Organizace cvičení, úvod do statistiky a deskriptivní statistika

Supplementum ke cvičení 4ST210 Statistika pro finance

Lubomír Štěpánek[1, 2]

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

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

Obsah

Online složka předmětu

- prezentace a další materiály ke cvičení jsou dostupné na

https://github.com/LStepanek/4ST210_Statistika_pro_finance

Organizace předmětu

- cvičící
  - Ing. MUDr. Lubomír Štěpánek
- email

  lubomir.stepanek@vse.cz
- konzultační hodiny
  - v NB366 vždy v úterý mezi 11:00–12:00, po předchozí emailové domluvě i jindy

Podklady a materiály

- oficiální vzorcovník, oficiální statistické tabulky a výčet důležitých kapitol z doporučené literatury je zde

https://kstp.vse.cz/predmety-4st201-4st210-a-4st204/

- tabulky i vzorcovník je vhodné nosit s sebou vytisklý na cvičení
- prezentace ze cvičení a vzorové příklady budou v online složce předmětu, tedy

https://github.com/LStepanek/4ST210_Statistika_pro_finance

Doporučená literatura

Hindls, Richard, Markéta Arltová, Stanislava Hronová, Ivana Malá, Luboš Marek, Iva Pecáková a Řezanková Hana. *Statistika v ekonomii*. Praha: Professional Publishing, 2018. ISBN: 978-80-88260-09-7.

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

Cíle předmětu

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

Náplň předmětu

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

[1] Richard Hindls, Markéta Arltová, Stanislava Hronová, Ivana Malá, Luboš Marek, Iva Pecáková a Řezanková Hana. *Statistika v ekonomii*. Praha: Professional Publishing, 2018. ISBN: 978-80-88260-09-7

Hodnocení předmětu

- v průběhu semestru lze získat až $2 \times 20 = 40$ bodů za průběžné testy
- bodové rozložení je následující

typ testování	maximální možný bodový zisk
první průběžný test	20
druhý průběžný test	20
závěrečný test	60
Σ	100

Průběžné testy

- za každý z průběžných testů je možné získat maximálně 20 bodů, celkem tedy maximálně 40 bodů
- každý z průběžných testů se obvykle skládá právě ze tří početních příkladů
- na každý průběžný test je oficiálně 45 minut (na mých cvičeních však 45–50 minut)
- u průběžných testů je povoleno používat
  - neprogramovatelný kalkulátor,
  - MS Excel®
  - oficiální vzorcovník (bez vlastních poznámek)
  - oficiální statistické tabulky
- pro připuštění k závěrečnému testu je nutné získat v součtu za oba průběžné testy **alespoň 16 bodů**

Průběžné testy

- první průběžný test budeme psát pravděpodobně 6. vyučovací týden, tj. 24. března 2020
- druhý průběžný test budeme psát pravděpodobně 11. vyučovací týden, tj. 28. dubna 2020

Použitá matematická notace

- součet $n \in \mathbb{N}$ čísel $x_1, x_2, \ldots, x_n$ značíme též symbolem (velká) sigma, $\sum_{i=1}^{n} x_i$, tedy

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

MS Excel®

=SUMA($x_1 : x_n$)

Použitá matematická notace

- součin $n \in \mathbb{N}$ čísel $x_1, x_2, \ldots, x_n$ značíme též symbolem (velké) pí, $\prod_{i=1}^{n} x_i$, tedy

$$\prod_{i=1}^{n} x_i = x_1 \cdot x_2 \cdot \cdots \cdot x_n$$

MS Excel®

=SOUČIN($x_1 : x_n$)

Použitá matematická notace

- součin všech přirozených čísel $1, 2, \ldots, n-1, n$ obvykle značíme $n!$ a čteme „en faktoriál", tedy

$$n! = n \cdot (n-1) \cdot \cdots \cdot 2 \cdot 1$$

MS Excel®

=FAKTORIÁL(n)

Použitá matematická notace

- počet všech možností, kolika lze z $n \in \mathbb{N} \cup \{0\}$ prvků vybrat $k \in \mathbb{N} \cup \{0\}$ prvků tak, že $k \leq n$ a nezáleží na pořadí výběru, obvykle značíme $\binom{n}{k}$, a platí

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

▸ MS Excel®

=KOMBINACE($n; k$)

Příklad

(i) Najděme $x \in \mathbb{N}$ takové, aby

$$\binom{x}{2} = 10.$$

(ii) Najděme $x \in \mathbb{N}$ takové, aby

$$\binom{x}{3} = 455.$$

(iii) Najděme $x \in \mathbb{N}$ takové, aby

$$x! = 362880.$$

Příklad

- Jirka má právě sedm různých triček, čtvery kalhoty a pět párů bot.
  (i) Kolika navzájem různými způsoby může vytvořit svůj outfit?
  (ii) Kolik by potřeboval triček, aby měl každý den v roce originální[2] outfit?

[2] Ve smyslu jiný outfit než během kteréhokoliv ostatního dne v roce.

