Organizace cvičení, úvod do statistiky a deskriptivní statistika

—

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

20. září 2019

Obsah

Online složka předmětu

- prezentace a další materiály ke cvičení jsou dostupné na

https://github.com/LStepanek/4ST201_Statistika

Organizace předmětu

- cvičící
  - Ing. MUDr. Lubomír Štěpánek
- email

lubomir.stepanek@vse.cz

- konzultační hodiny
  - v NB366 vždy v pátek 11:00–12:00, po předchozí emailové domluvě i jindy

Cíle předmětu

- smyslem je uvést studenty do deskriptivní statistiky, dále do teorie pravděpodobnosti a induktivní statistiky

Náplň předmětu

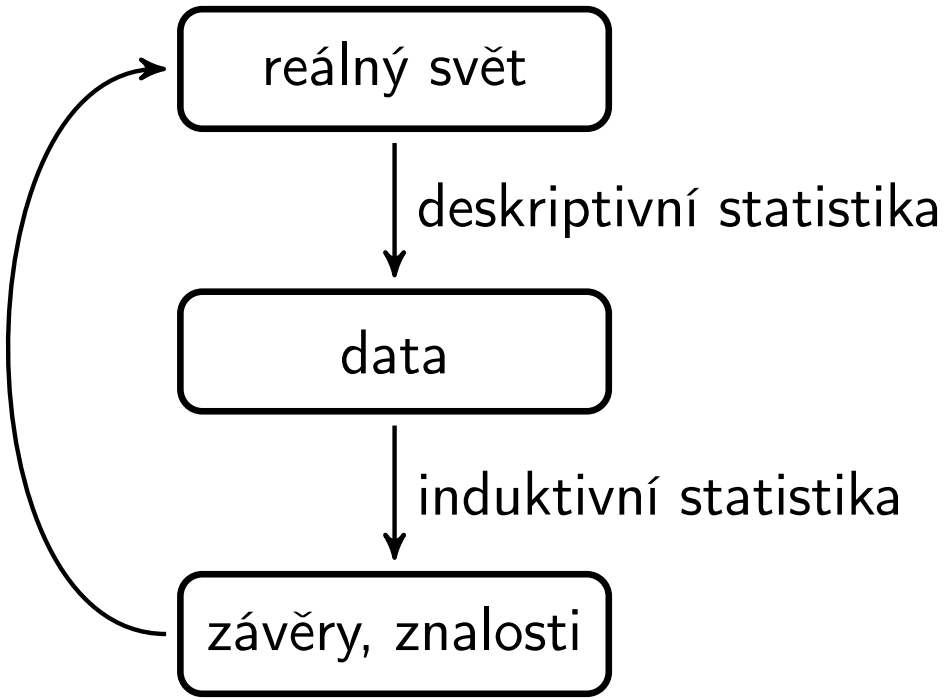
- bude procvičena látka na úrovni učebnice *Statistika pro ekonomy*[1]
- probírané okruhy
  - úvod do statistiky
  - deskriptivní statistika
  - pravděpodobnost
  - induktivní statistika
  - testování hypotéz
  - korelační a regresní analýza
  - indexní analýza

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

Dělení statistiky

- deskriptivní statistika
  - popisuje data, ale nedělá na nich žádné „velké“ závěry
- induktivní statistika
  - pozoruje konkrétní data a vyvozuje z nich obecné závěry, ovšem s udáním stupně jejich spolehlivosti

Vzájemný vztah deskriptivní a induktivní statistiky



Pojem *statistický znak, veličina*

- statistický znak, veličina
  - měřitelná (veličina) či jinak zjistitelná (znak) charakteristika našeho zájmu
  - např. tělesná výška, pohlaví, mzda, apod.

Pojem *statistická jednotka*

- statistická jednotka
  - základní atomický prvek zájmu, u nějž lze měřit nebo jinak získat hodnotu statistického znaku či veličiny
  - např. student, pacient, stát, molekula, apod.

Pojem *statistický soubor*

- statistický soubor
  - množina statistických jednotek (prvků statistického souboru)
  - např. třída žáků, kohorta pacientů, apod.

Vztah statistického znaku (veličiny), jednotky a souboru

- každá statistická jednota (prvek) statistického souboru má svou hodnotu[2] určitého zkoumaného statistického znaku či veličiny (jde-li o měřitelný znak)
- např. *ve školní třídě změříme tělesnou výšku každého žáka*
  - *školní třída* je statistický soubor
  - *žáci* jsou statistické jednotky (prvky)
  - *tělesná výška* je statistická veličina

[2] ta může eventuálně chybět nebo být neznámá (missing value)

Intermezzo

- měříme tělesné hmotnosti v kohortě pacientů-diabetiků na interním oddělení
- určeme, co je v takovém případě
  - statistickým znakem, resp. veličinou
  - statistickou jednotkou
  - statistickým souborem

Cíle deskriptivní statistiky

- cílem je popsat soubor dat
  - číselně (resp. tabulkou)
  - graficky
- popisné číselné ukazatele i grafické přístupy se liší, pokud jde
  - o kvantitativní statistický znak (veličinu)
  - o kvalitativní statistický znak

Kvantitativní znak (veličina)

- je vyjádřen číslem (a obvykle s jednotkou), kdy s číselnou hodnotou je smysluplné provádět aritmetické operace
- číslo tedy nenese pouze „katalogizační“ význam
- někdy též označován jako *numerický* typ dat

Dělení kvantitativního znaku (veličiny)

- dle spojitosti číselných hodnot
  - *spojitý* – hodnoty nabývají reálných čísel, nebo je na ně lze převést nějakou bijekcí
    - např. hmotnost, výška atd.
  - *diskrétní* – hodnoty jsou oddělená čísla obvykle ve smyslu počet či pořadí
    - např. počty pacientů atd.
- dle měřítka
  - *intervalová stupnice* – lze si smysluplně odpovědět, o kolik se dvě hodnoty liší, ale ne kolikrát
    - např. °C, datumy atd.
  - *poměrová stupnice* – lze si smysluplně odpovědět, o kolik se dvě hodnoty liší i kolikrát se liší
    - např. °K

Kvalitativní znak

- je vyjádřen obvykle slovně
- pokud vyjádřen číslem, pak nese pouze „katalogizační“ význam a není smysluplné s ním provádět aritmetické operace
- někdy též označován jako *kategorický* typ dat

Dělení kvalitativního znaku

- dle měřítka
  - *nominální stupnice* – dvě či více vzájemně se vylučujících, rovnocenných tříd, které nelze uspořádat na číselné ose
    - např. pohlaví {muž, žena}
    - rodinný stav muže {svobodný, ženatý, rozvedený, vdovec, registrovaný}
  - *ordinální stupnice* – kategorie je možné uspořádat vzestupně/sestupně, lze si smysluplně odpovědět, která hodnota je větší než jiná (ale ne o kolik, natož kolikrát)
    - např. pořadí v závodu, grade tumoru {1, 2, 3, 4} atd.

Intermezzo

- určete typ znaku a stupnice u následujících příkladů
  - procentuální úspěšnost v testu v souboru studentů jednoho kruhu [%]
  - soubor všech červencových dní jednoho roku (1., 2., . . . , 31.)
  - soubor čísel všech autobusů projíždějících zastávkou Kajetánka (174, 180, . . .)
  - soubor mutací genu CFTR (F508del, . . .)
  - bolest hodnocená pomocí VAS [0–10]
  - počet porodů v jedné porodnici za jednu noc
  - staging kolorektálního karcinomu {1, 2, 3, 4}

Popis kvantitativního znaku

- např. tělesná výška, glykémie, výše mzdy, atd.
- číselně
  - poloha (*center*)
    - aritmetický průměr, medián, modus
  - variabilita (*spread*)
    - rozpětí (min-max), směrodatná odchylka, rozptyl
  - tvar (*shape*)[3]
    - šikmost, špičatost
- graficky
  - krabicový diagram (boxplot)
  - histogram

