Variabilita a její míry, úvod do pravděpodobnosti

—

Supplementum ke cvičení 4ST201 Statistika

Lubomír Štěpánek[1, 2]

[1]Oddělení biomedicínské statistiky
Ústav biofyziky a informatiky
1. lékařská fakulta
Univerzita Karlova, Praha

ČVUT FBMI

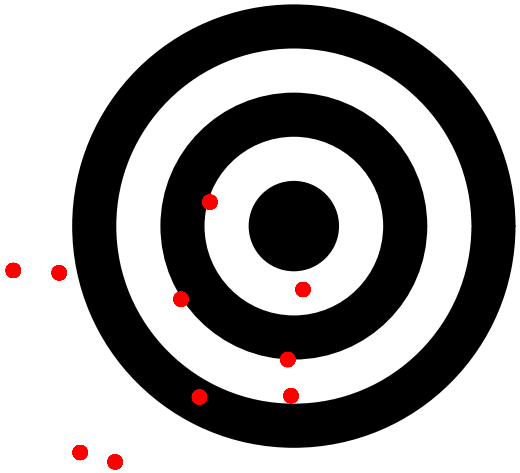
[2]Katedra biomedicínské informatiky
Fakulta biomedicínského inženýrství
České vysoké učení technické v Praze

4. října 2019

Obsah

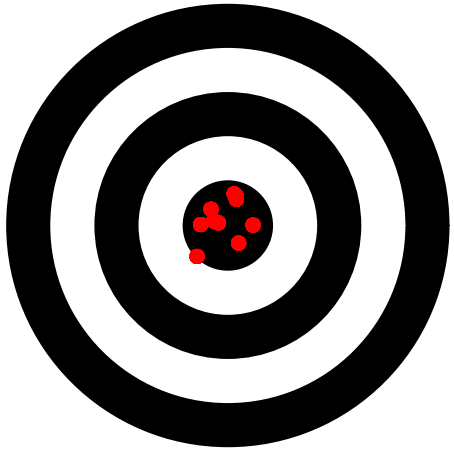
Intuitivní pohled na variabilitu (a střední hodnotu)