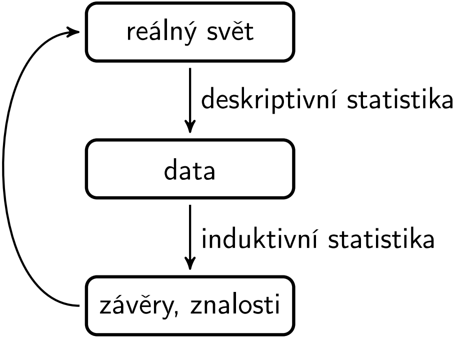
Příklad

- Uvažujme číslice $1, 2, 3, \ldots, 6, 7$.
  (i) Kolika navzájem různými způsoby z nich můžeme sestavit trojciferné číslo, pokud se číslice mohou opakovat?
  (ii) Kolika navzájem různými způsoby z nich můžeme sestavit trojciferné číslo, pokud se číslice nemohou opakovat?
  (iii) Pro kolik takových trojciferných čísel bude platit, že první jejich cifra je menší než druhá a zároveň druhá jejich cifra je menší než třetí?

Dělení statistiky

- deskriptivní statistika
  - popisuje data, ale nedělá na nich žádné „velké“ závěry
- induktivní statistika
  - pozorujte 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[3] 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

[3] 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
- např.
  - tělesná výška, hmotnost, stupně Celsia, skóre z testu, směnné kurzy, HDP daného státu atd.
  - počet kandidátů, počet pacientů, počet dětí v rodinách daného státu, věk probanda atd.

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
- např.
  - pohlaví {muž, žena}, rodinný stav muže {svobodný, ženatý, rozvedený, vdovec, registrovaný} atd.
  - pořadí v závodu, grade tumoru {1, 2, 3, 4} atd.

Příklad

- určeme typ znaku u následujících příkladů
  - procentuální úspěšnost v testu v souboru studentů jednoho kruhu [%]
  - soubor všech červencových dní jednoho roku (1., 2., . . . , 31.)
  - směnný kurz USD–CZK k danému datu
  - soubor čísel všech tramvají projíždějících zastávkou Husinecká
  - počet porodů v jedné porodnici za jednu noc
  - staging kolorektálního karcinomu {1, 2, 3, 4}

Popis kvantitativního znaku

- např. tělesná výška, glykémie, výše mzdy, sázkový kurz atd.
- číselně (**center – spread – shape!**)
  - mírou polohy (*center*)
    - aritmetický průměr, geometrický průměr, harmonický průměr, medián, modus, kvantil
  - mírou variability (*spread*)
    - rozpětí (min-max), směrodatná odchylka, rozptyl, kvantil
  - tvar (*shape*)[4]
    - šikmost, špičatost
- graficky
  - krabicový diagram (boxplot)
  - histogram

[4] lépe jen graficky

Aritmetický průměr

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

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

MS Excel®

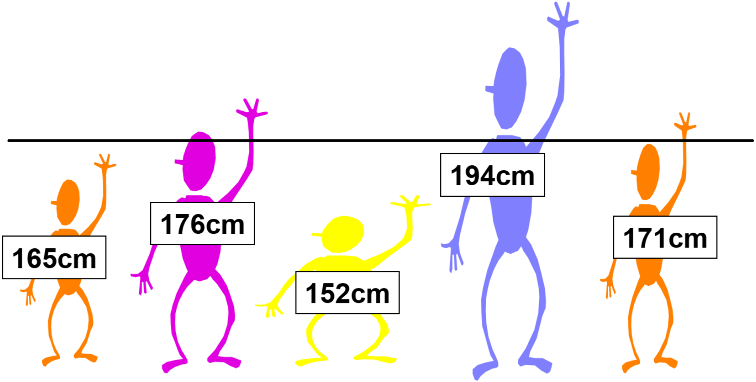
=PRŮMĚR($x_1 : x_n$)

- pro n hodnot $x_1, x_2, \dots, x_n$ tvořících soubor m čísel takových, že je mezi nimi číslo x_1 právě m_1-krát, číslo x_2 právě m_2-krát, ..., číslo x_n právě m_n-krát a je $m_1 + m_2 + \cdots + m_n = m$, spočítáme jejich aritmetický průměr jako

$$\bar{x} = \frac{m_1 x_1 + m_2 x_2 + \cdots + m_n x_n}{m} = \frac{1}{m} \sum_{i=1}^{n} m_i x_i$$