[3] lépe jen graficky

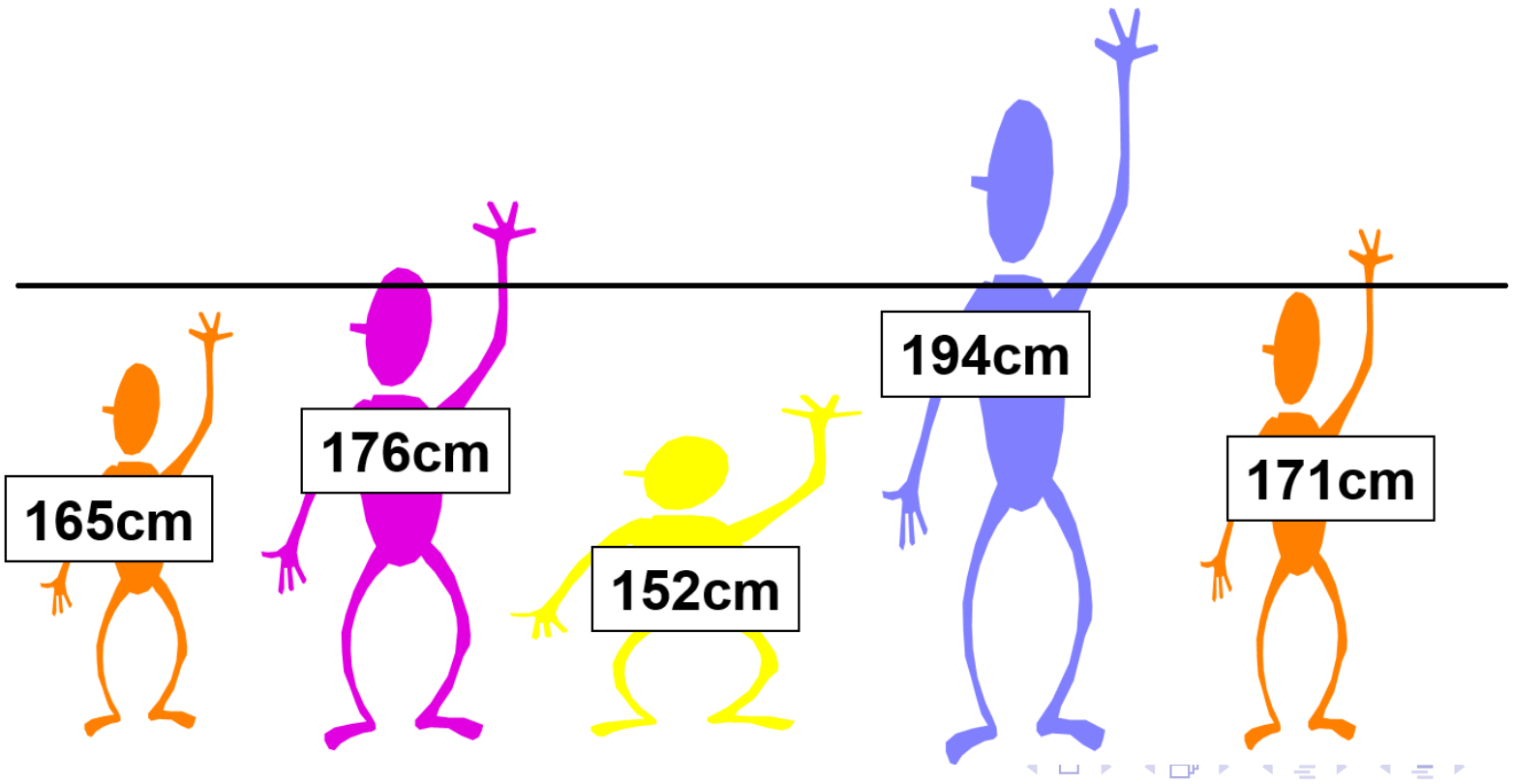
Aritmetický průměr

- pro n čísel $x_1, x_2, \ldots, x_n$ spočítáme jejich aritmetický průměr jako

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

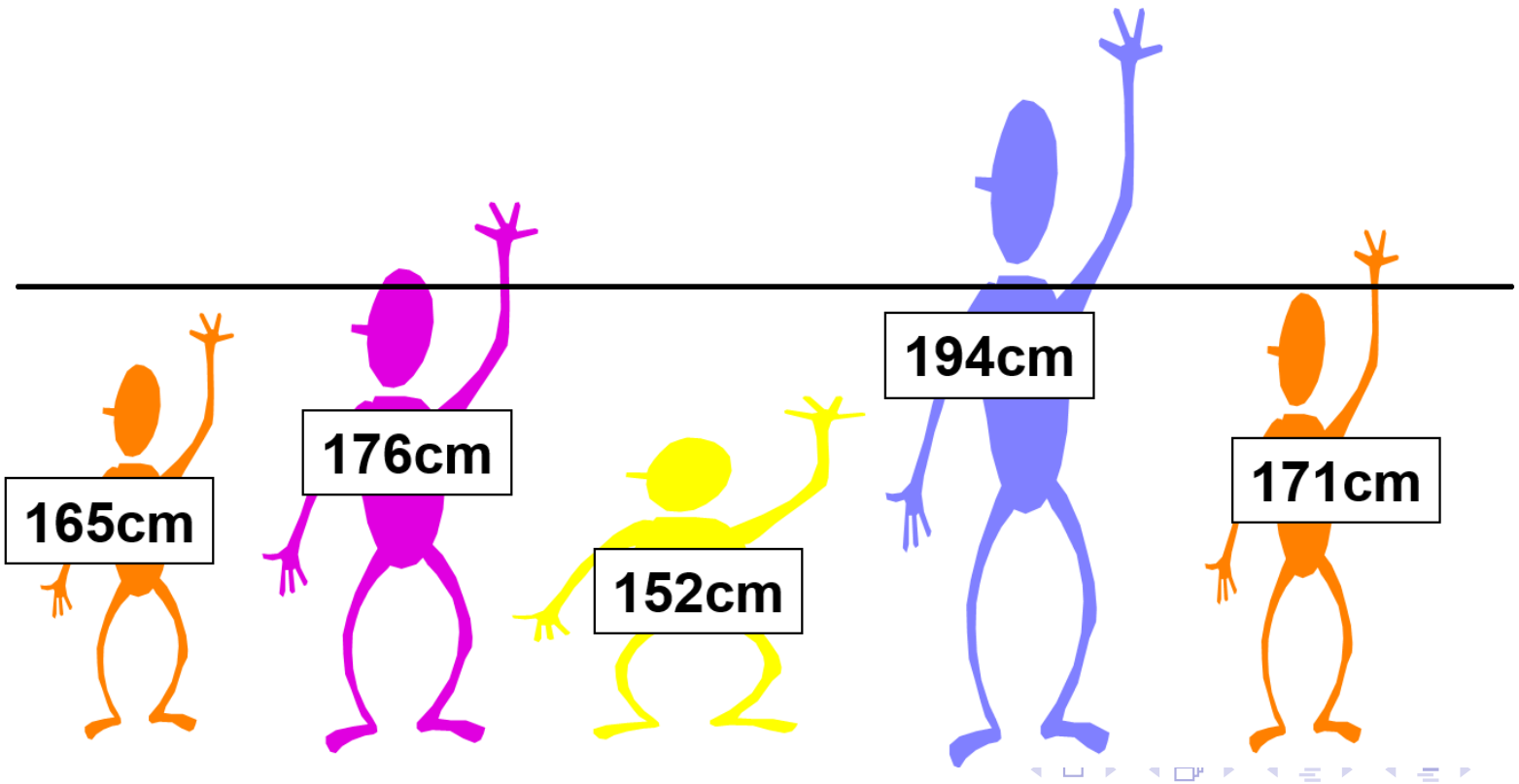
Intermezzo

- určeme aritmetický průměr z následujícího souboru tělesných výšek



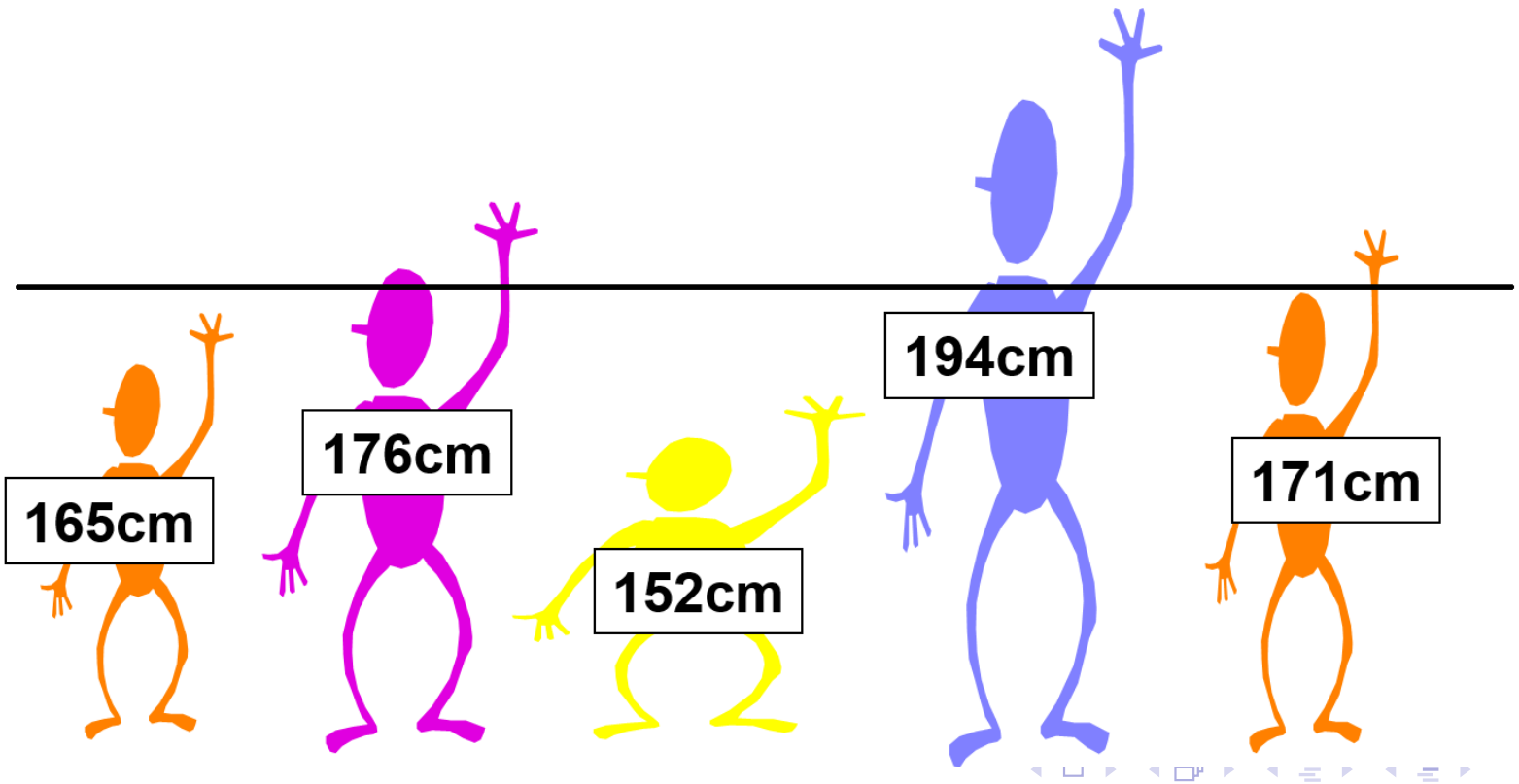
Intermezzo

- určeme aritmetický průměr z následujícího souboru tělesných výšek
- $\bar{x} = \frac{165+176+152+194+171}{5} \doteq 171{,}6\ [\mathrm{cm}]$



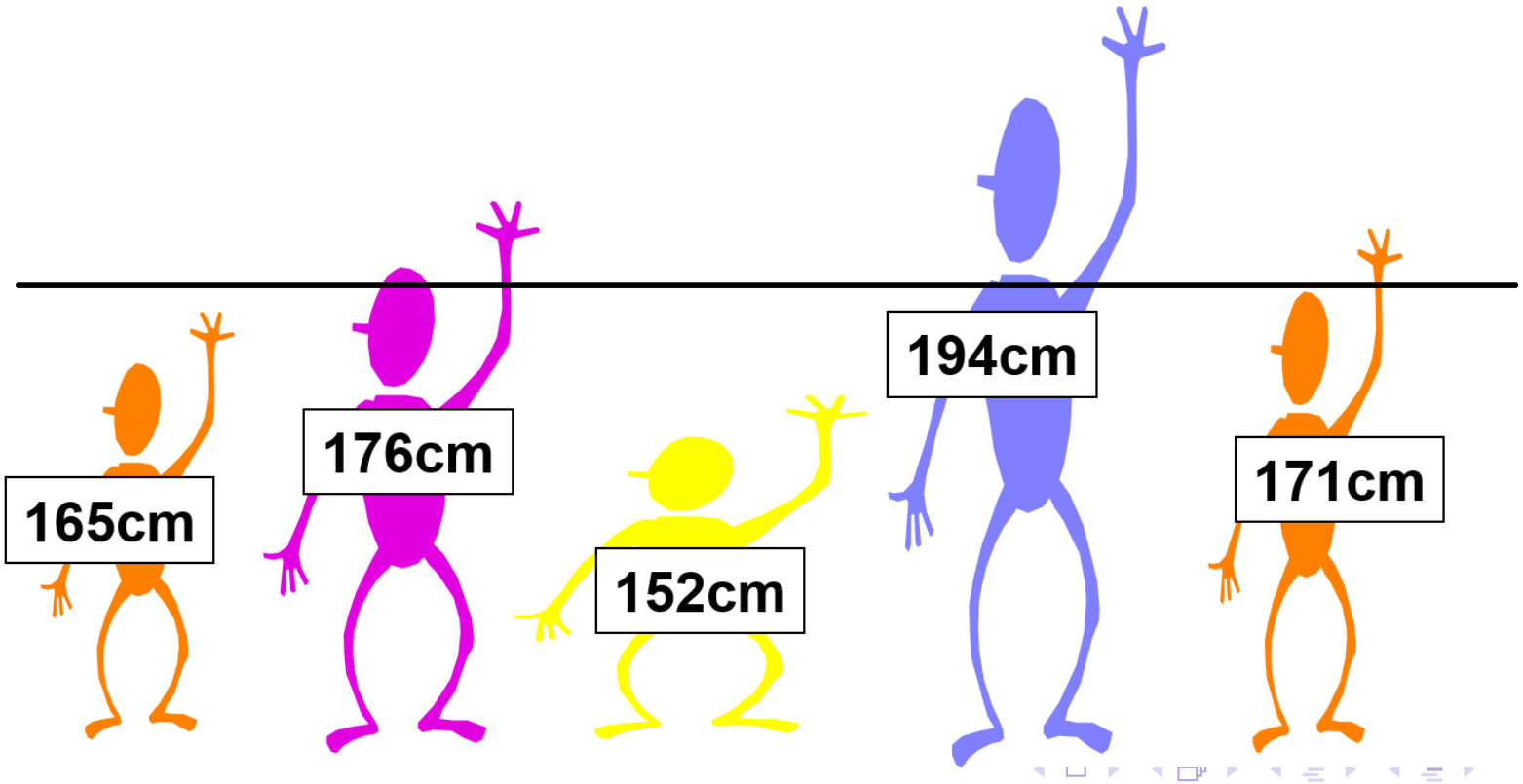
Intermezzo

- určeme aritmetický průměr z následujícího souboru tělesných výšek
- $\bar{x} = \frac{165+176+152+194+171}{5} \doteq 171{,}6\ [\mathrm{cm}]$
- kolik navzájem různých průměrů může mít jeden soubor čísel?



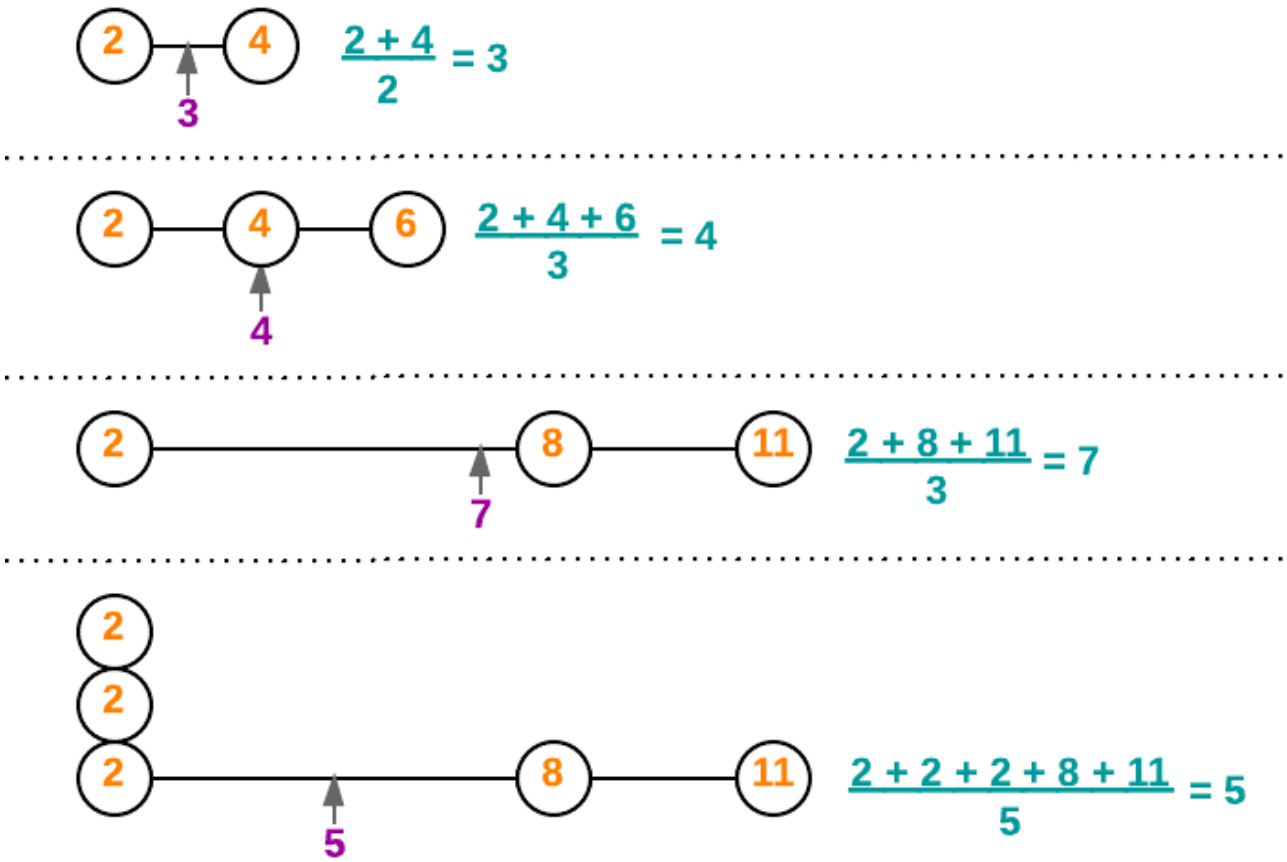
Intermezzo

- určeme aritmetický průměr z následujícího souboru tělesných výšek
- $\bar{x} = \frac{165+176+152+194+171}{5} \doteq 171{,}6\ [\mathrm{cm}]$
- kolik navzájem různých průměrů může mít jeden soubor čísel?
- pouze jeden



Geometrická interpretace aritmetického průměru

- pokud zavěsíme n jednogramových závaží na pozice čísel $x_1, x_2, \ldots, x_n$ pravítka, hodnota průměru $\bar{x}$ je v těžišti soustavy

