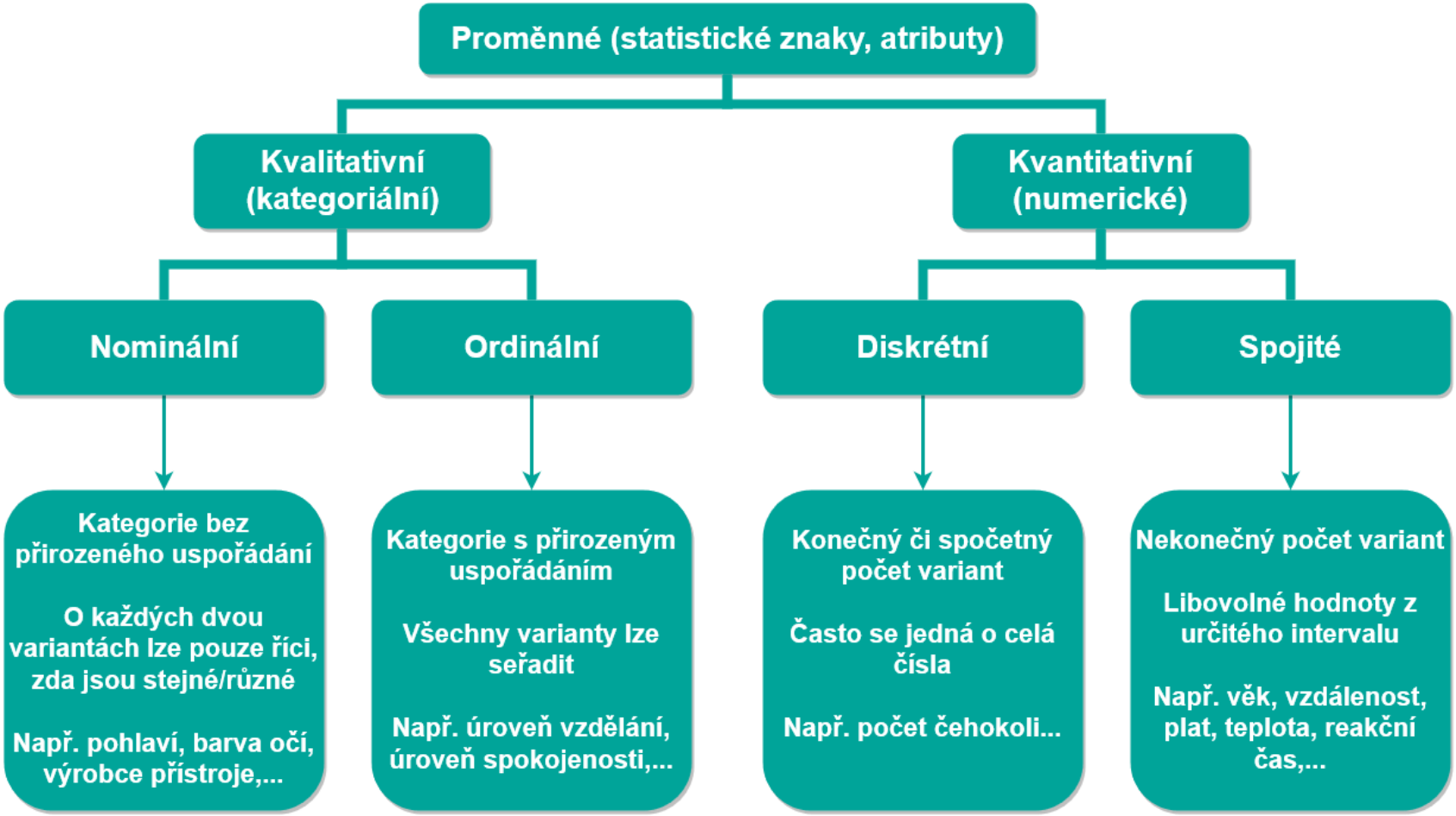
Základy statistiky

Exploratční analýza dat

Adéla Vrtková, Martina Litschmannová

adela.vrtkova@vsb.cz, martina.litschmannova@vsb.cz

VŠB TECHNICKÁ UNIVERZITA OSTRAVA | FAKULTA ELEKTROTECHNIKY A INFORMATIKY | KATEDRA APLIKOVANÉ MATEMATIKY


Proměnné (statistické znaky, atributy)
Kvalitativní (kategoriální)
Kvantitativní (numerické)
Nominální
Ordinální
Diskrétní
Spojité
Kategorie bez přirozeného uspořádání
O každých dvou variantách lze pouze říci, zda jsou stejné/různé
Např. pohlaví, barva očí, výrobce přístroje,...
Kategorie s přirozeným uspořádáním
Všechny varianty lze seřadit
Např. úroveň vzdělání, úroveň spokojenosti,...
Konečný či spočetný počet variant
Často se jedná o celá čísla
Např. počet čehokoli...
Nekonečný počet variant
Libovolné hodnoty z určitého intervalu
Např. věk, vzdálenost, plat, teplota, reakční čas,...

EDA pro kvalitativní proměnnou

- tabulky četností
  - absolutní a relativní četnosti
  - popř. kumulativní abs. a rel. četnosti pro ordinální proměnnou
- sloupcové grafy
- výsečové grafy

Tabulka rozdělení četnosti		
Varianty x_i	**Absolutní četnosti** n_i	**Relativní četnosti** p_i
x_1	n_1	$p_1 = n_1/n$
x_2	n_2	$p_2 = n_2/n$
$\vdots$	$\vdots$	$\vdots$
x_k	n_k	$p_k = n_k/n$
Celkem	$n = \sum_i n_i$	**1**

Tabulka rozdělení četnosti		
Věková kategorie	**Absolutní četnosti**	**Relativní četnosti (%)**
pod 35 let	77	37,37864
35 až 50 let	85	41,26214
nad 50 let	44	21,35922
Celkem	**206**	**100**

Tabulka rozdělení četnosti		
Věková kategorie	**Absolutní četnosti**	**Relativní četnosti (%)**
pod 35 let	77	37,37864
35 až 50 let	85	41,26214
nad 50 let	44	21,35922
Celkem	**206**	**100**

Jak zaokrouhlit relativní četnost?

Tabulka rozdělení četnosti		
Věková kategorie	**Absolutní četnosti**	**Relativní četnosti (%)**
pod 35 let	77	37,4
35 až 50 let	85	41,3
nad 50 let	44	21,4
Celkem	**206**	**100,1**

Tabulka rozdělení četnosti		
Věková kategorie	**Absolutní četnosti**	**Relativní četnosti (%)**
pod 35 let	77	37,4
35 až 50 let	85	41,3
nad 50 let	44	21,4
Celkem	**206**	**100,1**

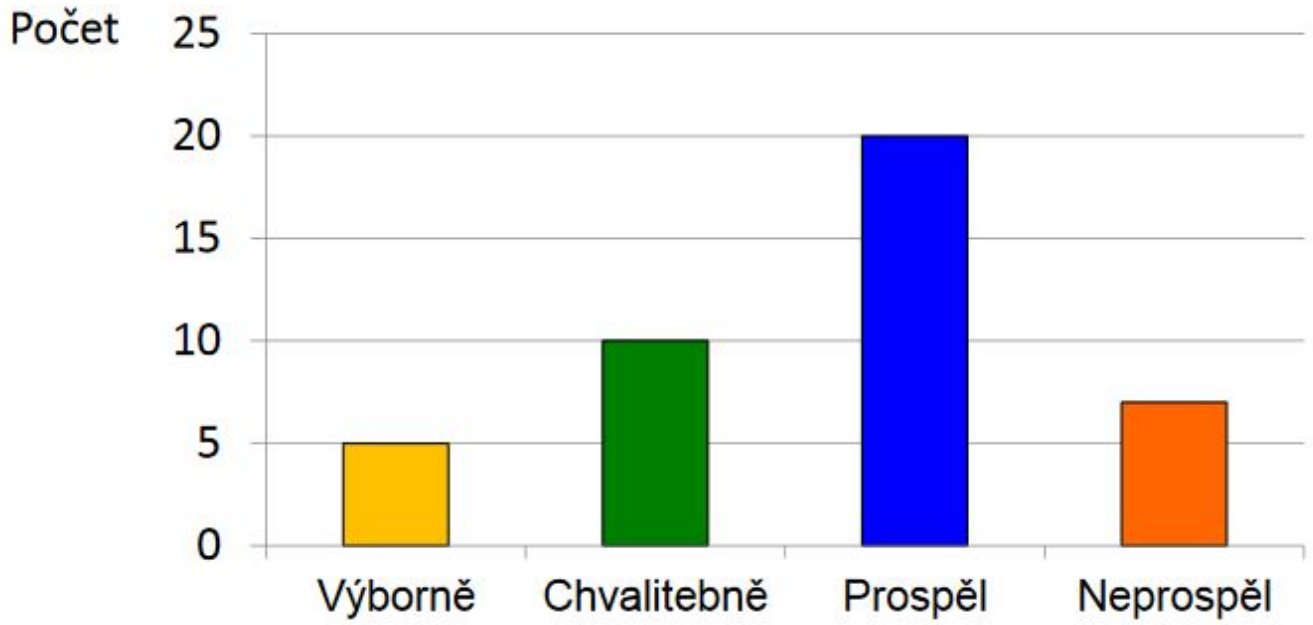
Pozor na zaokrouhlovací chybu!

Tabulka rozdělení četnosti		
Věková kategorie	**Absolutní četnosti**	**Relativní četnosti (%)**
pod 35 let	77	37,4
35 až 50 let	85	41,3
nad 50 let	44	21,3
Celkem	**206**	**100,0**