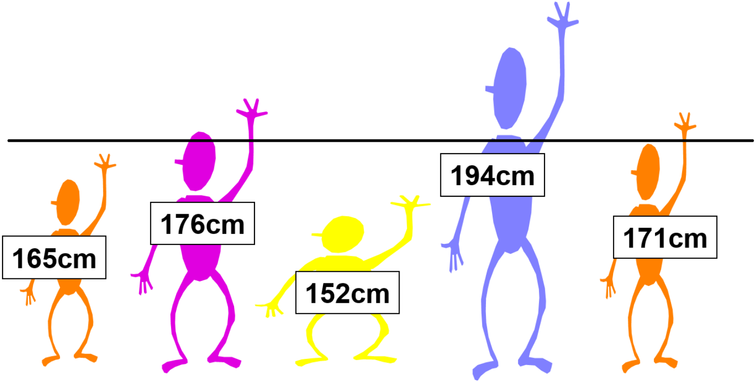
Příklad

- určeme aritmetický průměr z následujícího souboru tělesných výšek



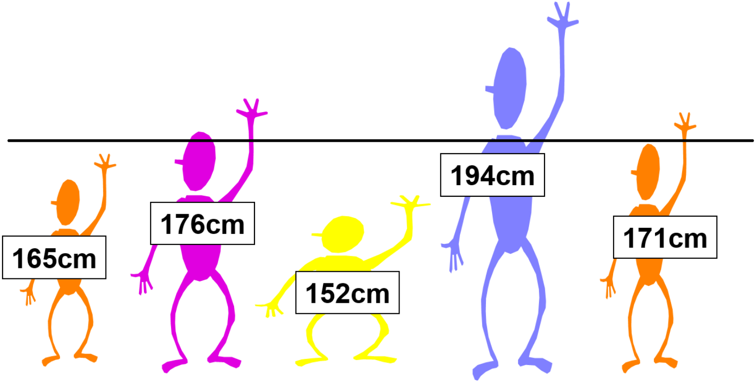
Příklad

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



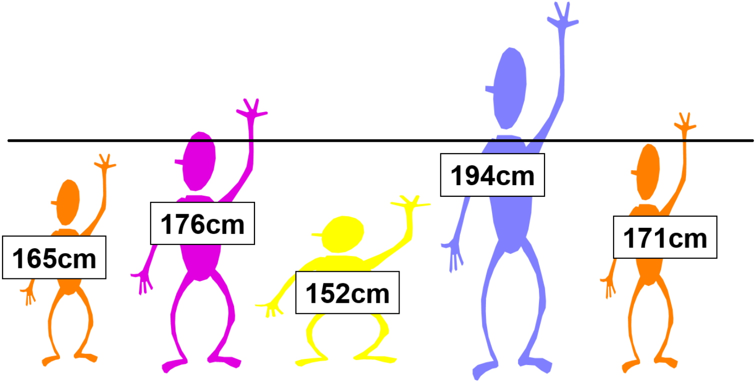
Příklad

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



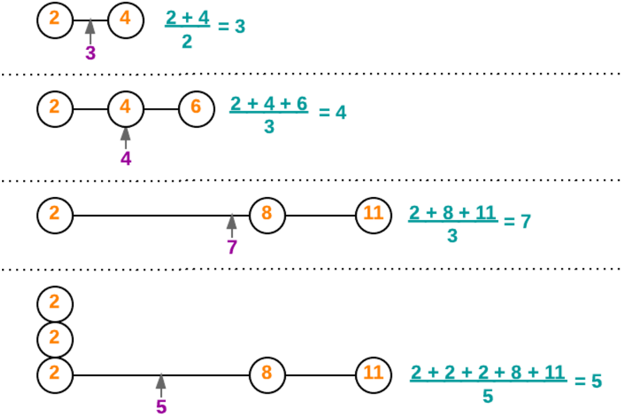
Příklad

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



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

Geometrická interpretace aritmetického průměru

- pokud zavěsíme n jednogramových závaží na pozice čísel $x_1, x_2, \dots, x_n$ pravítka, hodnota průměru $\bar{x}$ je v těžišti soustavy
- *Důkaz.* Označme polohu těžiště na pravítku x_t, dále hmotnost závaží je $m = 1$ [g] a gravitační konstanta je $g \doteq 10$ [m/s^2]. Zřejmě, pokud jde o těžiště, pak je soustava je v rovnovážném klidu a musí platit zákon zachování hybnost. Hybnost je u soustavy v rovnovážném klidu nulová, tedy

$$
\begin{aligned}
(x_1 - x_t)mg + (x_2 - x_t)mg + \cdots + (x_n - x_t)mg &= 0 \\
(x_1 - x_t) + (x_2 - x_t) + \cdots + (x_n - x_t) &= 0 \\
x_1 + x_2 + \cdots + x_n - nx_t &= 0 \\
x_1 + x_2 + \cdots + x_n &= nx_t \\
\frac{x_1 + x_2 + \cdots + x_n}{n} &= x_t.
\end{aligned}
$$

□

Příklad

- V souboru šestnácti čísel je jejich aritmetických průměr roven $10{,}3$. Jak se změní jejich aritmetický průměr, pokud
  - zvýšíme každé z čísel o $5{,}2$?
  - zvýšíme každé z čísel třikrát?
  - zvýšíme polovinu čísel o $7{,}2$ a zbytek čísel zmenšíme o $7{,}2$?
  - zvýšíme polovinu čísel o $3{,}1$ a druhou polovinu čísel zmenšíme o $1{,}1$?

Příklad

- Ve třídě 8. A jedné základní školy bylo původně 32 žáků. Určitý počet žáků, jejichž průměrný inteligenční kvocient je 120, byl z kázeňských důvodů přesunut ze třídy 8. A do jiné třídy dané školy. Díky tomu vzrostl průměrný inteligenční kvocient zbylých žáků v 8. A ze 127 na 128. Určete, kolik žáků bylo přesunuto ze třídy 8. A do jiné zmíněné třídy.