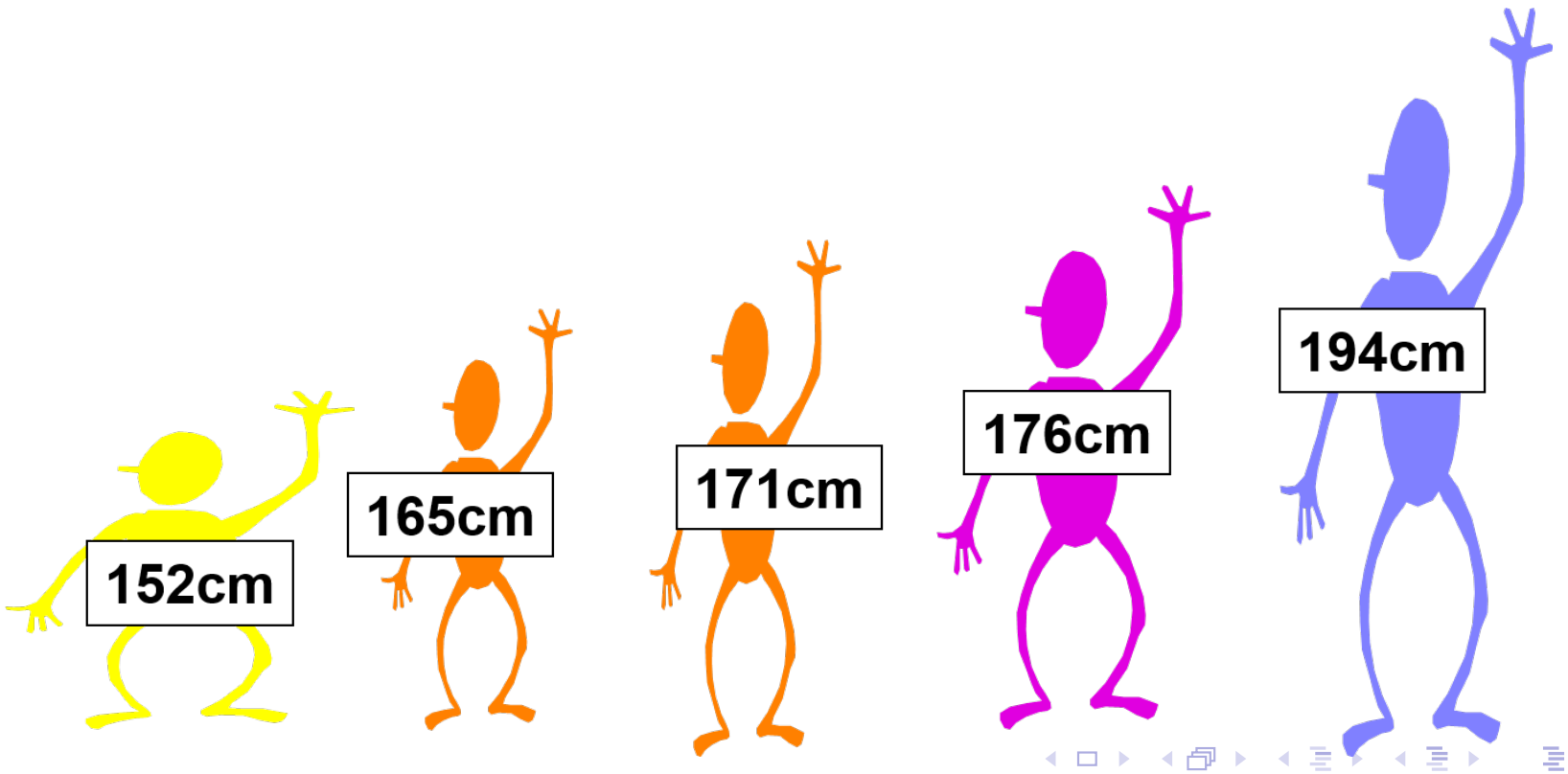


Medián

- medián je „prostřední“ prvek, zhruba polovina hodnot je větší než medián a zbylá polovina hodnot je menší než medián
- pro n čísel $x_1, x_2, \ldots, x_n$ zjistíme jejich medián tak, že
  - (i) čísla seřadíme vzestupně
  - (ii) medián $\tilde{x}$ je prostřední hodnota (pro n liché), resp. aritmetický průměr z „prostředních“ dvou hodnot (pro n sudé)

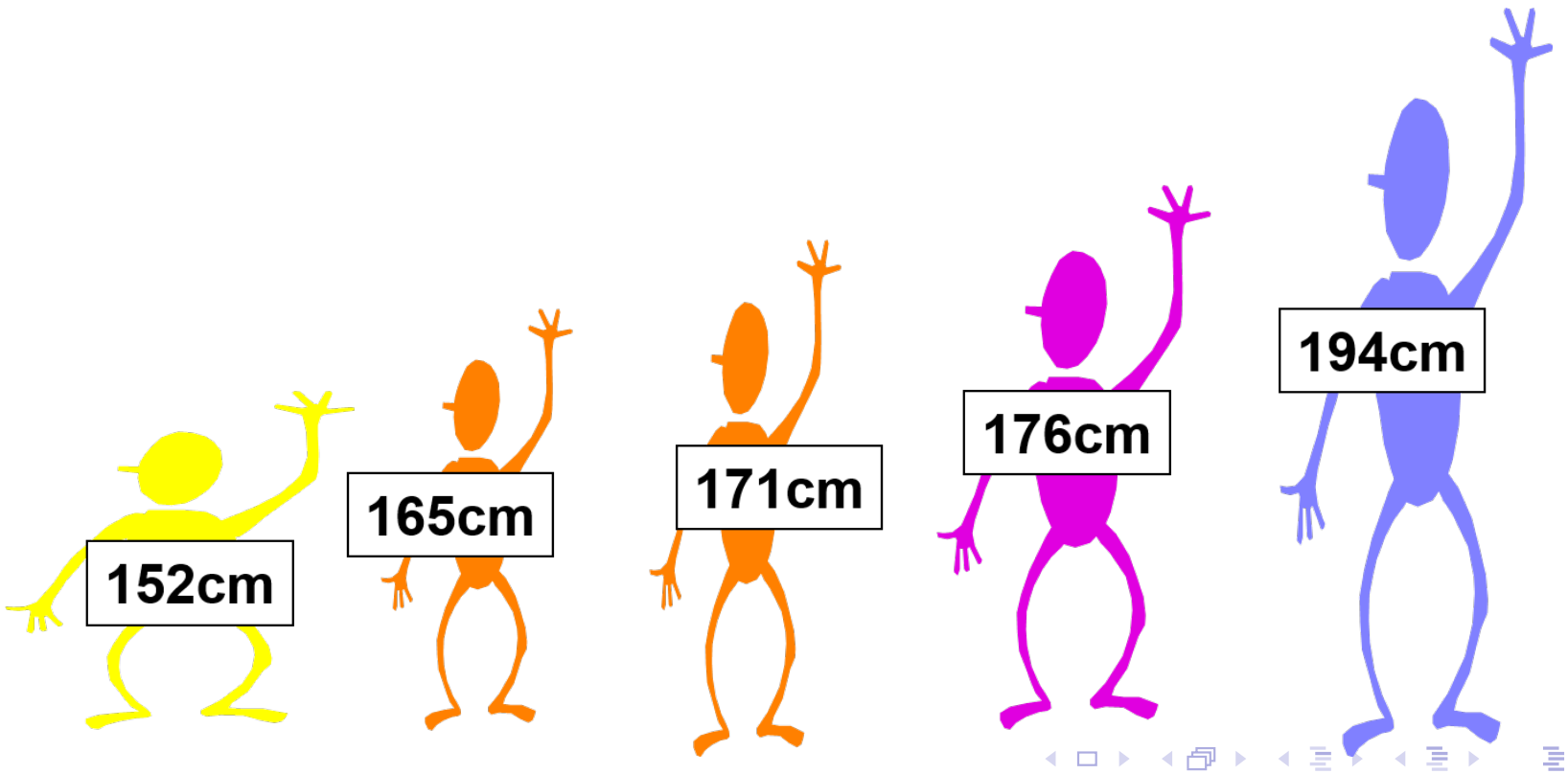
Intermezzo

- určeme medián z následujícího souboru tělesných výšek



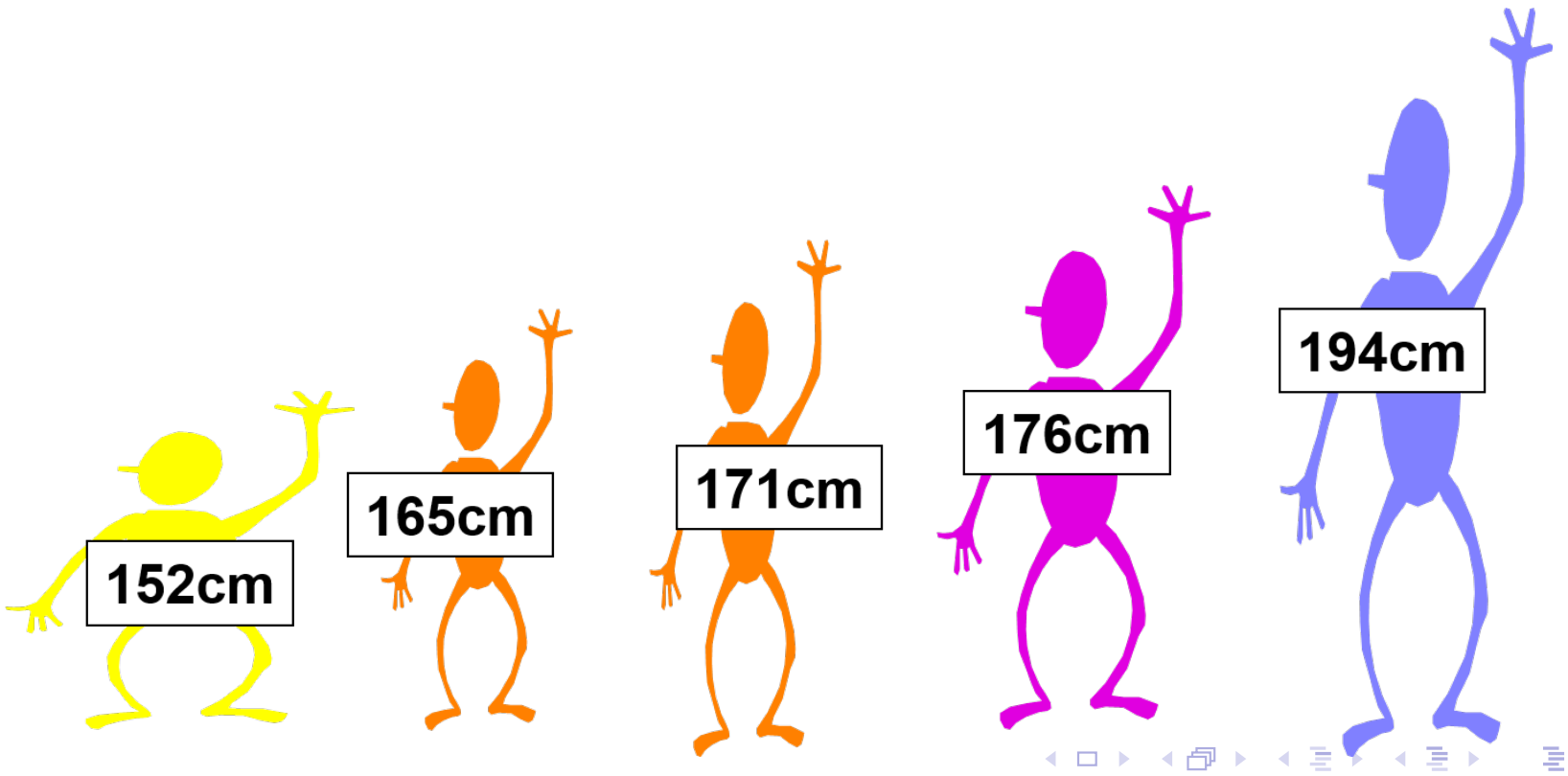
Intermezzo

- určeme medián z následujícího souboru tělesných výšek
- $\tilde{x} = 171\ [\mathrm{cm}]$



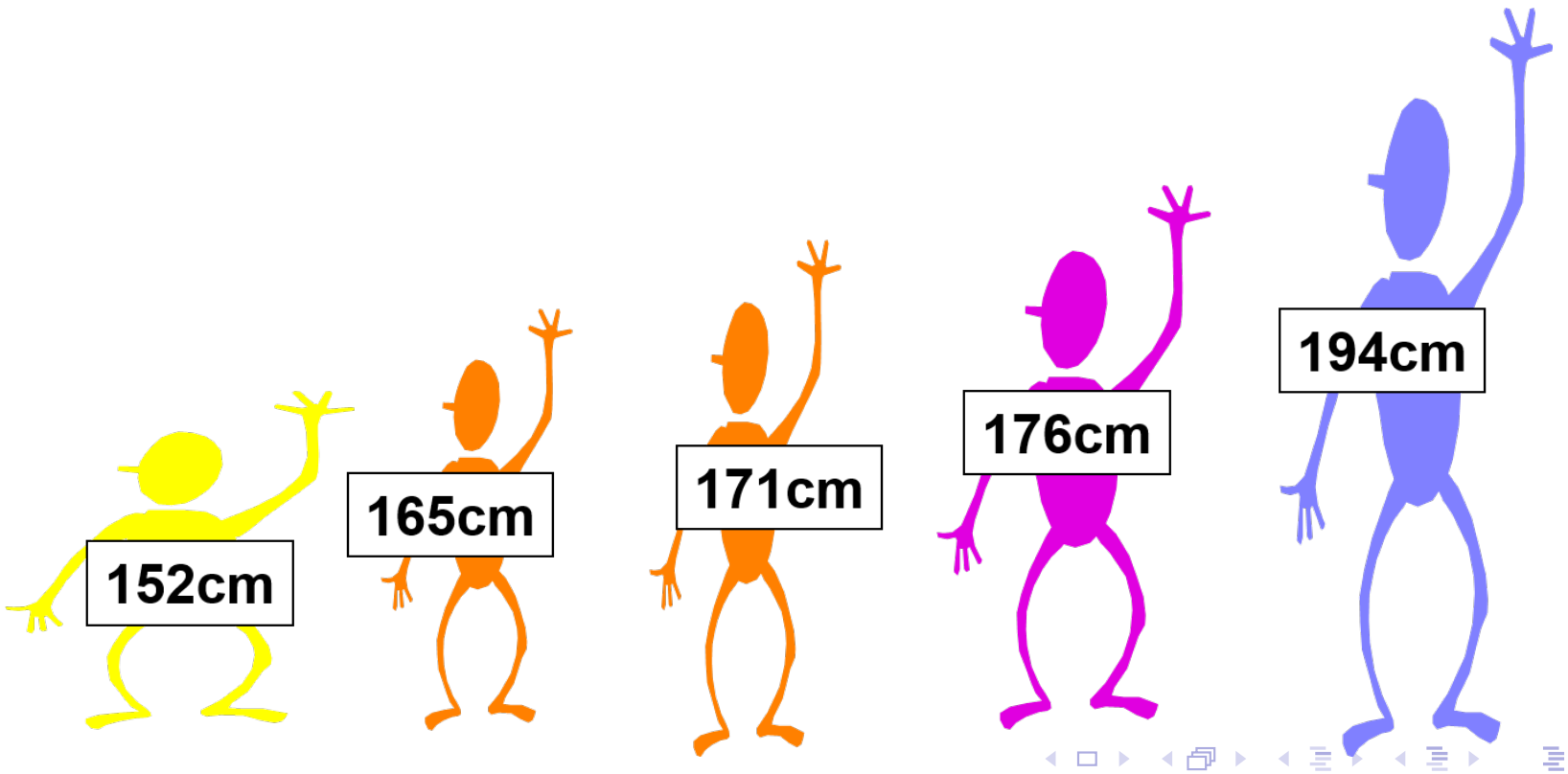
Intermezzo

- určeme medián z následujícího souboru tělesných výšek
- $\tilde{x} = 171\ [\mathrm{cm}]$
- kolik navzájem různých mediánů může mít jeden soubor čísel?



