MÁME DATA – A CO DÁL?

(1. ČÁST)

Martina Litschmannová, Adéla Vrtková

Obsah:

- Co je to statistika?
- Jak provést statistické šetření?
- Jak zapsat výsledky šetření?
  (datová matice vs. jiné formy zápisu)
- Exploratorní (popisná) analýza kategoriálních dat

Co je to statistika?

Google – 47.10^6 odkazů (čeština), $1{,}4.10^9$ odkazů (angličtina)

- Uspořádaný datový soubor (statistika přístupů na web. stránky, statistika střel na branku, statistika nehodovosti, ekonomické statistiky, …)
  - ✓ Český statistický úřad, Real Time Statistics Project

- Teoretická disciplína, která se zabývá metodami sběru a analýzy dat (matematická statistika vs. aplikovaná statistika)

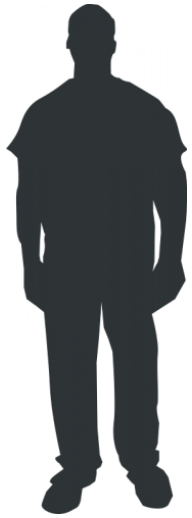
- Číselný údaj „syntetizující“ vlastnosti datových souborů (četnost, průměr, rozptyl, …)

Co vypovídá statistika o jednotlivci?

Lukáš Pavlásek
(jednotlivec)

skaut

tanečník



občan ČR

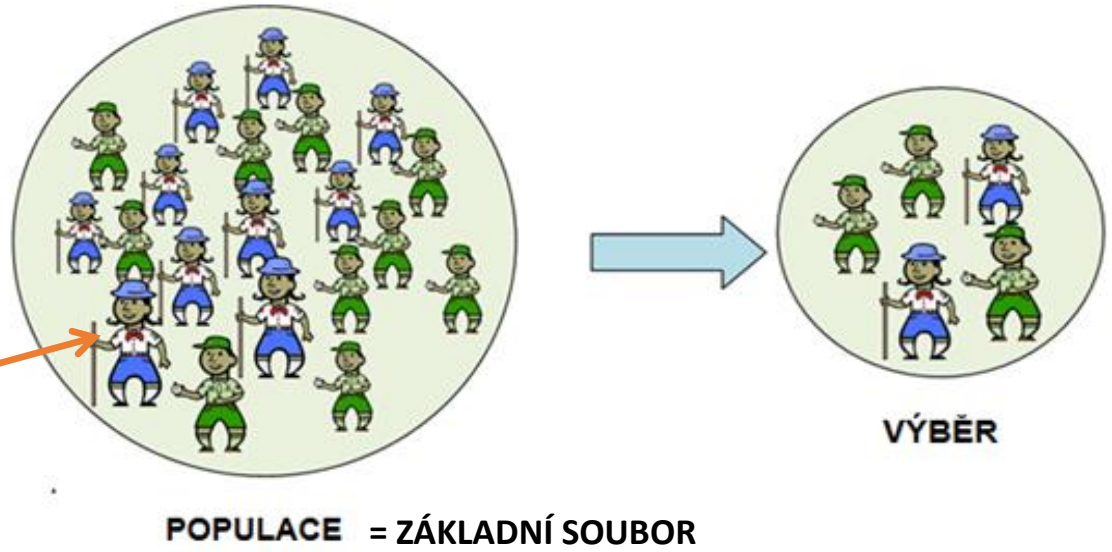
- Statistika nezkoumá jednotlivce jako individualitu, ale jako anonymního nositele některého znaku (činnosti, vlastnosti).
- Statistika je nauka o hromadných jevech.

Jak provést statistické šetření?

úplné šetření

výběrové šetření

statistická jednotka



REPREZENTATIVNÍ výběr

statistické znaky – údaje, které u statistických znaků sledujeme (např. váha, výška, IQ, …)

Jak zapsat výsledky statistického šetření?

Datová matice (standardní datový formát)

ID	kapacita po 100 cyklech	výrobce
1	1780.4	A
2	1751.4	A
3	1743.5	B
4	1727.4	B
5	1728.8	C
6	1767.5	C
7	1838.7	D
8	1734.1	A
9	1688.8	D

- Každý **řádek** matice obsahuje **údaje o jedné statistické jednotce**.
- **V prvním sloupci** (nebo jako popisky řádků) se obvykle uvádí **identifikační číslo** statistické jednotky (důležité pro jednoznačné spárování s konkrétní statistickou jednotkou, zejména při poskytování anonymizovaných dat zpracovateli).

Jak zapsat výsledky statistického šetření?

Jiná forma zápisu

Kapacita po 100 cyklech			
Výrobce A	**Výrobce B**	**Výrobce C**	**Výrobce D**
1780,4	1654,2	1663,3	1668,4
1751,4	1663,1	1641,1	1641,9
1743,5	1633,3	1621,5	1620
1727,4	1642,2	1610,7	1685,8
1728,8	1656,7		1610,5

Nevýhody:

- Obtížnější analýza pomocí statistického software.
- Chybí jednoznačná identifikace příslušných statistických jednotek.

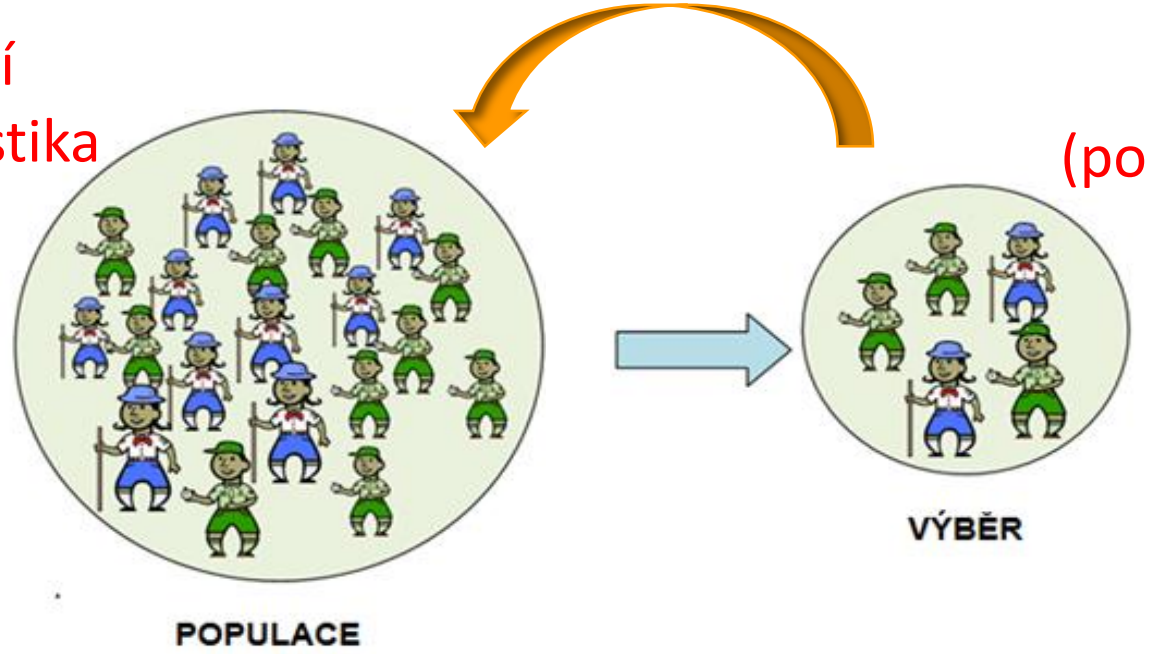
Před vlastní analýzou je velmi vhodné převést data do datové matice.

Jak analyzovat data?

Statistická indukce

Exploratorní
(popisná) statistika

Exploratorní
(popisná) statistika



Exploratorní analýza dat

Grafická prezentace a uspořádání dat do názornější formy a jejich popis několika málo hodnotami, které by obsahovaly co největší množství informací obsažených v původním souboru.

Typy statistických znaků (proměnných)

Nominální
- varianty jsou ve formátu text nebo číselný kód
- o každých dvou variantách lze říci, zda jsou různé
- např. škola, fakulta, obor, výrobce, …
- Další dělení: dichotomické (alternativní), vícekategoriální (množné)

Ordinální (pořadová)
- varianty jsou ve fomátu text, datum nebo číslo
- u každých dvou variant lze stanovit jejich pořadí
- např. úroveň vzdělání, známka (A, B, …, E), úroveň spokojenosti, …

Kvalitativní

Intervalové (rozdílové)
- varianty jsou v číselném formátu
- u každých dvou variant lze určit jejich pořadí a rozdíl
- např. teplota ve °C, chyba měření, …

Poměrové
- varianty jsou v číselném formátu (pouze kladná čísla + nulový bod)
- u každých dvou variant lze určit jejich pořadí, rozdíl a podíl (poměr)
- např. teplota v K, velikost chyby měření, …

Kvantitativní
(numerické, kardinální)
Další dělení: diskrétní, spojité

EDA pro kvalitativní proměnnou

Číselné charakteristiky

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:	$n_1 + n_2 + \cdots + n_k = n$	1

+ Modus (název nejčetnější varianty)

Číselné charakteristiky

TABULKA ROZDĚLENÍ ČETNOSTI		
Typ pasažéra	**Absolutní četnosti**	**Relativní četnosti (%)**
Muž	77	37,37864
Žena	85	41,26214
Dítě	44	21,35922
Celkem:	206	100,00000

1% … 2,06 osob
0,00001% … 0,0000206 osob
0,1% … 0,206 osob

Jak zaokrouhlovat relativní četnost?

Číselné charakteristiky

TABULKA ROZDĚLENÍ ČETNOSTI		
Typ pasažéra	**Absolutní četnosti**	**Relativní četnosti (%)**
Muž	77	37,4
Žena	85	41,3
Dítě	44	21,4
Celkem:	206	100,1

POZOR
na zaokrouhlovací chybu!

Číselné charakteristiky

TABULKA ROZDĚLENÍ ČETNOSTI		
Typ pasažéra	**Absolutní četnosti**	**Relativní četnosti (%)**
Muž	77	37,4
Žena	85	41,3
Dítě	44	21,3
Celkem:	206	100,0

Dopočet do 100%!

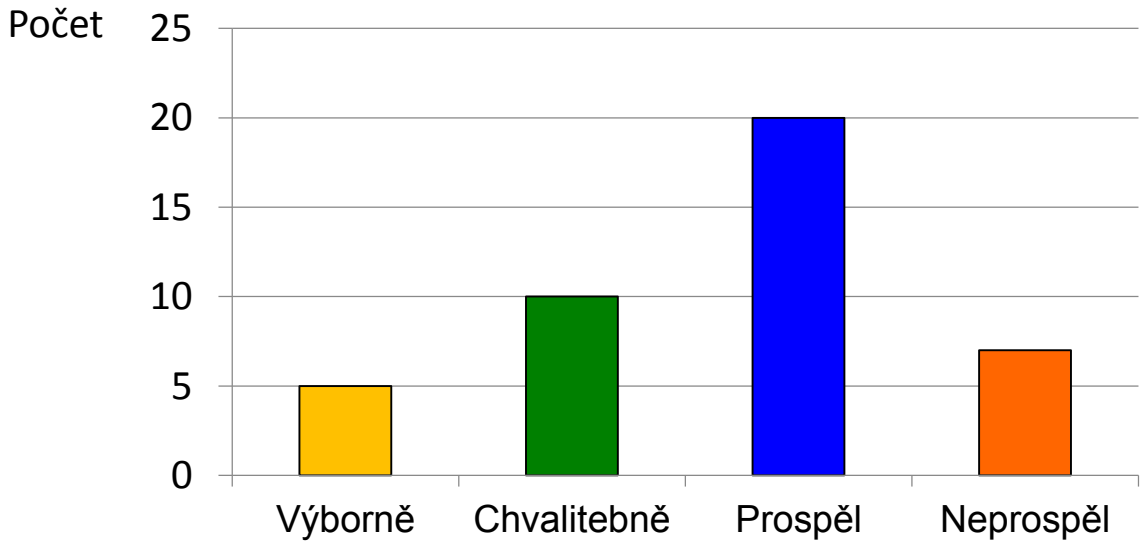
Číselné charakteristiky

TABULKA ROZDĚLENÍ ČETNOSTI		
Typ pasažéra	**Absolutní četnosti**	**Relativní četnosti (%)**
Muž	?	37,4
Žena	?	41,3
Dítě	?	21,3
Celkem:	206	100,0

Relativní četnosti uvádějme vždy pouze jako doplněk absolutních četností, nikoliv samostatně!