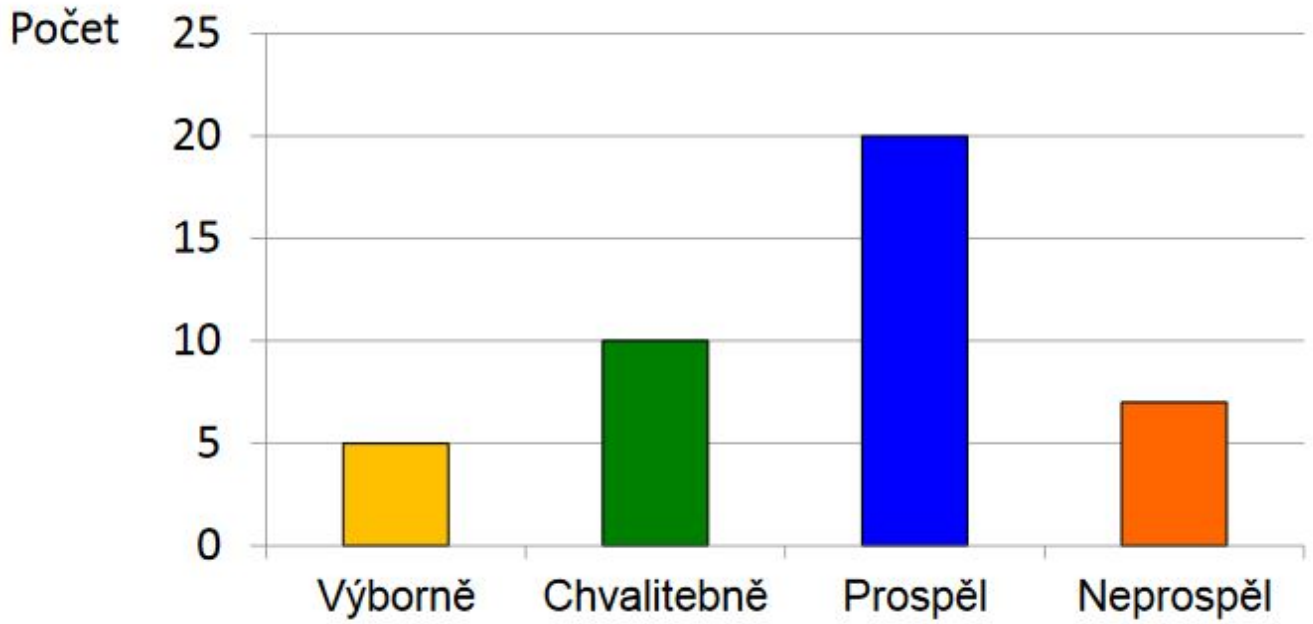
Vizualizace

- sloupcové grafy
- výsečové grafy

Sloupcový graf

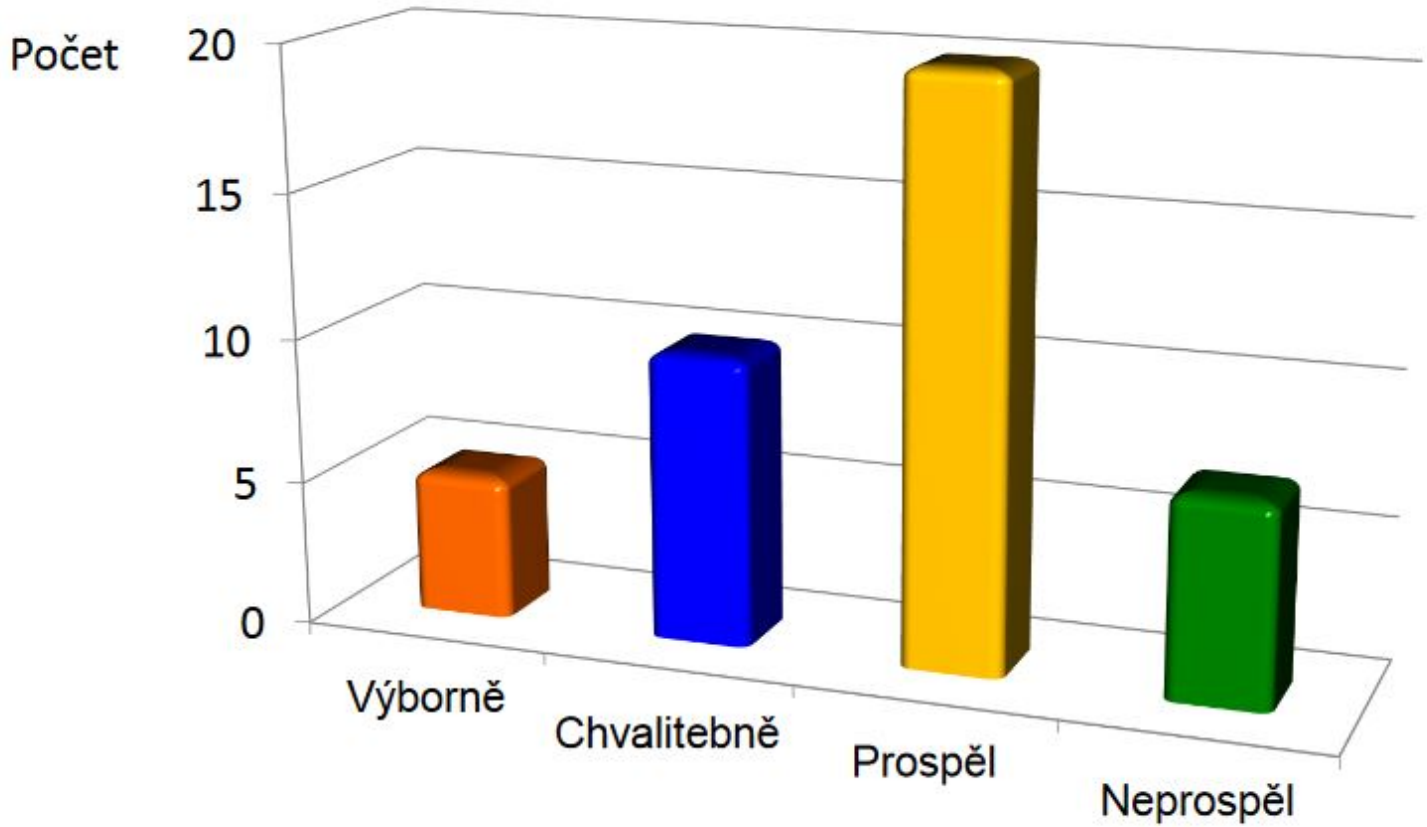


Sloupcový graf

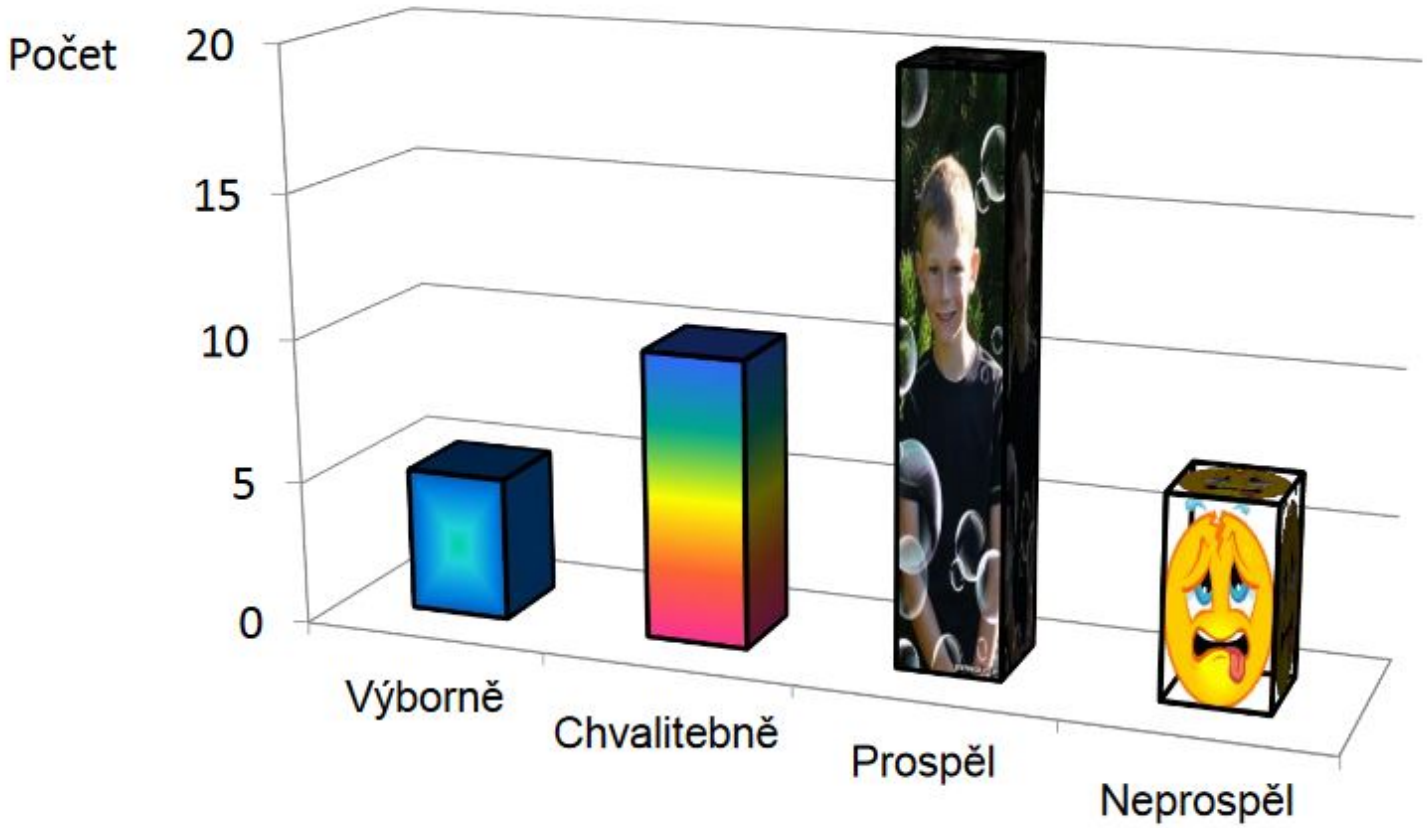


Návrhy na vylepšení?