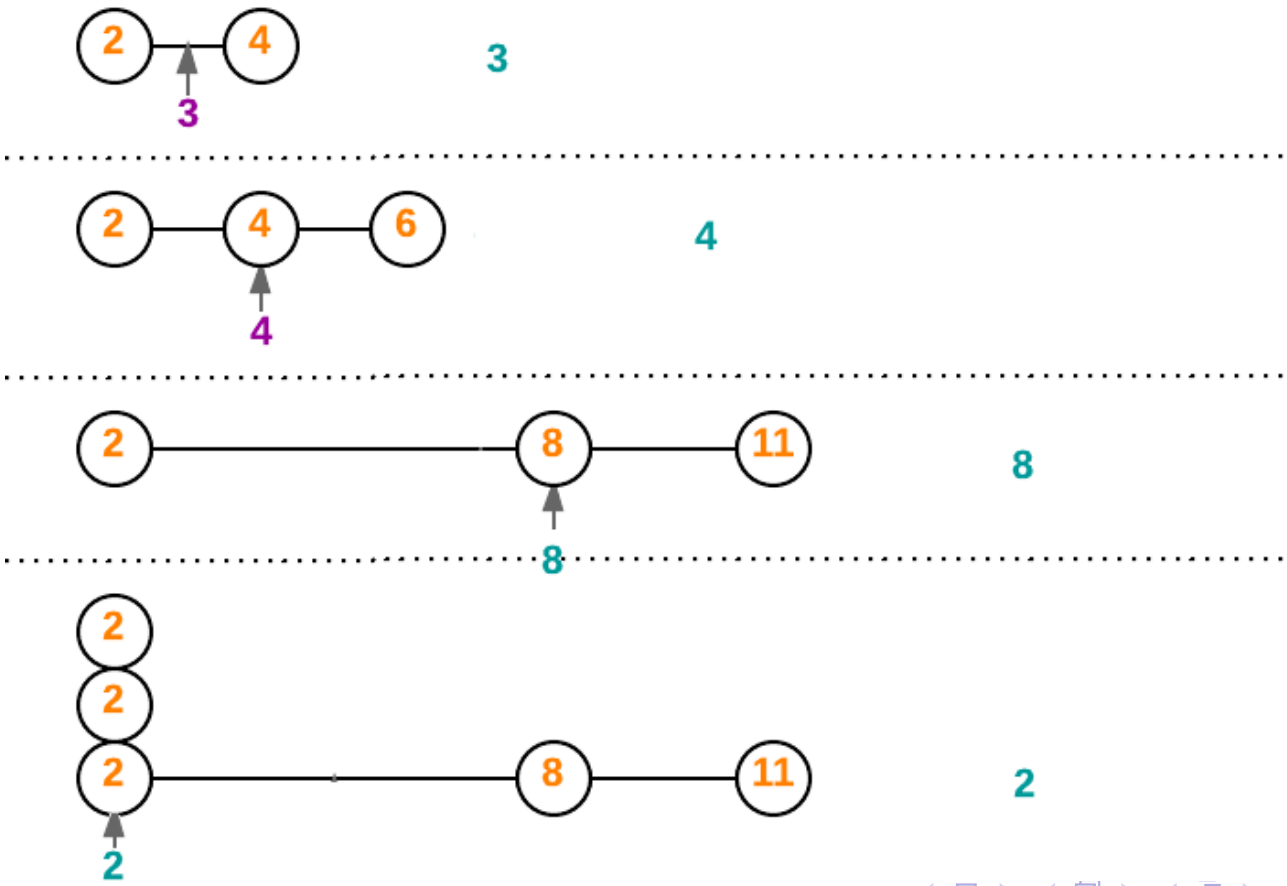
Intermezzo

- určeme medián z následujícího souboru tělesných výšek
- $\tilde{x} = 171\ [\mathrm{cm}]$
- kolik navzájem různých mediánů může mít jeden soubor čísel?
- pouze jeden



Geometrická interpretace mediánu

- pokud na pravítku vyznačíme pozice čísel $x_1, x_2, \ldots, x_n$, hodnota mediánu $\tilde{x}$ má od všech vyznačených bodů nejmenší možný součet vzdáleností

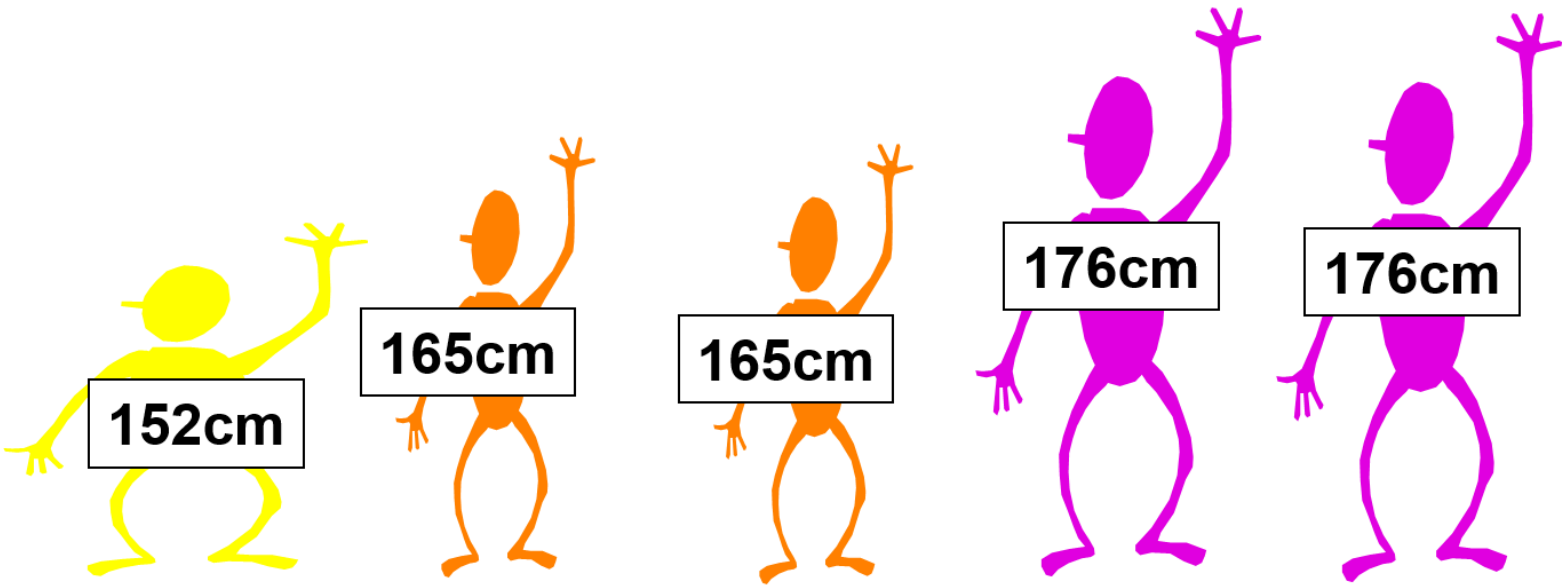


Modus

- modus je hodnota statistického znaku, který se v souboru čísel vyskytuje nejčastěji
  - pozor, modem není četnost takového prvku, tj. v souboru $\{10, 11, 11, 12\}$ je modem hodnota 11, nikoliv 2

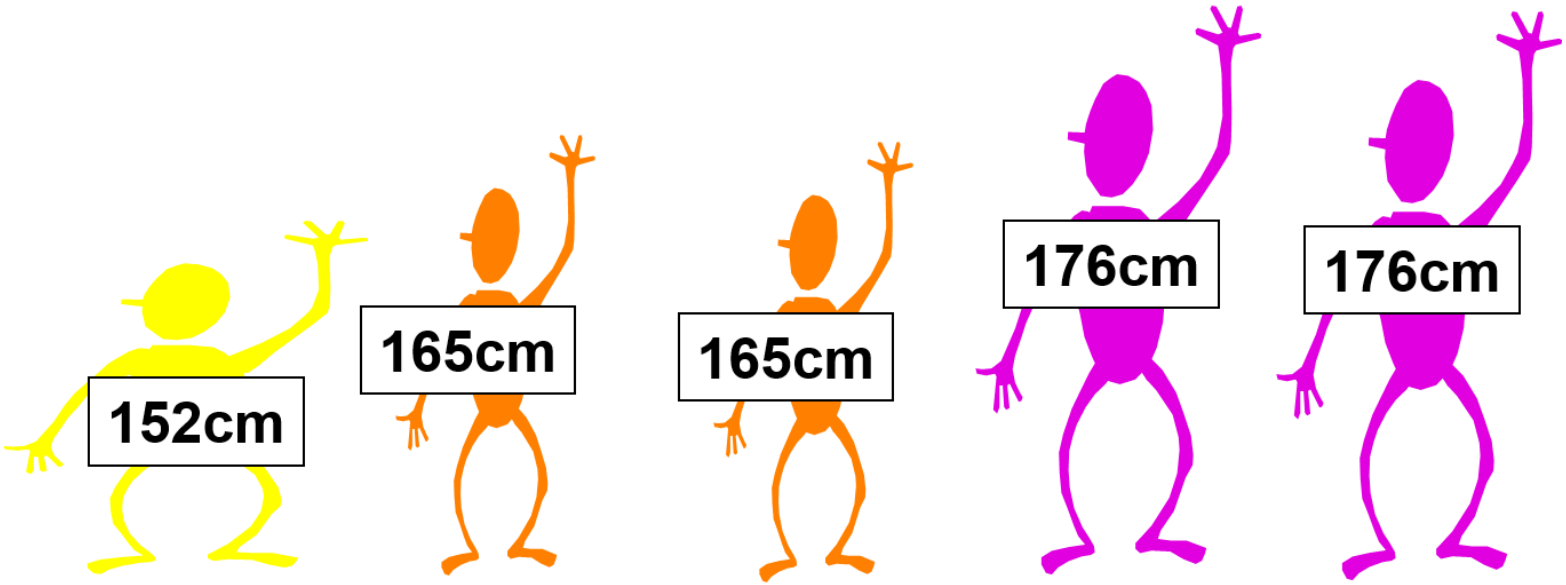
Intermezzo

- určeme modus z následujícího souboru tělesných výšek



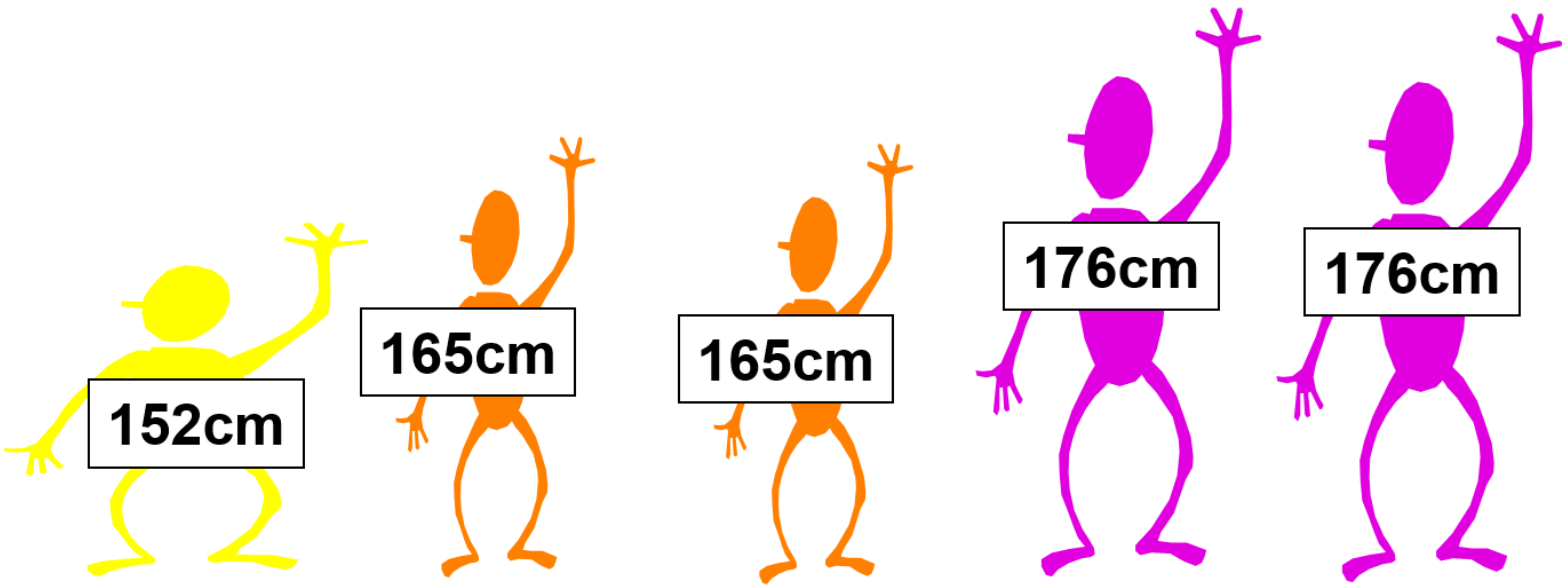
Intermezzo

- určeme modus z následujícího souboru tělesných výšek
- $\hat{x} = \{165; 176\}$ [cm]



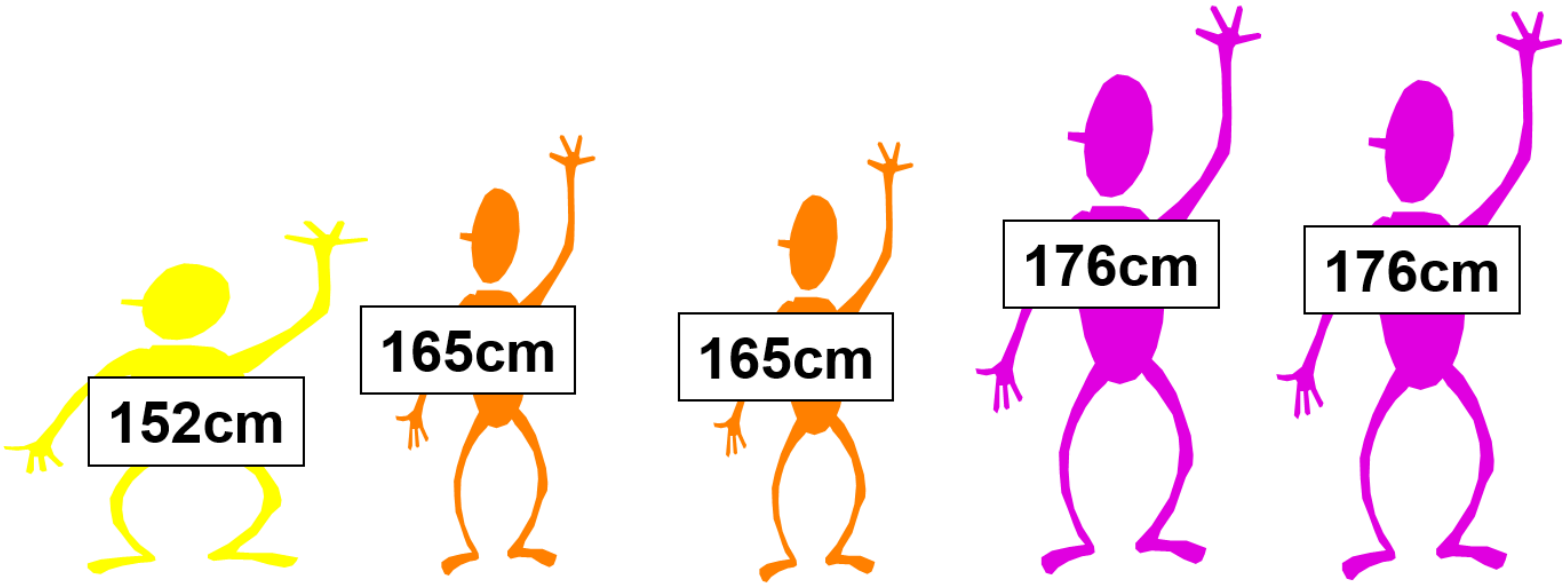
Intermezzo

- určeme modus z následujícího souboru tělesných výšek
- $\hat{x} = \{165; 176\}\ [\mathrm{cm}]$
- kolik navzájem různých modů může mít jeden soubor čísel?



