

Organizace cvičení, úvod do statistiky a deskriptivní statistika

—
Supplementum ke cvičení 4ST210 Statistika pro finance

Lubomír Štěpánek^{1, 2}



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



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

(2020) 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.

Online složka předmětu

- prezentace a další materiály ke cvičení jsou dostupné na

<https://github.com/LStepanek/4ST201> 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

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
 - indukativní statistika
 - testování hypotéz
 - korelační a regresní analýza
 - časové řady
 - indexní analýza

¹Richard 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, \dots, 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®

$$\text{SUMA}(x_1 : x_n)$$

Použitá matematická notace

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

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

► MS Excel®

SOUČIN($x_1 : x_n$)

Použitá matematická notace

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

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

► 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, čtyři 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 originální² outfit?

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

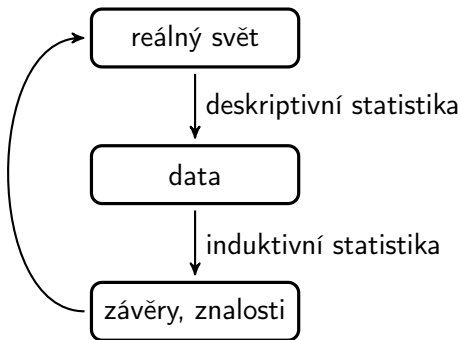
Příklad

- Uvažujme číslce 1, 2, 3, ..., 6, 7.
 - (i) Kolika navzájem různými způsoby z nich můžeme sestavit trojciferné číslo, pokud se číslce mohou opakovat?
 - (ii) Kolika navzájem různými způsoby z nich můžeme sestavit trojciferné číslo, pokud se číslce 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
 - 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 hodnotu³ 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

³ta může eventuálně chybět nebo být neznámá (missing value)

Intermezzo

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

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

- určíme 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}

Aritmetický průměr

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

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

► MS Excel®

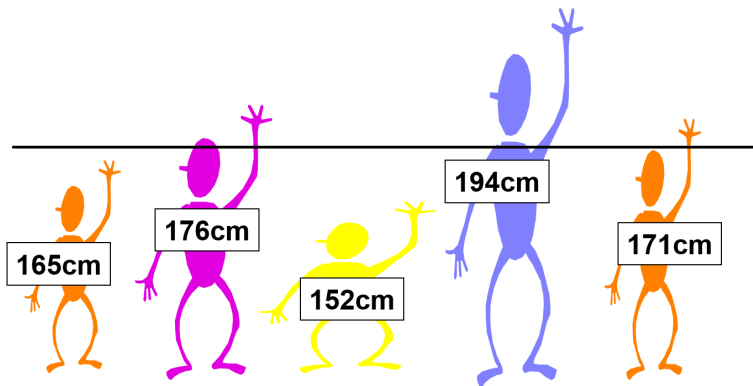
PRŮMĚR($x_1 : x_n$)

- 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, \dots , číslo x_n právě m_n -krát a je $n_1 + n_2 + \dots + m_n = m$, spočítáme jejich aritmetický průměr jako

