Variabilita a její míry, úvod do pravděpodobnosti

Supplementum ke cvičení 4ST210 Statistika pro finance

Lubomír Štěpánek^{1, 2}



 Oddělení biomedicínské statistiky Ústav biofyziky a informatiky
 lékařská fakulta
 Univerzita Karlova, Praha



²Katedra biomedicínské informatiky Fakulta biomedicínského inženýrství České vysoké učení technické v Praze

(2019) Lubomír Štěpánek, CC BY-NC-ND 3.0 (CZ)



Dílo lze dále svobodně šířit, ovšem s uvedením původního autora a s uvedením původní licence. Dílo není možné šířit komerčně ani s ním jakkoliv jinak nakládat pro účely komerčního zisku. Dílo nesmí být jakkoliv upravováno. Autor neručí za správnost informací uvedených kdekoliv v předložené práci, přesto vynaložil nezanedbatelné úsilí, aby byla uvedená fakta správná a aktuální, a práci sepsal podle svého nejlepšího vědomí a svých "nejlepších" znalostí problematiky.

Obsah

- Opakování
- Míry variability
- 3 Vlastnosti a rozklad rozptylu
- Úvod do pravděpodobnosti
- 6 Literatura



•000

 Bezpečnostní agentura SAFETY a.s. má 216 zaměstnanců a skládá se ze dvou dceřiných společností. V první dceřiné společnosti je průměrná měsíční mzda 21 650 Kč a v druhé 24 800 Kč. Průměrná mzda za celý holding je 23 650 Kč. Kolik zaměstnanců pracuje ve druhé dceřiné společnosti?





Opakování

• V bytovém komplexu je celkem 78 domácností, z nichž 34 nemá žádné parkovací místo v podzemních garážích, 30 domácností má jedno parkovací místo, 6 domácností má dvě parkovací místa, 5 domácností má tři parkovací místa a 3 domácnosti mají dokonce čtyři parkovací místa. Jaký průměrný počet parkovacích míst připadajících na domácnost? Sestavme tabulku absolutních a relativních četností pro počet parkovacích míst, včetně kumulativních protějšků.



Opakování

0000

V hudebním tělese je rozložení věku jeho hráčů následující (v letech),

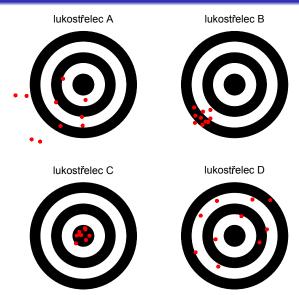
22, 82, 27, 43, 19, 47, 41, 34, 34, 42, 35, 39.

Určeme

- (i) mediánový věk hudebníků v tělese.
- (ii) modální věk hudebníků v tělese.
- (iii) první a třetí kvartil věku hudebníků v tělese.
- (iv) 80-tý percentil věku hudebníků v tělese.

- První dělník je schopen vyhloubit výkop za 8 hodin, druhý dělník za 6 hodin a třetí dělník pak za 10 hodin.
 - (i) Za kolik hodin vyhloubí jeden výkop, pokud budou pracovat společně?
 - Za kolik průměrně hodin je vyhlouben výkop, pokud pracují všichni tří dělníci a každý pracuje na svých výkopech?

Intuitivní pohled na variabilitu (a střední hodnotu)







remed mer variability

- absolutní míry
 - varianční rozpětí
 - kvartilové rozpětí
 - dále decilové rozpětí, percentilové rozpětí
 - rozptyl
 - směrodatná odchylka
- relativní míry
 - variační koeficient



- varianční rozpětí (též zvané jako min-max statistika) je nejjednodušší měrou variability
- pro n čísel x_1, x_2, \ldots, x_n spočítáme jejich varianční rozpětí R jako

$$R = x_{(n)} - x_{(1)} = x_{\text{max}} - x_{\text{min}},$$

kde $x_{(n)}$ je n-té nejmenší a $x_{(1)}$ je první nejmenší číslo z čísel x_1, x_2, \ldots, x_n , tedy $x_{(n)} \equiv x_{\max}$ a $x_{(1)} \equiv x_{\min}$