Intermezzo

- určeme modus z následujícího souboru tělesných výšek
- $\hat{x} = \{165; 176\}$ [cm]
- kolik navzájem různých modů může mít jeden soubor čísel?
- alespoň jeden



Intermezzo

- určeme aritmetický průměr a medián u každého z obou následujícího souborů

$$\boldsymbol{x}_1 = \{1, 2, 3, 4, 5\} \qquad \boldsymbol{x}_2 = \{1, 2, 3, 4, 90\}$$

Intermezzo

- určeme aritmetický průměr a medián u každého z obou následujícího souborů

$$\boldsymbol{x}_1 = \{1, 2, 3, 4, 5\} \qquad \boldsymbol{x}_2 = \{1, 2, 3, 4, 90\}$$

-

$$\bar{x}_1 = \tilde{x}_1 = 3; \qquad \bar{x}_2 = 20;\ \tilde{x}_2 = 3$$

Intermezzo

- určeme aritmetický průměr a medián u každého z obou následujícího souborů

$$\boldsymbol{x}_1 = \{1, 2, 3, 4, 5\} \qquad \boldsymbol{x}_2 = \{1, 2, 3, 4, 90\}$$

- $$\bar{x}_1 = \tilde{x}_1 = 3; \qquad \bar{x}_2 = 20;\ \tilde{x}_2 = 3$$

- která z měr polohy (průměr, medián) lépe vyhovuje „asymetrickým“ datům?

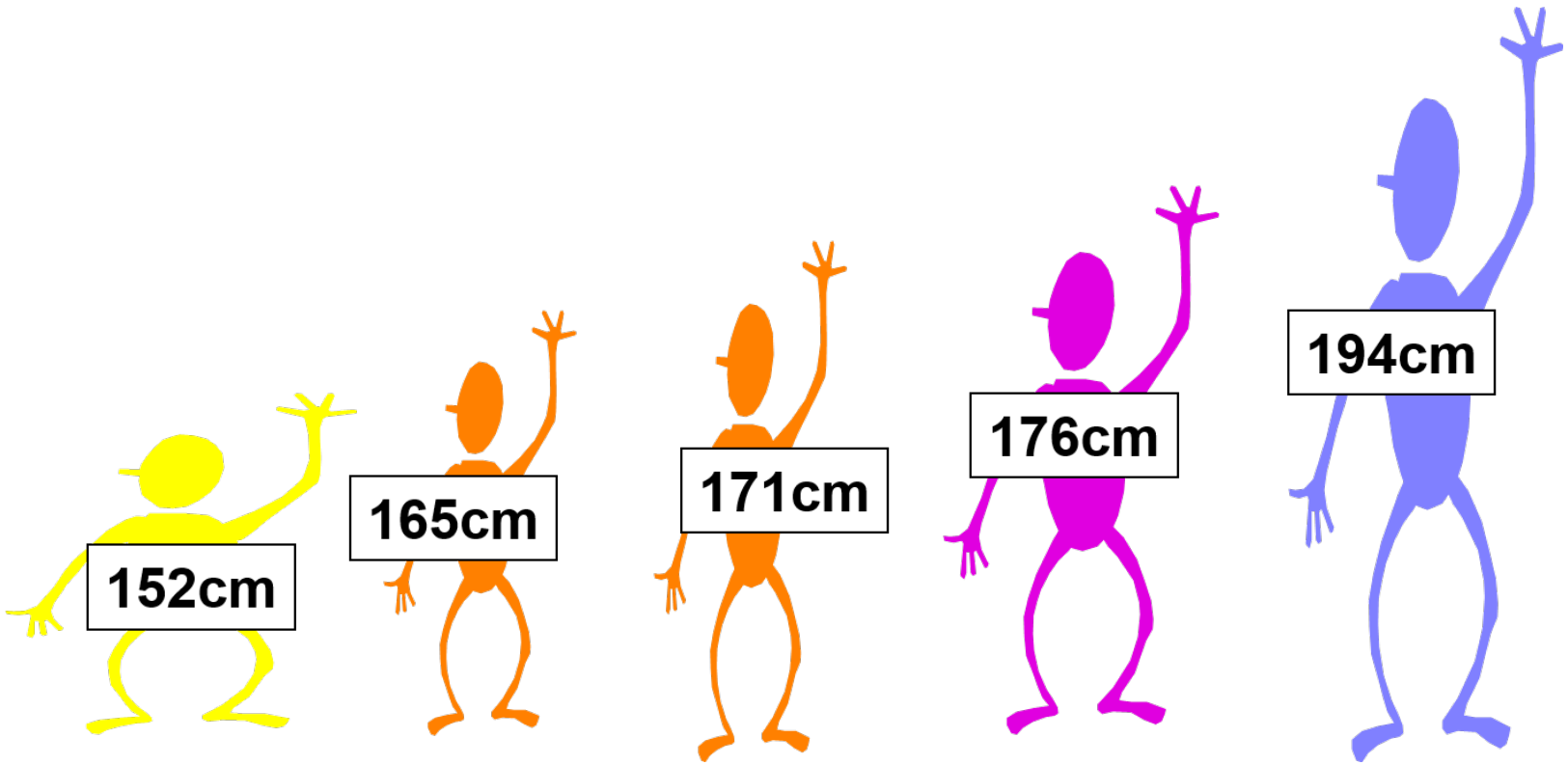
Rozpětí (min-max)

- rozpětí (min-max) je nejjednodušší měrou variability
- pro n čísel $x_1, x_2, \ldots, x_n$ spočítáme jejich rozpětí (min-max) jako

$$\text{min-max} = x_{\max} - x_{\min}$$

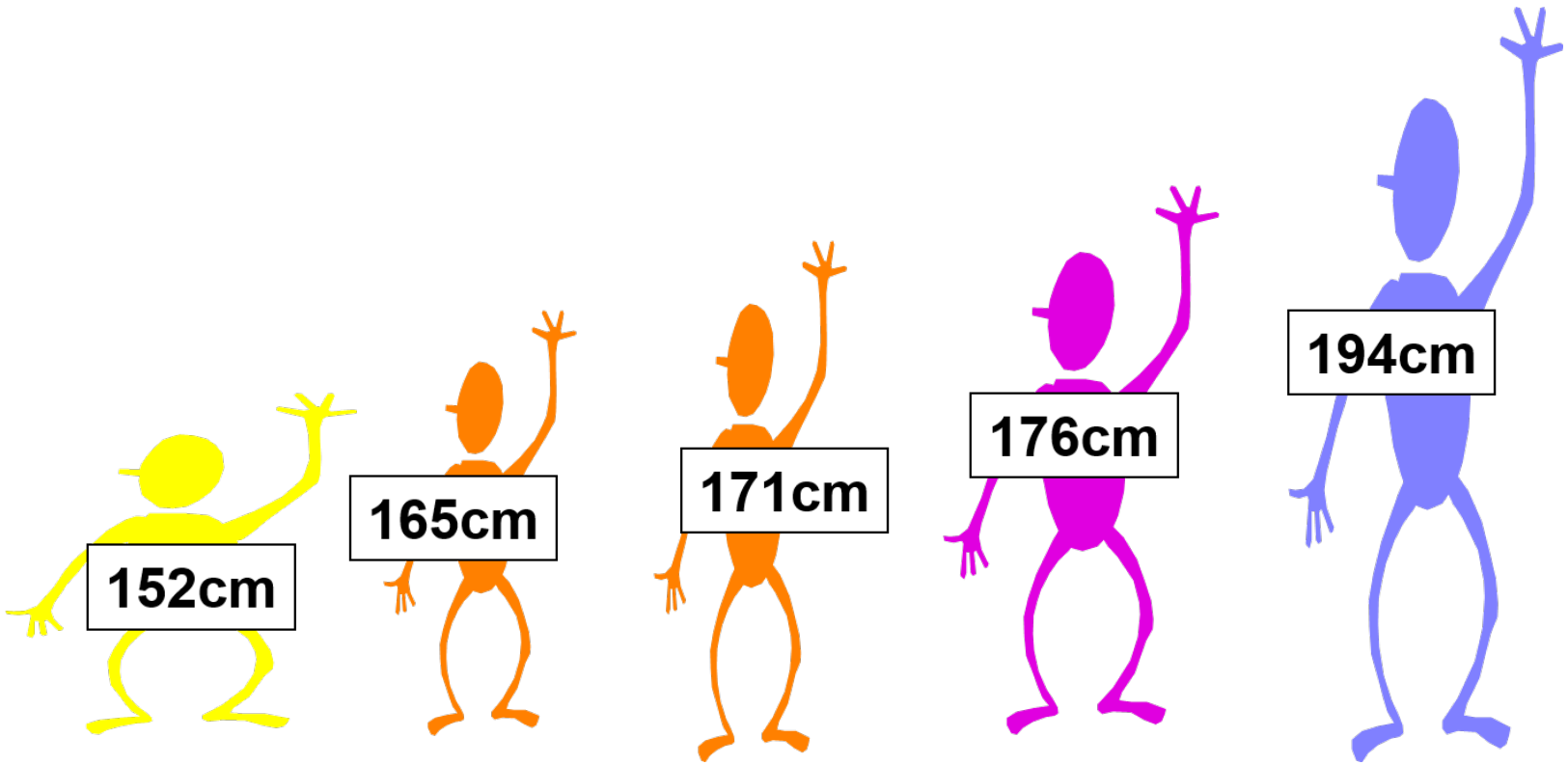
Intermezzo

- určeme rozpětí z následujícího souboru tělesných výšek



Intermezzo

- určeme rozpětí z následujícího souboru tělesných výšek
- $$\text{min-max} = x_{\max} - x_{\min} = 194 - 152 = 42 \text{ [cm]}$$



Směrodatná odchylka

- pro n čísel $x_1, x_2, \ldots, x_n$ spočítáme jejich směrodatnou odchylku jako

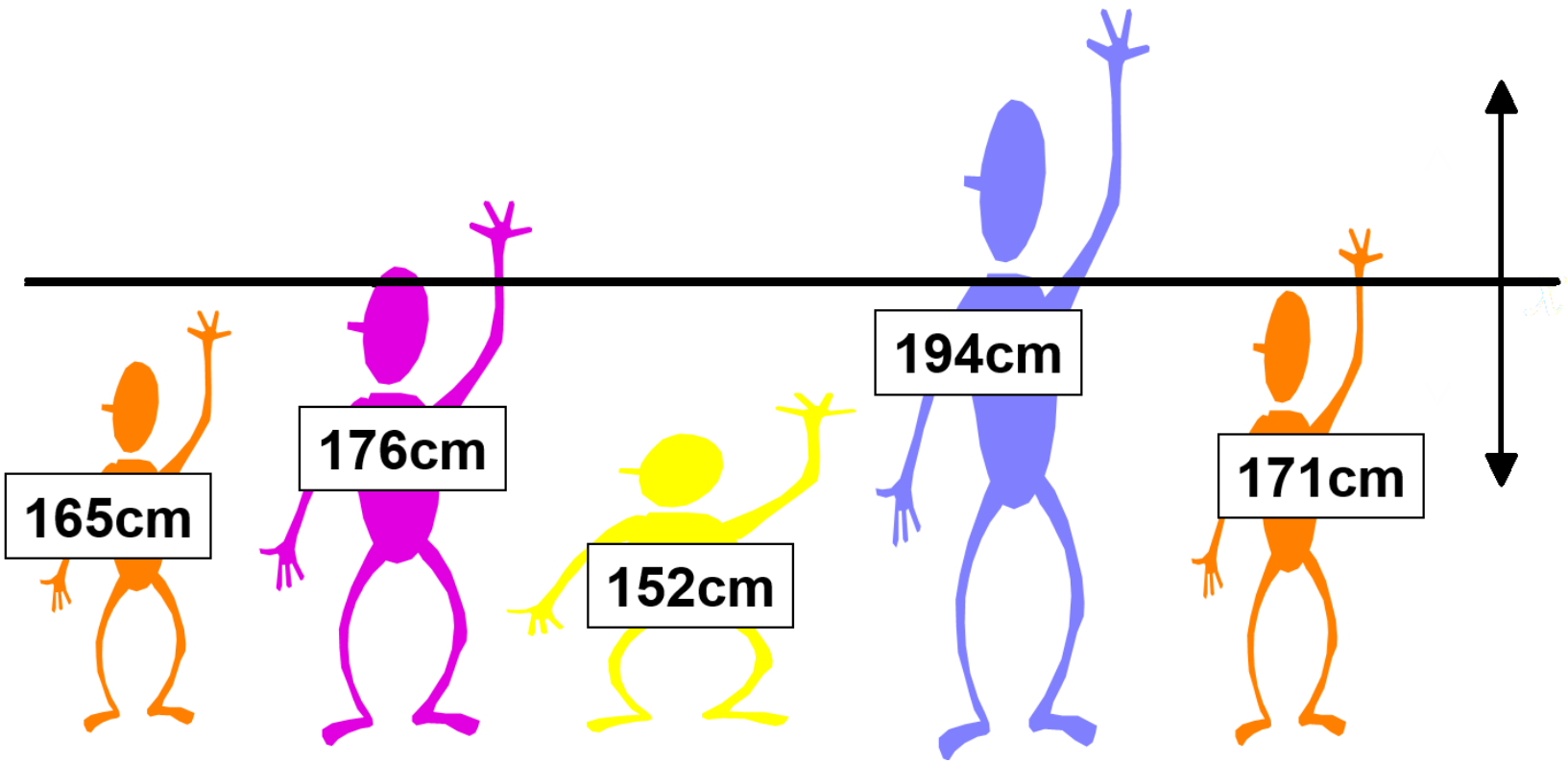
$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- pro stejných n čísel $x_1, x_2, \ldots, x_n$ spočítáme jejich rozptyl jako