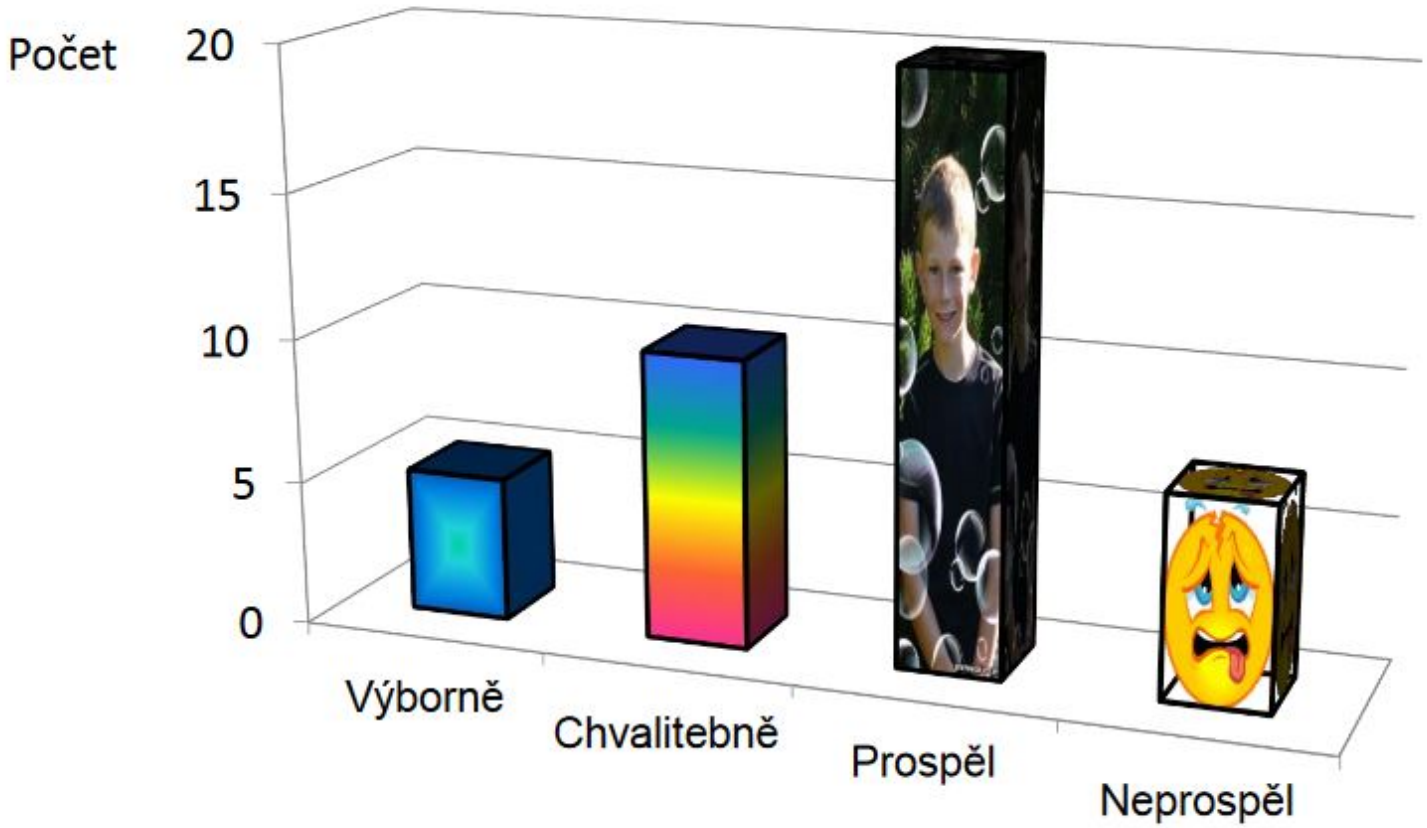
Sloupcový graf



Sloupcový graf

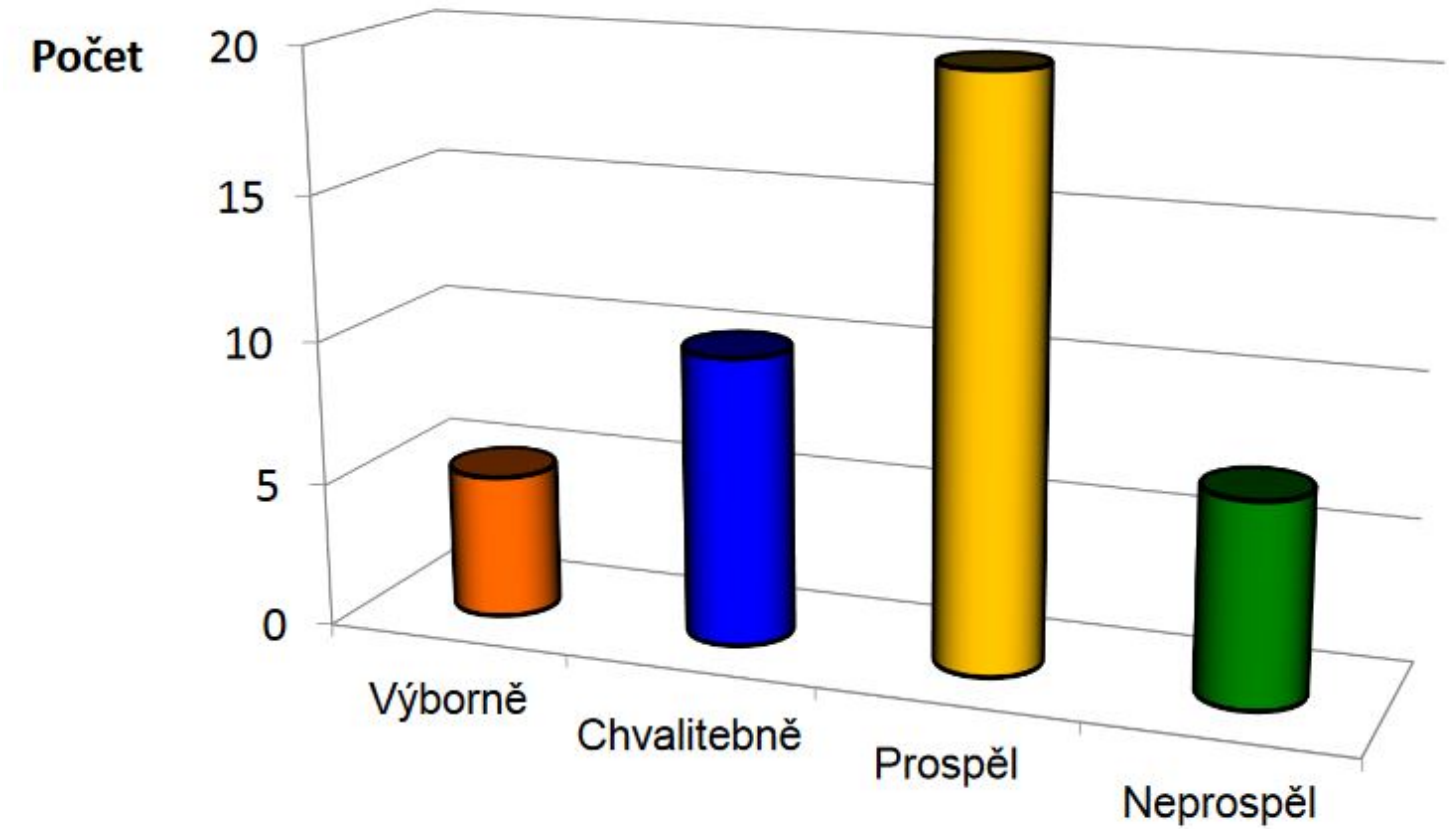


Sloupcový graf

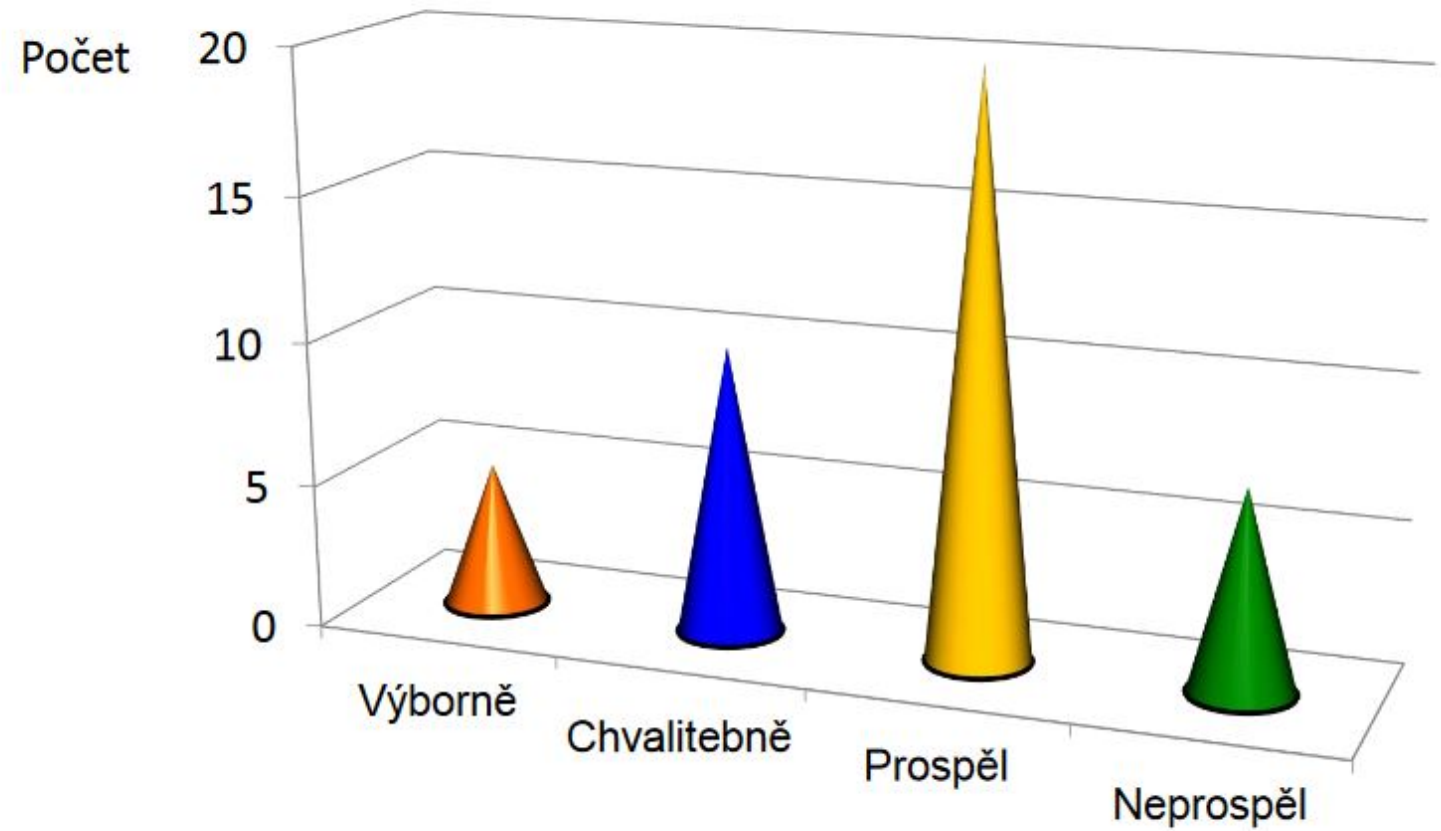


Nezapomínejme, že někdy „méně je více“.

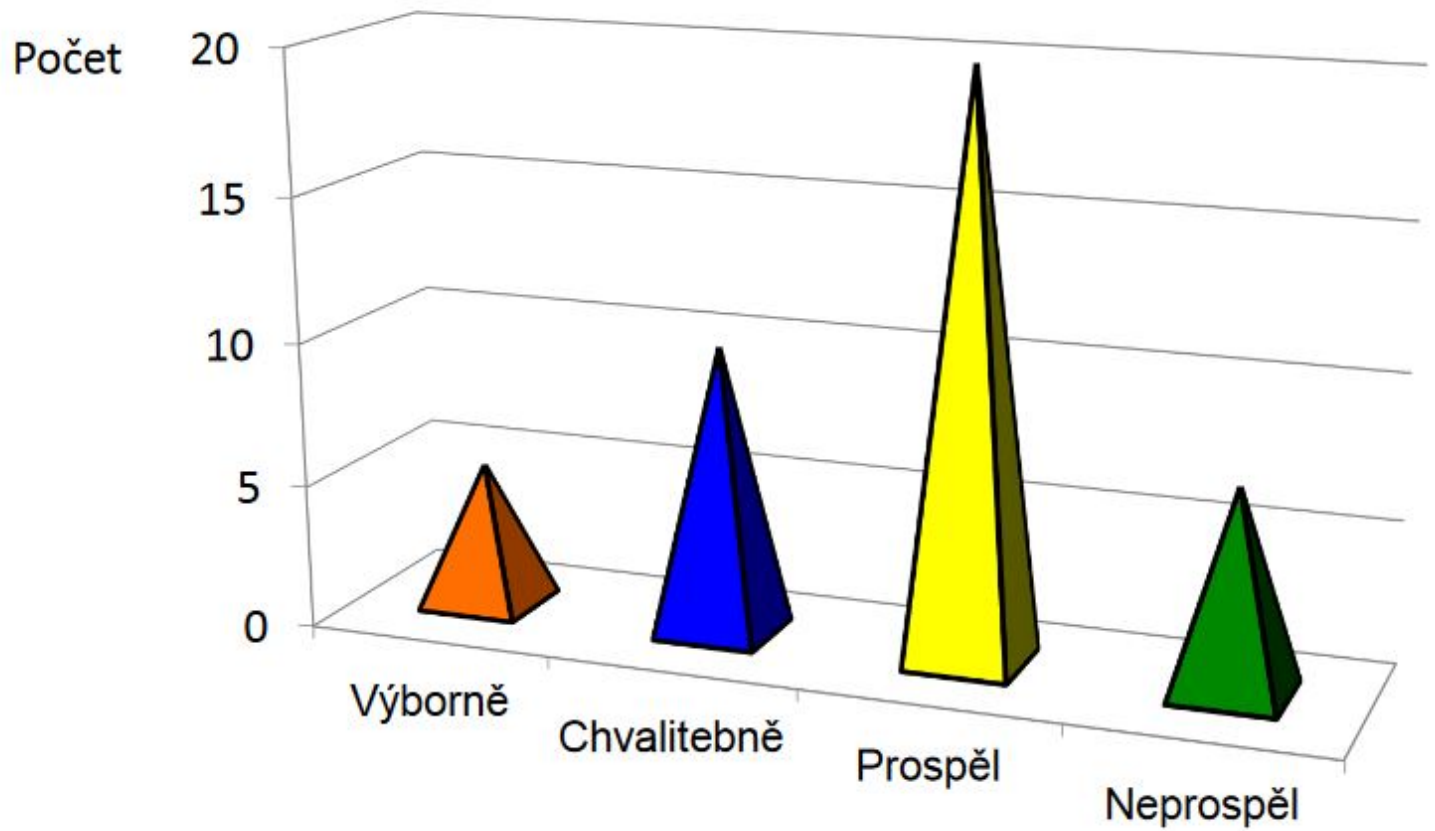
Okénko do historie aneb co Excel uměl...



Okénko do historie aneb co Excel uměl...



Okénko do historie aneb co Excel uměl...



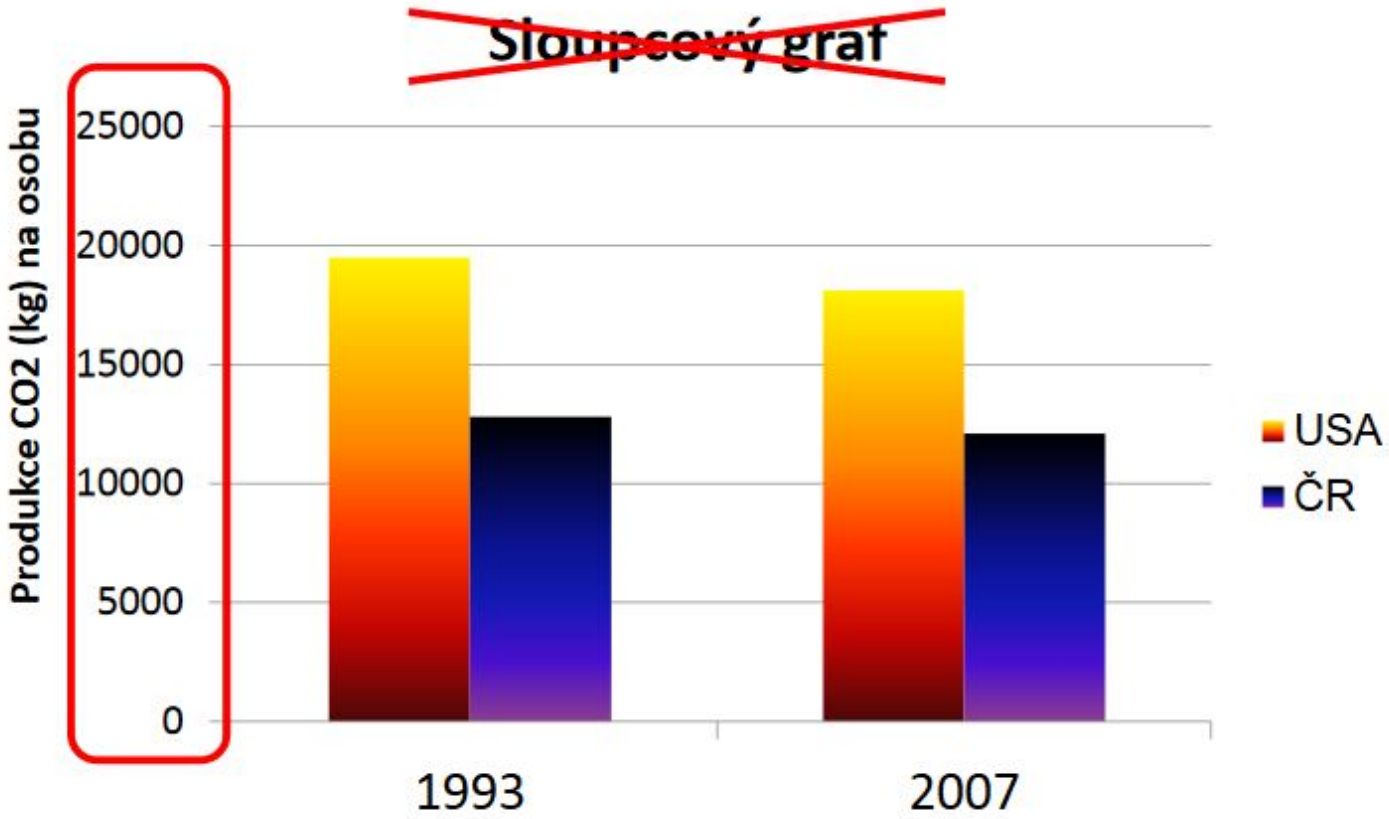
Pár tipů pro sloupcové grafy

- vyhnout se 3D grafům

„3D is great when it comes to James Cameron movies, but your data does not need to jump out at your viewers“
(zdroj: visage.co)

Pár tipů pro sloupcové grafy

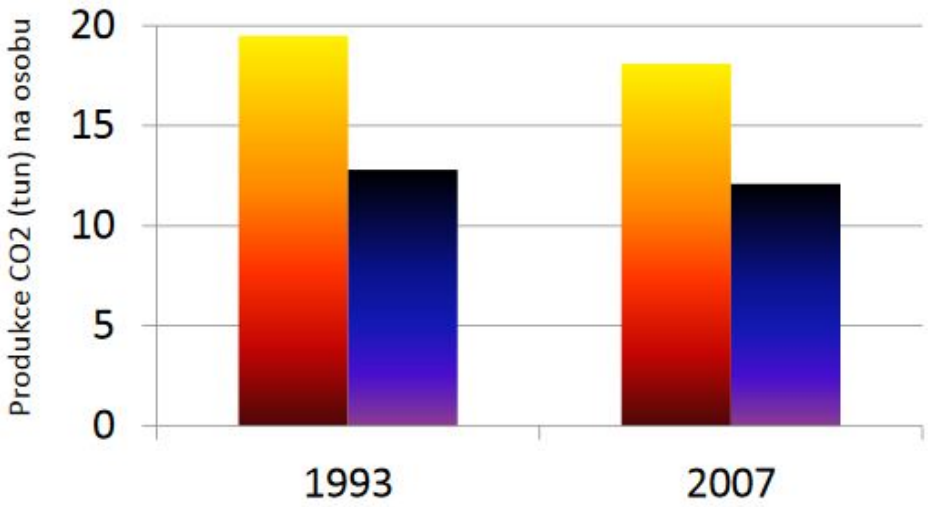
- neefektivní nuly a nadbytečné názvy a popisy

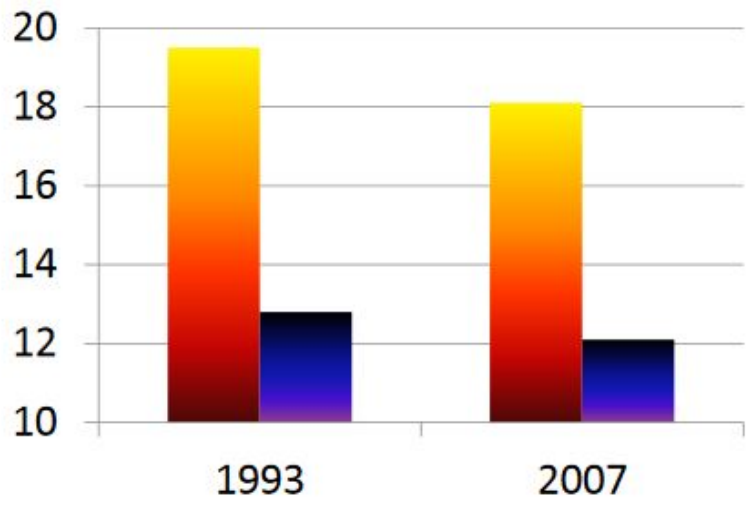


(zdroj dat: `http://en.wikipedia.org/wiki/List_of_countries_by_carbon_dioxide_emissions_per_capita`