Geometrický průměr

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

$$\bar{x}_G = \sqrt[n]{x_1 \cdot x_2 \cdot \cdots \cdot x_n} = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

MS Excel®

=GEOMEAN($x_1 : x_n$)

Příklad

- Během čtyř let cena akcie meziročně postupně vzrostla o 8 %, pak o 12 %, poté o 3 % a nakonec klesla o 1 %. Pokud by se během těchto čtyřech let měnila cena akcie meziročně vždy o konstantní procento, o kolik by to bylo?

Příklad

- Ukažme, že pokud se geometrický průměr původního množství n kladných čísel rovnal $\bar{x}_G$, pak aritmetický průměr logaritmů původních čísel je roven $\log \bar{x}_G$.

Harmonický průměr

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

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

MS Excel®

=HARMEAN($x_1 : x_n$)

- pro n nenulových hodnot $x_1, x_2, \ldots, x_n$ tvořících soubor m čísel takových, že je mezi nimi číslo x_1 právě m_1-krát, číslo x_2 právě m_2-krát, …, číslo x_n právě m_n-krát a je $m_1 + m_2 + \cdots + m_n = m$, spočítáme jejich harmonický průměr jako

$$\bar{x}_H = \frac{m}{\frac{m_1}{x_1} + \frac{m_2}{x_2} + \cdots + \frac{m_n}{x_n}} = \frac{m}{\sum_{i=1}^{n} \frac{m_i}{x_i}}$$

Příklad

- Loď pluje z říčního přístavu k ústí řeky rychlostí $v_1 > 0$, poté se vrací stejnou trasou zpět do přístavu rychlostí $v_2 > 0$. Určete průměrnou rychlost lodi na celkové trase z říčního přístavu k ústí řeky a zpět.

Příklad

- Totožná součástka se vyrábí na dvou automatech. Starší z nich vyrobí 1 kus každých 10 minut, nový každých 6 minut.
  (i) Jak dlouho trvá v průměru výroba jedné součástky?
  (ii) Jak dlouho trvá v průměru výroba jedné součástky, pracuje-li starší automat 5 hodin denně a nový 8 hodin denně?
  (iii) Jaká je týdenní produkce součástek, pracují-li oba stroje na maximum, tj. 8 hodin denně, 7 dní v týdnu?

Příklad

- Investor se rozhodl pravidelně investovat do určité akcie každý měsíc fixní částku 10 000 Kč po dobu 3 měsíců. Cena jedné akcie byla první měsíc 100 Kč, druhý měsíc 125 Kč a třetí měsíc 80 Kč. Jaká byla průměrná pořizovací cena jedné akcie?

Vztah mezi aritmetickým, geometrickým a harmonickým průměrem

- pro n nezáporných čísel $x_1, x_2, \ldots, x_n$ spočítejme jejich aritmetický průměr $\bar{x}$, geometrický průměr $\bar{x}_G$ a harmonický průměr $\bar{x}_H$
- pak platí

$$\bar{x}_H \leq \bar{x}_G \leq \bar{x}$$

  a rovnost nastává tehdy a jen tehdy, pokud je $x_1 = x_2 = \cdots = x_n$
- vztahu $\bar{x}_G \leq \bar{x}$ se někdy též říká AG nerovnost („aritmeticko-geometrická“)

Příklad

- Dokažme AG nerovnost pro n nezáporných čísel $x_1, x_2, \ldots, x_n$, pokud je $n = 2$.

Medián

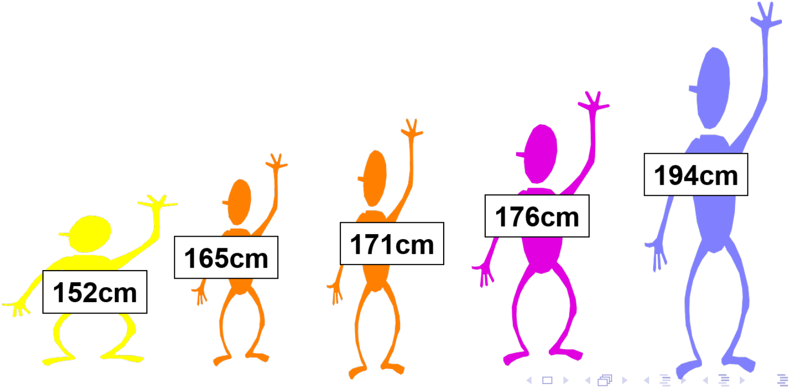
- medián je „prostřední" prvek, zhruba polovina hodnot je větší než medián a zbylá polovina hodnot je menší než medián
- pro n čísel $x_1, x_2, \ldots, x_n$ zjistíme jejich medián tak, že
  - (i) čísla seřadíme vzestupně, tj. pro původní vektor hodnot $(x_1, x_2, \ldots, x_n)^T$ získáme vektor hodnot $(x_{(1)}, x_{(2)}, \ldots, x_{(n)})^T$, kde pro $\forall i \in \{1, 2, \ldots, n\}$ je $x_{(i)}$ právě i-tá nejmenší hodnota ve vektoru $(x_1, x_2, \ldots, x_n)^T$
  - (ii) medián $\tilde{x}$ je prostřední hodnota (pro n liché), resp. aritmetický průměr z „prostředních" dvou hodnot (pro n sudé)