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

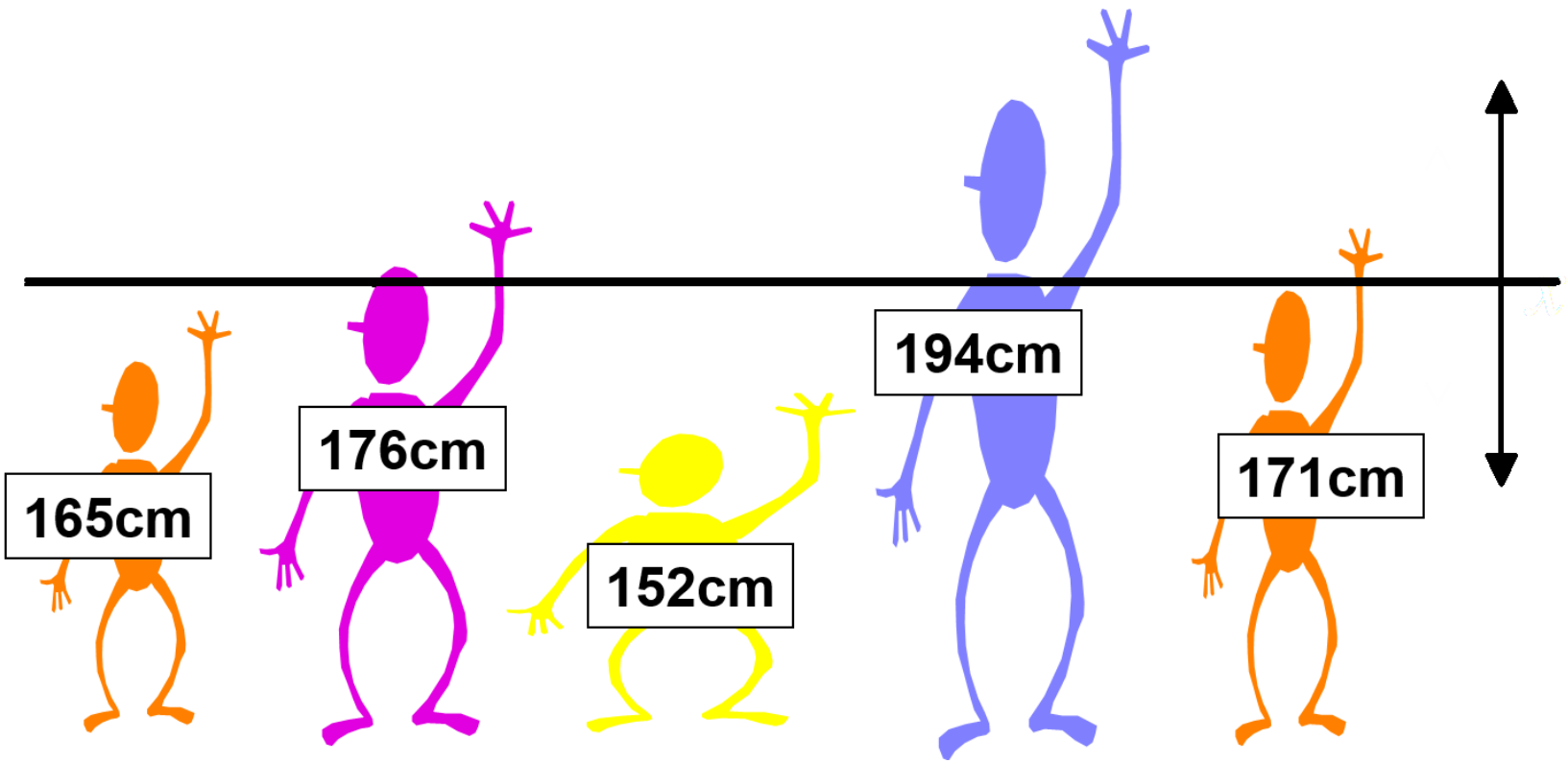
Intermezzo

- určeme směrodatnou odchylku a rozptyl z následujícího souboru tělesných výšek



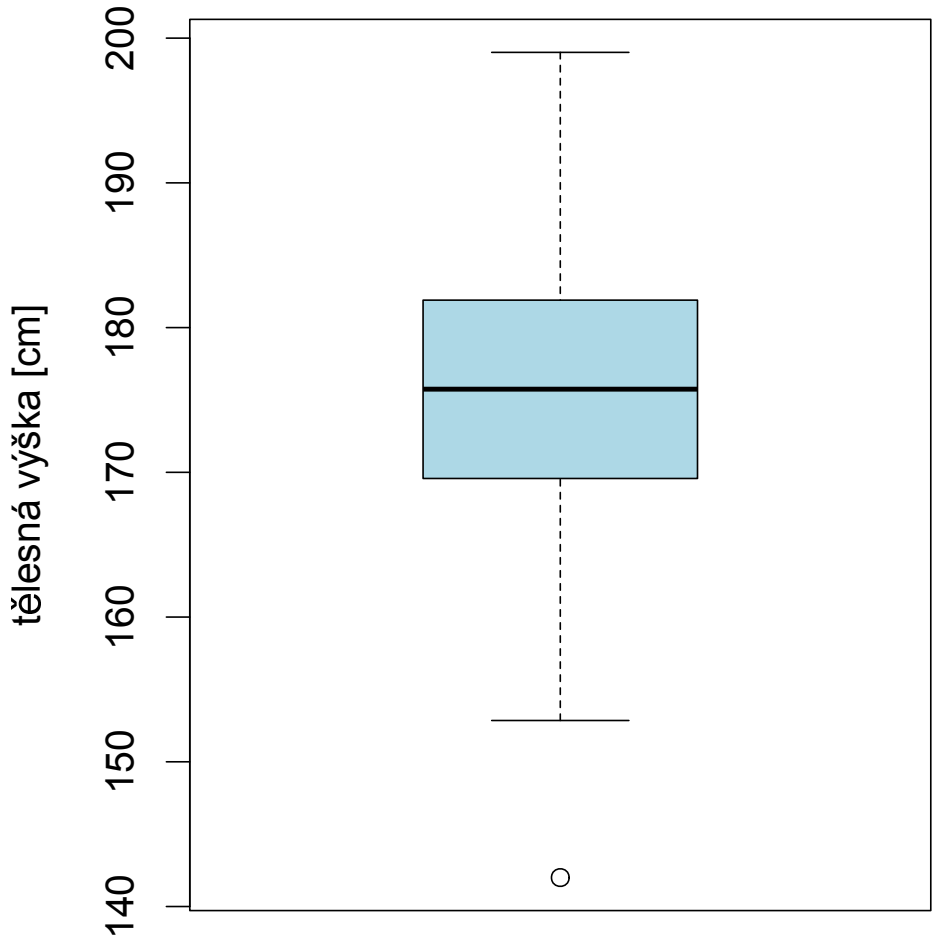
Intermezzo

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



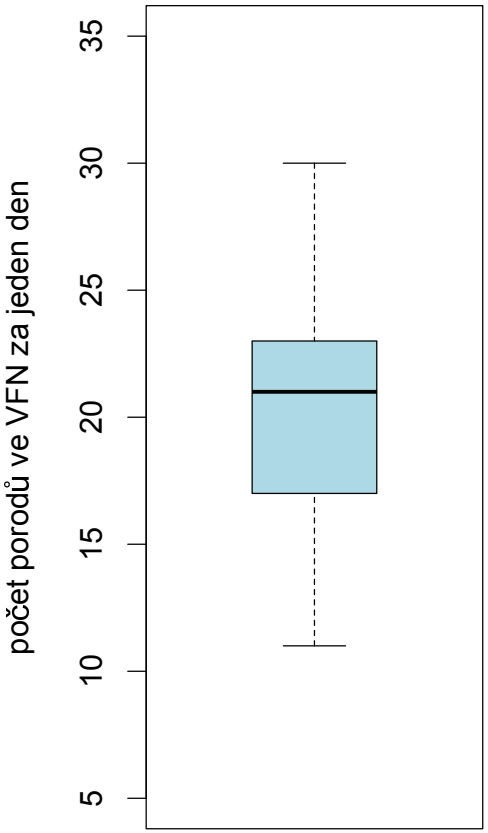
Krabicový diagram (boxplot)

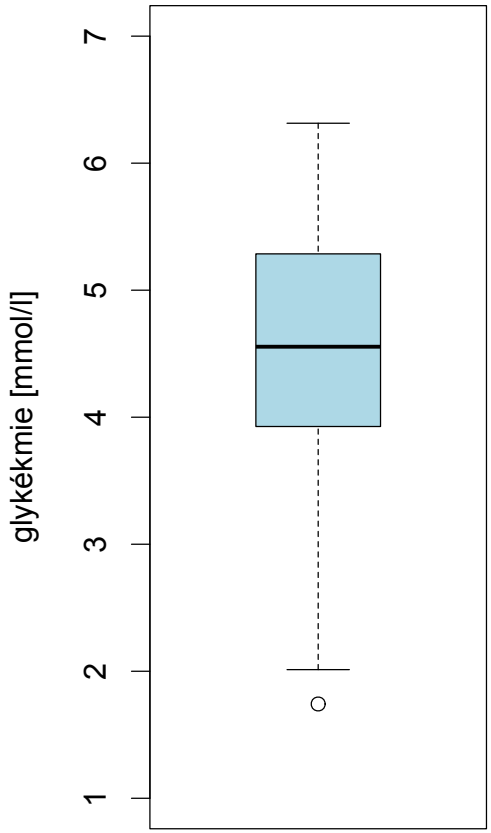
- vhodný pro kvantitativní znaky

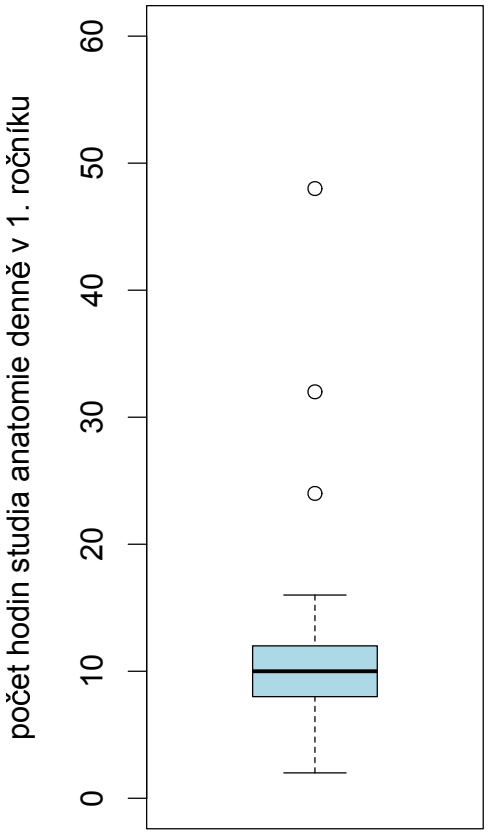


Intermezzo

- který z krabicových diagramů nedává smysl?