Pár tipů pro sloupcové grafy

- pozor na efekt změny rozsahu osy y

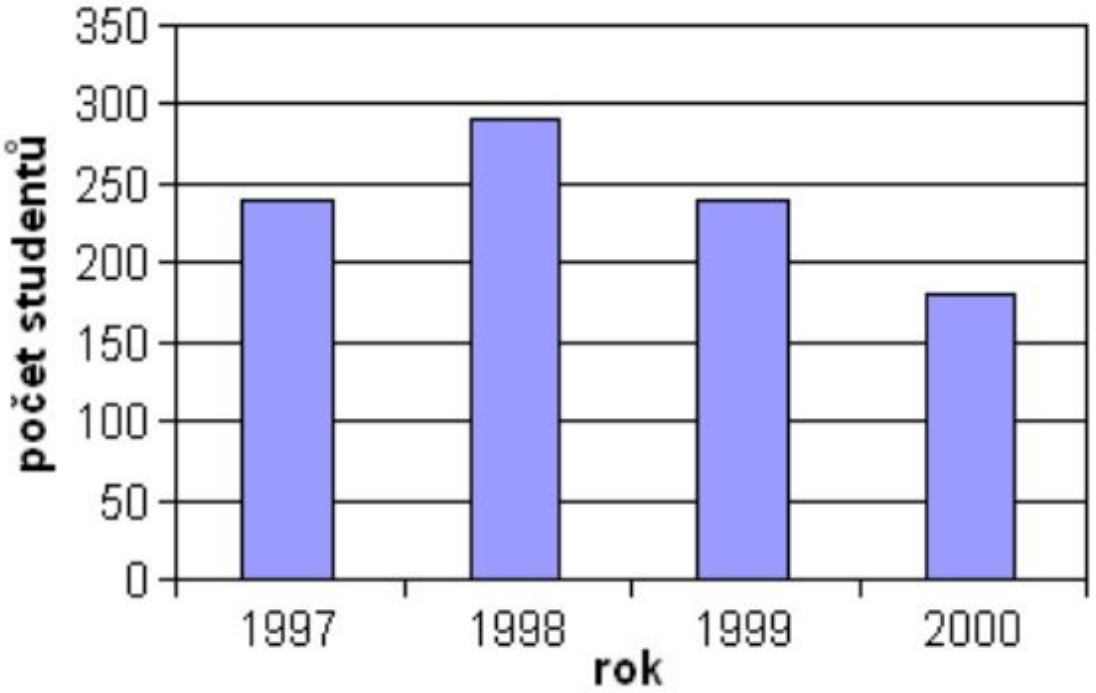




Testový příklad pro žáky 9. tříd

Určete pravdivost tvrzení:

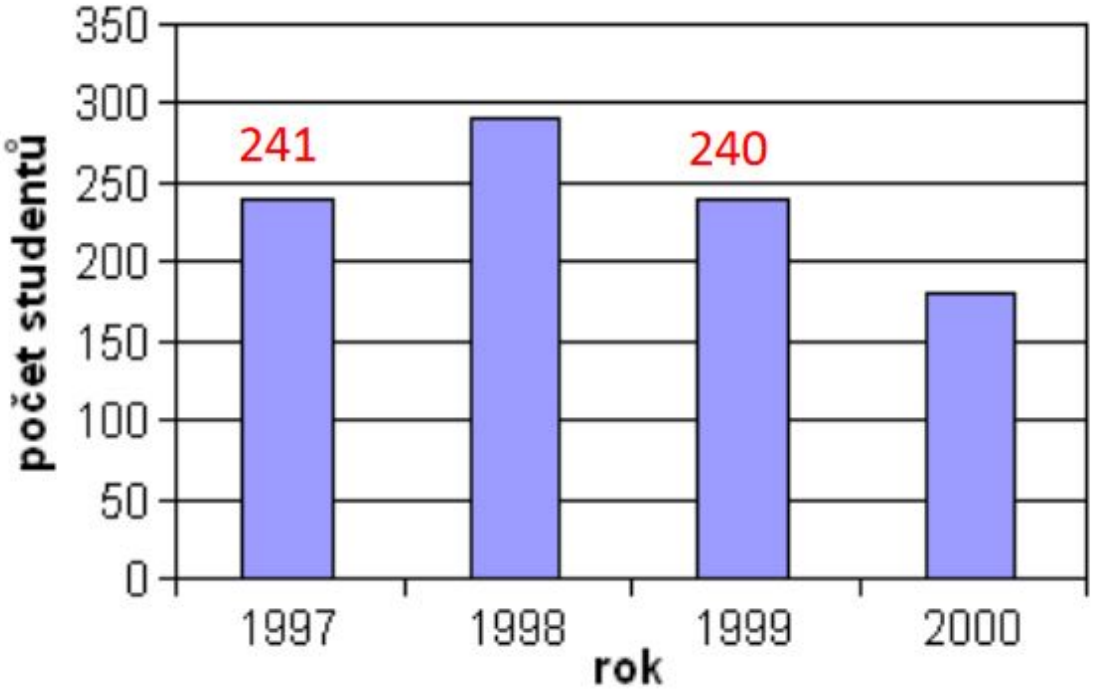
V žádných dvou letech nebyl počet studentů stejný.



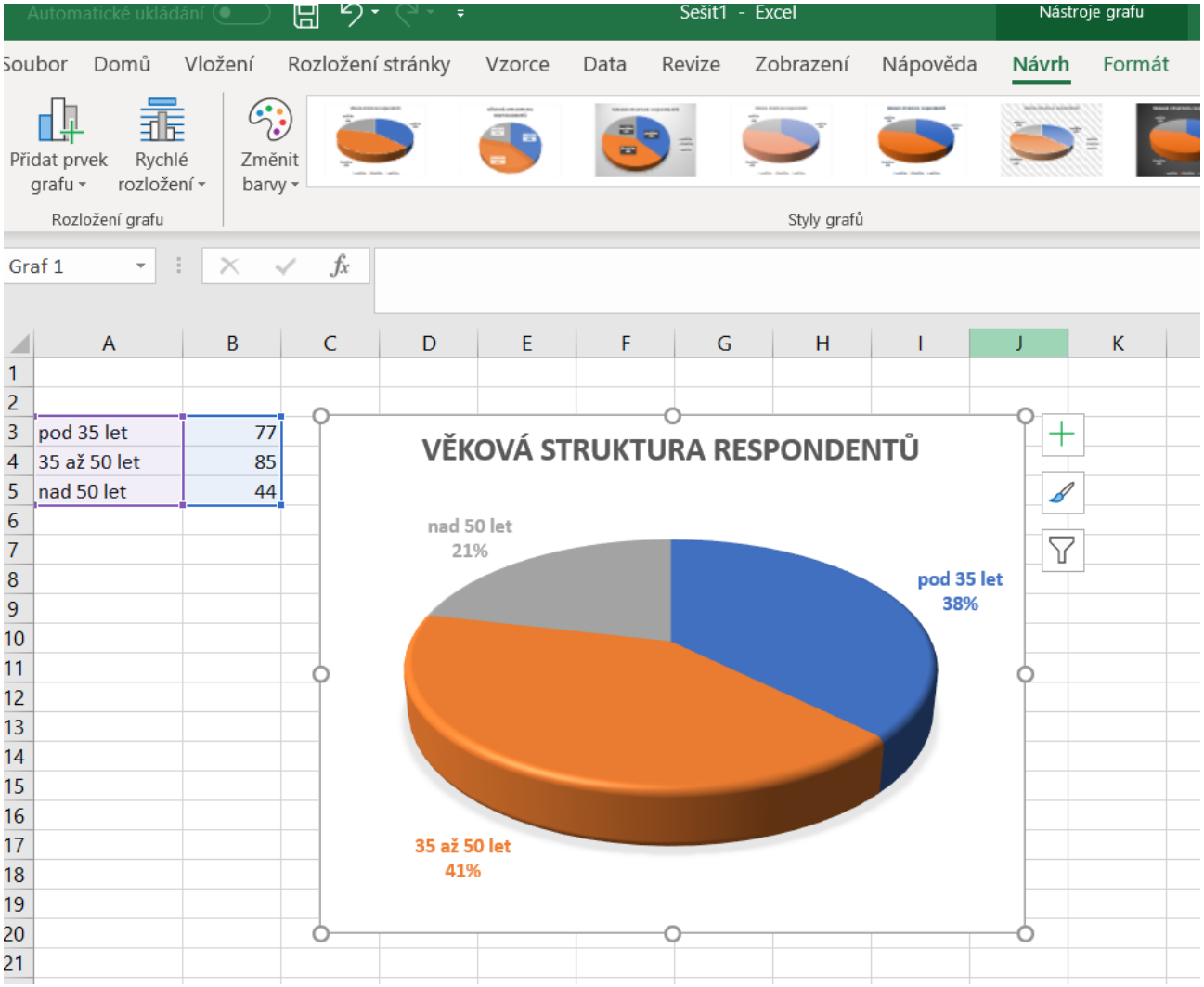
Testový příklad pro žáky 9. tříd

Určete pravdivost tvrzení:

V žádných dvou letech nebyl počet studentů stejný.

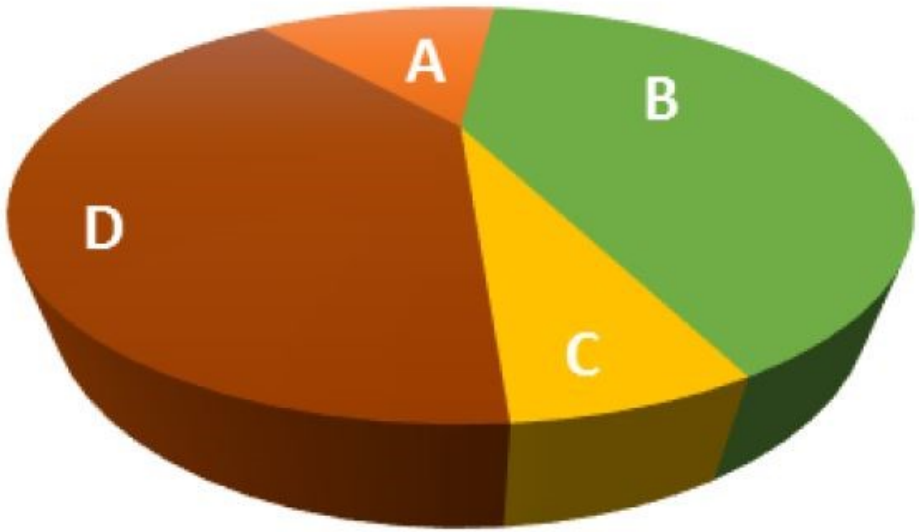


Výsečový graf



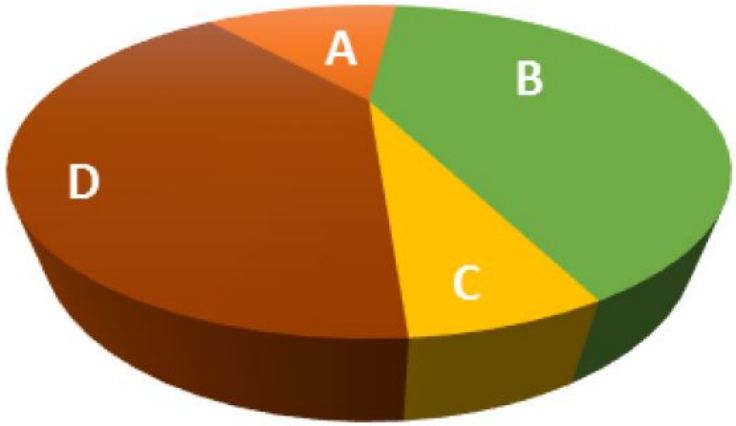
Výsečový graf

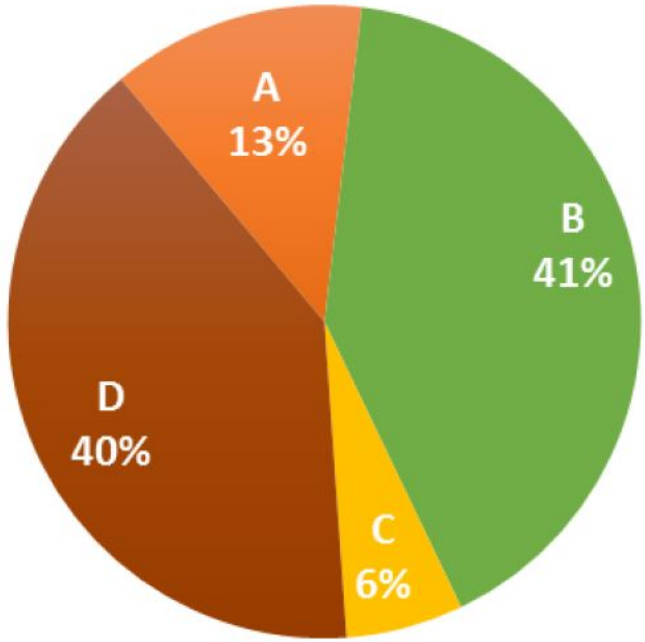
Jaký je poměr mezi velikostí výsečí A a C? Jaký je poměr mezi velikostí výsečí B a D?



Výsečový graf

Jaký je poměr mezi velikostí výsečí A a C? Jaký je poměr mezi velikostí výsečí B a D?