Ize snadno ukázat, že přibližně platí (tzv. pravidlo six sigma)

$$x_{(n)} - x_{(1)} = x_{\text{max}} - x_{\text{min}} \approx 6s_x,$$

kde s_x je výběrová směrodatná odchylka daného výběru

▶ MS Excel®

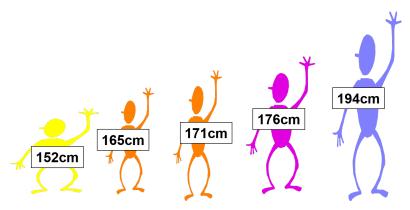
 $MAX(x_1:x_n) - MIN(x_1:x_n)$





Varianční rozpětí

• určeme varianční rozpětí z následujícího souboru tělesných výšek

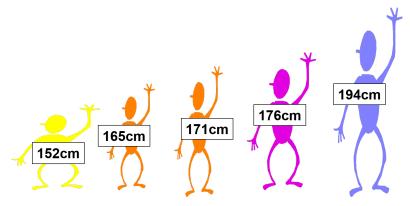


Varianční rozpětí

Opakování

určeme varianční rozpětí z následujícího souboru tělesných výšek

•
$$R = x_{(n)} - x_{(1)} = 194 - 152 = 42$$
 [cm]





Kvartilové rozpětí

Opakování

• kvartilové rozpětí nad souborem čísel x_1, x_2, \ldots, x_n je definováno iako

$$\tilde{x}_{0,75} - \tilde{x}_{0,25},$$

kde $\tilde{x}_{0.25}$ je první a $\tilde{x}_{0.75}$ třetí kvartil souboru čísel x_1, x_2, \ldots, x_n tedy obecně pro p-tý kvantil \tilde{x}_n je

$$\tilde{x}_p = \left\{ \begin{array}{c} x_{(\lfloor k \rfloor + 1)}, & \text{pro } k = np \notin \mathbb{N} \\ \\ \frac{1}{2} \left(x_{(k)} + x_{(k+1)} \right), & \text{pro } k = np \in \mathbb{N}, \end{array} \right.$$

kde $x_{(k)}$ je k-té nejmenší číslo mezi čísly x_1, x_2, \ldots, x_n , dále kde $0 \le p \le 1$ (zde p = 0.25 a p = 0.75) a kde |x| značí dolní celou část čísla x, tedy nejvyšší celé číslo takové, že nepřevýší x



Decilové rozpětí a percentilové rozpětí

• nad souborem čísel x_1, x_2, \ldots, x_n je definováno decilové rozpětí jako

$$\tilde{x}_{0,90} - \tilde{x}_{0,10}$$

a percentilové rozpětí jako

$$\tilde{x}_{0,99} - \tilde{x}_{0,01},$$

kde $\tilde{x}_{0.10}$ je první a $\tilde{x}_{0.90}$ devátý decil a $\tilde{x}_{0.01}$ je první a $\tilde{x}_{0.99}$ devětadevadesátý percentil souboru čísel x_1, x_2, \ldots, x_n

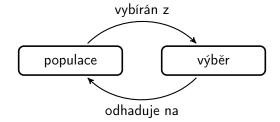
• tedy obecně 100p% rozpětí nad souborem čísel x_1, x_2, \ldots, x_n je definováno jako

$$\tilde{x}_{1-p} - \tilde{x}_p$$

tak, že \tilde{x}_p je p-tý kvantil a zde $0 \le p < 0.5$



- z populace je vybírán výběr
 - z charakteristik výběru jsou odhadovány charakteristiky populace (!)