Grafické znázornění

A) Sloupcový graf (bar chart)

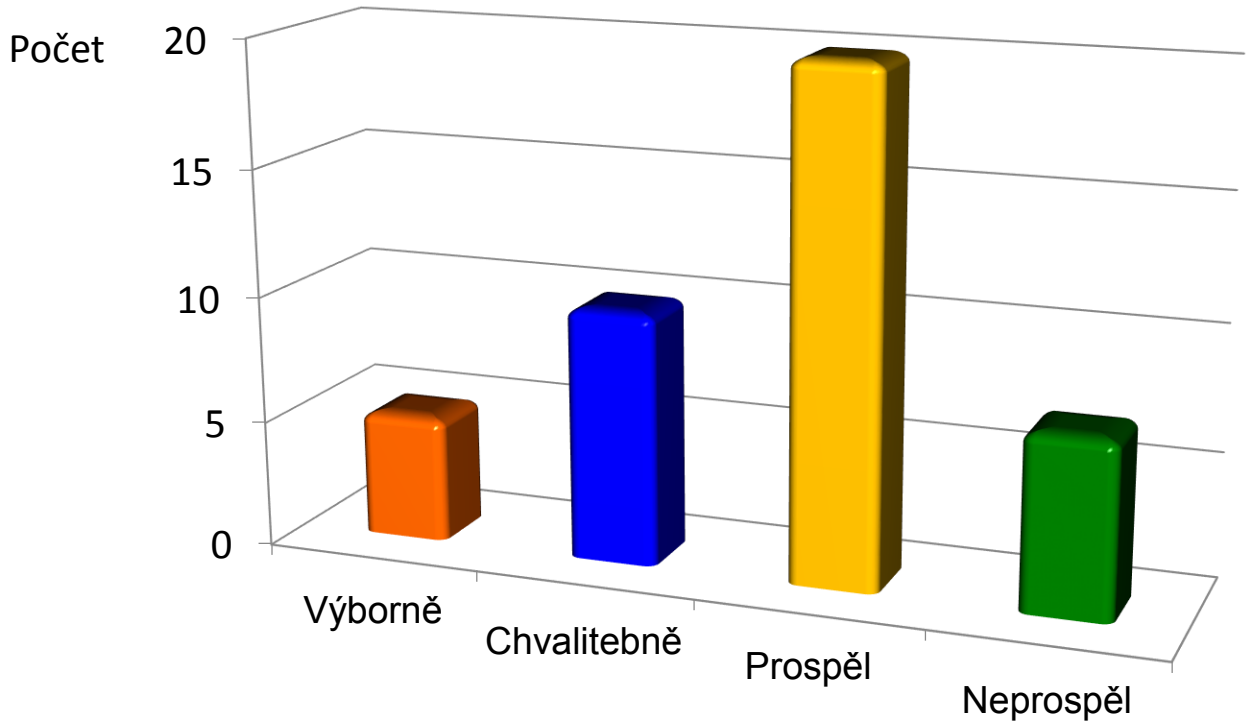


„…můžete vytvořit sloupcový graf a dodat mu zcela nový a přitažlivý vzhled“

http://office.microsoft.com/cs-cz/excel-help/prezentace-dat-ve-sloupcovem-grafu-HA010218663.aspx

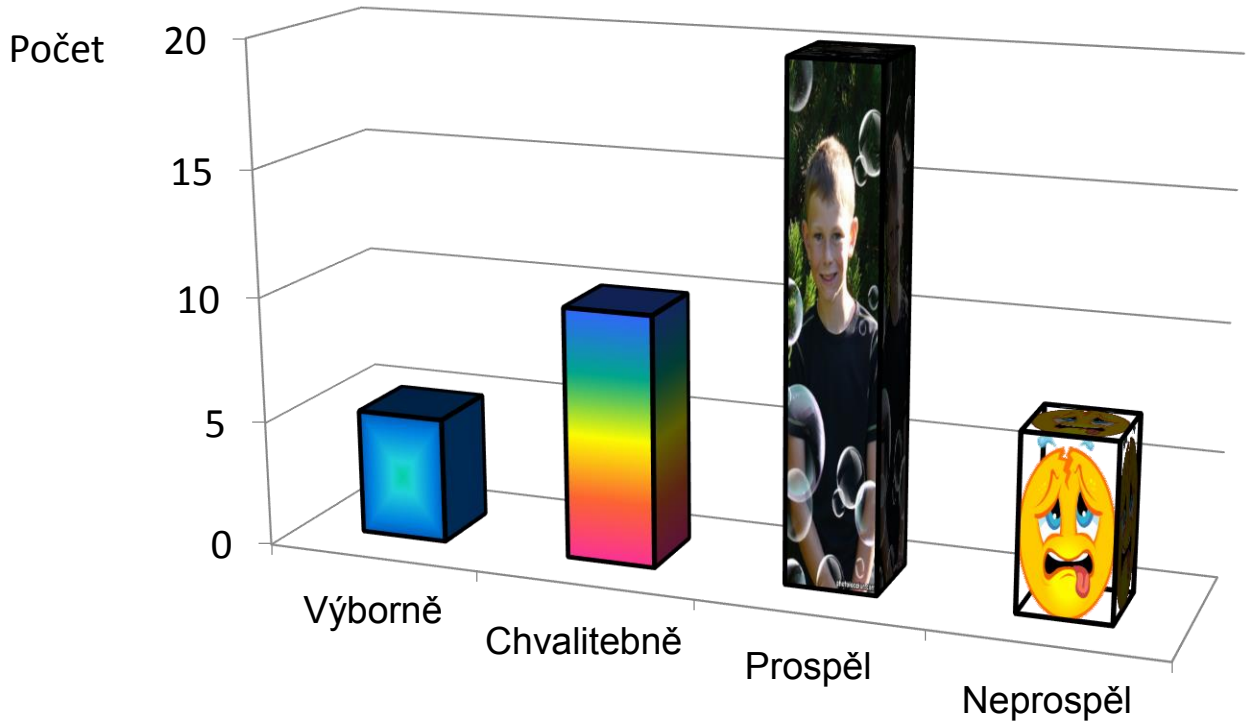
Grafické znázornění

A) Sloupcový graf (bar chart)



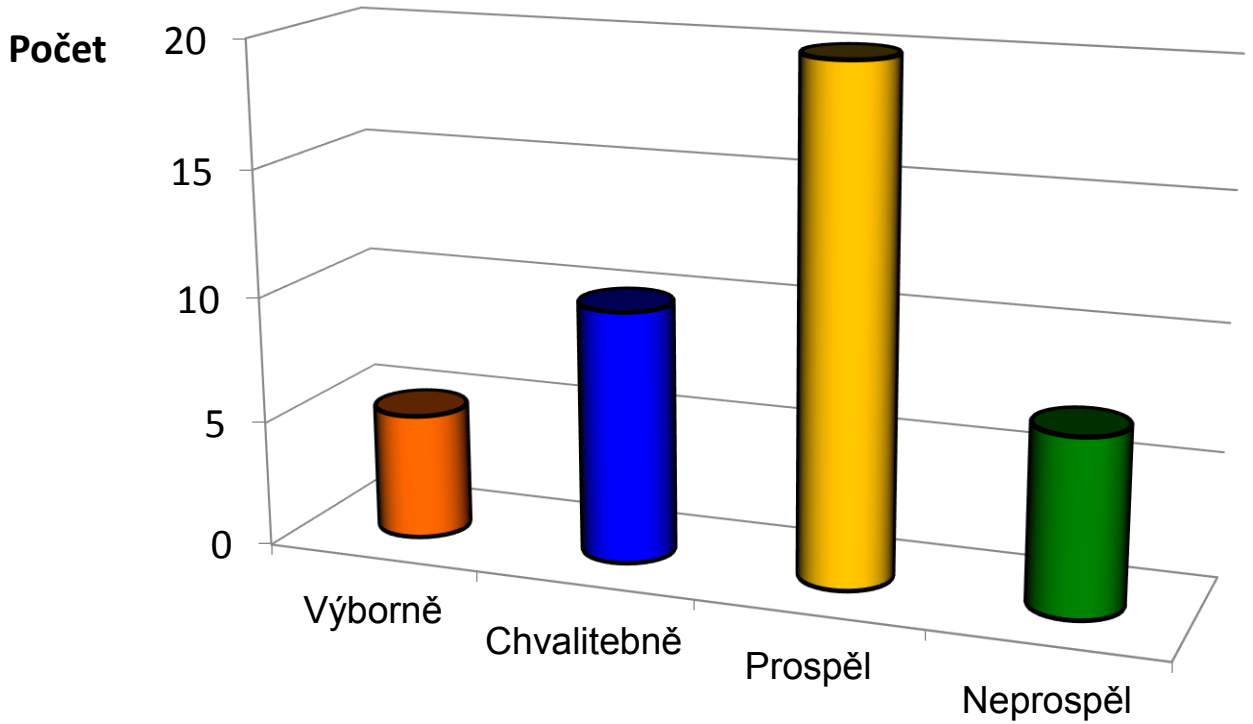
Grafické znázornění

A) Sloupcový graf (bar chart)