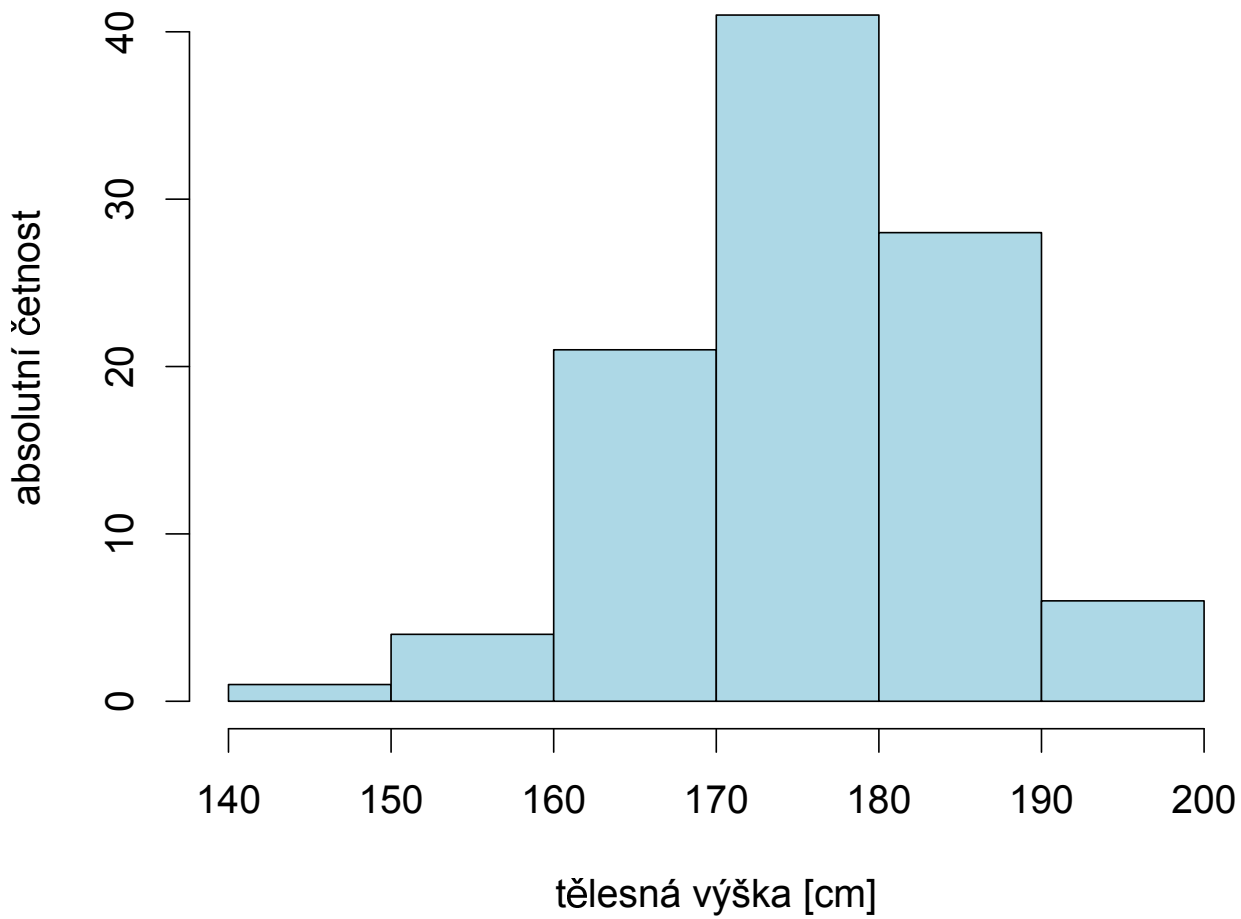






Histogram

- vhodný pro posouzení tvaru rozdělení hodnot

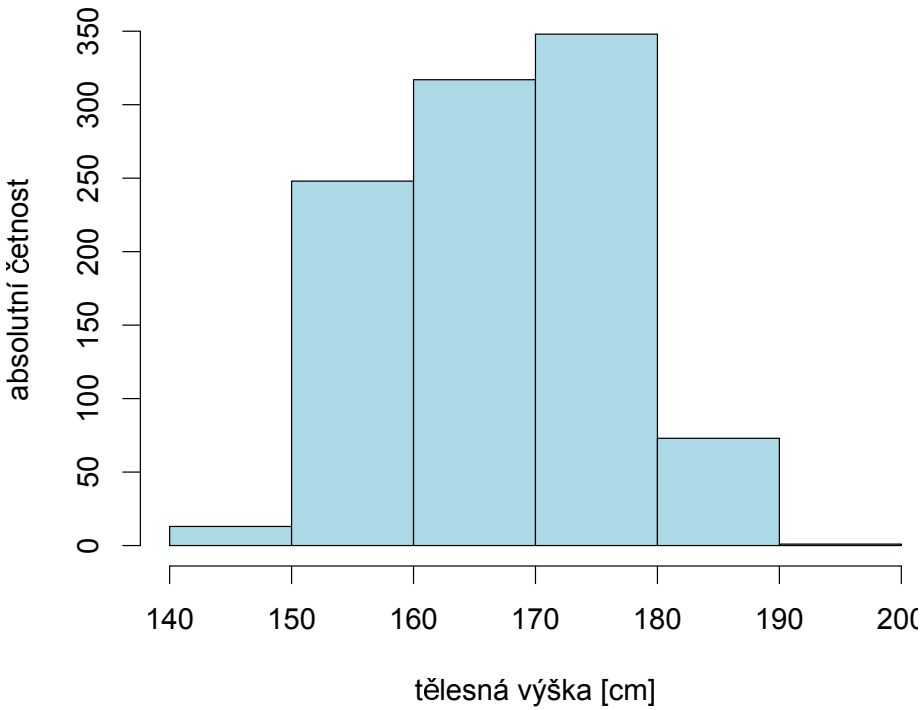


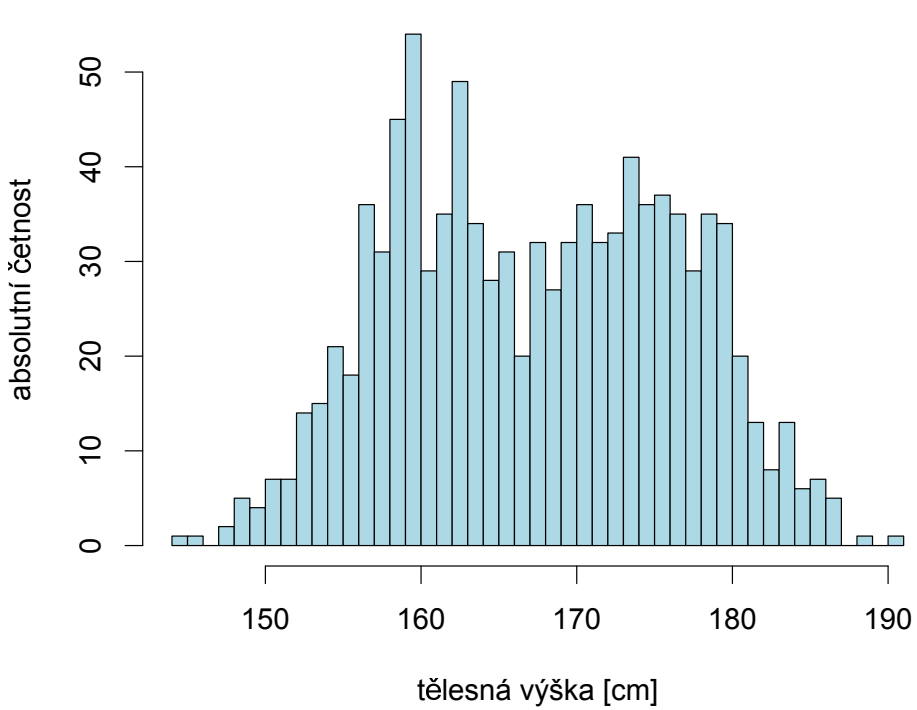
Počet intervalů v histogramu

- rozdílný počet intervalů histogramu mění „příběh“ dat!
- nejčastěji je počet intervalů k dán Sturgesovým pravidlem

$$k = \lceil \log_2 n \rceil,$$

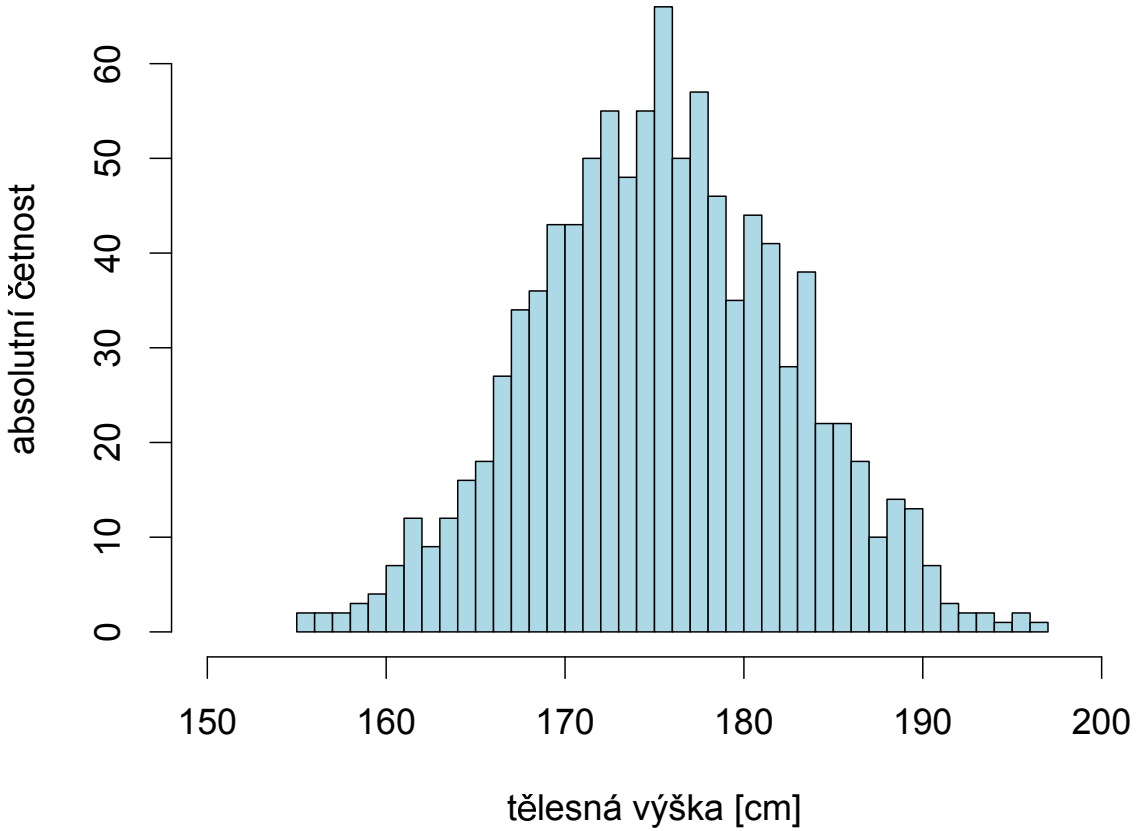
kde n je počet pozorování v souboru





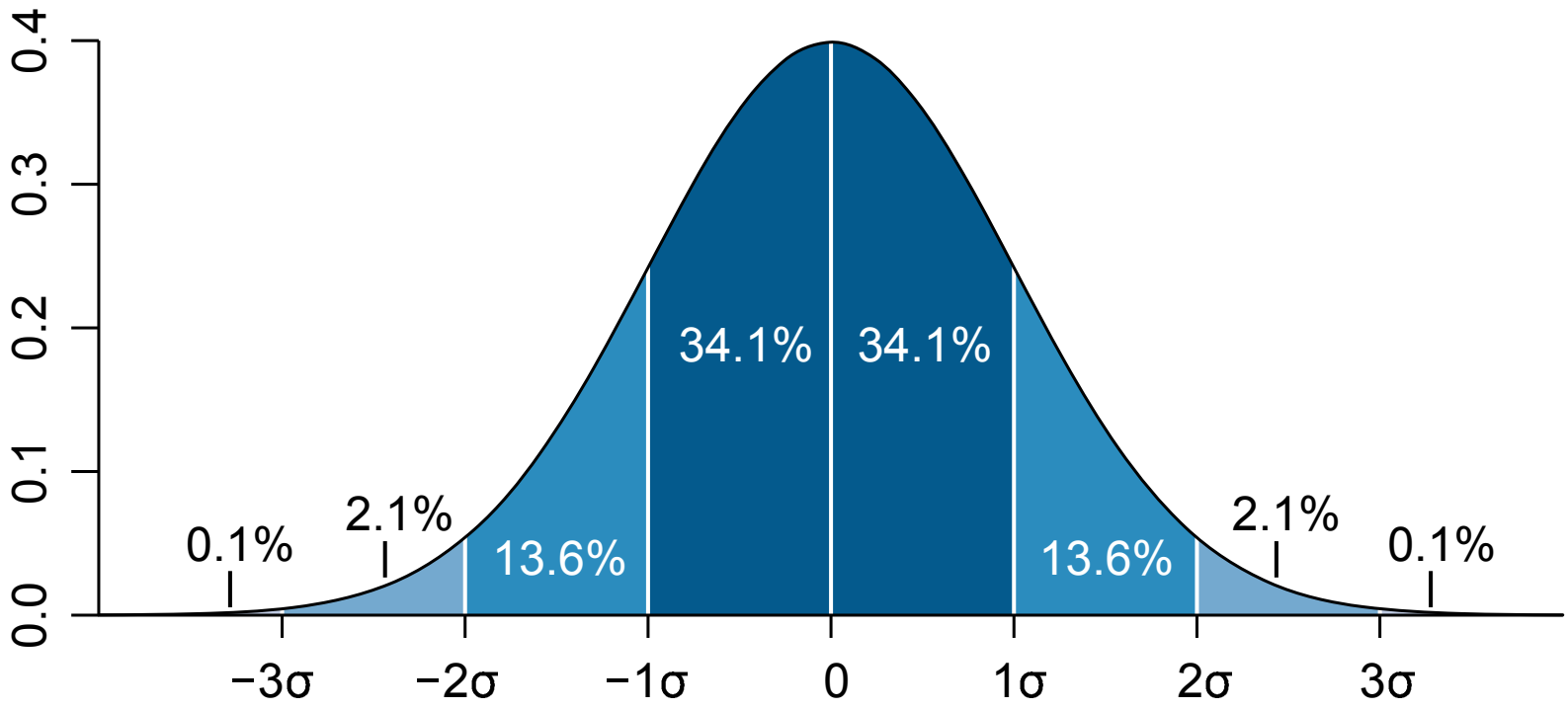
Normální rozdělení kvantitativního znaku

- lze odhadnout z histogramu



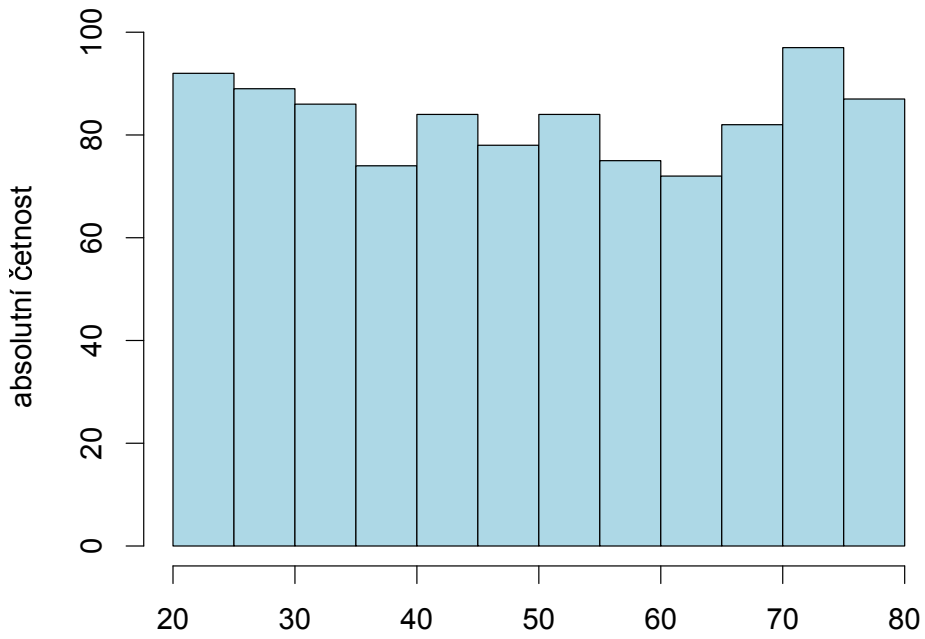
Vztah mezi polohou, variabilitou, tvarem a proporcí

- pokud je udržitelný předpoklad normálního rozložení, pak
  - v intervalu $\langle \bar{x} - s, \bar{x} + s \rangle$ leží asi 68 % hodnot
  - v intervalu $\langle \bar{x} - 2s, \bar{x} + 2s \rangle$ leží asi 95 % hodnot
  - v intervalu $\langle \bar{x} - 3s, \bar{x} + 3s \rangle$ leží asi 99,7 % hodnot