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

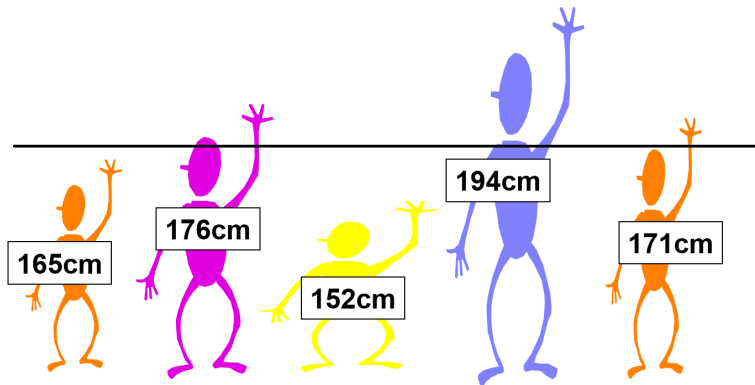
Příklad

- určíme aritmetický průměr z následujícího souboru tělesných výšek



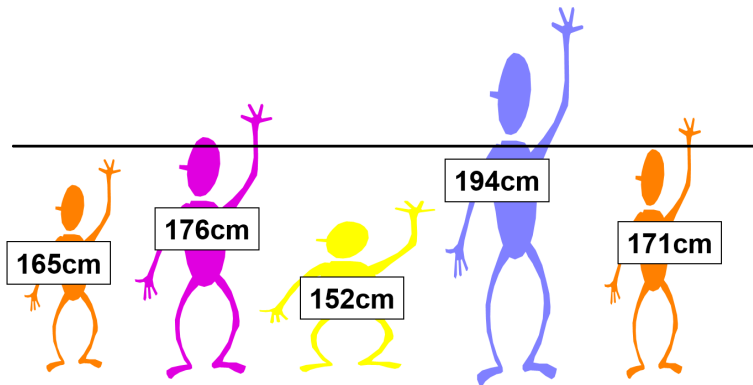
Příklad

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



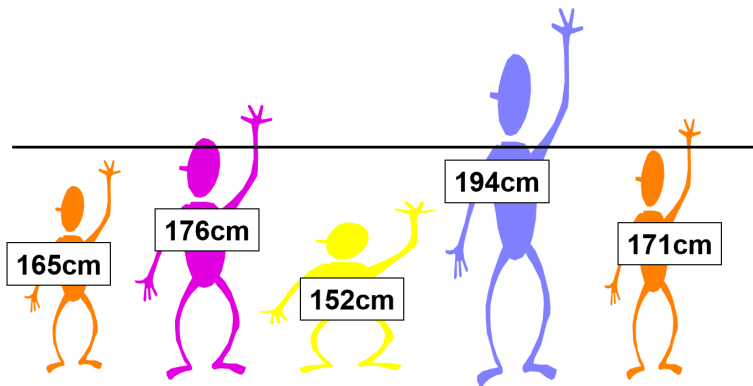
Příklad

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



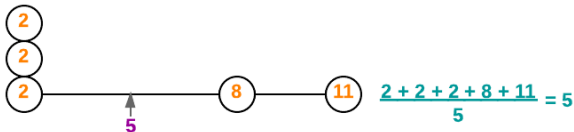
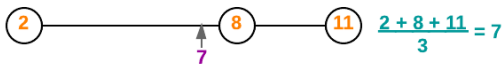
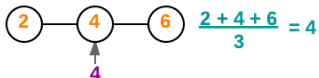
Příklad

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



Geometrická interpretace aritmetického průměru

- pokud zavěšíme n jednogramových závaží na pozice čísel x_1, x_2, \dots, 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

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

- 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, \dots, x_n spočítáme jejich harmonický průměr jako

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

► MS Excel®

=HARMEAN($x_1 : x_n$)

- pro n nenulových hodnot x_1, x_2, \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, \dots , číslo x_n právě m_n -krát a je $n_1 + n_2 + \dots + m_n = m$, spočítáme jejich harmonický průměr jako

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

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?

Vztah mezi aritmetickým, geometrickým a harmonickým průměrem

- pro n nezáporných čísel x_1, x_2, \dots, 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 $x_1 = x_2 = \dots = x_n$

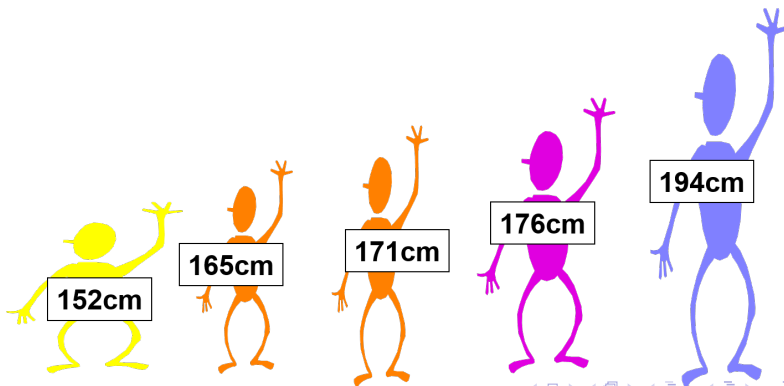
- vztahu $\bar{x}_G \leq \bar{x}$ se někdy též říká AG nerovnost („aritmeticko-geometrická“)