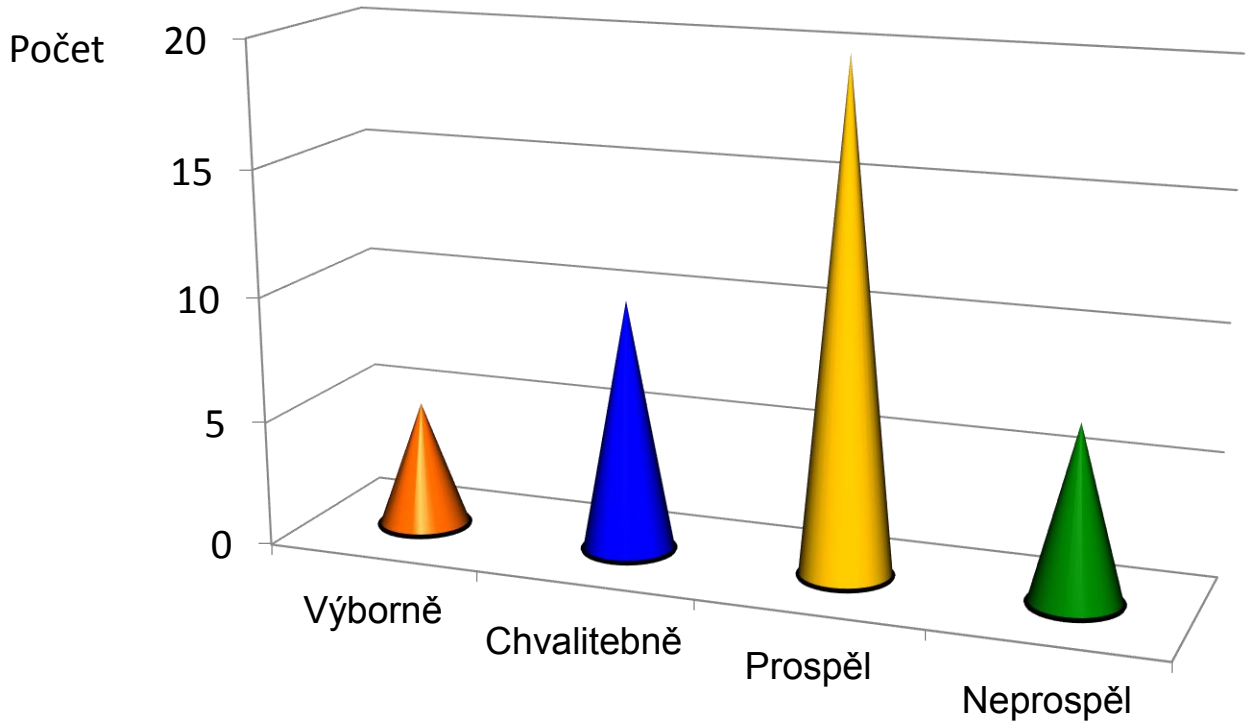
Grafické znázornění

A) Sloupcový graf (bar chart)



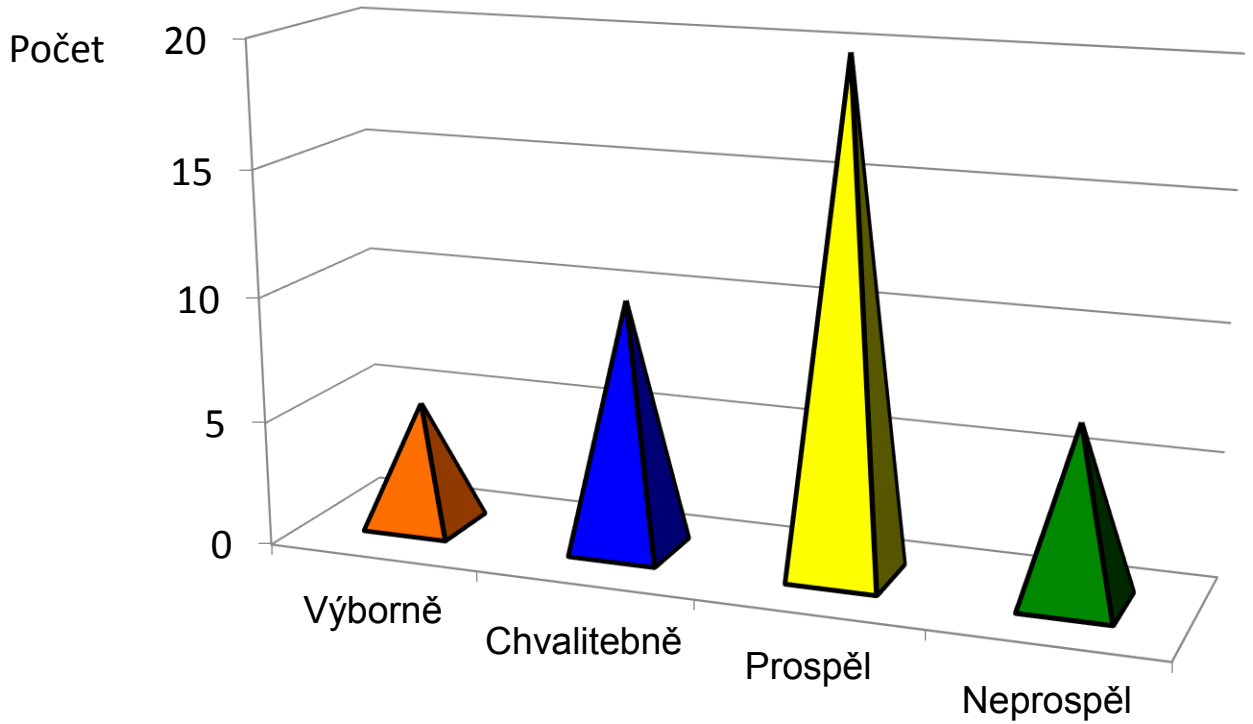
Grafické znázornění

A) Sloupcový graf (bar chart)



Grafické znázornění

A) Sloupcový graf (bar chart)



Grafické znázornění

A) Sloupcový graf (bar chart)

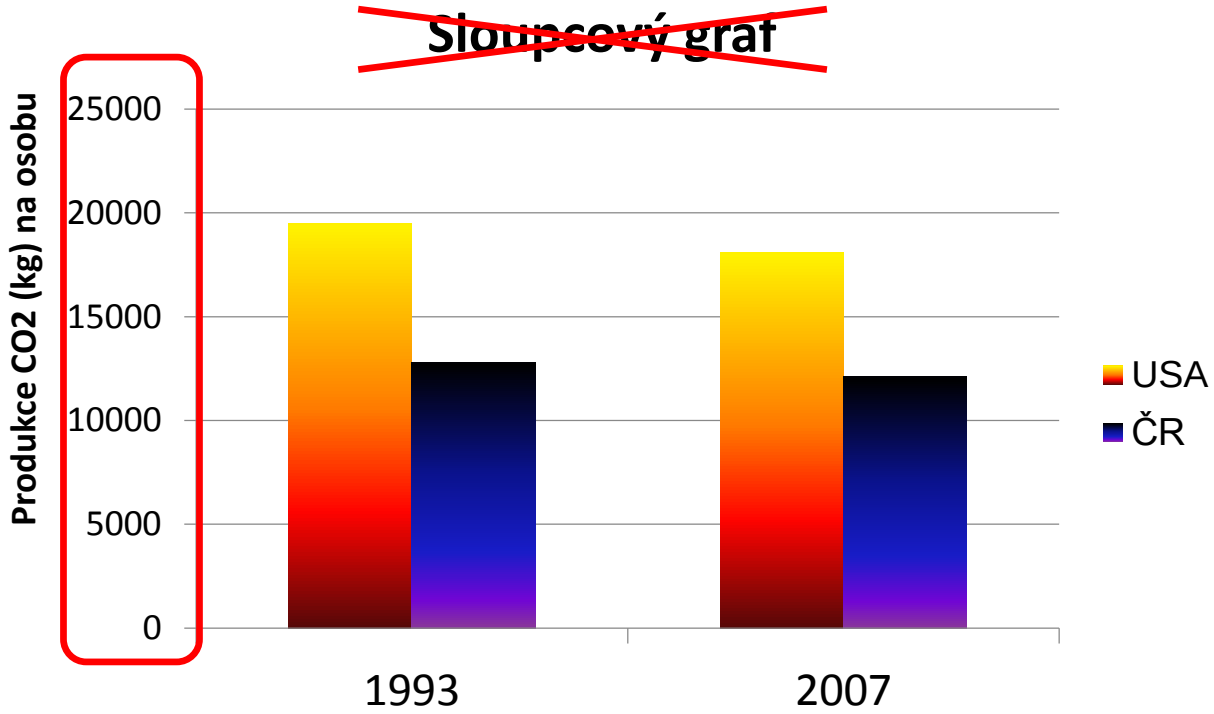
Na co si dát pozor?

- Subjektivně vnímáme plochu (objem), nikoliv výšku jednotlivých „sloupců“.

Grafické znázornění

A) Sloupcový graf (bar chart)

Na co si dát pozor?



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

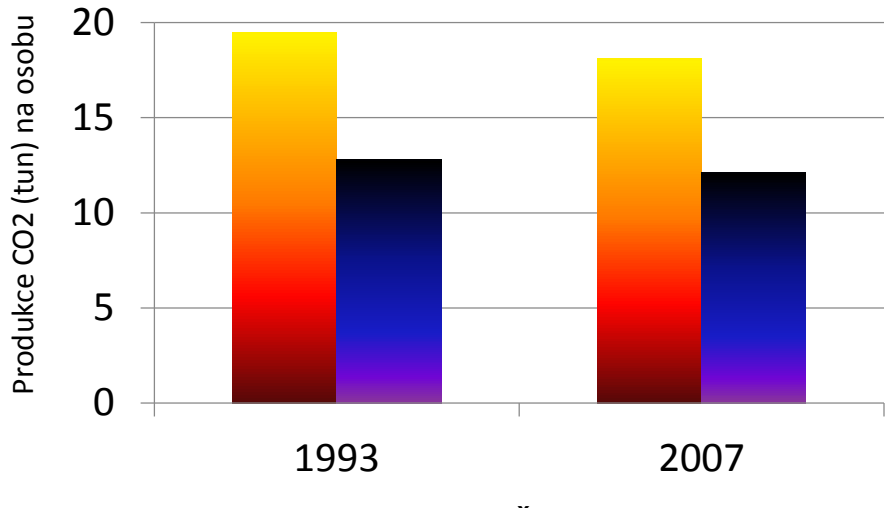
Grafické znázornění

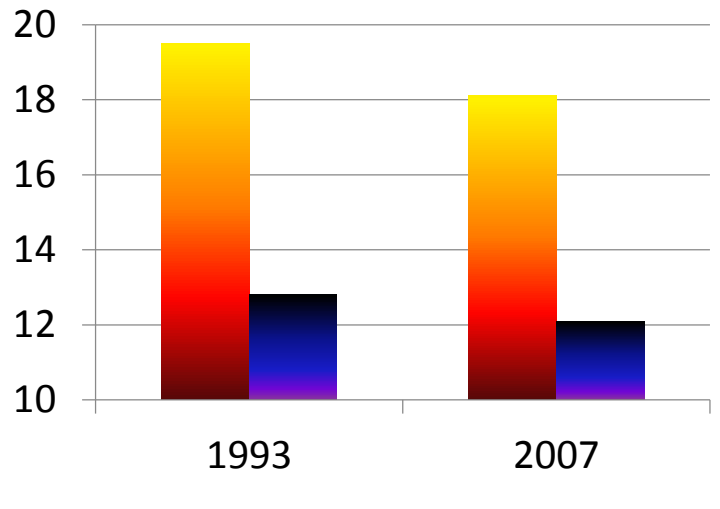
A) Sloupcový graf (bar chart)

Na co si dát pozor?