- předpokládejme, že soubor celé populace je tvořen právě n čísly x_1, x_2, \ldots, x_n
- pak pro těchto n čísel x_1, x_2, \ldots, x_n spočítáme populační rozptyl σ^2 (a současně výběrový rozptyl s_x^2) jako

$$\sigma^2 = s_x^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

a populační směrodatnou odchylku σ (a současně výběrovou směrodatnou odchylku s_x)

$$\sigma = s_x = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2},$$

kde $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ je populační průměr n čísel x_1, x_2, \dots, x_n

- předpokládejme, že máme výběr z populace, který je tvořen n čísly x_1, x_2, \ldots, x_n
- pak pro těchto n čísel x_1, x_2, \ldots, x_n lze spočítat "odhad"¹ populačního rozptylu σ^2 pomocí výběrového rozptylu jako

$$\sigma^{2} = \frac{n}{n-1} s_{x}^{2} = \frac{n}{n-1} \left(\frac{1}{n} \sum_{i=1}^{n} (x_{i} - \bar{x})^{2} \right),$$

podobně lze populační směrodatnou odchylku σ "odhadnout" pomocí výběrové směrodatné odchylky s_x jako

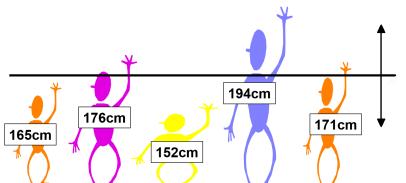
$$\sigma = \sqrt{\frac{n}{n-1}} s_x = \sqrt{\frac{n}{n-1}} \left(\sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2} \right),$$

kde $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ je výběrový průměr n čísel x_1, x_2, \ldots, x_n

¹pojmy "odhad" a "odhadnout" budeme zatím vnímat jen intuitivně Variabilita a její míry, úvod do pravděpodobnosti

Výpočet rozptylu a směrodatné odchylky

určeme směrodatnou odchylku a rozptyl z následujícího výběru tělesných výšek





Opakování

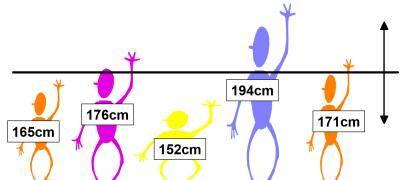
Literatura

Výpočet rozptylu a směrodatné odchylky

 určeme směrodatnou odchylku a rozptyl z následujícího výběru tělesných výšek

•
$$\sigma = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2} \doteq 15.4 \text{ [cm]}; \quad \sigma^2 \doteq 237.2 \text{ [cm}^2]$$

•
$$s_x = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} \doteq 13.8 \text{ [cm]}; \quad s_x^2 \doteq 189.8 \text{ [cm}^2]$$





- Určeme, jak se změní výběrový rozptyl a výběrová směrodatná odchylka, pokud se všechny hodnoty ve výběru
 - (i) zmenší o pět.
 - (ii) zvětší dvakrát.



Variační koeficient

Opakování

• pro n čísel x_1, x_2, \ldots, x_n spočítáme jejich variační koeficient v_x jako

$$v_x = \frac{s_x}{\bar{x}} = \frac{\sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}}{\frac{1}{n} \sum_{i=1}^n x_i}$$

 Vlivem ekonomických událostí vzrostla průměrná cena letenek v určité oblasti o 10 %, zatímco rozptyl ceny těchto letenek vzrostl o 46,41 %. Určeme, jak se změnil variační koeficient ceny těchto letenek.



• pro n čísel x_1, x_2, \ldots, x_n (tvořících populaci) spočítáme populační rozptyl jako

$$s_x^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2,$$

kde $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ je populační průměr n čísel x_1, x_2, \dots, x_n

snadno nahlédneme, že

$$s_x^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

$$= \frac{1}{n} \sum_{i=1}^n (x_i^2 - 2x_i \bar{x} + \bar{x}^2)$$

$$= \frac{1}{n} \left(\sum_{i=1}^n x_i^2 - 2\sum_{i=1}^n x_i \bar{x} + \sum_{i=1}^n \bar{x}^2 \right)$$