Vztah mezi polohou, variabilitou, tvarem a proporcí

- pokud **není** udržitelný předpoklad normálního rozložení, pak
  - v intervalu $\langle \bar{x} - 1s, \bar{x} + 1s \rangle$ nemusí ležet žádné hodnoty
  - v intervalu $\langle \bar{x} - 2s, \bar{x} + 2s \rangle$ leží alespoň 75 % hodnot
  - v intervalu $\langle \bar{x} - 3s, \bar{x} + 3s \rangle$ leží alespoň 88,9 % hodnot
- (vychází z Chebysevovy nerovnosti)



Popis kvalitativního znaku

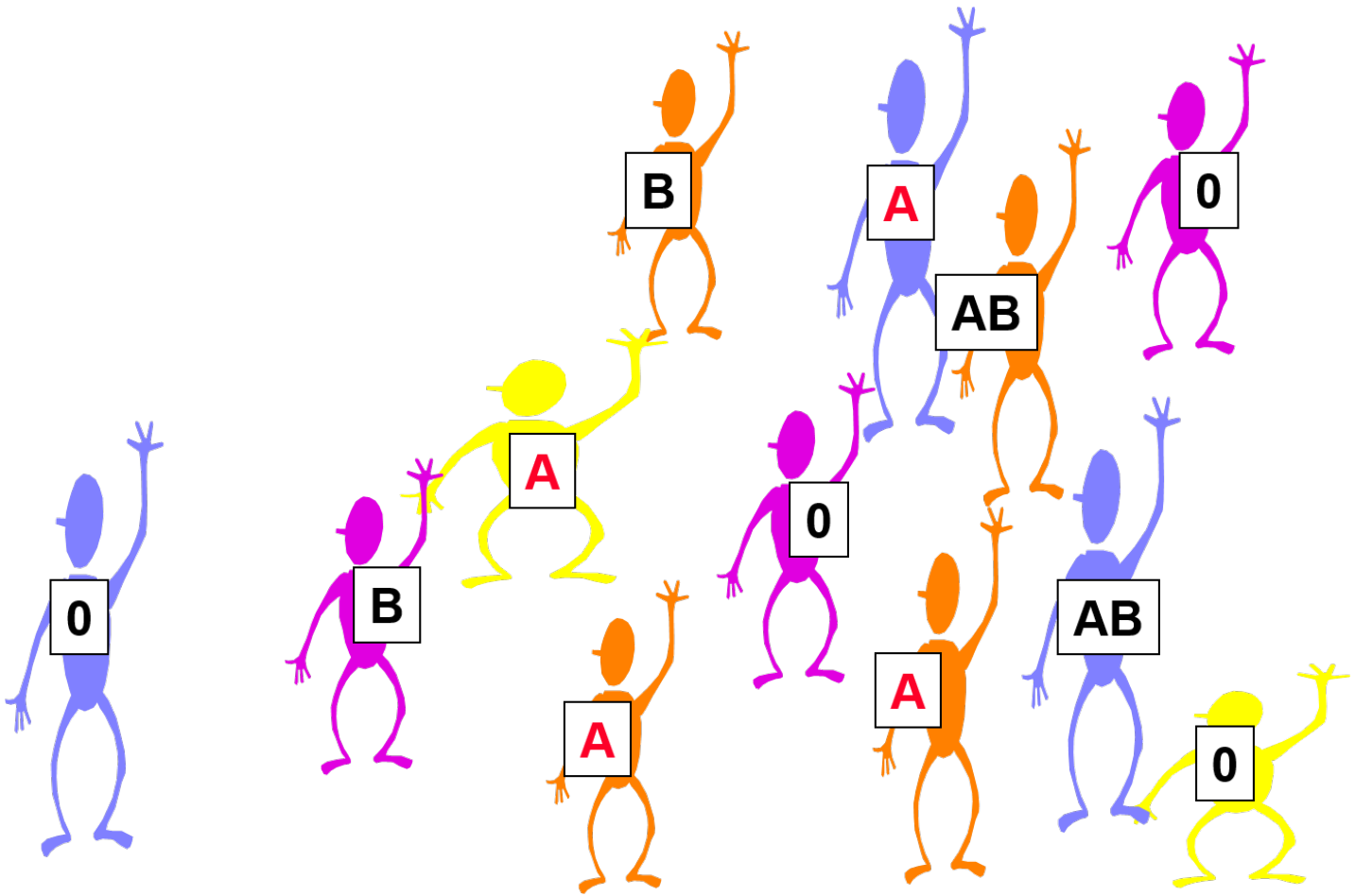
- např. krevní skupiny, grading tumoru, pohlaví, atd.
- číselně
  - absolutní, relativní četnosti
- graficky
  - koláčový diagram

Četnost

- *absolutní* četnost n_k kategorie k se rovná počtu jednotek souboru, jejichž statistický znak odpovídá kategorii k
- *relativní* četnost π_k kategorie k je podíl absolutní četnosti kategorie k a celkového rozsahu souboru

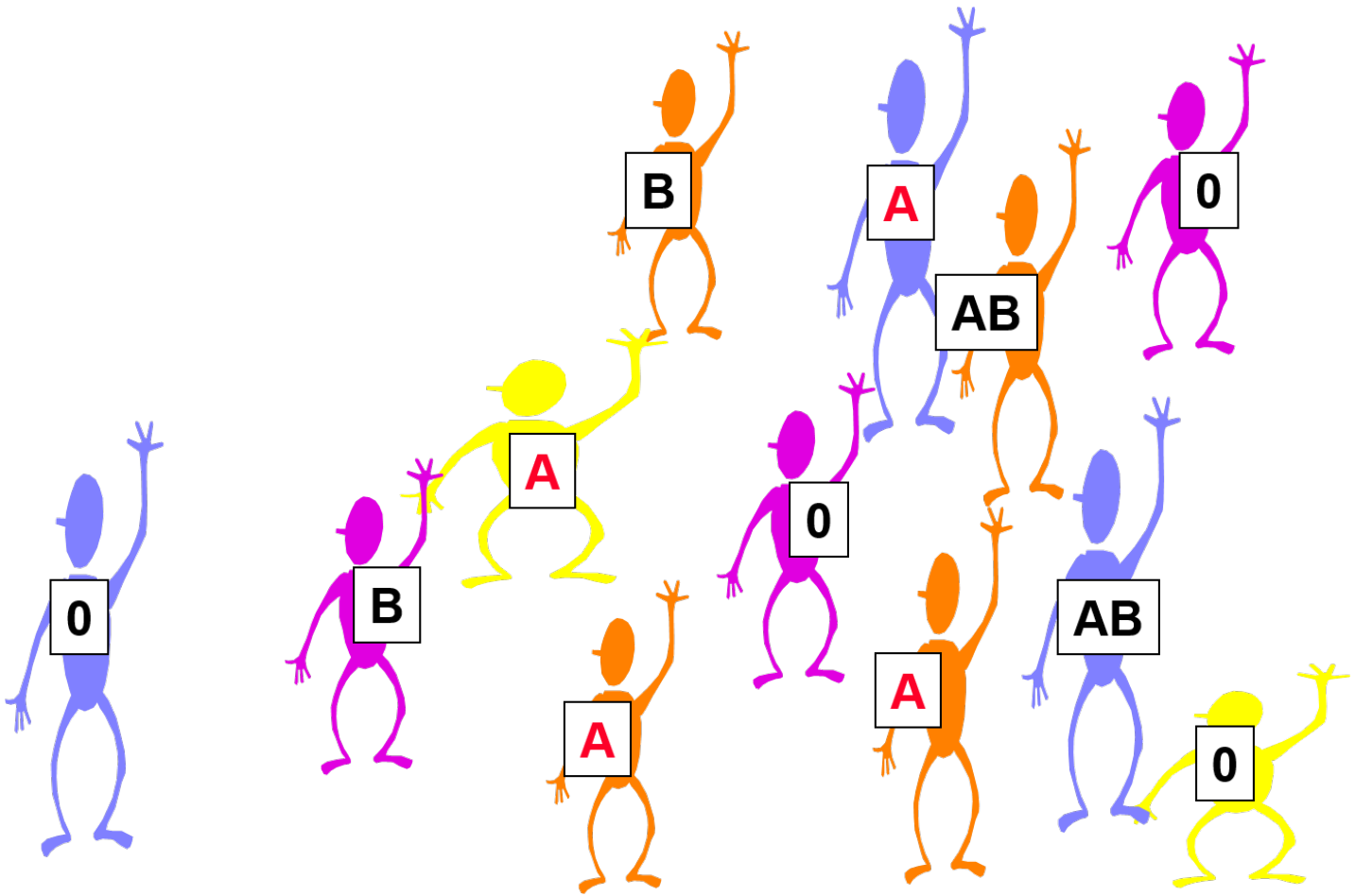
Intermezzo

- určeme absolutní a relativní četnost krevní skupiny A



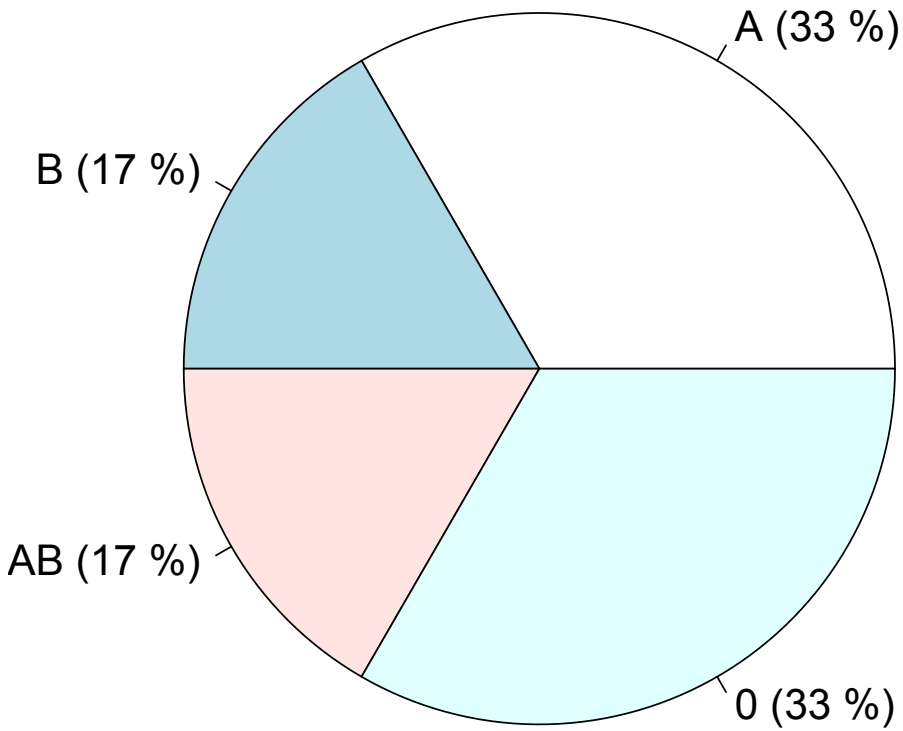
Intermezzo

- určeme absolutní a relativní četnost krevní skupiny A
- $n_A = 4; \quad \pi_A = \frac{4}{12} = \frac{1}{3}$



Koláčový diagram

- vhodný pro kvalitativní znaky k vyjádření četností jejich kategorií



Motivace

- ve výběru sto lidí je průměrná výška 175 cm a směrodatná odchylka je 10 cm
- jaká je s 95 % pravděpodobností průměrná výška populace?

Pojem *populace*

- **populace := základní soubor**
- úplná množina (statistický soubor) všech prvků (statistických jednotek), které spojuje určitá vlastnost a o kterých se snažíme statisticky něco zjistit
- prvky dány výčtem (je-li rozsah populace konečný), nebo společnou vlastností všech prvků (je-li rozsah populace nekonečný i konečný)
- rozsah konečně velké populace obvykle značíme N (u nekonečně velké populace $N \to \infty$)
- např. {T. G. Masaryk, E. Beneš, . . . , V. Klaus, M. Zeman}, {všichni dosavadní prezidenti českého státu}, {všichni obyvatelé Evropy}, apod.

Pojem *výběr*

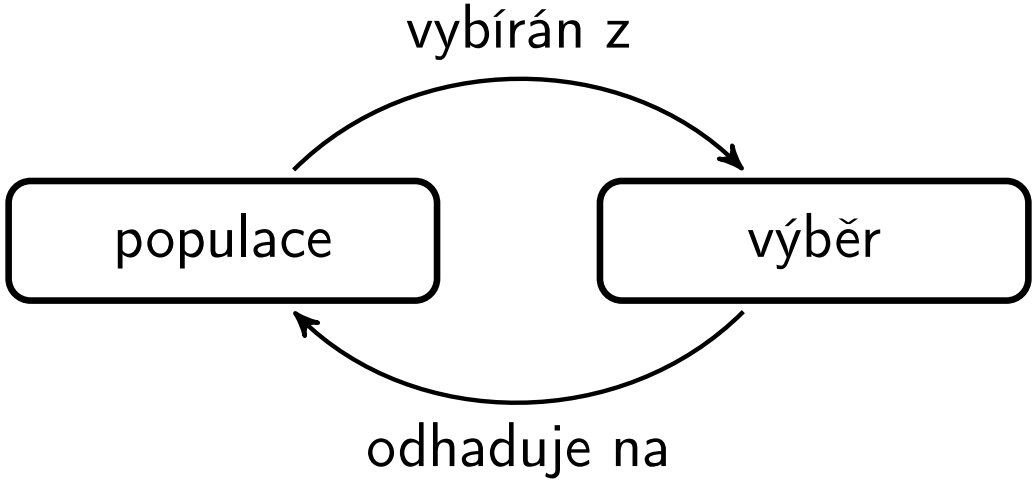
- vyšetřit celou populaci v praxi takřka nemožné
- nekonečně velké populace nelze celkově šetřit už z principu
- výběr := statistický soubor, obsahuje vybrané prvky z populace; je tedy podmnožinou populace
- výběr pořizujeme metodou náhodného, či záměrného výběru
- cílem získat reprezentativní výběr (vystihuje vlastnosti populace), nikoliv selektivní výběr

Reprezentativní výběr

- takový výběr, z kterého je induktivními metodami možné usuzovat na vlastnosti „mateřské“ populace
- pořizujeme *záměrným*, či *náhodným* výběrem
  - záměrný výběr – opírá se o expertízu, zatížen subjektivitou
  - náhodný výběr – náhodné, nezávislé vybírání prvků populace do výběru

Vztah populace a výběru

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



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