MS Excel®

=MEDIAN($x_1 : x_n$)

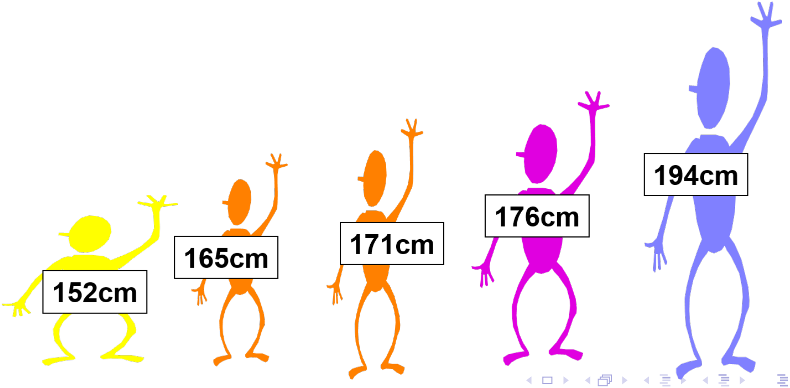
Příklad

- určeme medián z následujícího souboru tělesných výšek



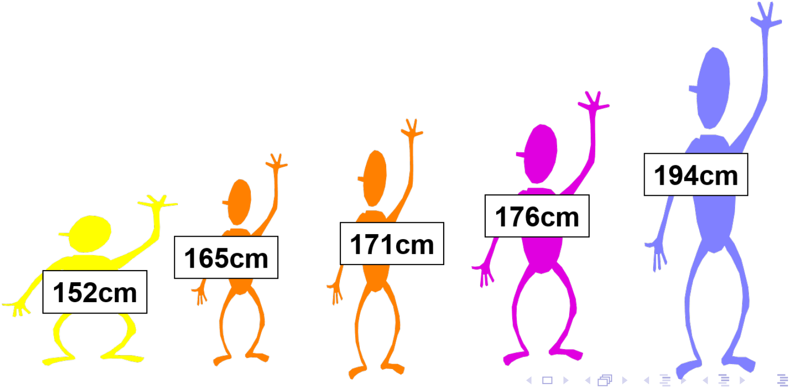
Příklad

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



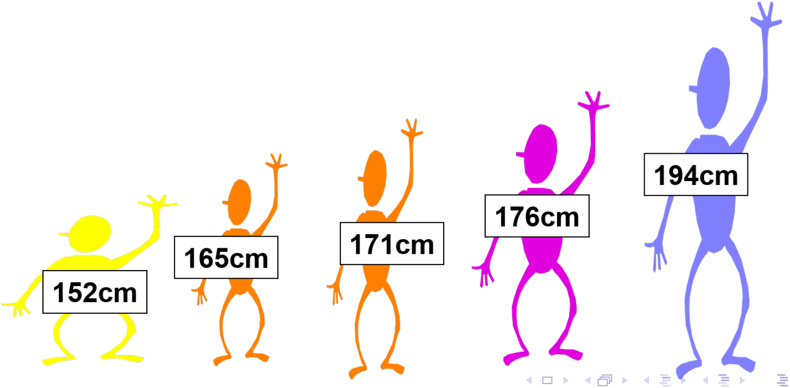
Příklad

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



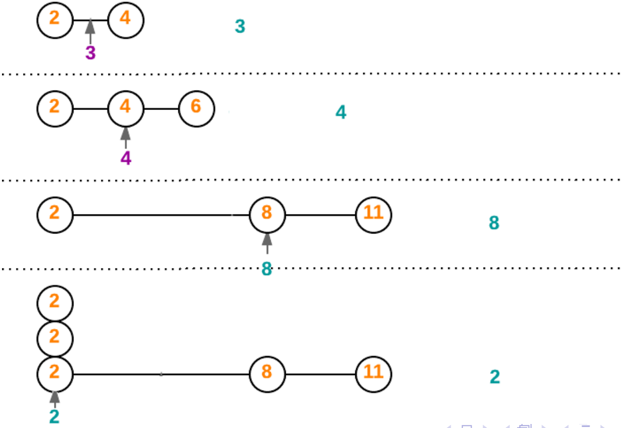
Příklad

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



p-tý kvantil

- nad souborem čísel $x_1, x_2, \dots, x_n$ se p-tý kvantil značí $\tilde{x}_p$ a je definován jako
$$\tilde{x}_p = \begin{cases} x_{(\lfloor k \rfloor + 1)}, & \text{pro } k = np \notin \mathbb{N} \\ \frac{1}{2}\left(x_{(k)} + x_{(k+1)}\right), & \text{pro } k = np \in \mathbb{N}, \end{cases}$$
kde $x_{(k)}$ je k-té nejmenší číslo mezi čísly $x_1, x_2, \dots, x_n$, dále kde $0 \leq p \leq 1$ a kde $\lfloor x \rfloor$ značí dolní celou část čísla x, tedy nejvyšší celé číslo takové, že nepřevýší x; tedy např. $\lfloor 4{,}3 \rfloor = 4$, $\lfloor 1 \rfloor = 1$ a $\lfloor -2{,}8 \rfloor = -3$
- pravděpodobnost, že hodnota v dané souboru nepřevýší hodnotu $\tilde{x}_p$, je rovna p, tedy
$$x \in (x_1, x_2, \dots, x_n)^T \implies P(x \leq \tilde{x}_p) = p$$
- zhruba $100p$ % hodnot v souboru $(x_1, x_2, \dots, x_n)^T$ je menších než nebo rovných hodnotě $\tilde{x}_p$