Výpočetní tvar (populačního) rozptylu

$$s_x^2 = \frac{1}{n} \left(\sum_{i=1}^n x_i^2 - 2 \sum_{i=1}^n x_i \bar{x} + \sum_{i=1}^n \bar{x}^2 \right)$$

$$= \frac{1}{n} \sum_{i=1}^n x_i^2 - 2 \bar{x} \frac{1}{n} \sum_{i=1}^n x_i + \frac{1}{n} \sum_{i=1}^n \bar{x}^2$$

$$= \overline{x^2} - 2 \bar{x} \bar{x} + \frac{1}{n} n \bar{x}^2$$

$$= \overline{x^2} - 2 \bar{x}^2 + \bar{x}^2$$

$$= \overline{x^2} - \bar{x}^2$$

$$= \frac{1}{n} \sum_{i=1}^n x_i^2 - \left(\frac{1}{n} \sum_{i=1}^n x_i \right)^2$$

• tvar $s_x^2 = \overline{x^2} - \bar{x}^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - \left(\frac{1}{n} \sum_{i=1}^n x_i\right)^2$ nazýváme výpočetním tvarem rozptylu

Opakování

• V zemi Statlandia žijí dva kouzelníci, jeden vždy mluví pravdu a druhý vždy lže. Mladší z nich nám řekl, že ve všech ovčích stádech Statlandie je čtverec průměru počtu ovcí 128,6 a průměr čtverců počtu ovcí 115,4. Starší nám řekl, že je to naopak. Který z nich určitě lhal? Mladší, nebo starší?



- budiž populace tvořena k podskupinami o četnostech n_1, n_2, \ldots, n_k tak, že $\sum_{i=1}^k n_i = n$ a že j-tá podskupina má průměr \bar{x}_j pro $\forall i \in \{1, 2, \dots, k\}$
- celkový průměr je zřejmě $\bar{x} = \frac{1}{n} \sum_{i=1}^k n_i \bar{x}_i$ a průměr kvadrátů je $\overline{x^2} = \frac{1}{n} \sum_{i=1}^k n_i x_i^2$
- pak rozptyl spočítáme s výhodou pomocí výpočetního tvaru

$$s_x^2 = \overline{x^2} - \bar{x}^2 = \frac{1}{n} \sum_{j=1}^k n_j x_j^2 - \left(\frac{1}{n} \sum_{j=1}^k n_j x_j\right)^2$$



- budiž populace tvořena k podskupinami o četnostech n_1, n_2, \ldots, n_k tak, že $\sum_{i=1}^k n_i = n$ a že j-tá podskupina má průměr \bar{x}_j pro $\forall i \in \{1, 2, \dots, k\}$
- pak lze ukázat, že rozptyl lze rozložit na dva sčítance

$$s_{x}^{2} = \frac{1}{n} \sum_{j=1}^{k} \sum_{i=1}^{n_{j}} (x_{i,j} - \bar{x})^{2} = \frac{1}{n} \left(\sum_{j=1}^{k} \sum_{i=1}^{n_{j}} (x_{i,j} - \bar{x}_{j})^{2} + \sum_{j=1}^{k} n_{j} (\bar{x}_{j} - \bar{x})^{2} \right)$$

$$= \frac{1}{n} \left(\sum_{j=1}^{k} \sum_{i=1}^{n_{j}} (x_{i,j} - \bar{x}_{j})^{2} \right) + \frac{1}{n} \left(\sum_{j=1}^{k} n_{j} (\bar{x}_{j} - \bar{x})^{2} \right)$$

$$= \frac{1}{n} \left(\sum_{j=1}^{k} n_{j} s_{x,j}^{2} \right) + \frac{1}{n} \left(\sum_{j=1}^{k} n_{j} (\bar{x}_{j} - \bar{x})^{2} \right)$$
vnitroskupinová variabilita
meziskupinová variabilita

meziskupinová variabilita

Odvození rozkladu rozptylu

$$s_x^2 = \frac{1}{n} \sum_{j=1}^k \sum_{i=1}^{n_j} (x_{i,j} - \bar{x})^2$$

$$= \frac{1}{n} \sum_{j=1}^k \sum_{i=1}^{n_j} ((x_{i,j} - \bar{x}_j) + (\bar{x}_j - \bar{x}))^2$$

$$= \frac{1}{n} \sum_{j=1}^k \sum_{i=1}^{n_j} ((x_{i,j} - \bar{x}_j)^2 + 2(x_{i,j} - \bar{x}_j)(\bar{x}_j - \bar{x}) + (\bar{x}_j - \bar{x})^2)$$

$$= \frac{1}{n} \sum_{j=1}^k \left(\sum_{i=1}^{n_j} (x_{i,j} - \bar{x}_j)^2 + 2(\bar{x}_j - \bar{x}) \sum_{i=1}^{n_j} (x_{i,j} - \bar{x}_j) + n_j(\bar{x}_j - \bar{x})^2 \right)$$