Příklad

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

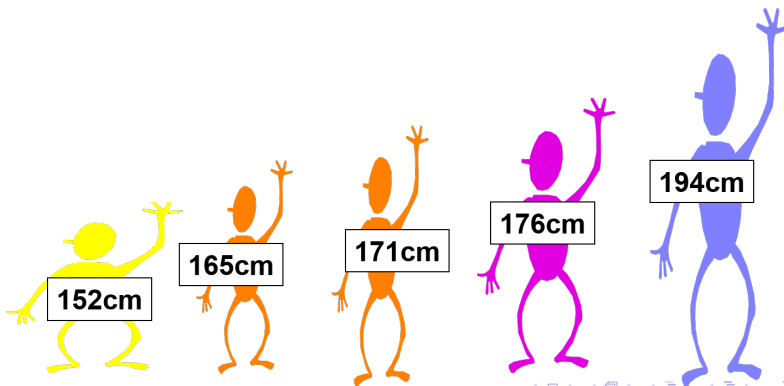
Příklad

- určíme medián z následujícího souboru tělesných výšek



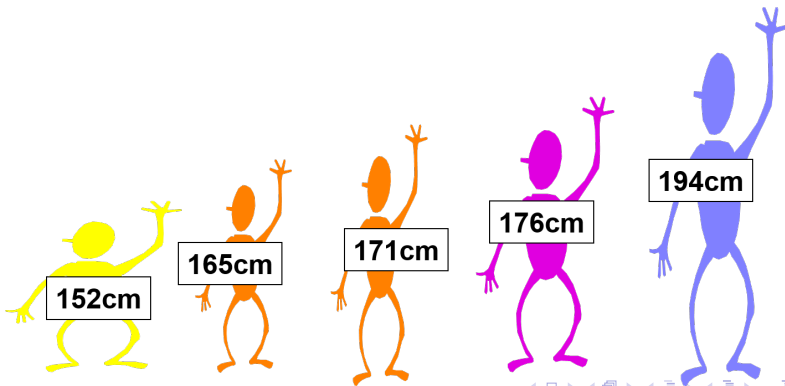
Příklad

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



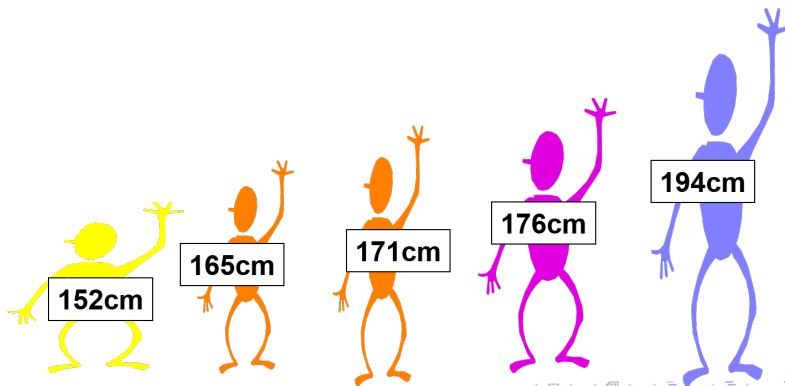
Příklad

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



Příklad

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



Příklad

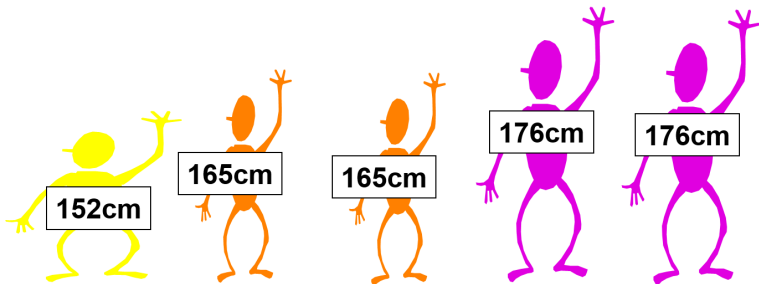
- V souboru hodnot $x = (1, 3, 2, 2, 4, 1, 4, 2, 2, 5, 1, 2)^T$ nalezněme
 - (i) hodnotu kvantilu $\tilde{x}_{0,25}$.
 - (ii) hodnotu kvantilu $\tilde{x}_{0,45}$.