p-tý kvantil pro vybrané hodnoty $0 \leq p \leq 1$

p	p-tý kvantil	triviální název
$0{,}25$	$\tilde{x}_{0{,}25}$	první kvartil
$0{,}50$	$\tilde{x}_{0{,}50}$	medián
$0{,}75$	$\tilde{x}_{0{,}75}$	třetí kvartil

Příklad

- V souboru hodnot $\boldsymbol{x} = (1, 3, 2, 2, 4, 1, 4, 2, 2, 5, 1, 2)^T$ nalezněme
  - (i) hodnotu kvantilu $\tilde{x}_{0,25}$.
  - (ii) hodnotu kvantilu $\tilde{x}_{0,45}$.

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
- pro n hodnot $x_1, x_2, \ldots, x_n$ tvořících soubor m čísel takových, že je mezi nimi číslo x_1 právě m_1-krát, číslo x_2 právě m_2-krát, ..., číslo x_n právě m_n-krát a je $m_1 + m_2 + \cdots + m_n = m$, určíme jejich modus jako

$$\hat{x} = \underset{i \in \{1,2,\ldots,n\}}{\arg\max} \{m_i\}$$

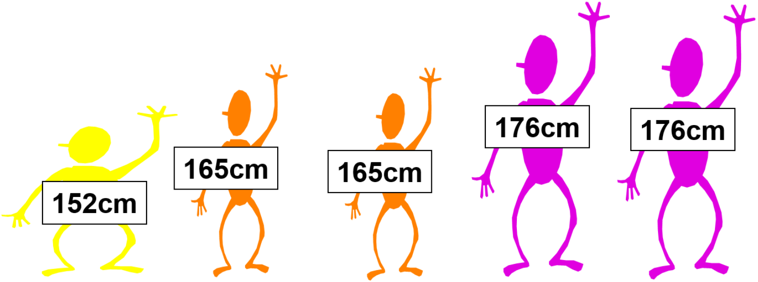
- všimněme si, že takových hodnot $\hat{x}$ ale může být více než jedna

MS Excel®

=MODE($x_1 : x_n$)