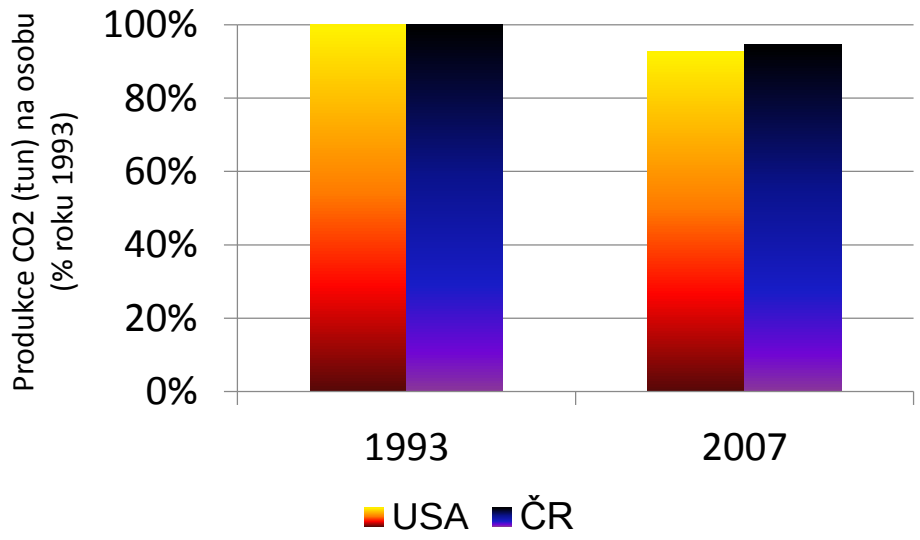
- Subjektivně vnímáme plochu (objem), nikoliv výšku jednotlivých „sloupců“.
- Nadbytečné názvy grafu, legendy, …
- Neefektivní nuly

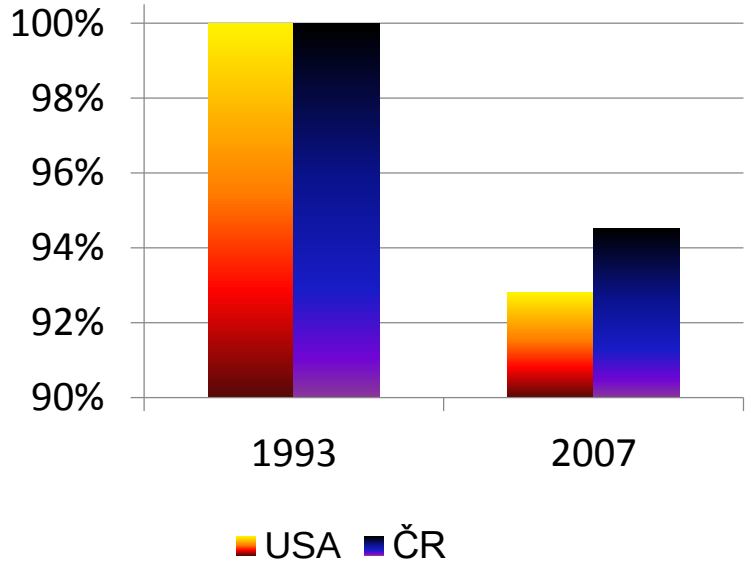
A na co ještě?

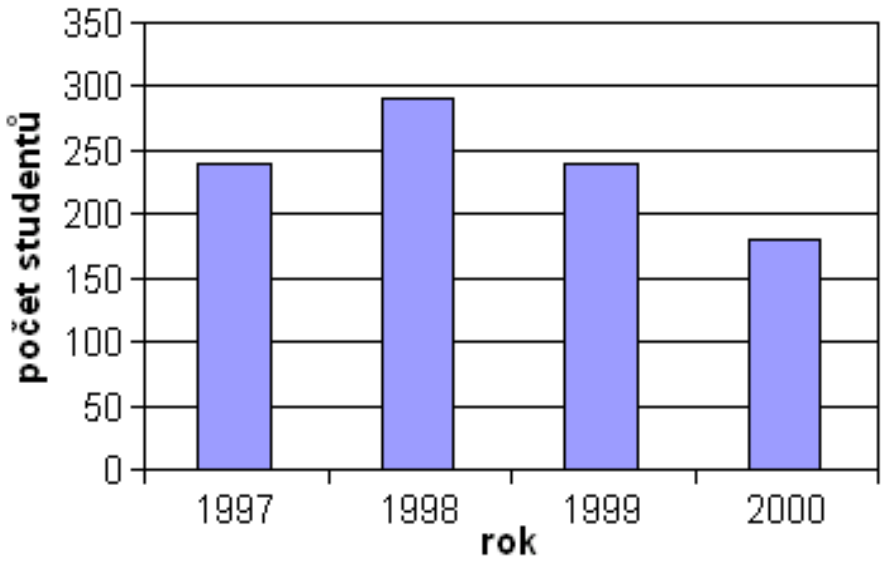




Který z grafů je „správný“?



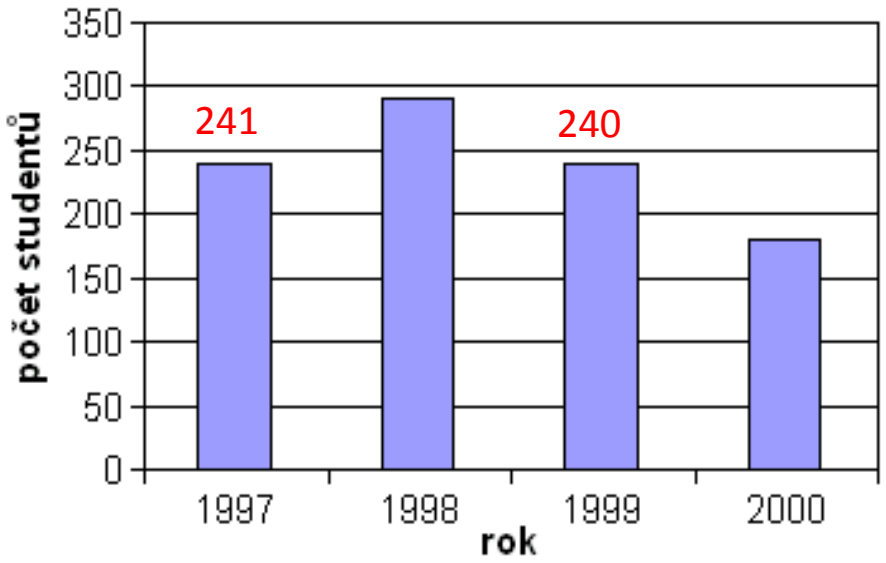




?

Určete pravdivost tvrzení:
V žádných dvou letech nebyl počet studentů stejný.

Zdroj: Testové příklady určené žákům 9. tříd.



Určete pravdivost tvrzení:
V žádných dvou letech nebyl počet studentů stejný.

Zdroj: Testové příklady určené žákům 9. tříd.

Grafické znázornění

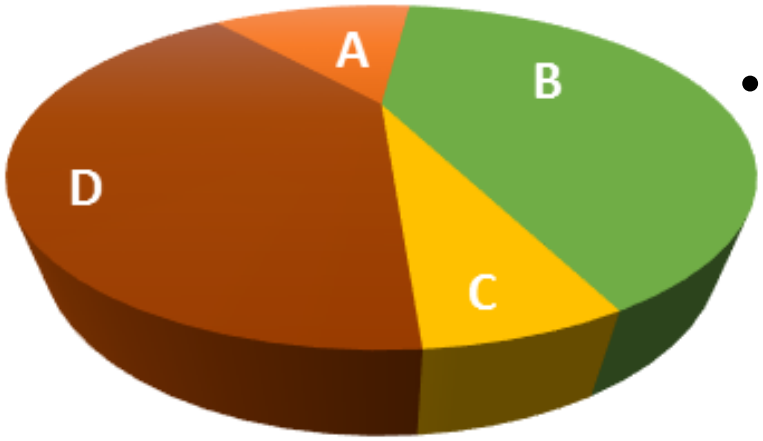
A) Sloupcový graf (bar chart)

Na co si dát pozor?

- Subjektivně vnímáme plochu (objem), nikoliv výšku jednotlivých „sloupců“.
- Nadbytečné názvy grafu, legendy, …
- Neefektivní nuly
- Informativní hodnota grafu

Grafické znázornění

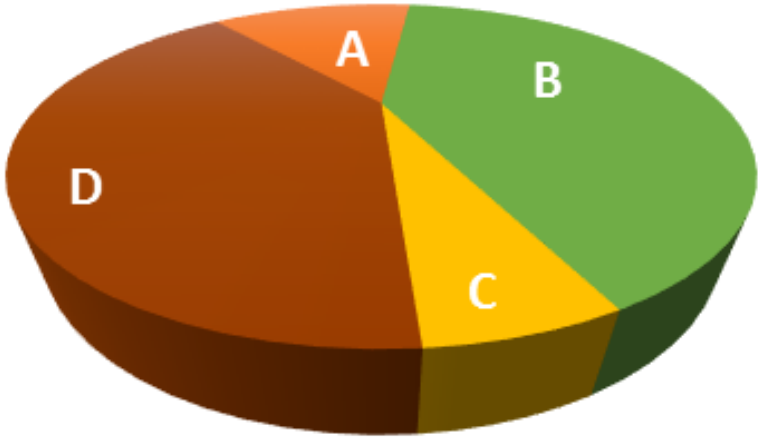
B) Výsečový graf – koláčový graf (pie chart)

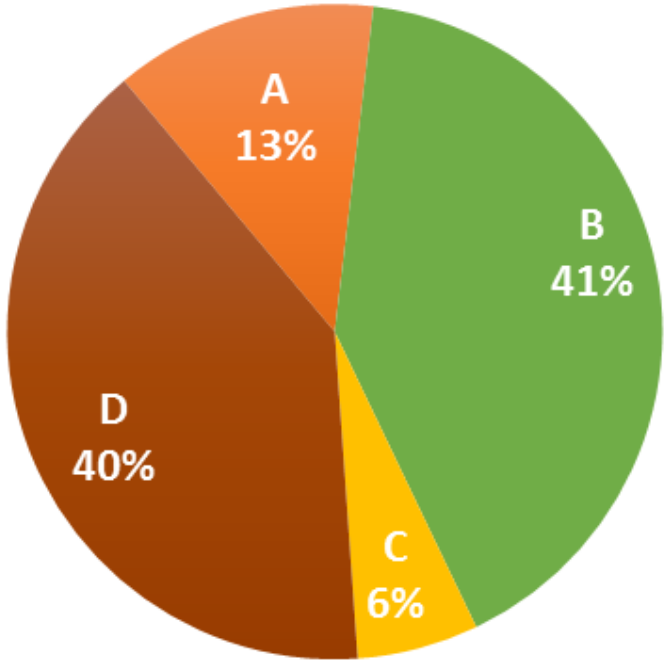


- Jaký je poměr mezi velikostí výsečí A a C?
- Jaký je poměr mezi velikostí výsečí B a D?

Grafické znázornění

B) Výsečový graf – koláčový graf (pie chart)





Grafické znázornění

B) Výsečový graf – koláčový graf (pie chart)

Na co si dát pozor?

Anketa

Souhlasíte s tím, že všichni akademičtí pracovníci VŠB – Technické univerzity Ostrava by měli povinně absolvovat kurz Analýza dat?

TAKHLE NE!!!

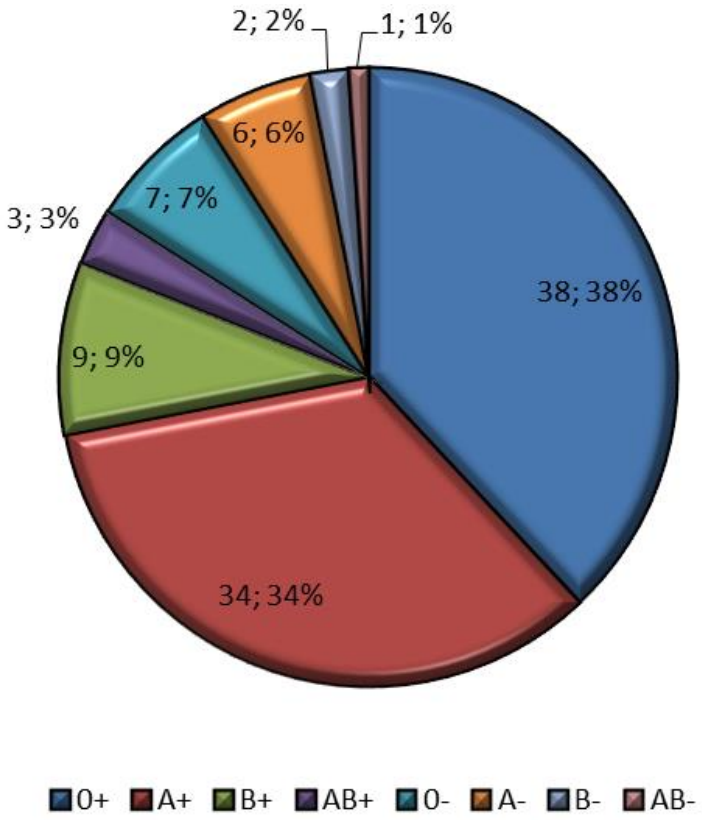
Grafické znázornění

B) Výsečový graf – koláčový graf (pie chart)

Na co si dát pozor?

