu}{\sqrt{\frac{S^2}{10}}}$$

studentovo t -rozdělení s 9 stupni volnosti. Právě této veličiny využijeme při testování našich hypotéz. Za platnosti H_0 je $\mu = 2$, můžeme tedy zkoumat, zda výběrová hodnota veličiny T spočtená jako

$$t = \frac{\bar{x} - 2}{\sqrt{\frac{s^2}{10}}}$$

spadá do *oboustranného 95% intervalu spolehlivosti pro veličinu s rozdělením $t(9)$* . Podrobněji řečeno, zajímá nás, zdali

$$t \in (-t_{0,975,9}; t_{0,975,9}), \text{ neboli } |t| \leq t_{0,975,9}.$$

Pokud ano, nulovou hypotézu nezamítáme. Podotkneme, že uvedená podmínka lze zapsat též jako

$$|\bar{x} - 2| \leq t_{0,975,9} \cdot \sqrt{\frac{s^2}{10}}.$$

V našem případě můžeme numericky spočítat $t = -1,25$, což je hodnota, která spadá do 95% intervalu spolehlivosti pro veličinu s rozdělením $t(9)$, který se určí zcela analogicky jako v 8c a je přibližně $(-2,26; 2,26)$. Nulovou hypotézu tedy ani v tomto případě nezamítáme.