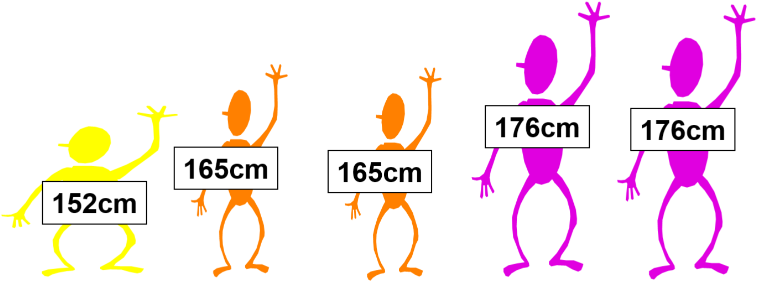
Příklad

- určeme modus z následujícího souboru tělesných výšek



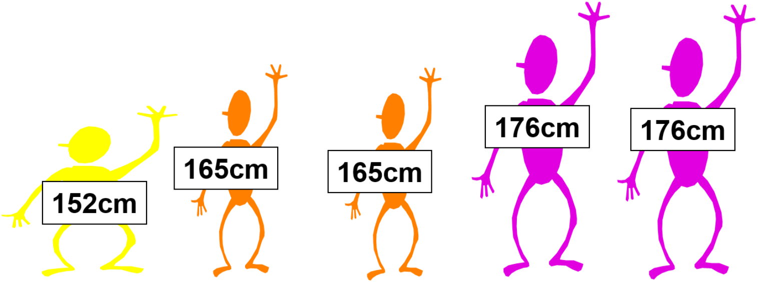
Příklad

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



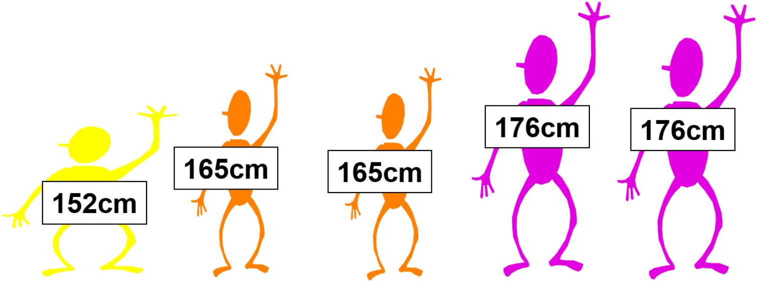
Příklad

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



Příklad

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



Příklad

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

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

Příklad

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

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

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

Příklad

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

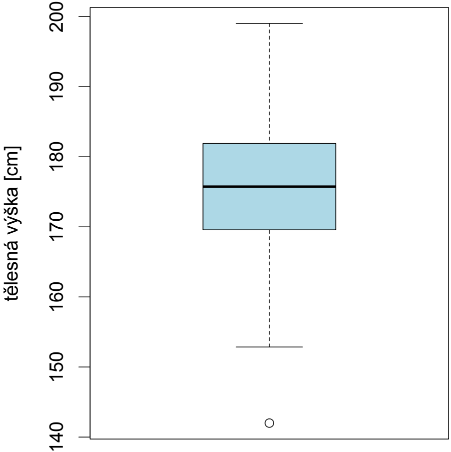
$$\boldsymbol{x}_1 = \{1, 2, 3, 4, 5\} \qquad \boldsymbol{x}_2 = \{1, 2, 3, 4, 90\}$$

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

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

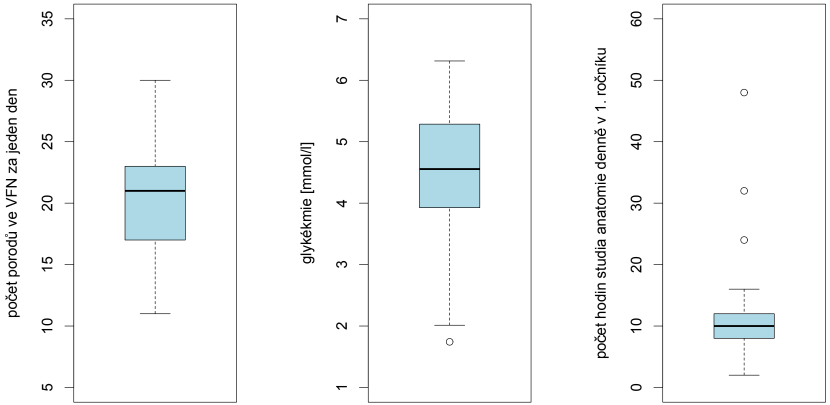
Krabicový diagram (boxplot)

- vhodný pro kvantitativní znaky



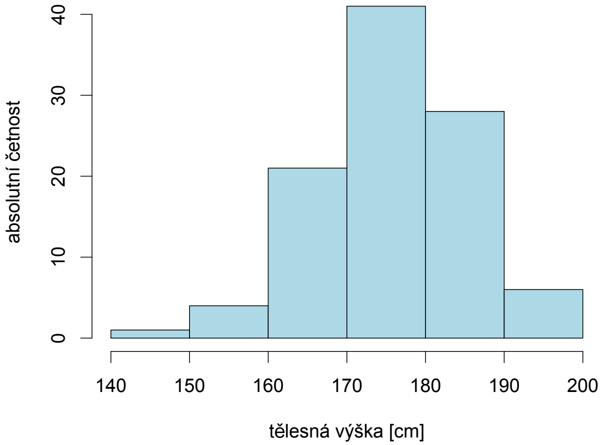
Příklad

- který z krabicových diagramů nedává smysl?



Histogram

- vhodný pro posouzení tvaru rozdělení hodnot

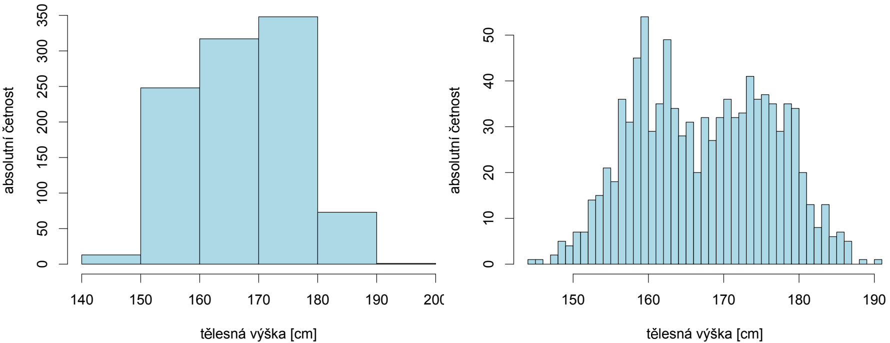


Počet intervalů v histogramu

- 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



Popis kvalitativního znaku

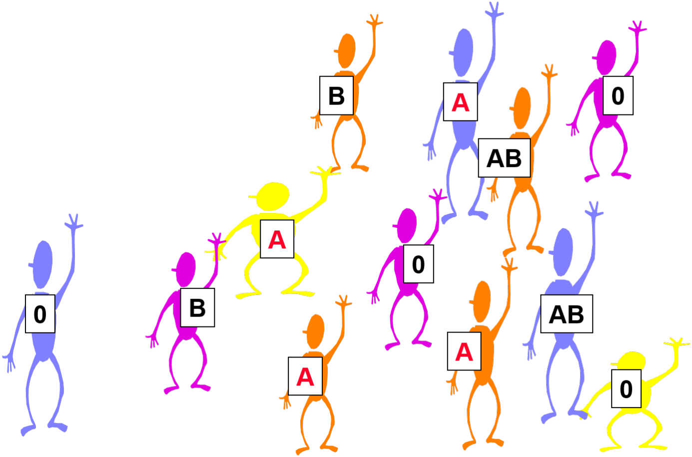
- např. krevní skupiny, grading tumoru, pohlaví, atd.
- číselně
  - absolutní, relativní četnosti
- graficky
  - koláčový diagram