lukostřelec A

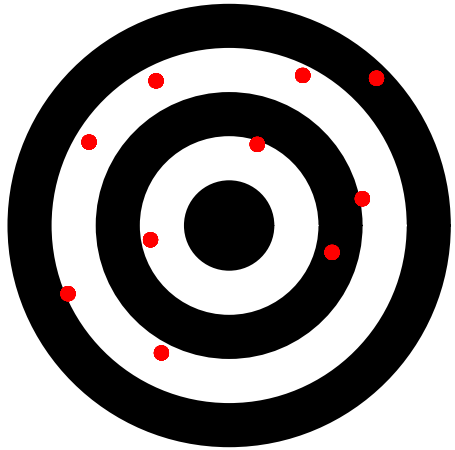


lukostřelec B

lukostřelec C



lukostřelec D



Přehled měr variability

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

Varianční rozpětí

- 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

$$R = x_{(n)} - x_{(1)},$$

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 $x_{(1)} \equiv x_{\min}$

MS Excel®

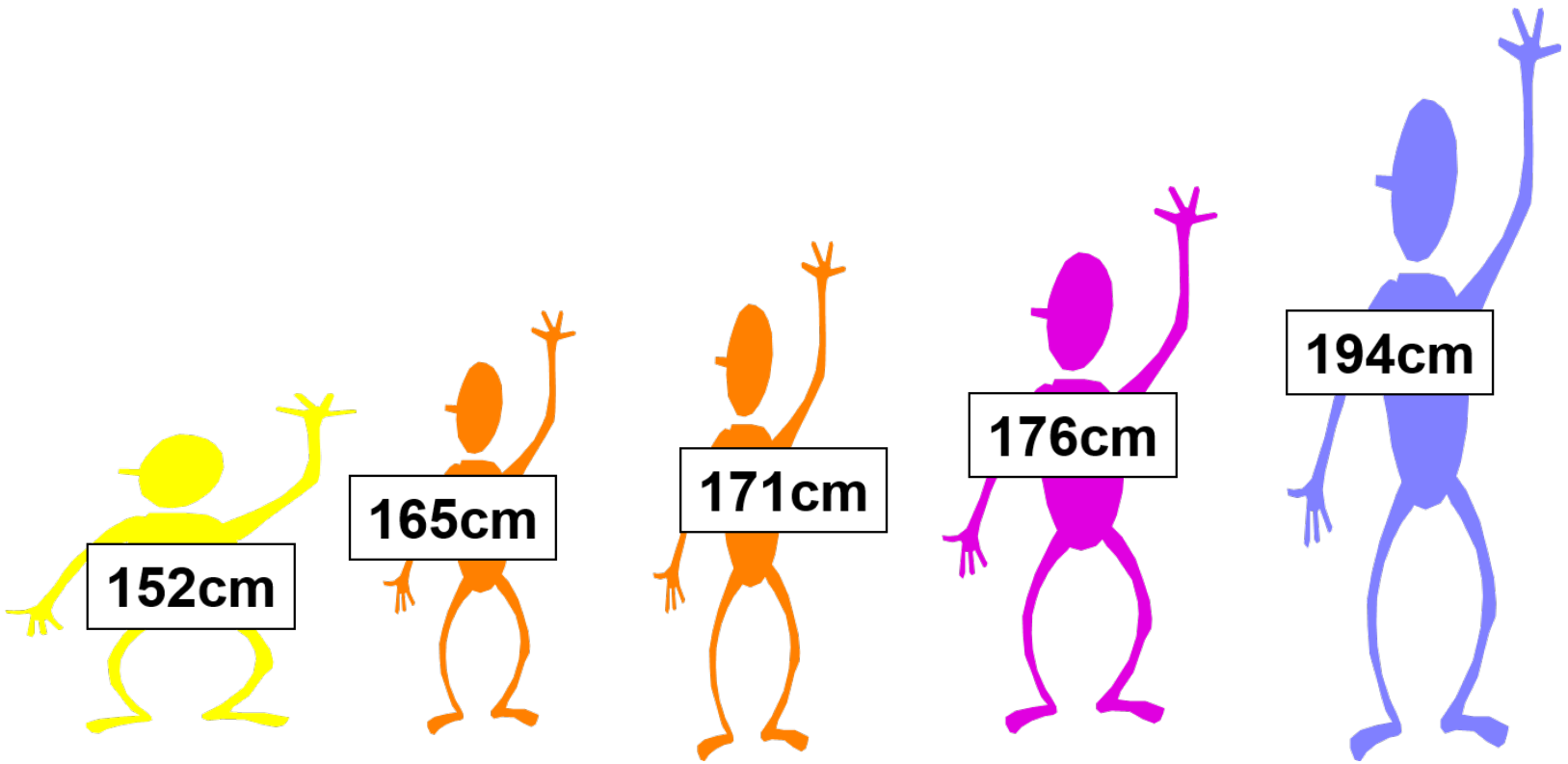
`MAX(`$x_1 : x_n$`) - MIN(`$x_1 : x_n$`)`