$s_x^2 = \frac{1}{n} \sum_{i=1}^k \left(\sum_{j=1}^{n_j} (x_{i,j} - \bar{x}_j)^2 + n_j (\bar{x}_j - \bar{x})^2 \right)$ $= \frac{1}{n} \left(\sum_{i=1}^{k} \sum_{j=1}^{n_j} (x_{i,j} - \bar{x}_j)^2 + \sum_{j=1}^{k} n_j (\bar{x}_j - \bar{x})^2 \right)$ $= \frac{1}{n} \left(\sum_{i=1}^{k} \sum_{j=1}^{n_j} (x_{i,j} - \bar{x}_j)^2 \right) + \frac{1}{n} \left(\sum_{j=1}^{k} n_j (\bar{x}_j - \bar{x})^2 \right)$ $= \frac{1}{n} \left(\sum_{j=1}^{k} n_j s_{x,j}^2 \right) + \frac{1}{n} \left(\sum_{j=1}^{k} n_j (\bar{x}_j - \bar{x})^2 \right)$ vnitroskupinová variabilita meziskupinová variabilita

- Obchodní řetězec odebírá určitý výrobek, jehož cena v průběhu roku sezónně kolísá, od dvou stálých dodavatelů A a B. Průměrná cena za celý rok od dodavatele A je 9 Kč, její směrodatná odchylka činí 2 Kč, výrobků od dodavatele A se nakoupilo 1000 kusů. U dodavatele B činí průměrná cena 10 Kč při směrodatné odchylce 1 Kč, nákup od dodavatele B byl 4000 kusů. Určeme
 - (i) variační koeficient vyjadřující variabilitu kolísání nákupní ceny během roku souhrnně za oba dva dodavatele dohromady.
 - (ii) zda se na celkové variabilitě nákupní ceny větší měrou podílí průběžné sezónní kolísání cen výrobku u jednotlivých dodavatelů v rámci roku, nebo zda jsou důležitější rozdíly mezi průměrnými cenami jednotlivých dodavatelů.

Opakování

• Soubor o šesti hodnotách má průměr 12 a rozptyl $4\frac{2}{3}$. Jak se změní průměr a rozptyl souboru, když do něj přibude hodnota 15?



- Soubor o šesti hodnotách má průměr 12 a rozptyl $4\frac{2}{3}$. Jak se změní průměr a rozptyl souboru, když do něj přibude hodnota 15?
- Řešení. $\bar{x}_{\text{nový}} \doteq 12,43, \ s_{x_{\text{nový}}}^2 \doteq 5,10$



Náhodný pokus (experiment) a náhodný jev

- náhodný pokus (experiment)
 - je děj, jehož výsledek se může při zopakování změnit i při zachování podmínek, závisí tedy na náhodě
 - např. hod kostkou, los z urny
- náhodný jev
 - je výsledek náhodného pokusu
 - obvykle se značí velkými písmeny A, B, C, \ldots, X, Y, Z
 - pravděpodobnost náhodného jevu A značíme P(A)
 - např. na kostce padne pět ok, z urny byla vytažena černá koule
- jistý jev
 - jev, který nastane vždy
 - např. na minci padne hlava, nebo orel
- nemožný jev
 - jev, který nenastane nikdy
 - např. na (laplaceovské) minci padne hrana





Klasická definice pravděpodobnosti

ullet pravděpodobnost jevu A je rovna podílu počtu případů m, které jsou jevu A příznivé, ku počtu n všech možným případů

$$P(A) = \frac{m}{n}$$

nutným předpokladem je, že všechny případy mohou nastat stejně často

Intermezzo

Opakování

ullet Jaká je pravděpodobnost jevu A, že na hrací kostce padne číslo větší než 2?