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

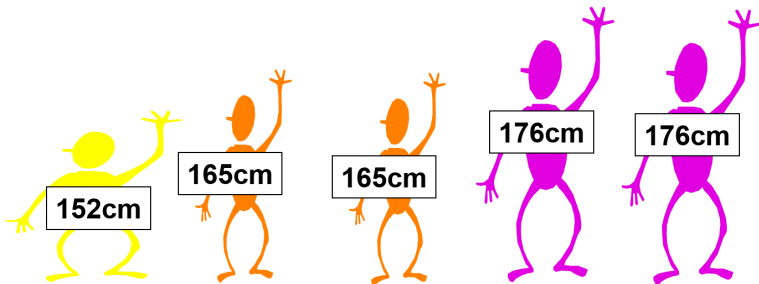
Příklad

- určeme modus z následujícího souboru tělesných výšek



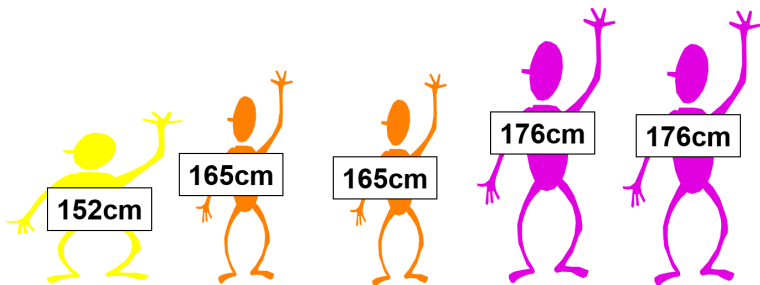
Příklad

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



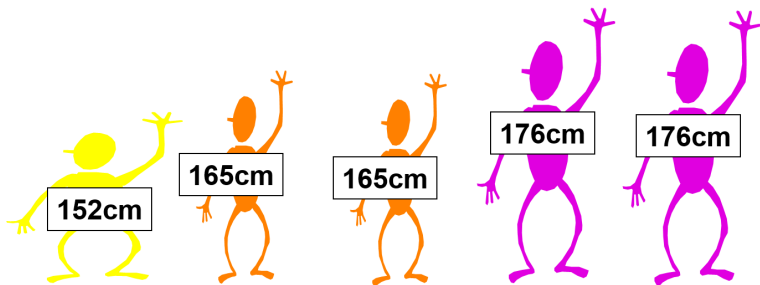
Příklad

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



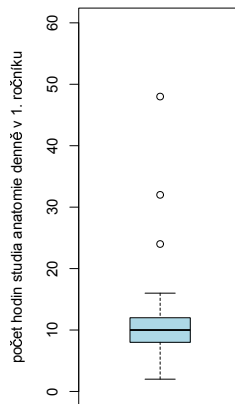
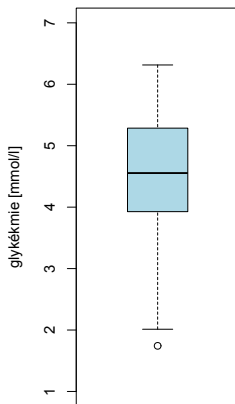
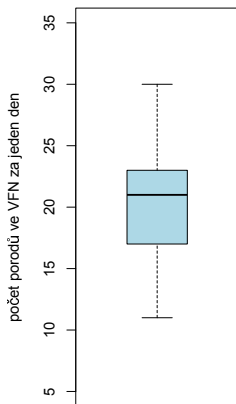
Příklad

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



Příklad

- který z krabicových diagramů nedává smysl?