R

`> max(c(`$x_1, x_2, \ldots, x_n$`)) - min(c(`$x_1, x_2, \ldots, x_n$`))`

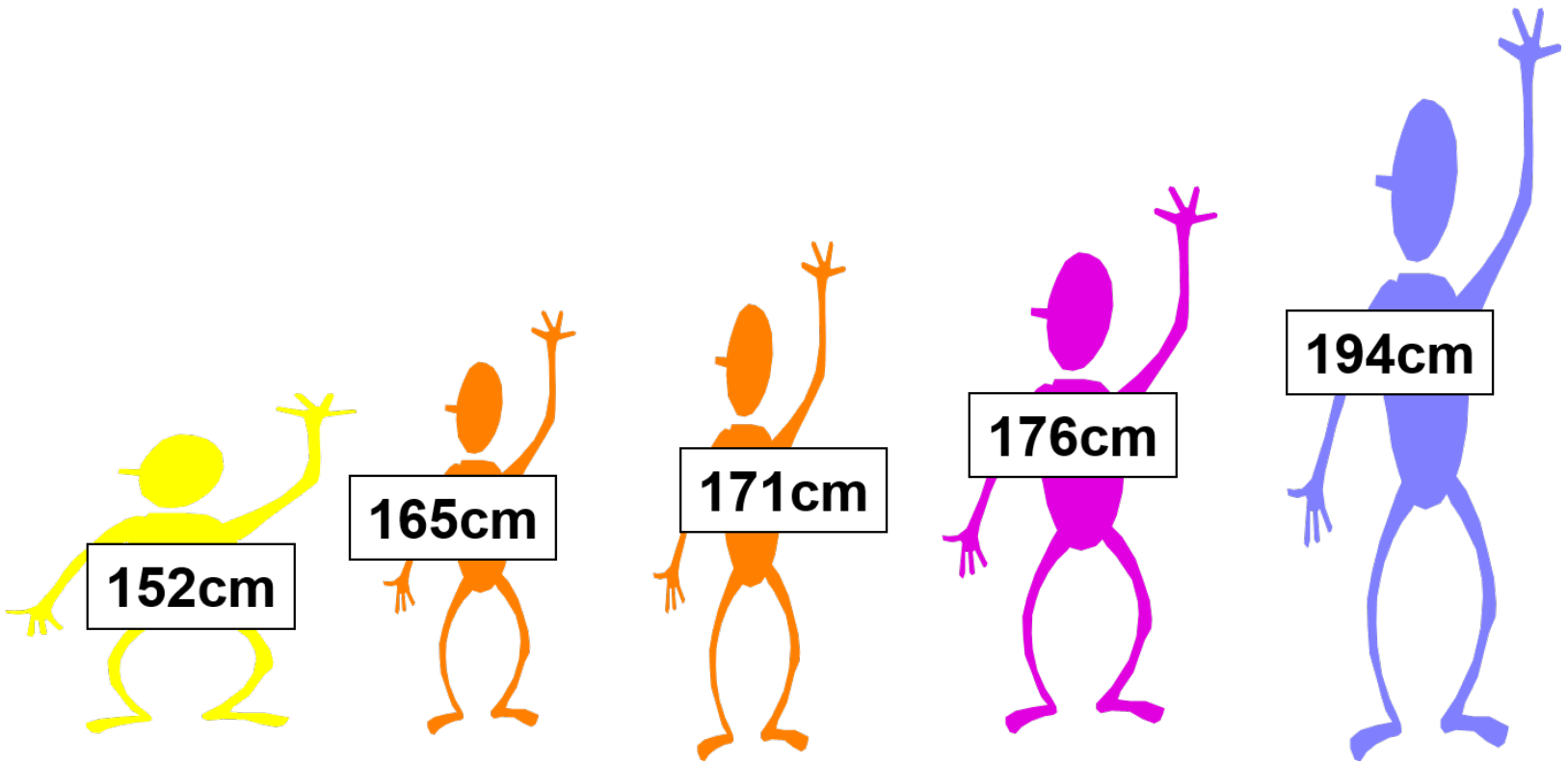
Varianční rozpětí

- určeme varianční rozpětí z následujícího souboru tělesných výšek



Varianční rozpětí

- určeme varianční rozpětí z následujícího souboru tělesných výšek
- $$R = x_{(n)} - x_{(1)} = 194 - 152 = 42 \text{ [cm]}$$



Kvartilové rozpětí

- kvartilové rozpětí nad souborem čísel $x_1, x_2, \ldots, x_n$ je definováno jako

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

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 kvintil $\tilde{x}_p$ je

$$\tilde{x}_p = \begin{cases} x_{(k+1)}, & \text{pro } k \neq np \\ \frac{1}{2}\left(x_{(k)} + x_{(k+1)}\right), & \text{pro } k = np, \end{cases}$$

kde $x_{(k)}$ je k-té nejmenší číslo mezi čísly $x_1, x_2, \ldots, x_n$ a $0 \leq p \leq 1$

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 kvintil, tedy

$$\tilde{x}_p = \begin{cases} x_{(k+1)}, & \text{pro } k \neq np \\ \frac{1}{2}\left(x_{(k)} + x_{(k+1)}\right), & \text{pro } k = np, \end{cases}$$

kde $x_{(k)}$ je k-té nejmenší číslo mezi čísly $x_1, x_2, \ldots, x_n$ a $0 \leq p \leq 1$