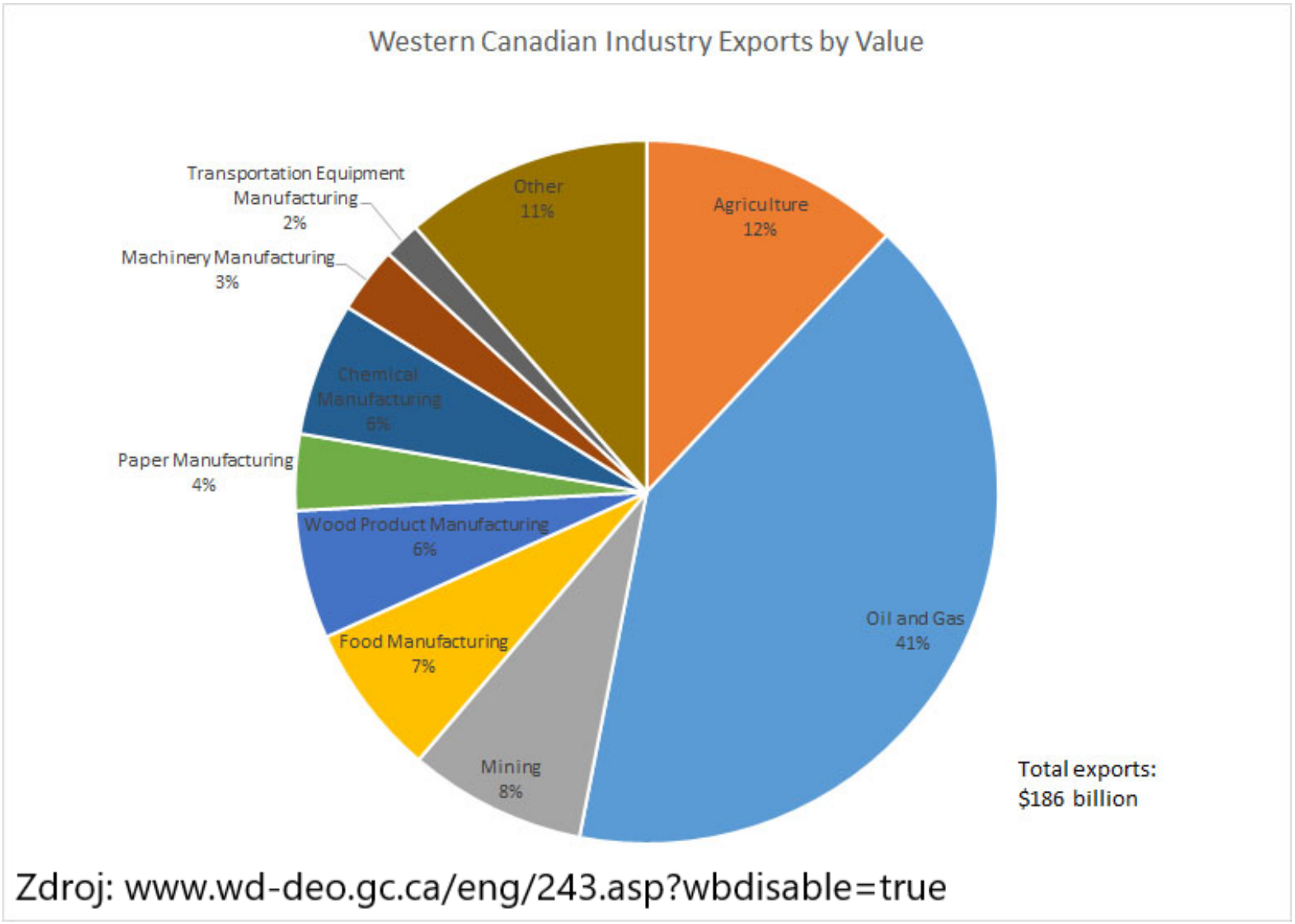


Výsečový graf

- citace z nápovědy ke statistickému softwaru R:

"Pie charts are a very bad way of displaying information. The eye is good at judging linear measures and bad at judging relative areas. A bar chart or dot chart is a preferable way of displaying this type of data."

Výsečový graf - jak to (ne)má vypadat



EDA pro kvantitativní proměnnou

- číselné charakteristiky
- krabicové grafy
- histogramy

Číselné charakteristiky

- míry polohy
- míry variability
- míry tvaru

Číselné charakteristiky

- míry polohy
- míry variability
- míry tvaru

Všechny míry lze vypočítat pro výběr i pro populaci (pokud jsou k dispozici data).

Míry polohy

- Míry polohy, také označované jako míry centrální tendence, cílí nejčastěji na odhad jakéhosi „středu“.
- Nejčastěji používané míry polohy:
  - aritmetický průměr,
  - medián,
  - modus,
  - minimum a maximum,
  - kvantily (kvartily, decily, percentily).

Míry polohy - aritmetický průměr

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Míry polohy - aritmetický průměr

- V malé vesnici někde v Americe žije 6 lidí, jejichž roční plat (v USD) je uveden níže.

  25 000, 27 000, 29 000, 35 000, 37 000, 38 000.

  Určete průměrný plat obyvatel této vesnice.

Míry polohy - aritmetický průměr

- V malé vesnici někde v Americe žije 6 lidí, jejichž roční plat (v USD) je uveden níže.

  25 000, 27 000, 29 000, 35 000, 37 000, 38 000.

  Určete průměrný plat obyvatel této vesnice.

31 830

Míry polohy - aritmetický průměr

- Do vesnice se přistěhoval Bill Gates, jehož roční příjem je 40 milionů dolarů.

  25 000, 27 000, 29 000, 35 000, 37 000, 38 000,
  40 000 000.

  Jaký je průměrný plat obyvatel této vesnice nyní?

Míry polohy - aritmetický průměr

- Do vesnice se přistěhoval Bill Gates, jehož roční příjem je 40 milionů dolarů.

  25 000, 27 000, 29 000, 35 000, 37 000, 38 000,
  40 000 000.

  Jaký je průměrný plat obyvatel této vesnice nyní?

  5 741 571

Míry polohy - aritmetický průměr

- Do vesnice se přistěhoval Bill Gates, jehož roční příjem je 40 milionů dolarů.

  25 000, 27 000, 29 000, 35 000, 37 000, 38 000,
  40 000 000.

  Jaký je průměrný plat obyvatel této vesnice nyní?

  5 741 571

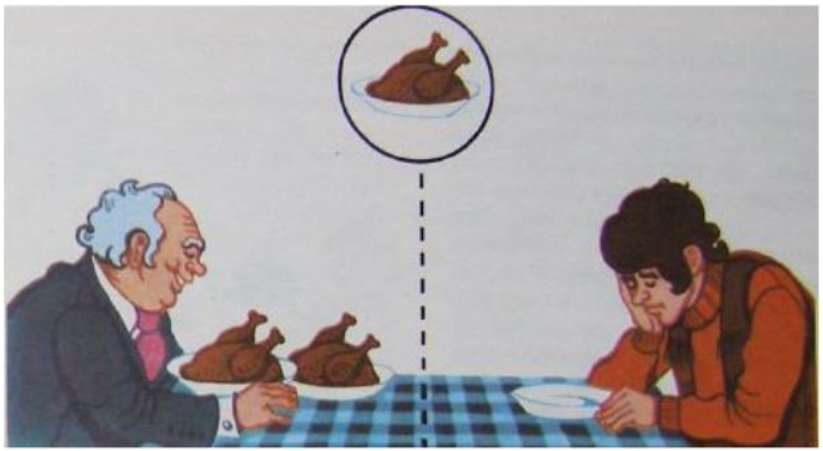
Aritmetický průměr je velice citlivý na odlehlé hodnoty!

Ošidný průměr

Matematik, fyzik a statistik střílejí na terč. Matematik střílí první a mine cíl kousek doleva. Fyzik střílí po něm a také mine cíl tentokrát kousek doprava. Statistik se začne radovat: „Hurá! Trefili jsme cíl!“.

Ošidný průměr

Matematik, fyzik a statistik střílejí na terč. Matematik střílí první a mine cíl kousek doleva. Fyzik střílí po něm a také mine cíl tentokrát kousek doprava. Statistik se začne radovat: „Hurá! Trefili jsme cíl!“.



Zdroj: SWOBODA, Helmut. Moderní statistika., 1977.

Míry polohy - kvantily

$100p\%$ kvantil $\tilde{x}_p$

- $100p$ % hodnot (dat) je menší nebo rovno hodnotě $\tilde{x}_p$
- např. 20 % hodnot je menší nebo rovno **20% kvantilu**, tudíž 80 % hodnot je nad ním

Medián je 50% kvantil.

Míry polohy - kvantily

- **Kvartily**
  - dolní kvartil = 25% kvantil ($\tilde{x}_{0,25}$)
  - medián = 50% kvantil ($\tilde{x}_{0,50}$)
  - horní kvartil = 75% kvantil ($\tilde{x}_{0,75}$)
- **Decily**
  - 1. decil = 10% kvantil ($\tilde{x}_{0,10}$)
  - 2. decil = 20% kvantil ($\tilde{x}_{0,20}$)
  - $\vdots$
  - 9. decil = 90% kvantil ($\tilde{x}_{0,90}$)
- **Percentily**
  - 1. percentil = 1% kvantil ($\tilde{x}_{0,01}$)
  - 2. percentil = 2% kvantil ($\tilde{x}_{0,02}$)
  - $\vdots$
  - 99. percentil = 99% kvantil ($\tilde{x}_{0,99}$)

Míry polohy - kvantily

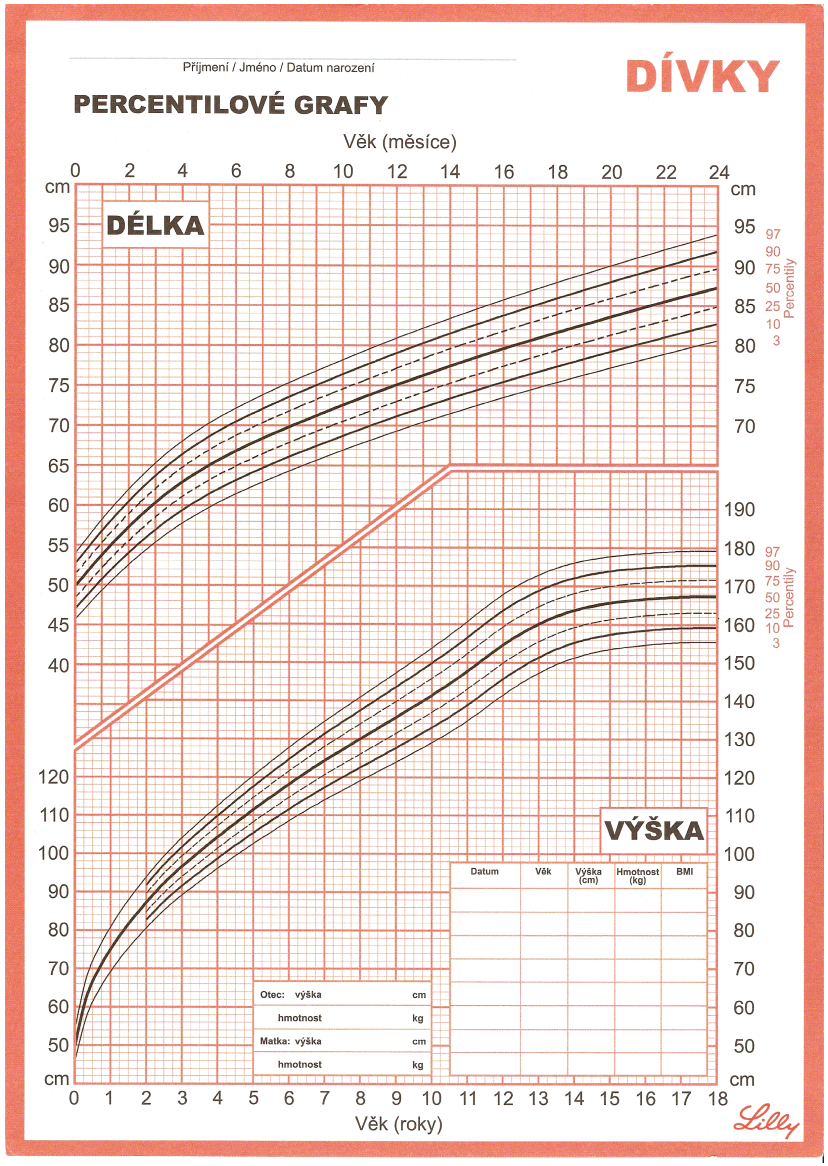
Jaké hodnoty reprezentují 0% a 100% kvantil?

Míry polohy - kvantily

Jaké hodnoty reprezentují 0% a 100% kvantil?

minimum a maximum

Kvantily v praxi



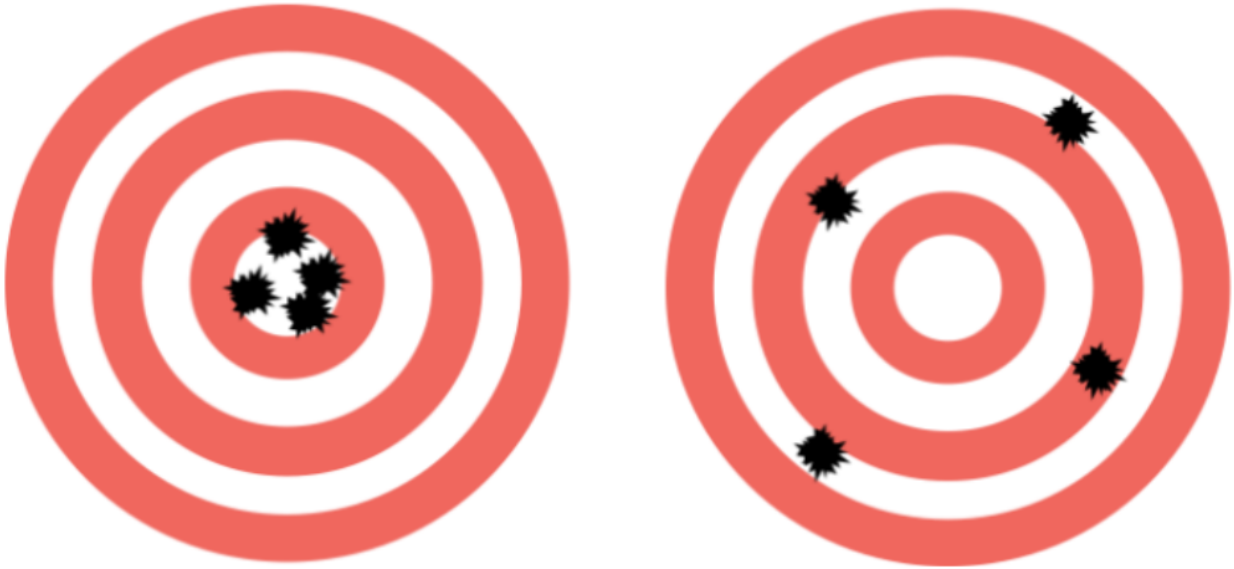
Kvantily v praxi

„80 % nemovitostí prodáme do 2 měsíců“
Zdroj: výloha Fincentrum Reality, 28. října 813/248, 70900 Ostrava

Míry variability

- Charakteristiky hodnotící rozptýlenost hodnot statistického souboru kolem nějaké míry polohy.
- Nejčastěji používané míry variability:
  - rozptyl
  - směrodatná odchylka
  - variační koeficient
  - rozpětí
  - mezikvartilové rozpětí

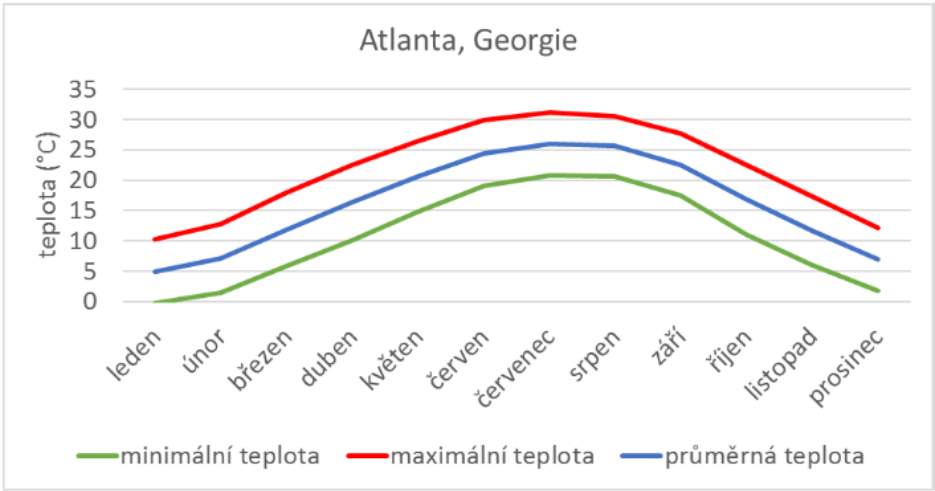
Míry variability - proč je potřebujeme?