- Neuvádění abs. četností, resp. celkového počtu respondentů v „blízkosti“ grafu
- Nadbytečné názvy grafu

Výskyt krevních skupin a Rh faktoru v USA



Krevní skupina	Rh faktor		Celkem
	Rh+	Rh-	
0	38	7	45
A	34	6	40
B	9	2	11
AB	3	1	4
Celkem	84	16	100

Procentuální zastoupení krevních skupin v populaci USA

Grafické znázornění

B) Výsečový graf – koláčový graf (pie chart)

Na co si dát pozor?

- Neuvádění abs. četností, resp. celkového počtu respondentů v „blízkosti“ grafu
- Nadbytečné názvy grafu, legendy, …
- Ne vždy je graf přehlednější než tabulka

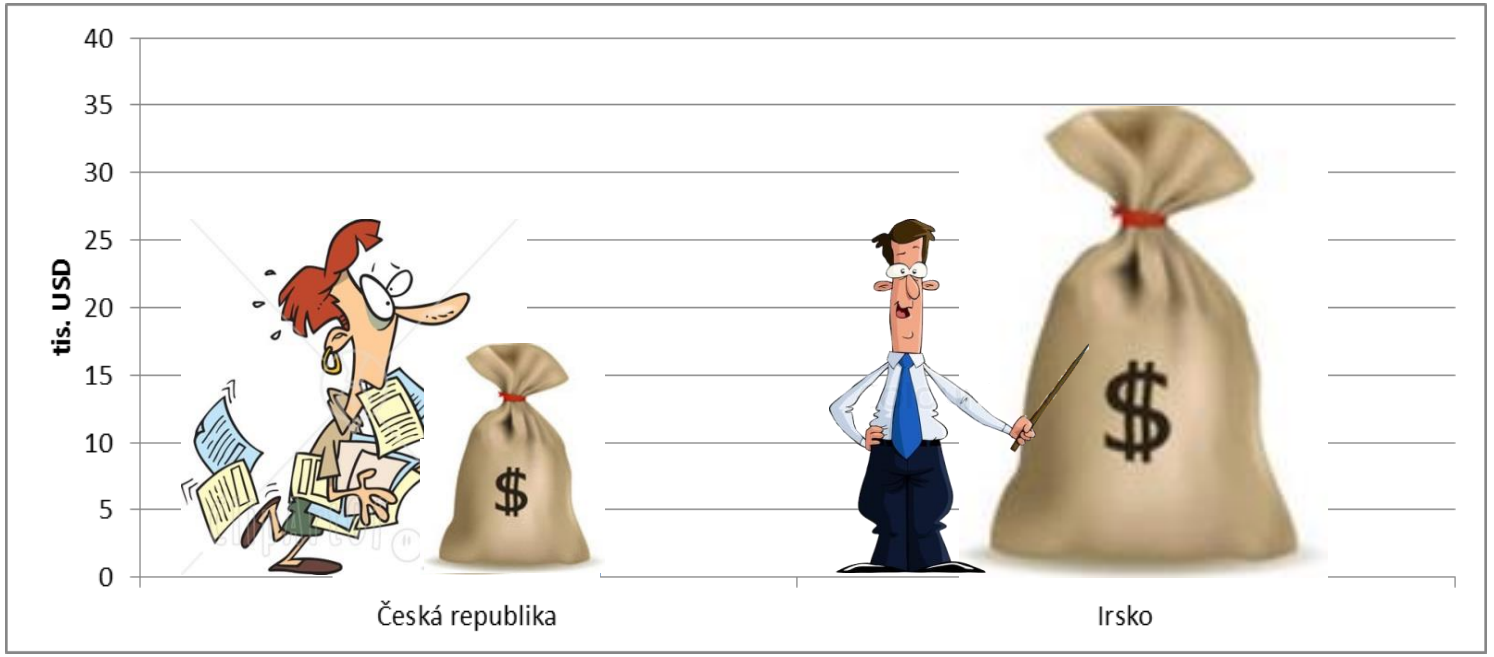
Grafické znázornění

A) Sloupcový graf (bar chart)

B) Výsečový graf – koláčový graf (pie chart)

C) Obrázkové grafy

Obrázkové grafy – užiteční pomocníci?



Srovnání průměrných ročních nástupních platů učitelů středních škol v ČR (17 244 $) a Irsku (34 604 $)

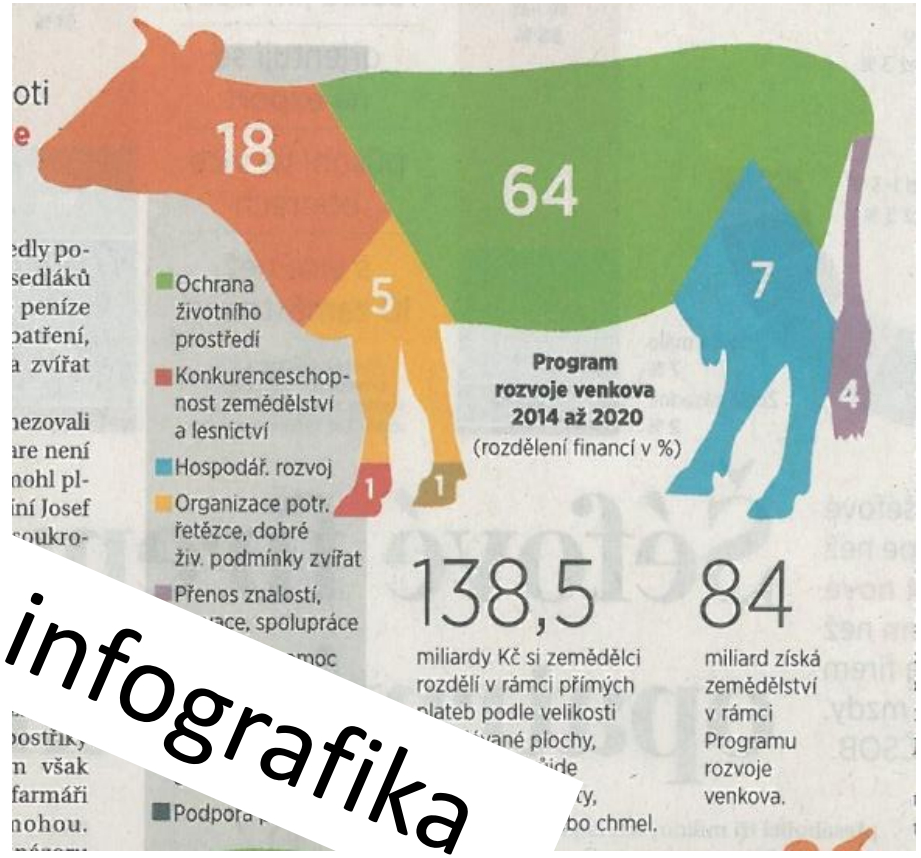
Obrázkové grafy – užiteční pomocníci?

Srovnání průměrných ročních nástupních platů učitelů středních škol

v ČR (17 244 $) a Irsku (34 604 $)

Několik praktických příkladů aneb „To přece bylo v novinách…“

Obrázkové grafy – užiteční pomocníci?

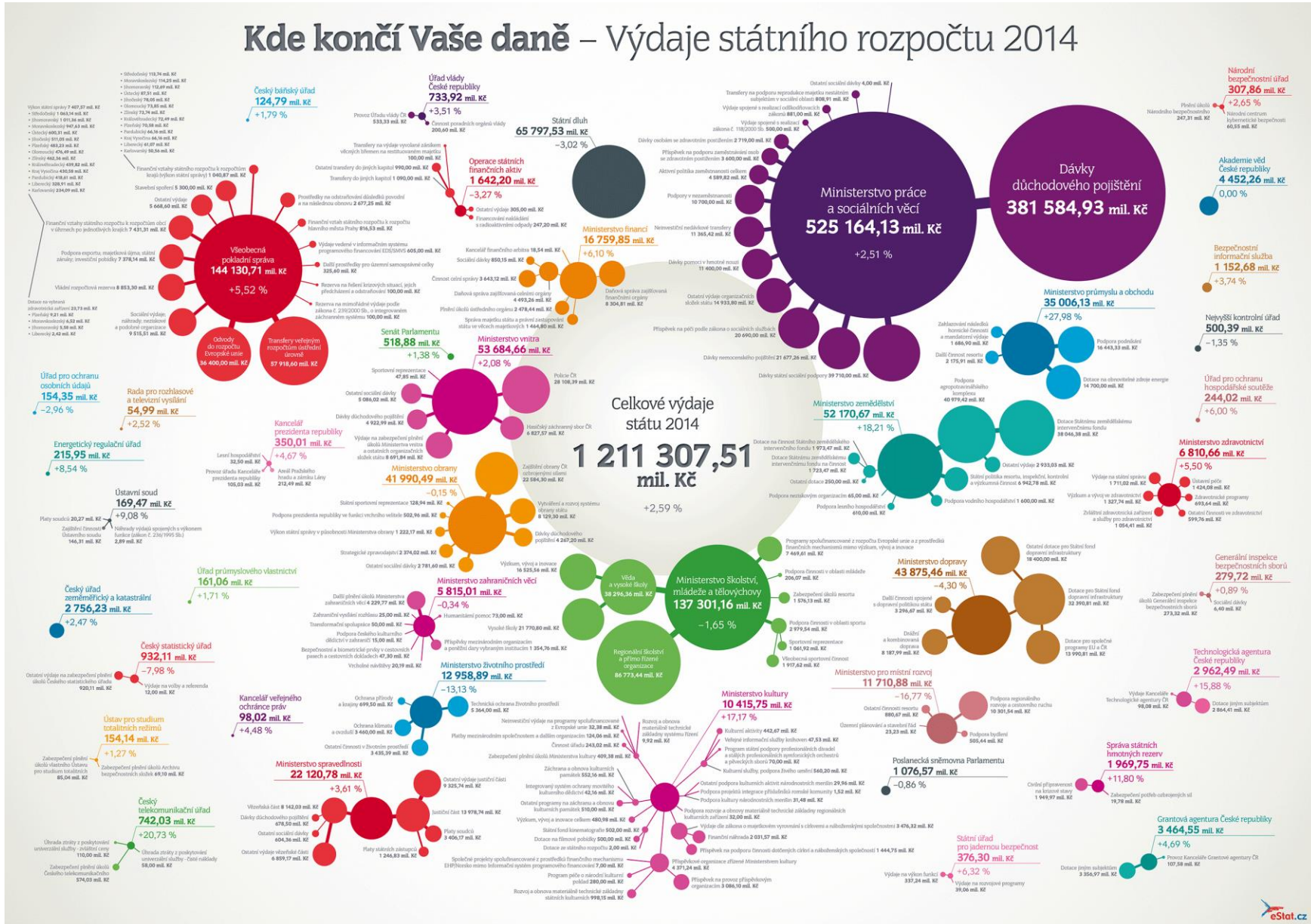


(Zdroj: Mf Dnes, 10. 7. 2014:

Zemědělci si rozdělí miliardy. Krávy a vepři se budou mít lépe.