Populační rozptyl a populační směrodatná odchylka

- 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 jako

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

a populační směrodatnou odchylku

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

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

Výběrový rozptyl a výběrová směrodatná odchylka

- 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 pomocí výběrového rozptylu jako

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

podobně lze populační směrodatnou odchylku „odhadnout“ pomocí výběrové směrodatné odchylky

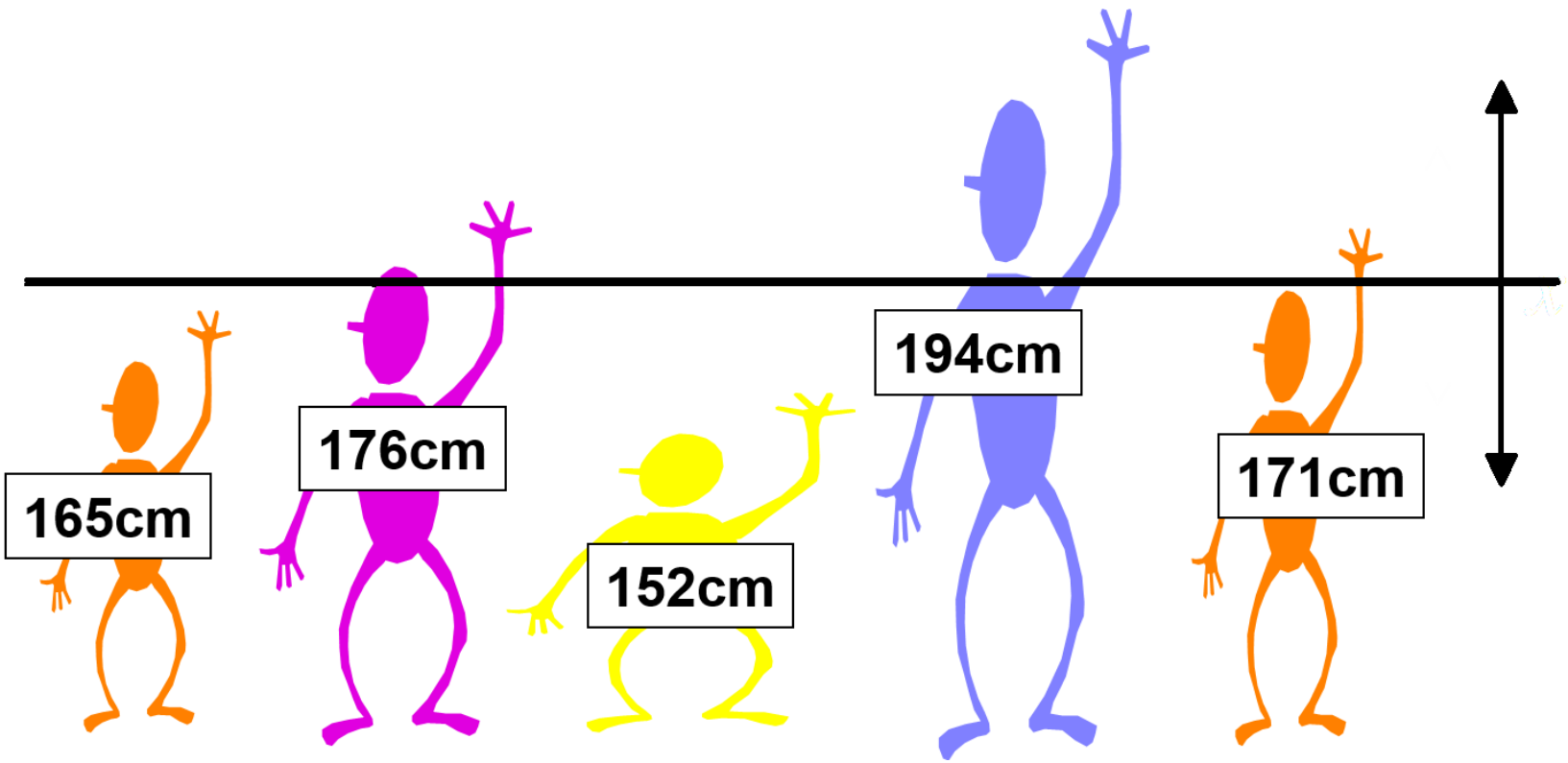
$$s_x = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2},$$