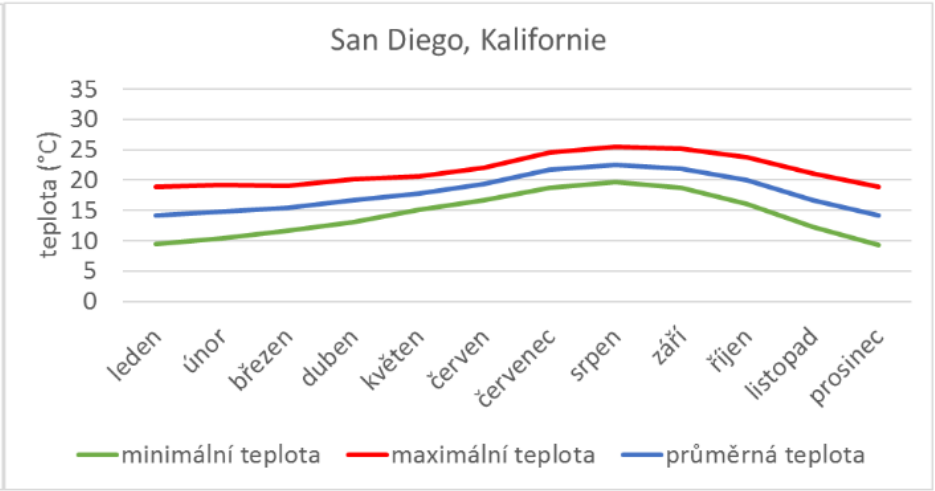


Zdroj: blackswanfarming.com

Míry variability - proč je potřebujeme?

- Průměrná teplota v Atlantě je 15,5°C.
- Průměrná teplota v San Diegu je 16,5°C.





Míry variability - (výběrový) rozptyl

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

Míry variability - (výběrový) rozptyl

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

Jednotky rozptylu jsou druhou mocninou jednotek analyzované proměnné!

Míry variability - (výběrová) směrodatná odchylka

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

K čemu je směrodatná odchylka dobrá?

- Čebyševova nerovnost: $\forall k > 0 : P(\mu - k\sigma < X < \mu + k\sigma) > 1 - \frac{1}{k^2}$

k	$P(\mu - k\sigma < X < \mu + k\sigma)$
1	> 0
2	$> 0,75$
3	$> 0,89$

- $\mu \cong \overline{x}$
- $\sigma \cong s$

Míry variability - variační koeficient

- Používá se většinou pro proměnné nabývající nezáporných hodnot.
- Umožňuje srovnání variability proměnných, které mají různé jednotky.

$$V = \frac{s}{|\overline{x}|}$$

- Obvykle vyjádřen v procentech. Čím je nižší, tím homogennější soubor.

$$V = \frac{s}{|\overline{x}|} \cdot 100 \quad (\%)$$

- Opatrně, je-li průměr blízký nule.

Míry variability - rozpětí

$$Rozpětí = Max - Min$$

Míry variability - mezikvartilové rozpětí

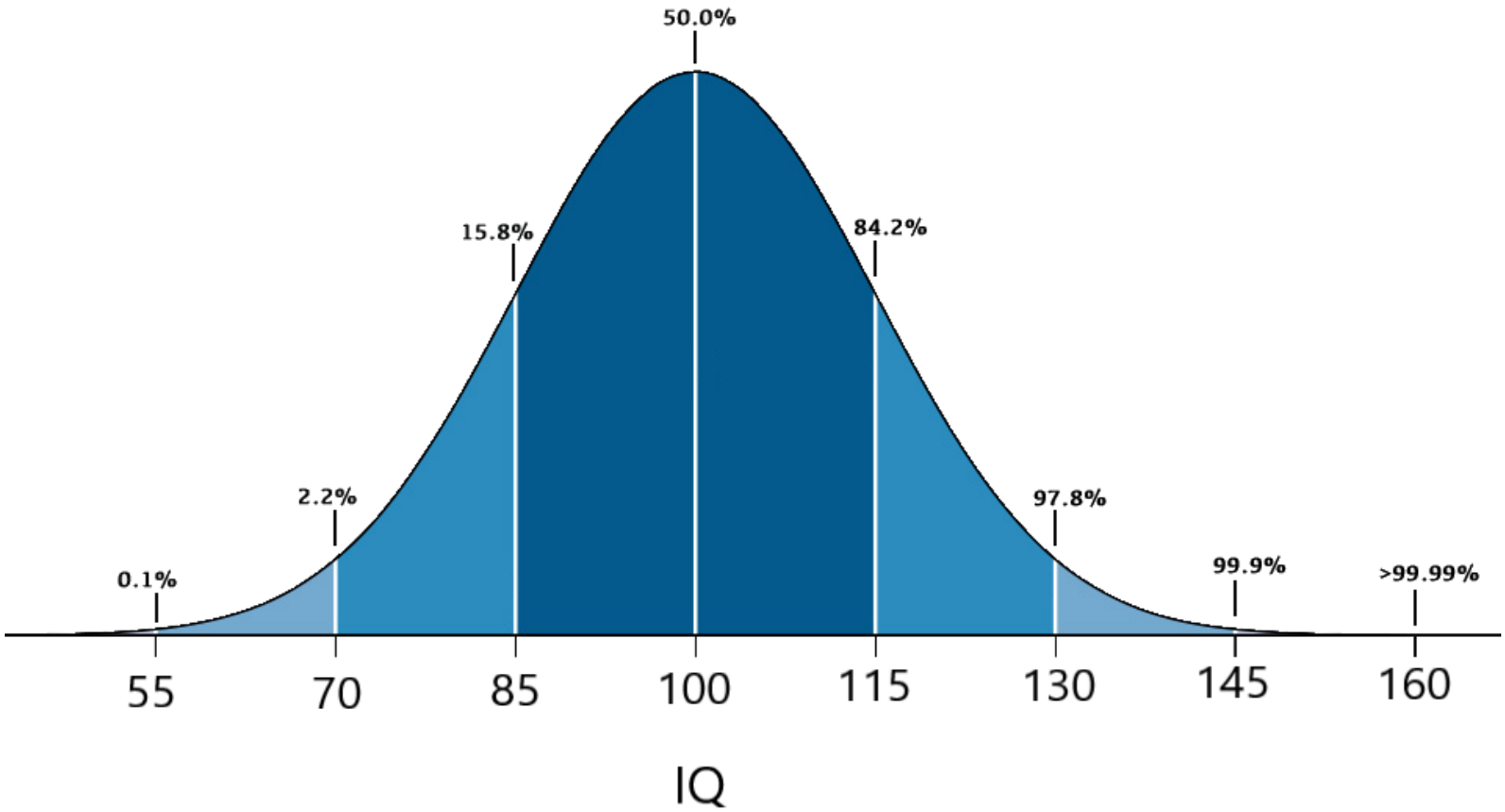
- Využívá se např. při identifikaci odlehlých pozorování.

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

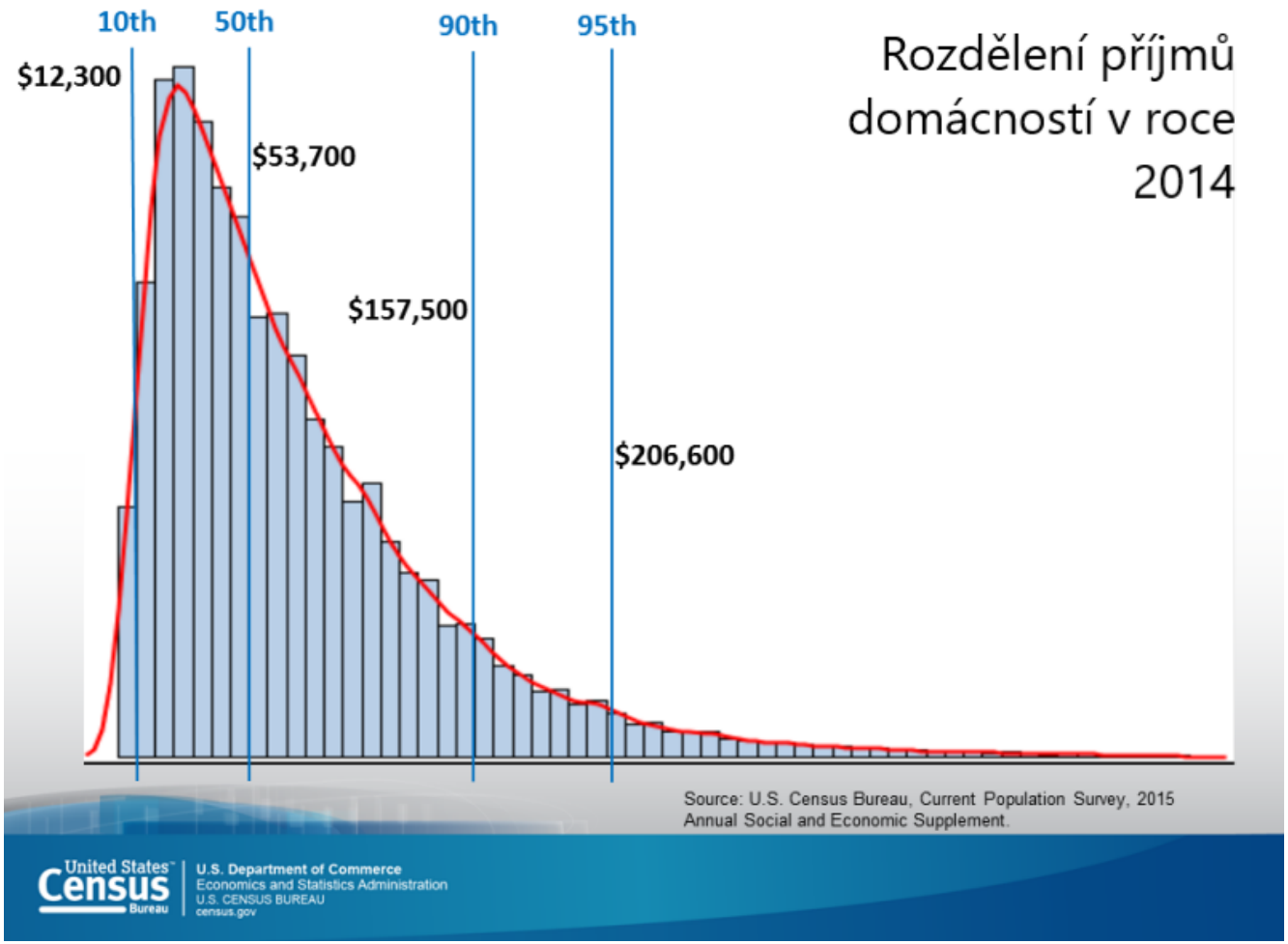
Míry tvaru

- Míry tvaru popisují symetrii dat a koncentraci dat kolem průměru.
- Slouží ke srovnání rozdělení dat s rozdělením normálním.
- Nejčastěji používané míry tvaru:
  - šikmost
  - špičatost

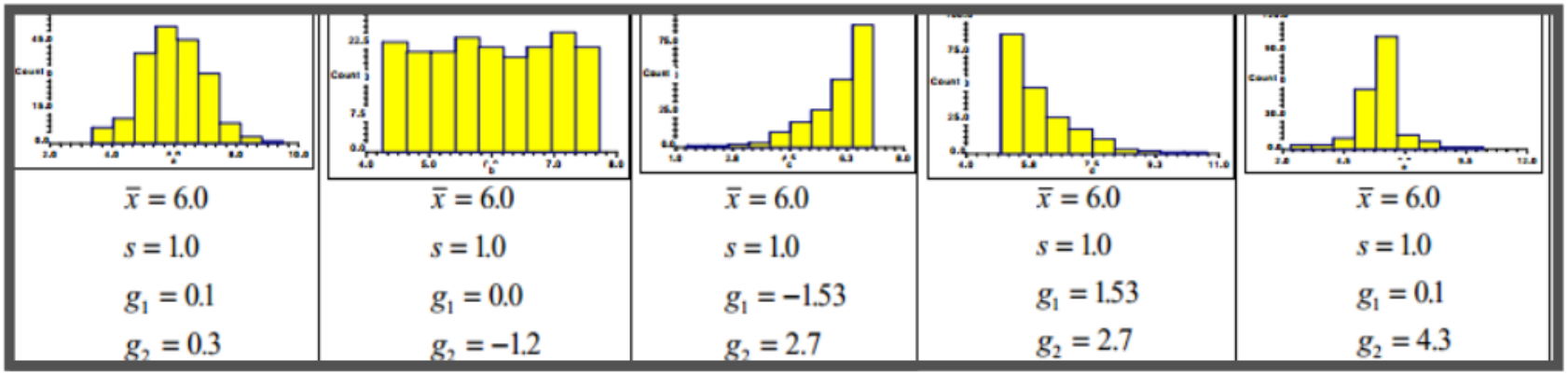
Míry tvaru



Míry tvaru


10th
50th
90th
95th
$12,300
$53,700
$157,500
$206,600
Rozdělení příjmů domácností v roce 2014
Source: U.S. Census Bureau, Current Population Survey, 2015 Annual Social and Economic Supplement.
United States Census Bureau
U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU
census.gov