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

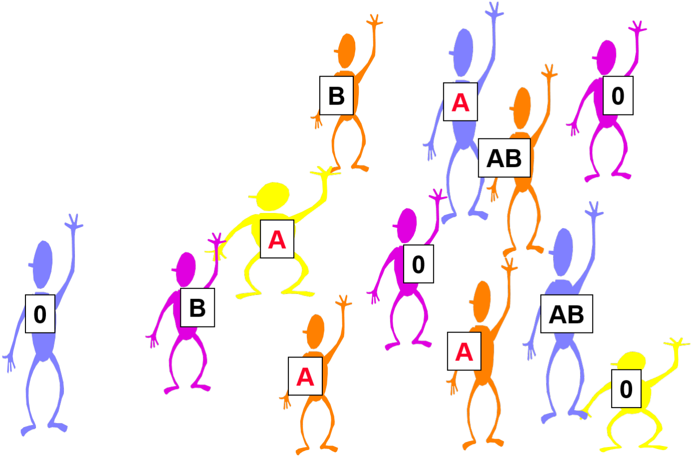
Příklad

- určeme absolutní a relativní četnost krevní skupiny A



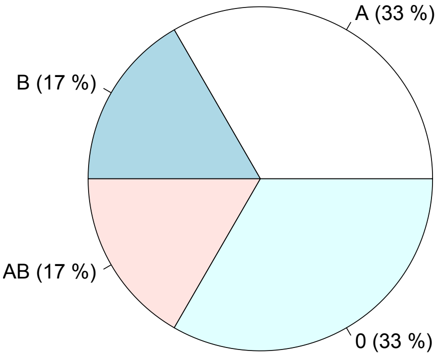
Příklad

- 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



Kumulativní četnost

- lze-li hodnoty kvalitativního znaku uspořádat vzestupně (či sestupně), jde-li tedy kvalitativní znak *ordinální*, má smysl definovat i kumulativní četnost
  - kumulativním ordinálním znakem může být např. platová tarifní třída (5, 6, . . . , 14, 15), stádium onkologické nemoci (I, II, III, IV), apod.
- *kumulativní absolutní* četnost $n_{k,\text{ kumul}}$ kategorie k se rovná počtu jednotek souboru, jejichž statistický znak odpovídá kategorii k a všem nižším kategoriím, tedy

$$n_{k,\text{ kumul}} = \sum_{i \in \{\text{kategorie } k \text{ a všechny nižší}\}} n_i$$

- *kumulativní relativní* četnost $\pi_{k,\text{ kumul}}$ kategorie k je podíl kumulativní absolutní četnosti kategorie k a celkového rozsahu souboru n, tedy

$$\pi_{k,\text{ kumul}} = \frac{n_{k,\text{ kumul}}}{n}$$

Vlastnosti kumulativní četnosti

- snadno nahlédneme, že kumulativní absolutní četnost $n_{\text{nejvyšší katerogie, kumul}}$ pro nejvyšší kategorii v daném uspořadatelném souboru je rovna jeho rozsahu n, neboť

$$\begin{aligned} n_{\text{nejvyšší katerogie, kumul}} &= \sum_{i \in \{\text{nejvyšší katerogie a všechny nižší}\}} n_i = \\ &= \sum_{i \in \{\text{všechny katerogie}\}} n_i = \\ &= n \end{aligned}$$

- obdobně kumulativní relativní četnost $\pi_{\text{nejvyšší katerogie, kumul}}$ pro nejvyšší kategorii v daném uspořadatelném souboru je rovna 1, neboť

$$\pi_{\text{nejvyšší katerogie, kumul}} = \frac{n_{\text{nejvyšší katerogie, kumul}}}{n} = \frac{n}{n} = 1$$

Příklad

- Na jisté katedře je věkový profil zaměstnanců takový, že šest z nich patří věkově do dekády $\langle 20, 30)$ let, čtyři z nich do dekády $\langle 30, 40)$ let sedm z nich do dekády $\langle 40, 50)$ let, dva z nich do dekády $\langle 50, 60)$ let a jeden z nich do dekády $\langle 60, 70)$ let. Určete absolutní a relativní četnosti všech dekád na dané katedře, současně určete i jejich kumulativní protějšky.

Literatura

Hindls, Richard, Markéta Arltová, Stanislava Hronová, Ivana Malá, Luboš Marek, Iva Pecáková a Řezanková Hana. *Statistika v ekonomii*. Praha: Professional Publishing, 2018. ISBN: 978-80-88260-09-7.

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