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

[1] pojmy „odhad“ a „odhadnou“ budeme zatím vnímat jen intuitivně

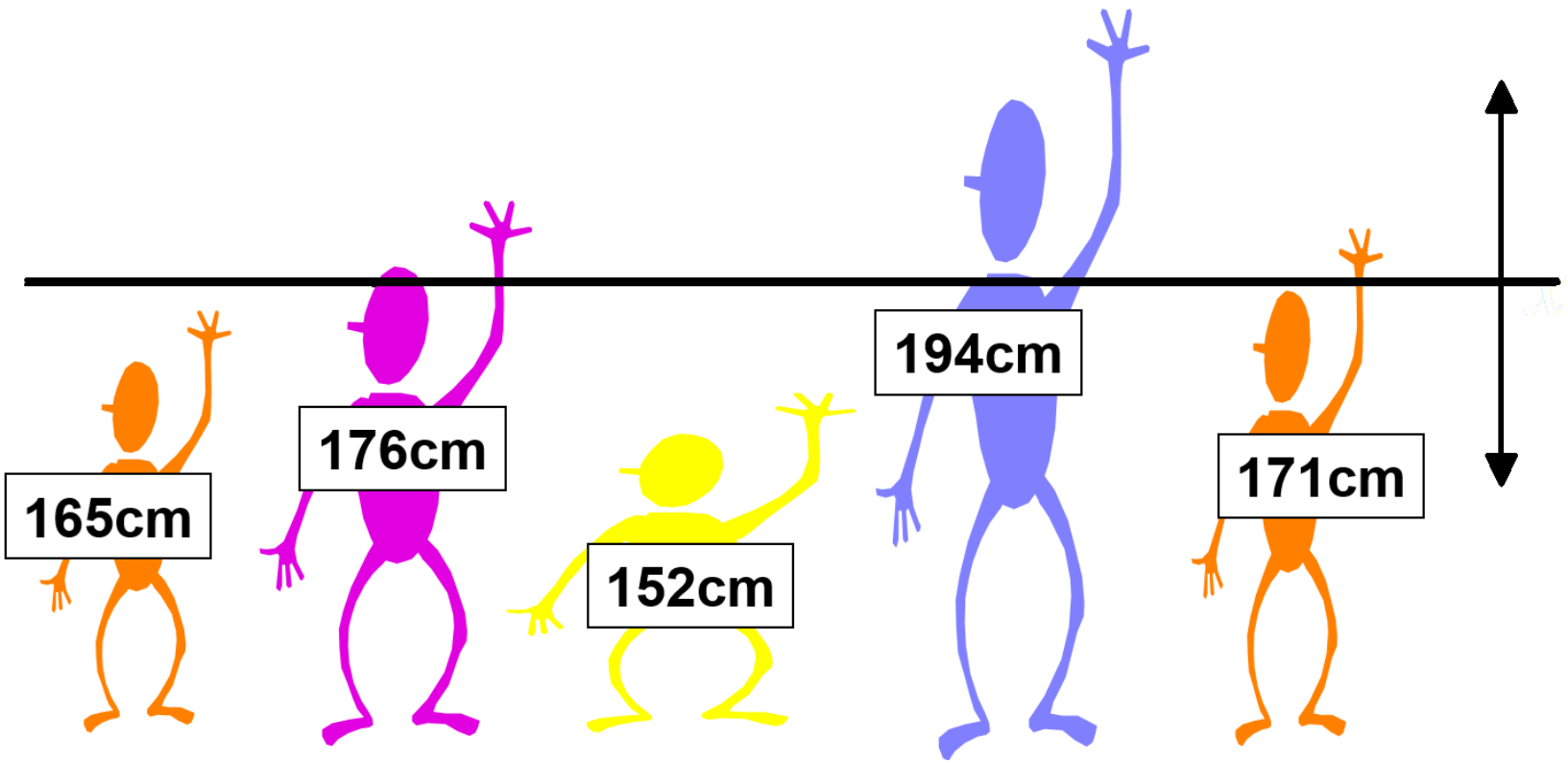
Výpočet rozptylu a směrodatné odchylky

- určeme směrodatnou odchylku a rozptyl z následujícího výběru tělesných výšek



Výpočet rozptylu a směrodatné odchylky

- určeme směrodatnou odchylku a rozptyl z následujícího výběru tělesných výšek
- $$s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \doteq 15{,}4\ [\mathrm{cm}]; \quad s_x^2 \doteq 237{,}2\ [\mathrm{cm}^2]$$