- Jaká je pravděpodobnost jevu A, že na hrací kostce padne číslo větší než 2?
- Řešení. $P(A)=\frac{|\mathrm{padne}\ 3,\ 4,\ 5\ \mathrm{nebo}\ 6\ \mathrm{ok}|}{6}=\frac{4}{6}=\frac{2}{3}$



Geometrická definice pravděpodobnosti

• pravděpodobnost jevu A je rovna podílu plochy S odpovídající případům, které jsou jevu A příznivé, ku ploše Ω odpovídající všem možným případům

$$P(A) = \frac{S}{\Omega}$$

zde již jednotlivé případy nemusí nutně nastat stejně často



Intermezzo

Opakování

• Z intervalu (0,1) náhodně vybereme dvě čísla x a y. Jaká je pravděpodobnost jevu, že $2y \le x^2$?

Množinové vztahy množin $\mathcal A$ a $\mathcal B$

\mathcal{A} je podmnožinou \mathcal{B}

Opakování

Pokud je $\forall a \in \mathcal{A} : a \in \mathcal{B}$, pak \mathcal{A} je podmnožinou \mathcal{B} , což značíme $\mathcal{A} \subset \mathcal{B}$.

Sjednocení množin \mathcal{A} a \mathcal{B}

Sjednocení množin \mathcal{A} a \mathcal{B} je taková množina $\mathcal{A} \cup \mathcal{B}$, že

$$\mathcal{A} \cup \mathcal{B} = \{ \forall \ x : x \in \mathcal{A} \lor x \in \mathcal{B} \}.$$

Průnik množin A a B

Průnik množin \mathcal{A} a \mathcal{B} je taková množina $\mathcal{A} \cap \mathcal{B}$, že

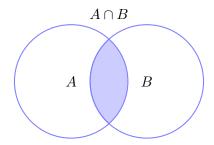
$$\mathcal{A} \cap \mathcal{B} = \{ \forall \ x : x \in \mathcal{A} \land x \in \mathcal{B} \}.$$

Asymetrický rozdíl množin A a B

Asymetrický rozdíl množin \mathcal{A} a \mathcal{B} je taková množina $\mathcal{A} - \mathcal{B}$, že

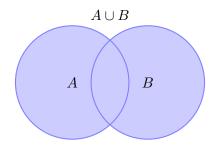
$$\mathcal{A} - \mathcal{B} = \{ \forall \ x : x \in \mathcal{A} \land x \notin \mathcal{B} \}.$$

• pravděpodobnost, že nastanou oba jevy A i B, značíme $P(A \cap B)$



edhocem jevu A a L

• pravděpodobnost, že nastane alespoň jeden z jevů A nebo B, značíme $P(A \cup B)$



nechť A a B jsou náhodné jevy, pak platí

$$0 \le P(A) \le 1$$

$$0 \le P(B) \le 1$$

• dále pokud A je podjevem B, tedy $A \subseteq B$

$$P(A) \le P(B)$$

vždy však

Opakování

$$P(A^{C}) = 1 - P(A)$$

 $P(A \cup B) = P(A) + P(B) - P(A \cap B),$

kde A^C je doplňkový jev k jevu A

 pokud jsou A a B vylučující se jevy (neslučitelné), pak $P(A \cap B) = 0$ a

$$P(A \cup B) = P(A) + P(B)$$

Sčítání a násobení pravděpodobnostní

- nechť A a B jsou náhodné jevy
- pak $P(A \cup B) = P(A) + P(B)$, pokud jsou A a B neslučitelné jevy
- a dále $P(A \cap B) = P(A) \cdot P(B)$, pokud jsou A a B nezávislé jevy

Intermezzo

Je možné, aby dva jevy byly neslučitelné a současně i nezávislé? Zkoumejme.



Děkuji za pozornost!

lubomir.stepanek@vse.cz lubomir.stepanek@lf1.cuni.cz lubomir.stepanek@fbmi.cvut.cz