„Úžasná infografika o výdajích státního rozpočtu České republiky v roce 2013“

Zdroj: http://www.estat.cz/zpravy/informace-k-projektum/kde-konci-vase-dane/



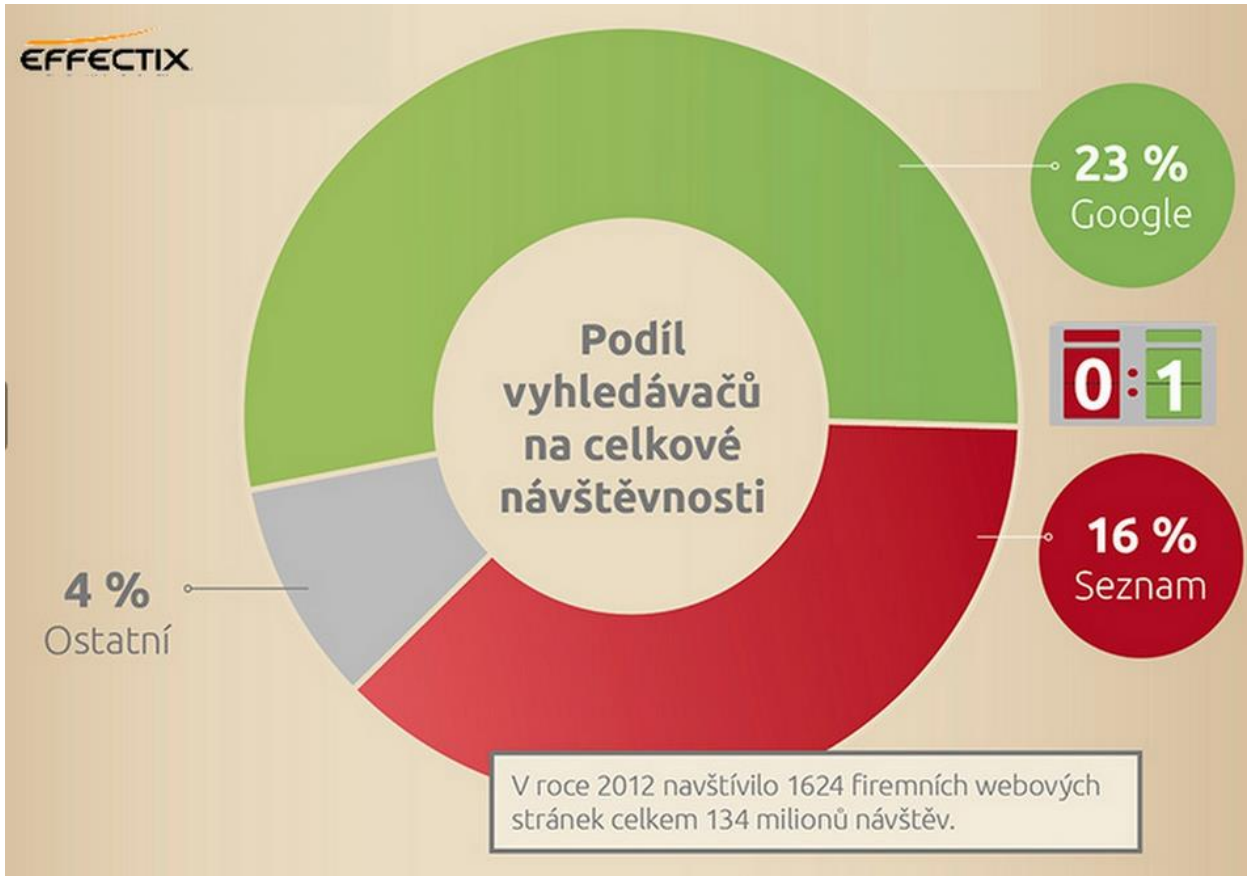
Zdroj: http://www.estat.cz/zpravy/informace-k-projektum/kde-konci-vase-dane/

Příklad s klobásou

Příklad s klobásou

2. místo: klobása z trhů na náměstí Míru (45 Kč)	
Obsah masa:	79,73 %
Obsah tuku:	24,94 %
Obsah bílkovin:	12,4 %
Obsah vody:	55,56 %
Obsah kolagenu:	2,6 %
Sójová bílkovina:	méně než 0,7 %

Souboj vyhledávačů

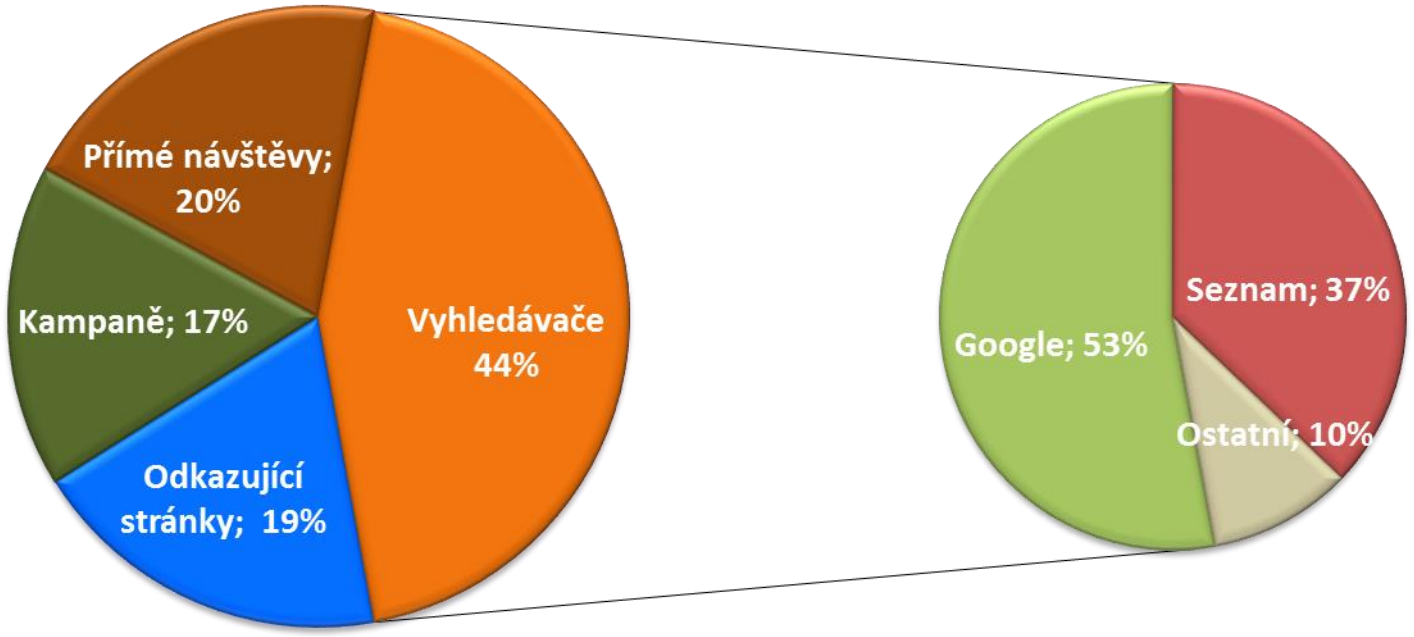


Zdroj: http://www.zive.cz/clanky/infografika-souboj-vyhledavacu-seznamcz-a-google/sc-3-a-167776/default.aspx

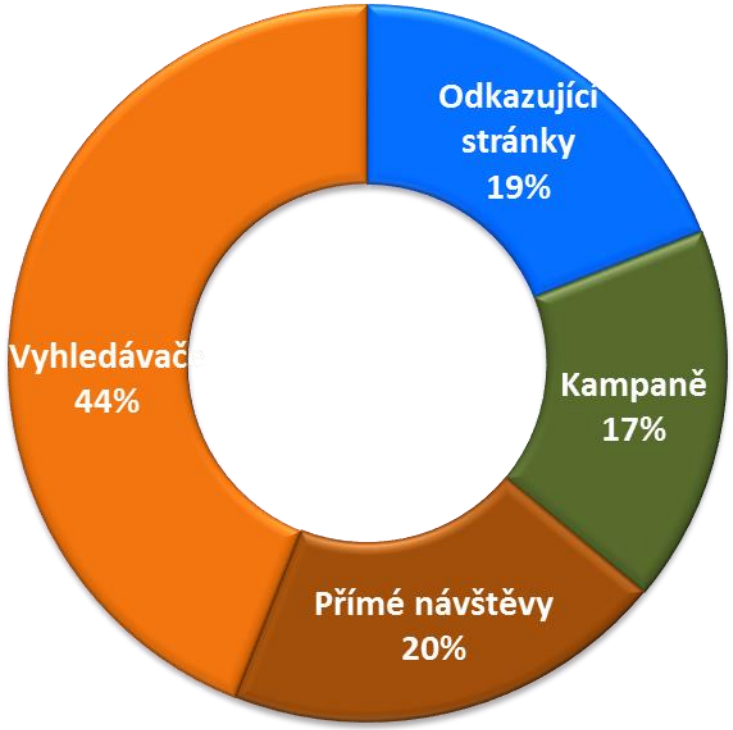
Souboj vyhledávačů

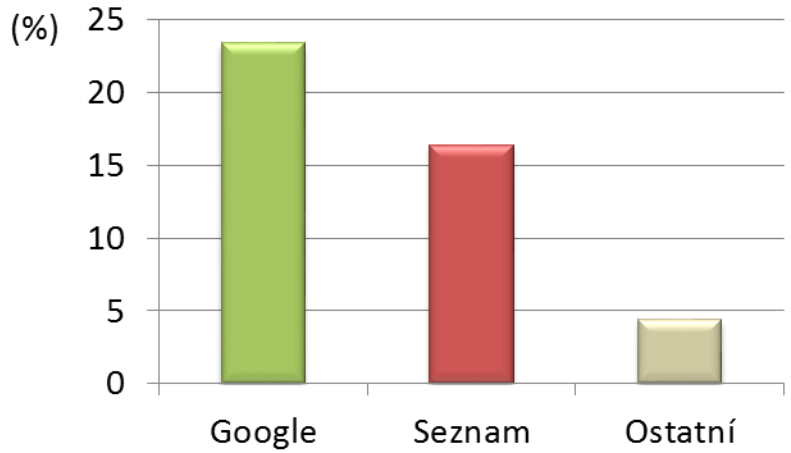
Zdroj: http://www.zive.cz/clanky/infografika-souboj-vyhledavacu-seznamcz-a-google/sc-3-a-167776/default.aspx