Variační koeficient

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

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

Výpočetní tvar (populačního) rozptylu

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

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

  kde $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ je populační průměr n čísel $x_1, x_2, \ldots, x_n$
- snadno nahlédneme, že

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

Výpočetní tvar (populačního) rozptylu

$$
\begin{aligned}
\sigma^2 &= \frac{1}{n}\left(\sum_{i=1}^{n} x_i^2 - 2\sum_{i=1}^{n} x_i\bar{x} + \sum_{i=1}^{n} \bar{x}^2\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - 2\bar{x}\frac{1}{n}\sum_{i=1}^{n} x_i + \frac{1}{n}\sum_{i=1}^{n} \bar{x}^2 \\
&= \overline{x^2} - 2\bar{x}\bar{x} + \frac{1}{n}n\bar{x}^2 \\
&= \overline{x^2} - 2\bar{x}^2 + \bar{x}^2 \\
&= \overline{x^2} - \bar{x}^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2
\end{aligned}
$$

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

Intermezzo

- 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?

Intermezzo

- 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?
- *Řešení.* Buď $\bar{x}$ průměrný počet ovcí na stádo a $\overline{x^2}$ průměrný čtverec počtu ovcí na stádo ve *Statlandii*. Protože je $\sigma^2 = \overline{x^2} - \bar{x}^2$, je i $\overline{x^2} - \bar{x}^2 = \sigma^2 \geq 0$, a tedy $\overline{x^2} - \bar{x}^2 \geq 0$, čili $\overline{x^2} \geq \bar{x}^2$. Proto musí být průměr čtverců počtu ovcí minimálně tak velký jako čtverec průměru počtu ovcí na stádo. Mladší kouzelník tedy lhal. □

Výpočet (populačního) rozptylu pomocí známých četností podskupin souboru (populace)

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

$$\sigma^2 = \frac{1}{n} \sum_{j=1}^{n} n_j x_j^2 - \left(\frac{1}{n} \sum_{j=1}^{n} n_j x_j \right)^2$$

Rozklad rozptylu

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

$$\begin{aligned}
\sigma^2 &= \frac{1}{n} \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{i,j} - \bar{x})^2 \\
&= \frac{1}{n} \left(\sum_{j=1}^{k} n_j (\bar{x}_j - \bar{x})^2 + \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{i,j} - \bar{x}_j)^2 \right) \\
&= \underbrace{\frac{1}{n} \left(\sum_{j=1}^{k} n_j (\bar{x}_j - \bar{x})^2 \right)}_{\text{meziskupinová variabilita}} + \underbrace{\frac{1}{n} \left(\sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{i,j} - \bar{x}_j)^2 \right)}_{\text{vnitroskupinová variabilita}}
\end{aligned}$$

Intermezzo

- 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?

Intermezzo

- 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.* $\bar{x}_{\text{nový}} \doteq 12{,}43$, $s^2_{x,\ \text{nový}} \doteq 5{,}10$ □

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

- 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

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

Intermezzo

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

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

- Z intervalu $\langle 0, 1 \rangle$ náhodně vybereme dvě čísla x a y. Jaká je pravděpodobnost jevu, že $2y \leq x^2$?