Míry tvaru

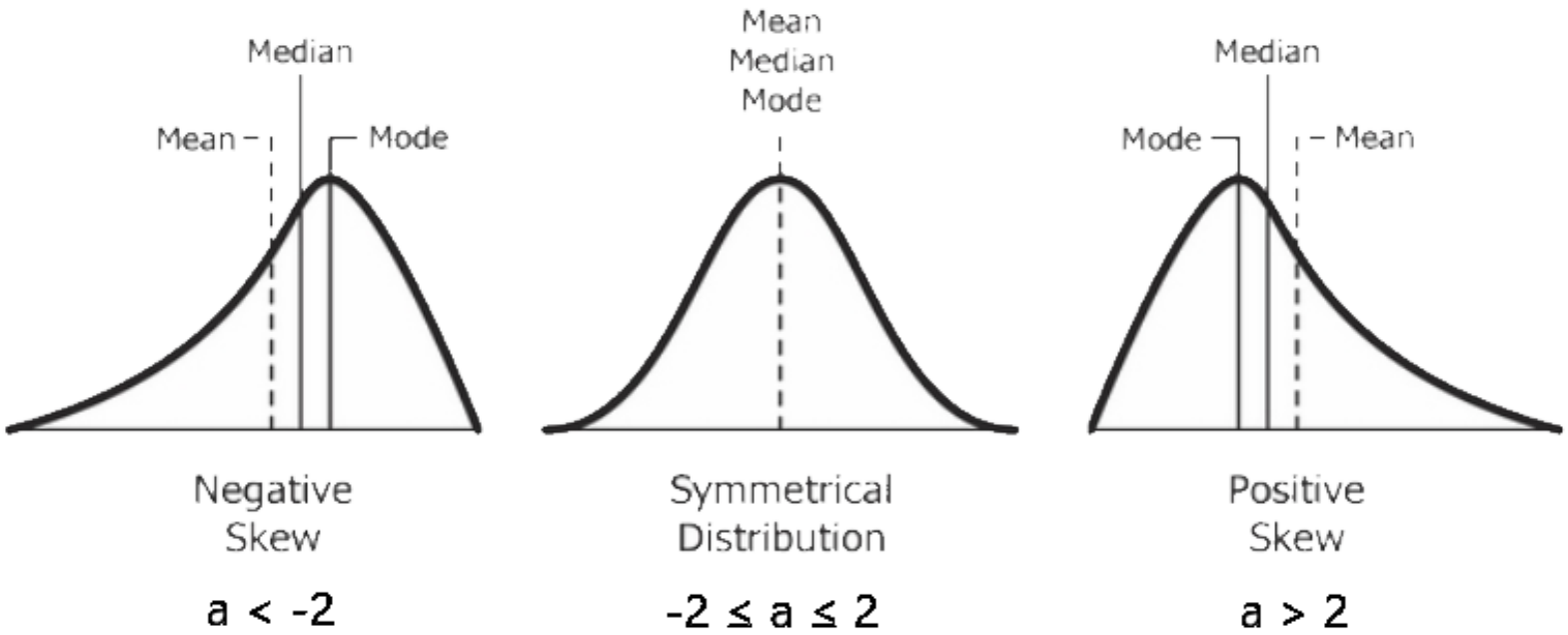


Zdroj: TVRDÍK, J.: Základy matematické statistiky, Ostravská univerzita, 2008

- Všechny ukázky mají stejný průměr a směrodatnou odchylku. Na první pohled je však vidět různý tvar rozdělení. Právě tyto rozdíly nám číselně dokážou popsat míry tvaru - šikmost a špičatost (na obrázku značeno g_1 a g_2).

Míry tvaru - šikmost

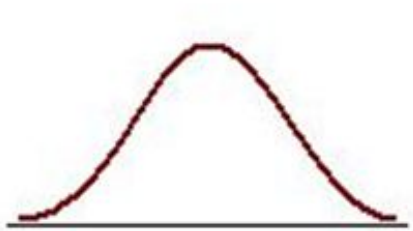
$$a = \frac{n}{(n-1)(n-2)} \cdot \frac{\sum_{i=1}^{n}(x_i - \overline{x})^3}{s^3}$$



Míry tvaru - špičatost

$$b = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \cdot \frac{\sum_{i=1}^{n}(x_i - \overline{x})^4}{s^4} - 3\frac{(n-1)^2}{(n-2)(n-3)}$$

b < -2



-2 ≤ b ≤ 2

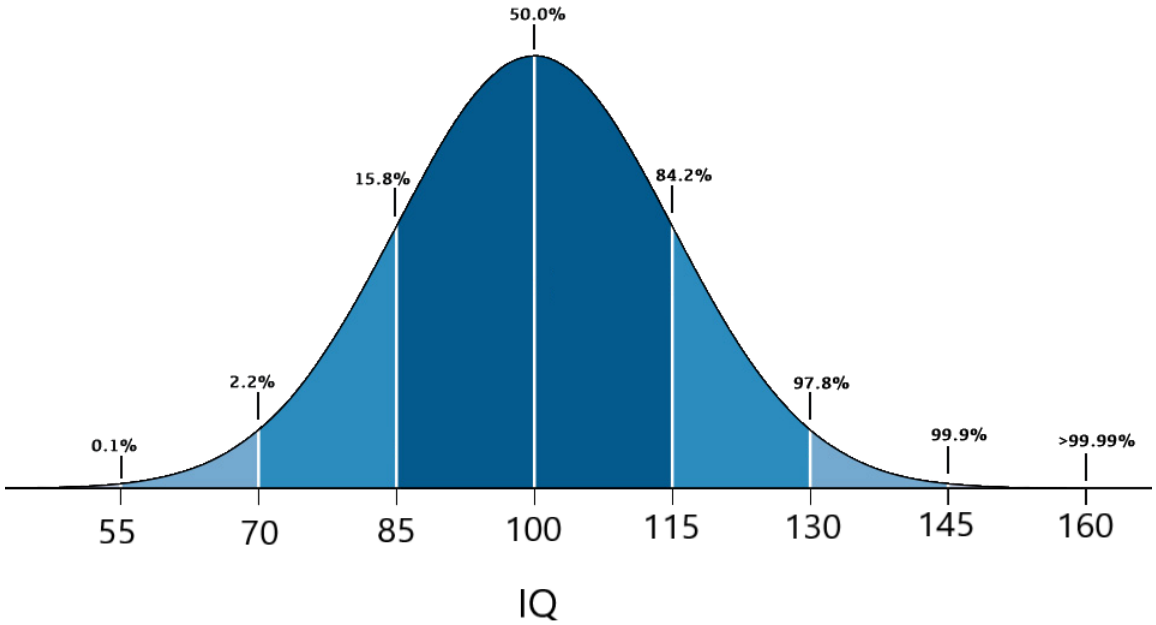
b > 2

Míry tvaru

Míry tvaru mohou být v různých softwarech různě definovány. Je důležité si definici vždy v nápovědě zkontrolovat, abychom výsledky správně interpretovali.

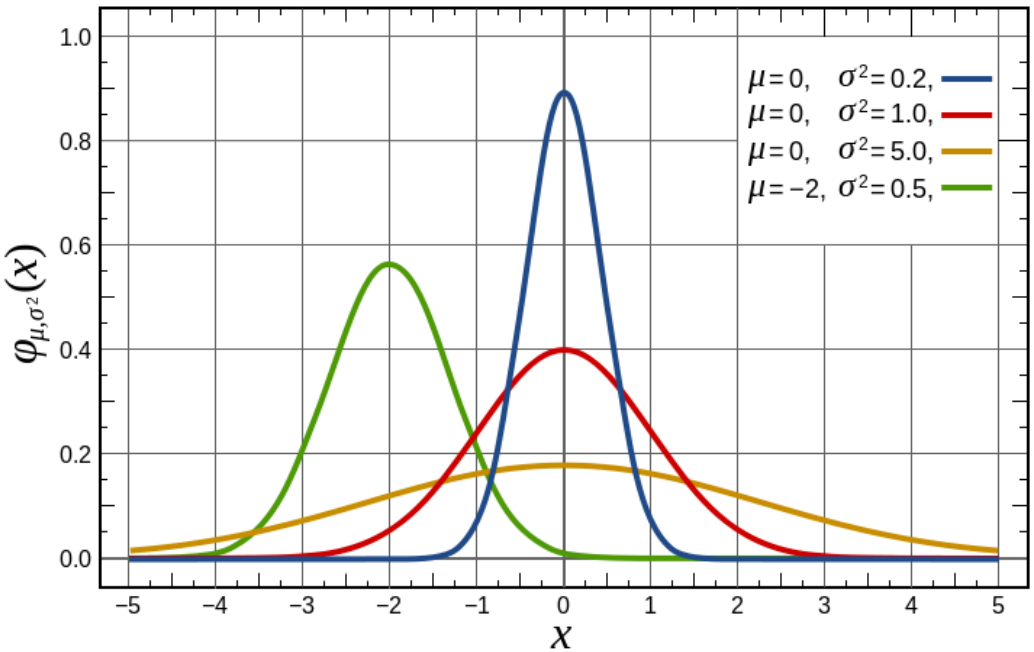
Normální rozdělení

- nazýváno také Gaussovské
- jedná se o jakési „chování“ dat, které má výhodné vlastnosti



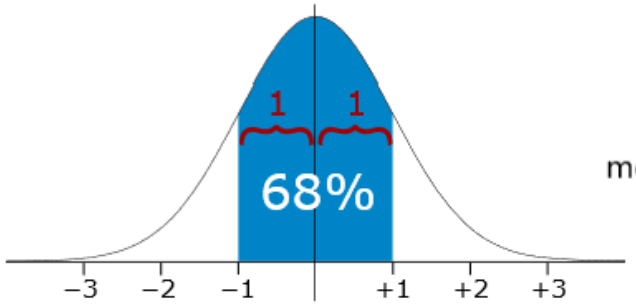
Normální rozdělení

- $N(\mu, \sigma^2)$ - μ (průměr), σ^2 (rozptyl)
- výběr pocházející z normálního rozdělení by měl mít šikmost a špičatost přibližně nulovou
- průměr = modus = medián

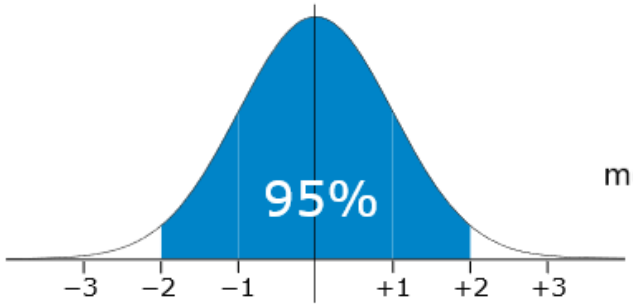


www.statisticshowto.com/probability-and-statistics/normal-distributions/

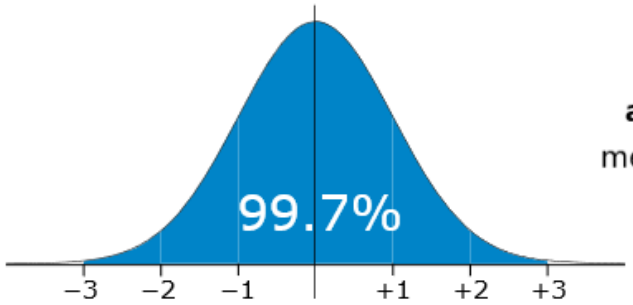
Normální rozdělení - empirické pravidlo 3 sigma



asi 68 % hodnot se nachází ve vzdálenosti menší než **1 směrodatná odchylka** od průměru



asi 95 % hodnot se nachází ve vzdálenosti menší než **2 směrodatné odchylky** od průměru



asi 99,7 % hodnot se nachází ve vzdálenosti menší než **3 směrodatné odchylky** od průměru

www.mathsisfun.com/data/standard-normal-distribution.html

Odlehlá pozorování

- Hodnoty, které se **mimořádně liší od ostatních hodnot** analyzované proměnné a tím ovlivňují např. vypovídací hodnotu průměru.
- Můžou se vyskytnou zcela **náhodně, chybou měření, lidskou chybou** apod.
- Známe-li příčinu odlehlosti a předpokládáme-li, že již nenastane, jsme oprávněni tato pozorování vyloučit.
- V ostatních případech je nutno zvážit, zda se vyloučením odlehlých pozorování nepřipravíme o důležité informace.

    - Metoda vnitřních/vnějších hradeb
    - Chauvenetovo kritérium
    - Mahalanobisova vzdálenost

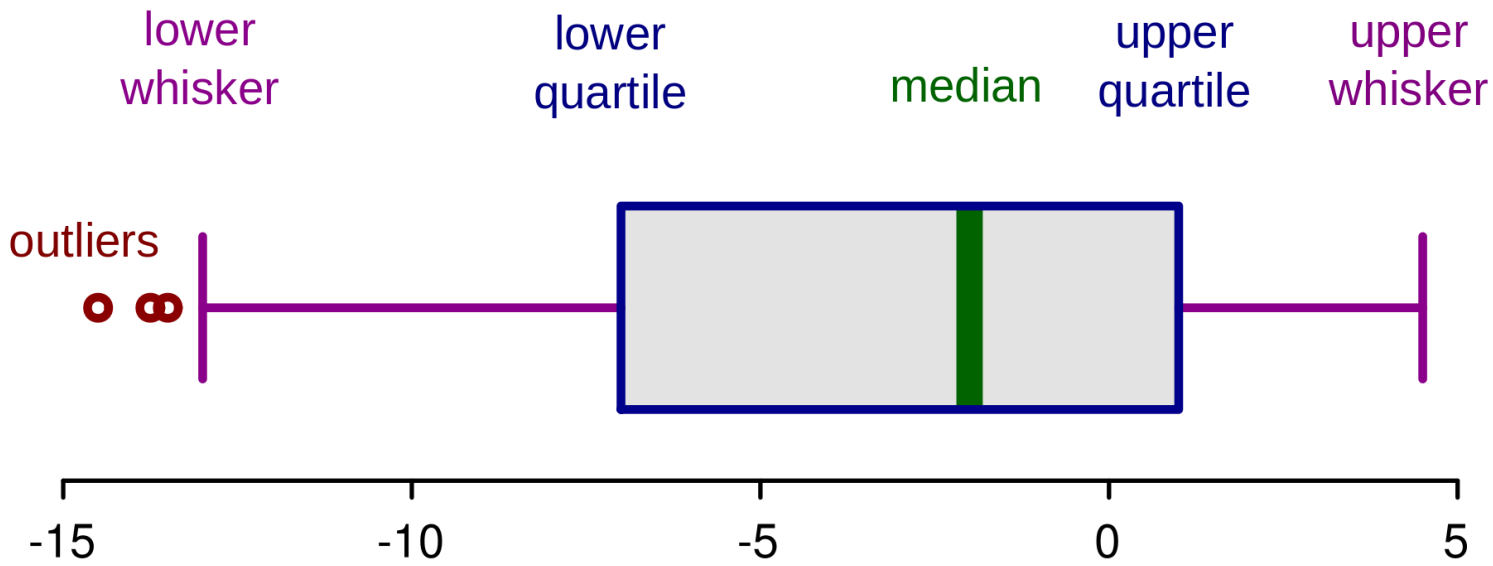
Metoda vnitřních hradeb

Je-li pozorování
menší než $\tilde{\mathbf{x}}_{\mathbf{0,25}} - \mathbf{1,5} \cdot \mathbf{IQR}$
nebo
větší než $\tilde{\mathbf{x}}_{\mathbf{0,75}} + \mathbf{1,5} \cdot \mathbf{IQR}$,
mělo by být označeno za odlehlé pozorování.