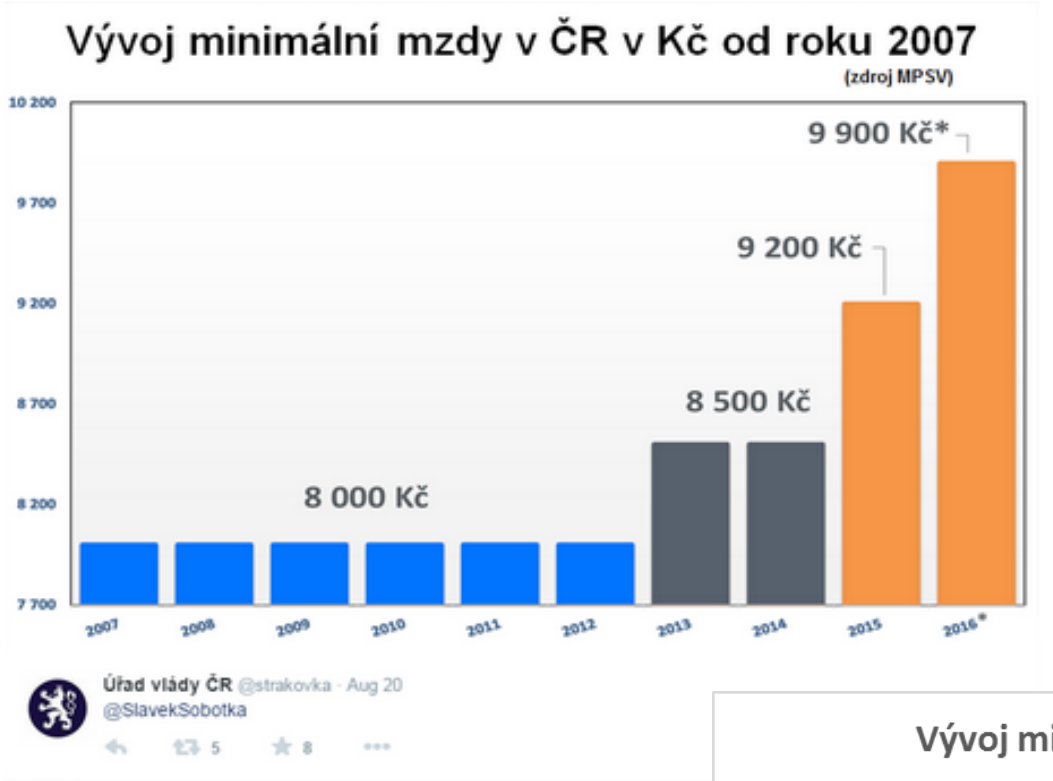
Jak výsledky šetření zobrazit správně?



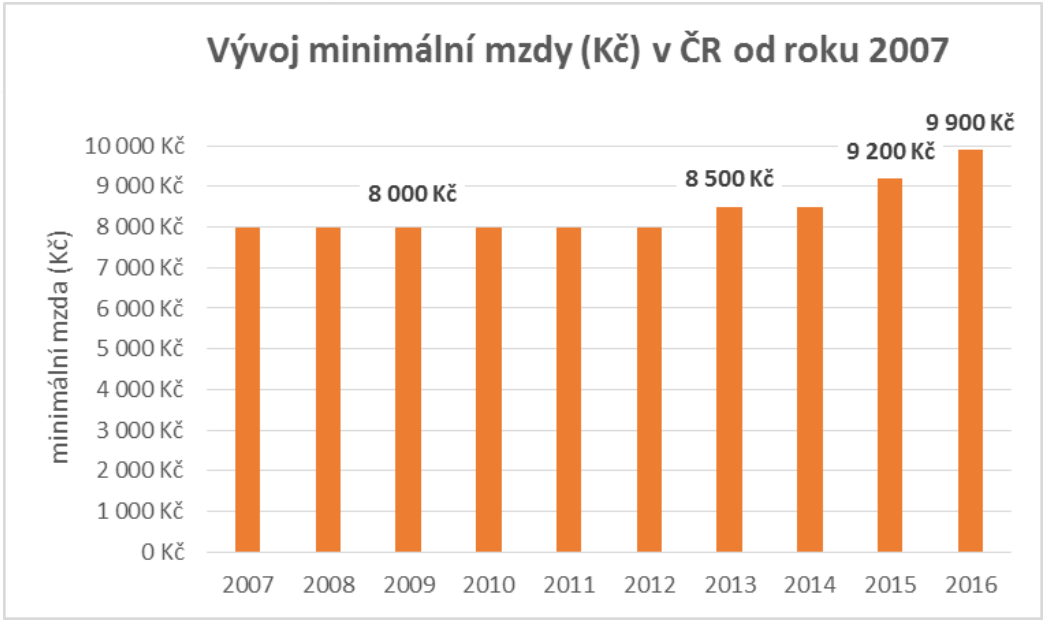
Jak výsledky šetření zobrazit správně?







***Zdroj**: Twitter @strakovka (20. srpna 2015)*



MÁME DATA – A CO DÁL?

(2. ČÁST)

Martina Litschmannová, Adéla Vrtková

Obsah:

- Exploratorní (popisná) analýza kvantitativních dat
  - Číselné charakteristiky (míry polohy, míry variability, míry šikmosti a špičatosti)
  - Odlehlá pozorování
  - Zaokrouhlování číselných charakteristik
  - Vizualizace kvantitativních dat

Typy statistických znaků (proměnných)

Nominální
- varianty jsou ve formátu text nebo číselný kód
- o každých dvou variantách lze říci, zda jsou různé
- např. škola, fakulta, obor, výrobce, …
- Další dělení: dichotomické (alternativní), vícekategoriální (množné)

Ordinální (pořadová)
- varianty jsou ve fomátu text, datum nebo číslo
- u každých dvou variant lze stanovit jejich pořadí
- např. úroveň vzdělání, známka (A, B, …, E), úroveň spokojenosti, …

Kvalitativní

Intervalové (rozdílové)
- varianty jsou v číselném formátu
- u každých dvou variant lze určit jejich pořadí a rozdíl
- např. teplota ve °C, chyba měření, …

Poměrové
- varianty jsou v číselném formátu (pouze kladná čísla + nulový bod)
- u každých dvou variant lze určit jejich pořadí, rozdíl a podíl (poměr)
- např. teplota v K, velikost chyby měření, …

Kvantitativní
(numerické, kardinální)
Další dělení: diskrétní, spojité

EDA pro kvantitativní proměnnou

Číselné charakteristiky

A) Míry polohy (úrovně)
B) Míry variability
C) Míry šikmosti a špičatosti

Míry polohy

- Odhadují skutečnou populační střední hodnotu na základě výběrového souboru.

- Patří mezi ně: výběrový aritmetický průměr, výběrový geometrický průměr, výběrový medián a modus.

- Dalšími mírami polohy, které se týkají popisu i polohy jiných hodnot než středních, jsou kvantily.

Ošidný průměr

Statistik,
který má hlavu v sauně a nohy v ledničce,
hovoří o příjemné průměrné teplotě.

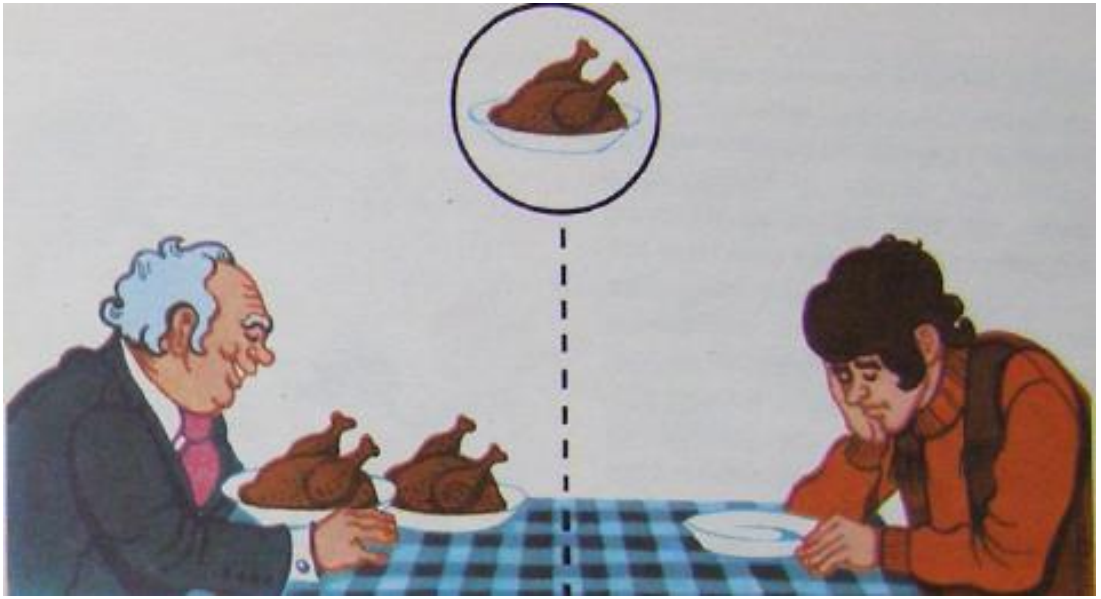
Autor neznámý

Aritmetický průměr

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

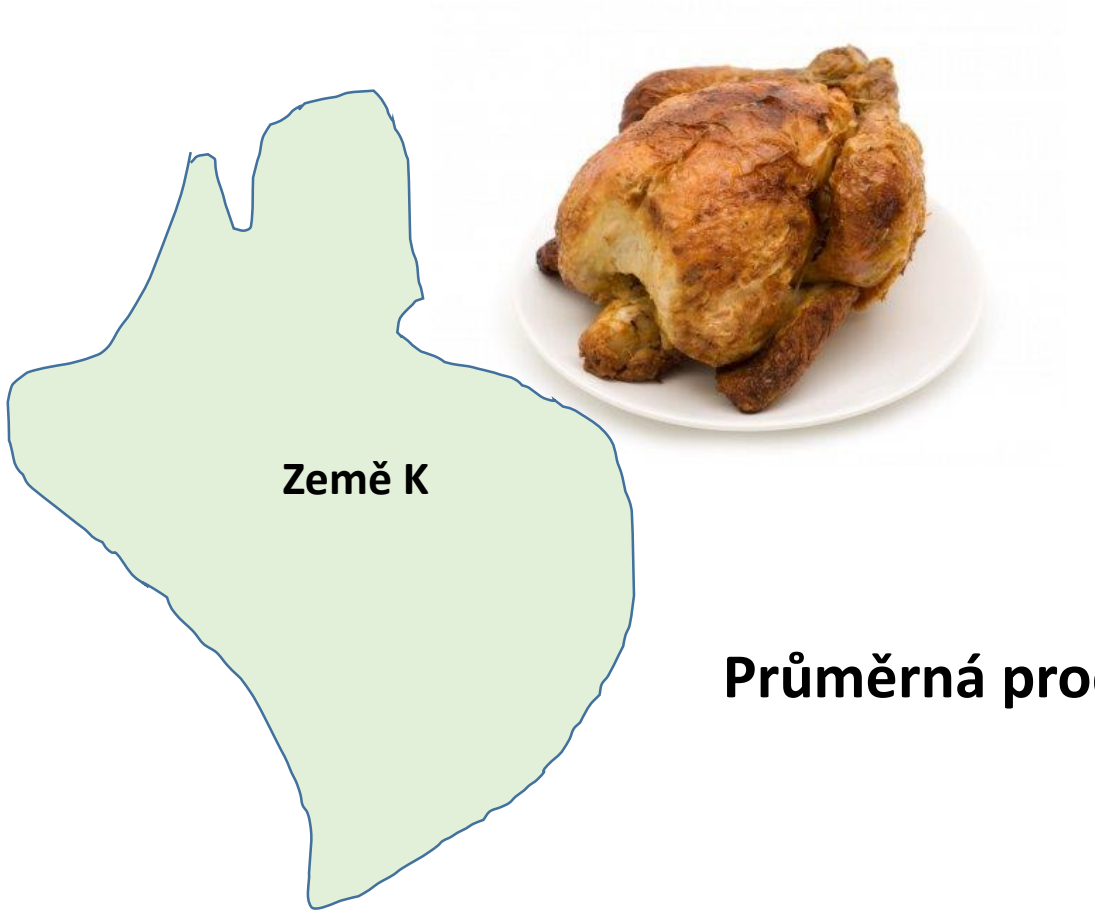
Pozor na ošidnost aritmetického průměru!

