

# Variabilita a její míry, rozklad variance

—  
Supplementum ke cvičení 4ST210 Statistika pro finance

Lubomír Štěpánek<sup>1, 2</sup>



<sup>1</sup>Oddělení biomedicínské statistiky  
Ústav biofyziky a informatiky  
1. lékařská fakulta  
Univerzita Karlova, Praha



<sup>2</sup>Katedra 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 kdekoli 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

- 1 Opakování
- 2 Míry variability
- 3 Vlastnosti a rozklad rozptylu
- 4 Literatura

# Příklad

- 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?

# Příklad

- 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.

# Příklad

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

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

Určeme

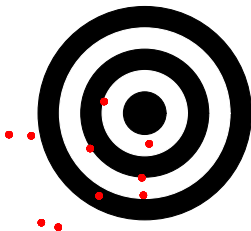
- (i) mediánový věk hudebníků v tělese.
- (ii) modální věk hudebníků v tělese.
- (iii) první a třetí kvartil věku hudebníků v tělese.
- (iv) 80-tý percentil věku hudebníků v tělese.

# Příklad

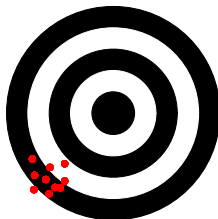
- První dělník je schopen vyhloubit výkop za 8 hodin, druhý dělník za 6 hodin a třetí dělník pak za 10 hodin.
  - (i) Za kolik hodin vyhloubí jeden výkop, pokud budou pracovat společně?
  - (ii) Za kolik průměrně hodin je vyhlouben výkop, pokud pracují všichni tři dělníci a každý pracuje na svých výkopech?

# Intuitivní pohled na variabilitu (a střední hodnotu)

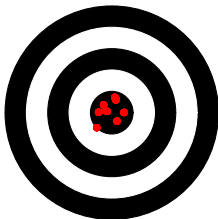
lukostřelec A



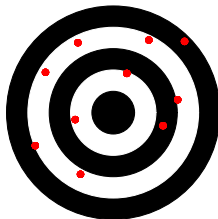
lukostřelec B



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  $x_1, x_2, \dots, x_n$  spočítáme jejich varianční rozpětí  $R$  jako

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

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

- lze snadno ukázat, že přibližně platí (tzv. *pravidlo six sigma*)

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

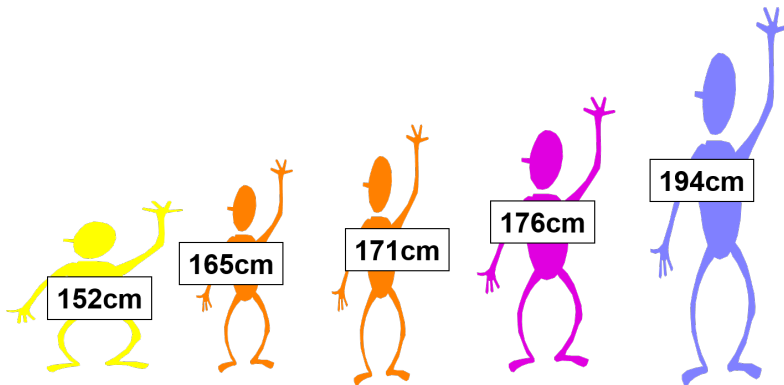
kde  $s_x$  je výběrová směrodatná odchylka daného výběru

► MS Excel®

$\text{MAX}(x_1 : x_n) - \text{MIN}(x_1 : x_n)$

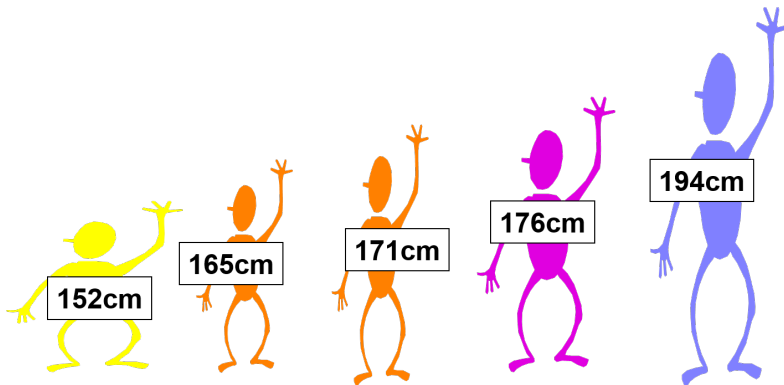
# Varianční rozpětí

- určeme varianční rozpětí z následujícího souboru tělesných výšek



# Varianční rozpětí

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



# Kvartilové rozpětí

- kvartilové rozpětí nad souborem čísel  $x_1, x_2, \dots, x_n$  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  $x_1, x_2, \dots, x_n$ , tedy obecně pro  $p$ -tý kvantil  $\tilde{x}_p$  je

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

kde  $x_{(k)}$  je  $k$ -té nejmenší číslo mezi čísly  $x_1, x_2, \dots, x_n$ , dále kde  $0 \leq p \leq 1$  (zde  $p = 0,25$  a  $p = 0,75$ ) a kde  $\lfloor x \rfloor$  značí dolní celou část čísla  $x$ , tedy nejvyšší celé číslo takové, že nepřevyší  $x$

# Decilové rozpětí a percentilové rozpětí

- nad souborem čísel  $x_1, x_2, \dots, x_n$  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  $x_1, x_2, \dots, x_n$