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

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

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

Průnik množin $\mathcal{A}$ a $\mathcal{B}$

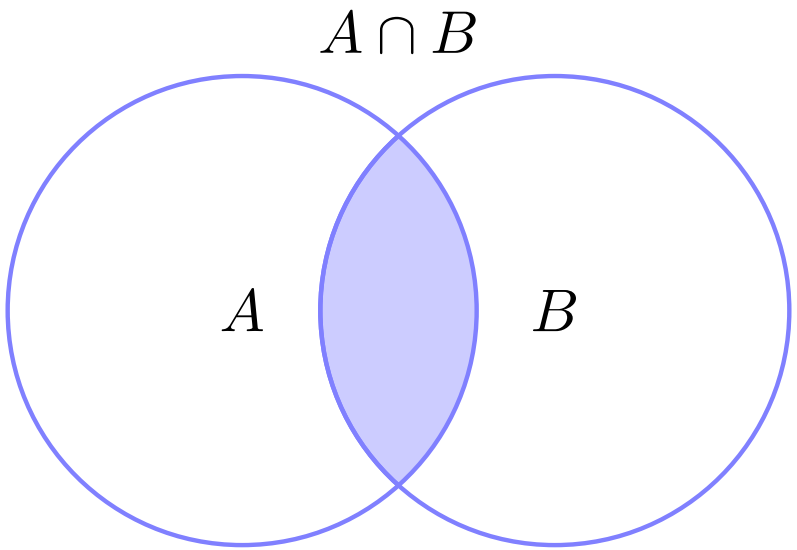
Průnik množin $\mathcal{A}$ a $\mathcal{B}$ je taková množina $\mathcal{A} \cap \mathcal{B}$, že $\mathcal{A} \cap \mathcal{B} = \{\forall\ x : x \in \mathcal{A} \wedge x \in \mathcal{B}\}$.

Asymetrický rozdíl množin $\mathcal{A}$ a $\mathcal{B}$

Asymetrický rozdíl množin $\mathcal{A}$ a $\mathcal{B}$ je taková množina $\mathcal{A} - \mathcal{B}$, že $\mathcal{A} - \mathcal{B} = \{\forall\ x : x \in \mathcal{A} \wedge x \notin \mathcal{B}\}$.

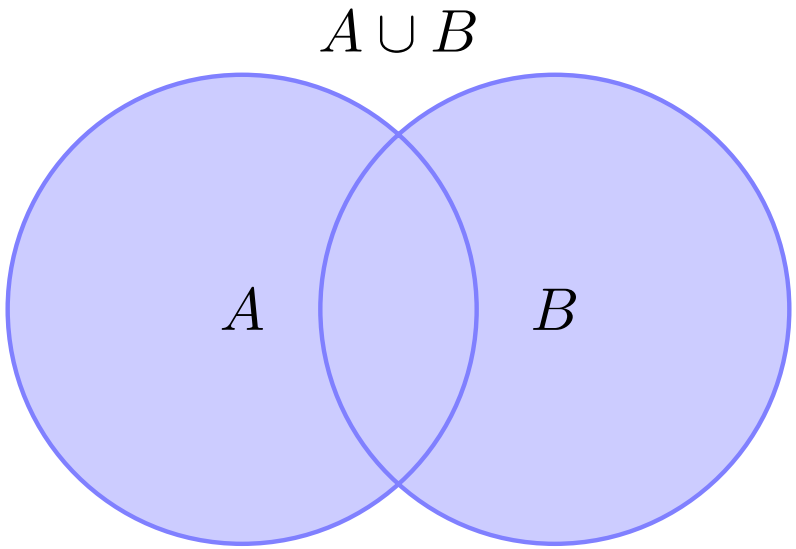
Průnik jevů A a B

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



Sjednocení jevů A a B

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



Vlastnosti pravděpodobnosti

- nechť A a B jsou náhodné jevy, pak platí

$$0 \leq P(A) \leq 1$$
$$0 \leq P(B) \leq 1$$

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

$$P(A) \leq P(B)$$

- vždy však

$$P(A^C) = 1 - P(A)$$
$$P(A \cup B) = P(A) + P(B) - P(A \cap B),$$

  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.

Literatura


Hindls, Richard, Stanislava Hronová, Jan Seger a Jakub Fischer. *Statistika pro ekonomy*. Praha: Professional Publishing, 2007. ISBN: 978-80-86946-43-6.


Marek, Luboš. *Statistika v příkladech*. Praha: Professional Publishing, 2015. ISBN: 978-80-7431-153-6.

Děkuji za pozornost!

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