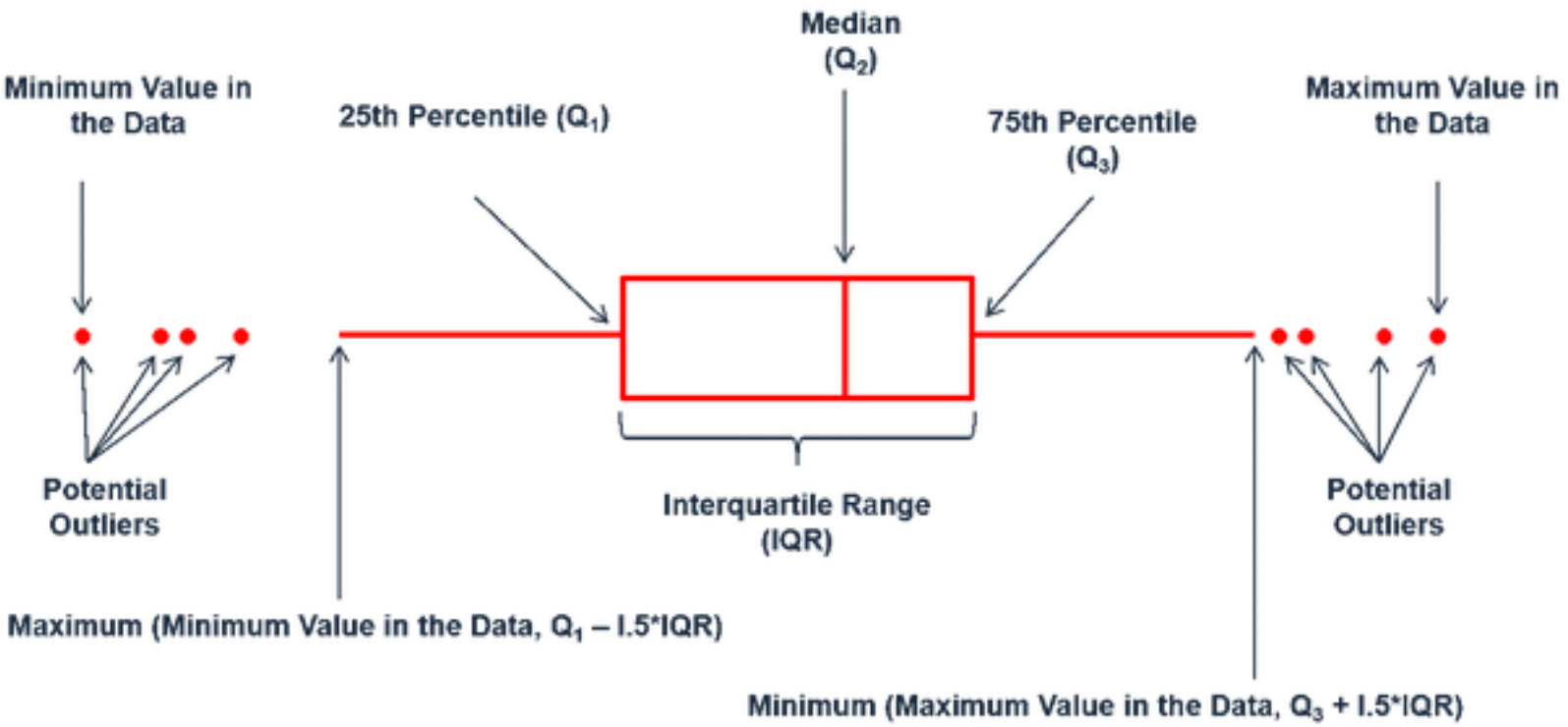
Vizualizace

- krabicové grafy (= boxploty)
- histogramy
- QQ-grafy

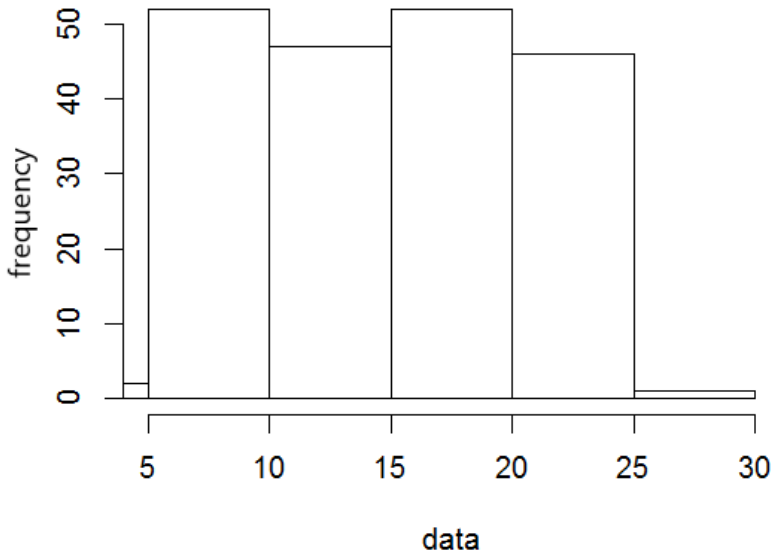
Krabicový graf

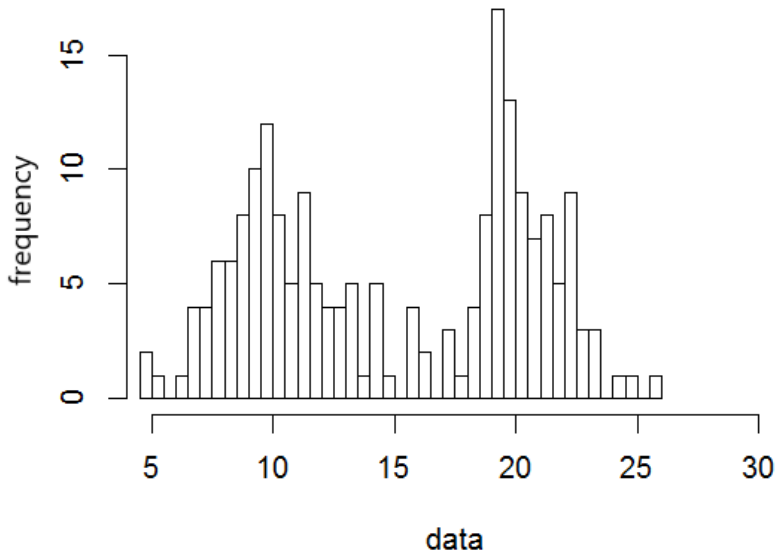


Krabicový graf

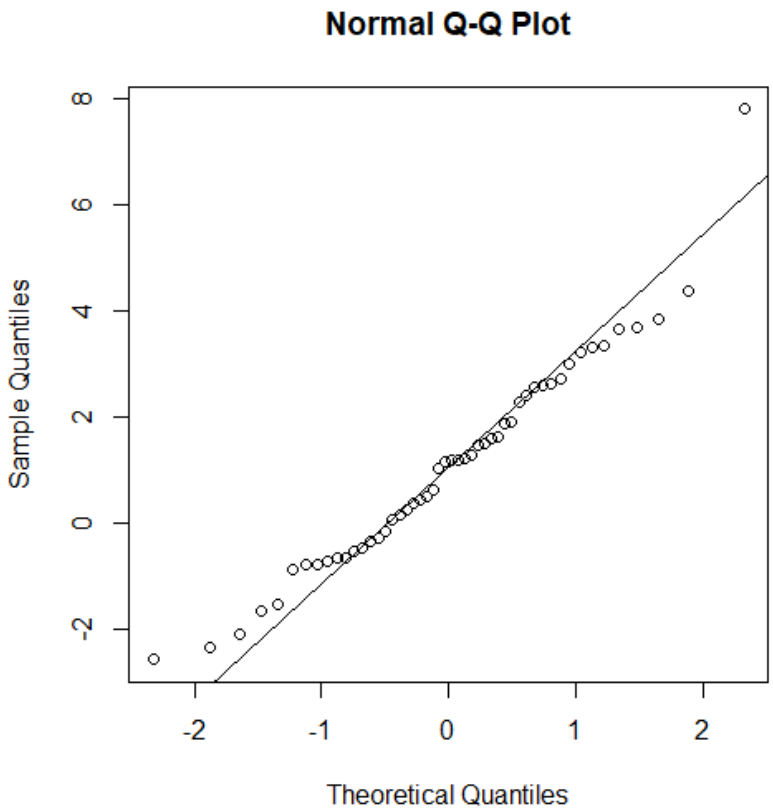


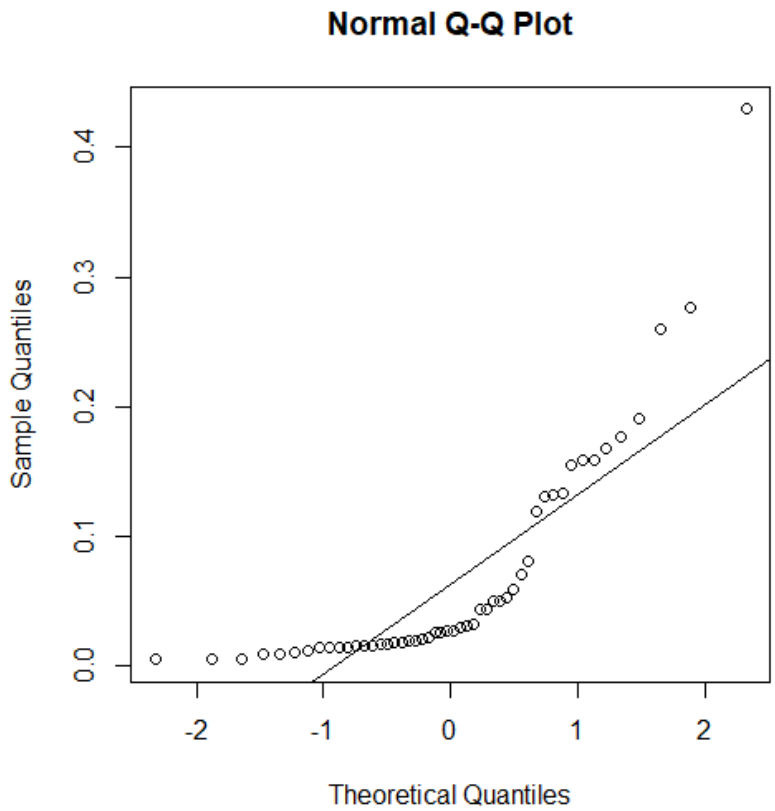
Histogram


frequency
0
10
20
30
40
50
5
10
15
20
25
30
data


frequency
0
5
10
15
5
10
15
20
25
30
data

QQ-graf


Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
-2
0
2
4
6
8
-2
-1
0
1
2


Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
0.0
0.1
0.2
0.3
0.4
-2
-1
0
1
2

Pravidla pro zaokrouhlování

Jak zaokrouhlit míry polohy? Jak zaokrouhlit směrodatnou odchylku?

BMI			
rozsah	průměr	medián	sm. odchylka
30	25,337	24,165	5,147

Pravidla pro zaokrouhlování

Zaokrouhlíme nahoru směrodatnou odchylku na určitý počet platných cifer podle pravidla (viz níže).

rozsah výběru	$\langle 2; 10\rangle$	$(10; 30\rangle$	$(30; 2000\rangle$	$(2000; \ldots$
počet platných cifer	1	2	3	4

BMI			
rozsah	průměr	medián	sm. odchylka
30	25,337	24,165	5,147

Pravidla pro zaokrouhlování

Zaokrouhlíme nahoru směrodatnou odchylku na určitý počet platných cifer podle pravidla (viz níže).

rozsah výběru	$\langle 2; 10\rangle$	$(10; 30\rangle$	$(30; 2000\rangle$	$(2000; \ldots$
počet platných cifer	1	2	3	4

BMI			
rozsah	průměr	medián	sm. odchylka
30	25,337	24,165	**5,2**

Pravidla pro zaokrouhlování

Zaokrouhlíme nahoru směrodatnou odchylku na určitý počet platných cifer podle pravidla a **míry polohy jsou zaokrouhleny na stejný počet desetinných míst, resp. na stejný řád**.

BMI			
rozsah	průměr	medián	sm. odchylka
30	25,337	24,165	**5,2**

Pravidla pro zaokrouhlování

Zaokrouhlíme nahoru směrodatnou odchylku na určitý počet platných cifer podle pravidla a **míry polohy jsou zaokrouhleny na stejný počet desetinných míst, resp. na stejný řád**.

BMI			
rozsah	průměr	medián	sm. odchylka
30	**25,3**	**24,2**	**5,2**

Pravidla pro zaokrouhlování

Zaokrouhlíme nahoru směrodatnou odchylku na určitý počet platných cifer podle pravidla a **míry polohy jsou zaokrouhleny na stejný počet desetinných míst, resp. na stejný řád**.

BMI			
rozsah	průměr	medián	sm. odchylka
30	**25,3**	**24,2**	**5,2**

Toto pravidlo lze aplikovat na všechny míry polohy s výjimkou minima a maxima, kde preferujeme hodnoty přímo z datového souboru.