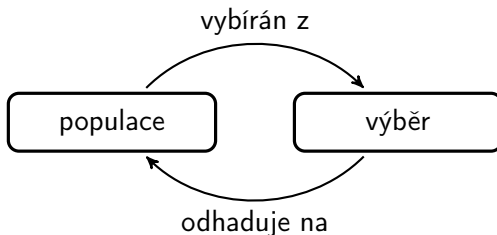
- tedy obecně  $100p\%$  rozpětí nad souborem čísel  $x_1, x_2, \dots, x_n$  je definováno jako

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

tak, že  $\tilde{x}_p$  je  $p$ -tý kvantil a zde  $0 \leq p < 0,5$

# Vztah populace a výběru

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



# Populační rozptyl a populační směrodatná odchylka

- předpokládejme, že soubor celé populace je tvořen právě  $n$  čísly  $x_1, x_2, \dots, x_n$
- pak pro těchto  $n$  čísel  $x_1, x_2, \dots, x_n$  spočítáme populační rozptyl  $\sigma^2$  (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  $\sigma$  (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  $x_1, x_2, \dots, x_n$



# Výběrový rozptyl a výběrová směrodatná odchylka

- předpokládejme, že máme výběr z populace, který je tvořen  $n$  čísly  $x_1, x_2, \dots, x_n$
- pak pro těchto  $n$  čísel  $x_1, x_2, \dots, x_n$  lze spočítat „odhad“<sup>1</sup> populačního rozptylu  $\sigma^2$  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  $\sigma$  „odhadnout“ pomocí výběrové směrodatné odchylky  $s_x$  jako

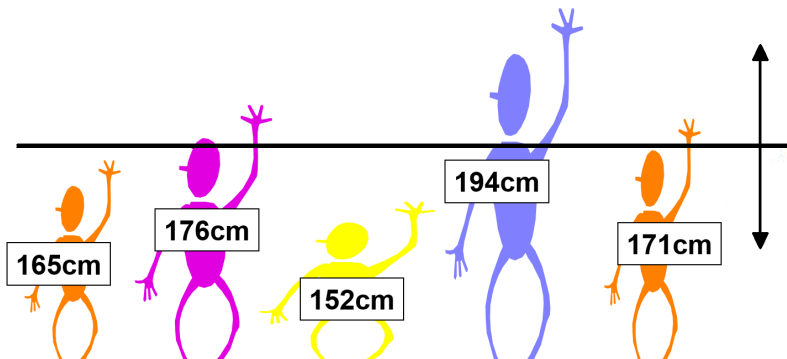
$$\sigma = \sqrt{\frac{n}{n-1}} s_x = \sqrt{\frac{n}{n-1}} \left( \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} \right),$$

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

<sup>1</sup>pojmy „odhad“ a „odhadnout“ budeme zatím vnímat jen intuitivně

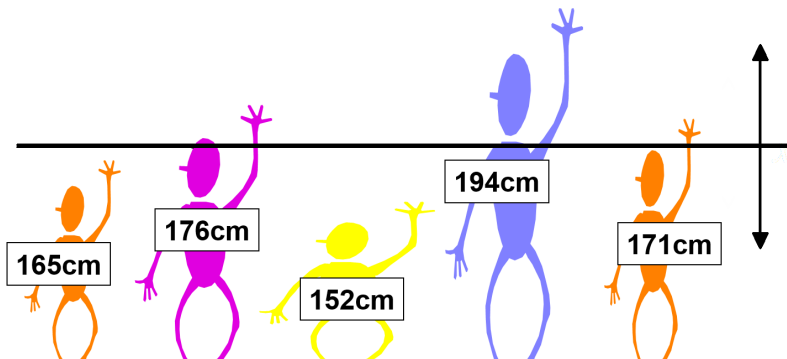
# Výpočet rozptylu a směrodatné odchylky

- určíme směrodatnou odchylku a rozptyl z následujícího výběru tělesných výšek



# Výpočet rozptylu a směrodatné odchylky

- určíme směrodatnou odchylku a rozptyl z následujícího výběru tělesných výšek
- $\sigma = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} \doteq 15,4 \text{ [cm]}; \quad \sigma^2 \doteq 237,2 \text{ [cm}^2\text{]}$
- $s_x = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} \doteq 13,8 \text{ [cm]}; \quad s_x^2 \doteq 189,8 \text{ [cm}^2\text{]}$



# Příklad

- 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

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

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

# Příklad

- 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.

# Výpočetní tvar (populačního) rozptylu

- pro  $n$  čísel  $x_1, x_2, \dots, x_n$  (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

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

# Výpočetní tvar (populačního) rozptylu

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

- 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



# Příklad

- 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?

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

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

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

# Rozklad rozptylu

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

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

# Odvození rozkladu rozptylu

$$\begin{aligned}s_x^2 &= \frac{1}{n} \sum_{j=1}^k \sum_{i=1}^{n_j} (x_{i,j} - \bar{x})^2 \\&= \frac{1}{n} \sum_{j=1}^k \sum_{i=1}^{n_j} ((x_{i,j} - \bar{x}_j) + (\bar{x}_j - \bar{x}))^2 \\&= \frac{1}{n} \sum_{j=1}^k \sum_{i=1}^{n_j} ((x_{i,j} - \bar{x}_j)^2 + 2(x_{i,j} - \bar{x}_j)(\bar{x}_j - \bar{x}) + (\bar{x}_j - \bar{x})^2) \\&= \frac{1}{n} \sum_{j=1}^k \left( \sum_{i=1}^{n_j} (x_{i,j} - \bar{x}_j)^2 + 2(\bar{x}_j - \bar{x}) \underbrace{\sum_{i=1}^{n_j} (x_{i,j} - \bar{x}_j)}_{=0} + n_j(\bar{x}_j - \bar{x})^2 \right)\end{aligned}$$

# Odvození rozkladu rozptylu

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

# Příklad

- 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.

# Příklad



- 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?

# Příklad

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



# 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