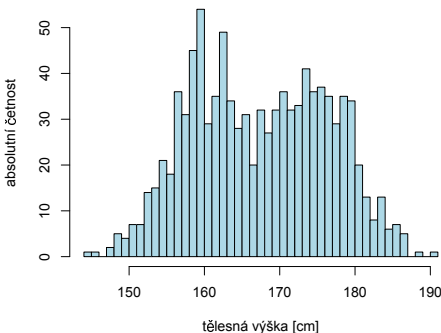
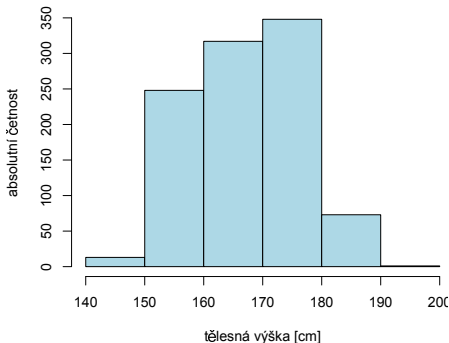


Počet intervalů v histogramu

- 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

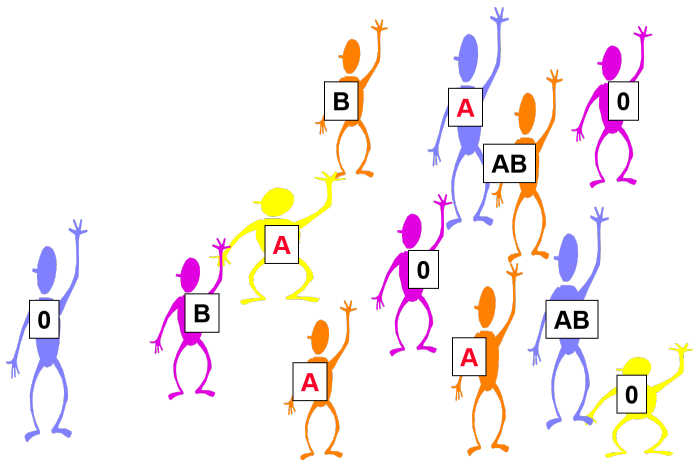


Četnost

- *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

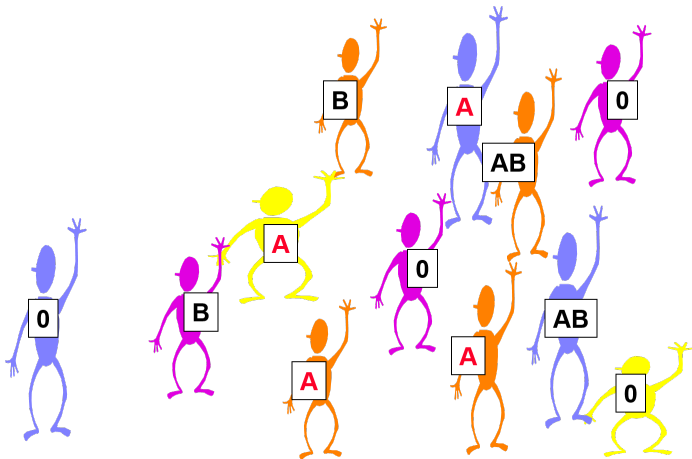
Příklad

- určeme absolutní a relativní četnost krevní skupiny A



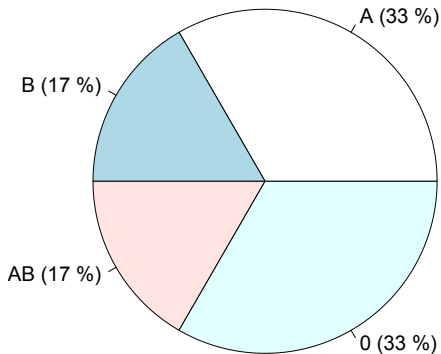
Příklad

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



Koláčový diagram

- vhodný pro kvalitativní znaky k vyjádření četností jejich kategorií



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