Ošidnost průměru



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

Ošidnost průměru



Průměrná produkce kuřat (na osobu):

1,0 (denně)

Ošidnost průměru

„Průměrná rodina má 2,2 dítěte.“

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

Průměr může nabývat hodnot, které nepatří do definičního oboru proměnné!

Ošidnost průměru

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

  $25 000 $27 000 $29 000
  $35 000 $37 000 $38 000

  Určete průměrný plat obyvatel této vesnice.

  ($31 830)

- Do vesnice se přistěhoval Bill Gates, jehož roční příjem je $40 000 000.

  $25 000 $27 000 $29 000
  $35 000 $37 000 $38 000 $40 000 000

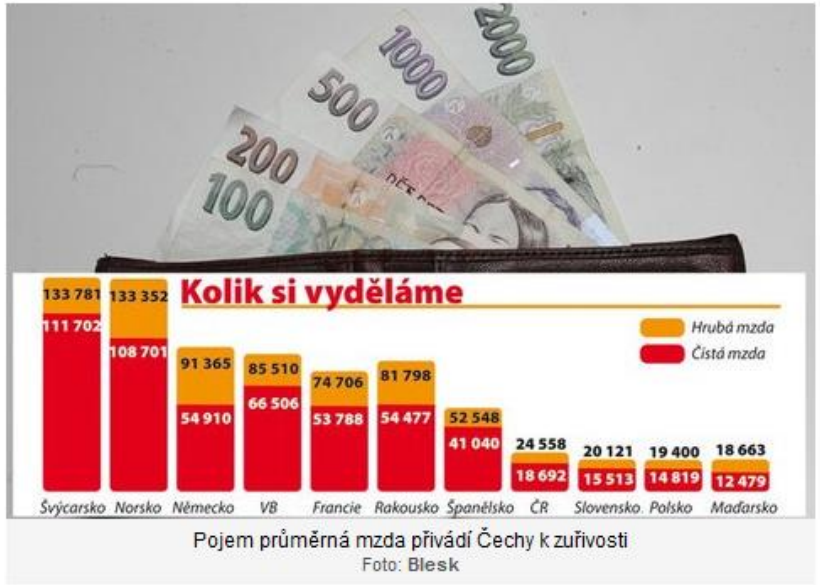
  Určete průměrný plat obyvatel této vesnice.

  ($5 741 571)

Průměr není rezistentní vůči odlehlým pozorováním!

Ošidnost průměru

Průměrná mzda přivádí Čechy k zuřivosti: Kdo z vás má 27 tisíc?



Pojem průměrná mzda přivádí Čechy k zuřivosti
Foto: Blesk

Zdroj: Blesk, 9.4.2013

Aritmetický průměr

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

Na co si dát pozor?

- Průměr není rezistentní vůči odlehlým pozorováním!
- Harmonický průměr (proměnné vyjadřující čas na jednotku výkonu, poměrná čísla)
- Geometrický průměr (tempa růstu)
  - Vážený průměr
- Průměrování dat na cirkulární škále
  Circular Statistics Toolbox

Výběrové kvantily

100p %-ní kvantil $\tilde{x}_p$

- odděluje 100p% menších hodnot od zbytku souboru

(100p% hodnot datového souboru je menších než toto číslo.)

Význačné výběrové kvantily

- Kvartily

  Dolní kvartil $\tilde{x}_{0,25}$

  Medián $\tilde{x}_{0,5}$

  Horní kvartil $\tilde{x}_{0,75}$

- Decily – $\tilde{x}_{0,1}; \tilde{x}_{0,2}; \ldots ; \tilde{x}_{0,9}$

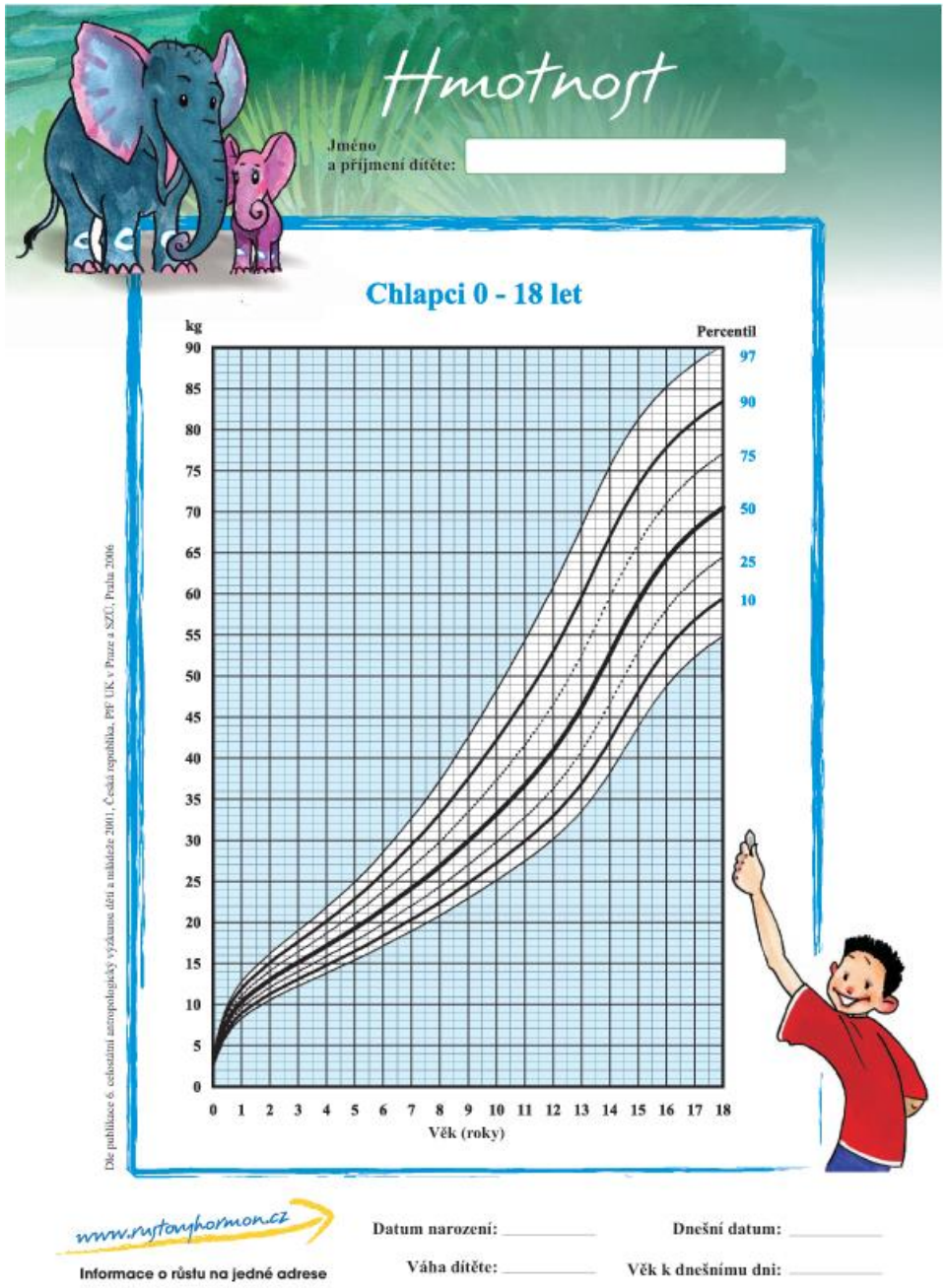
- Percentily – $\tilde{x}_{0,01}; \tilde{x}_{0,02}; \ldots; \tilde{x}_{0,03}$

- Minimum $\tilde{x}_{min}$ a Maximum $\tilde{x}_{max}$

Kde se s kvantily setkáme v praxi?

- vyhodnocení Národních srovnávacích zkoušek, …
- růstové grafy

Růstové grafy

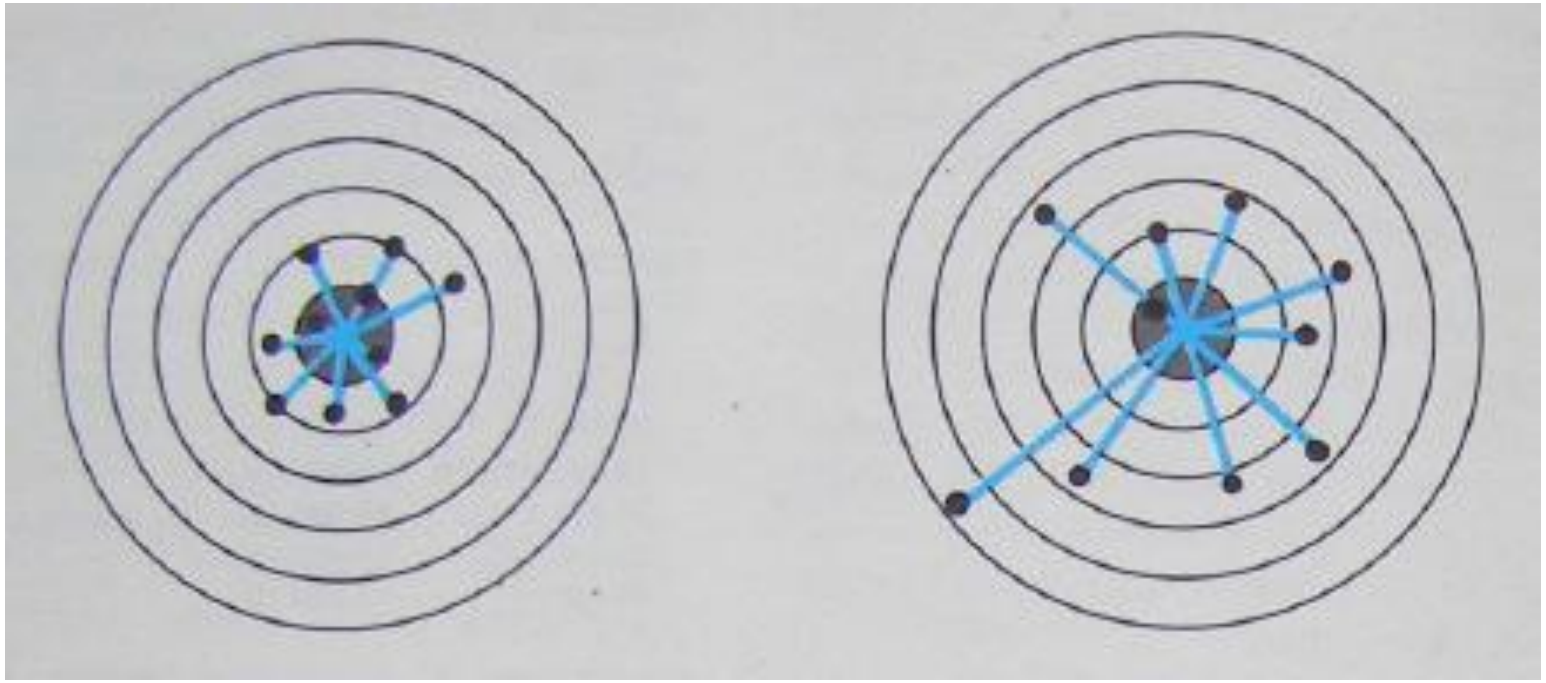


Míry variability

- Charakteristiky hodnotící rozptýlenost hodnot statistického souboru kolem nějaké míry polohy.

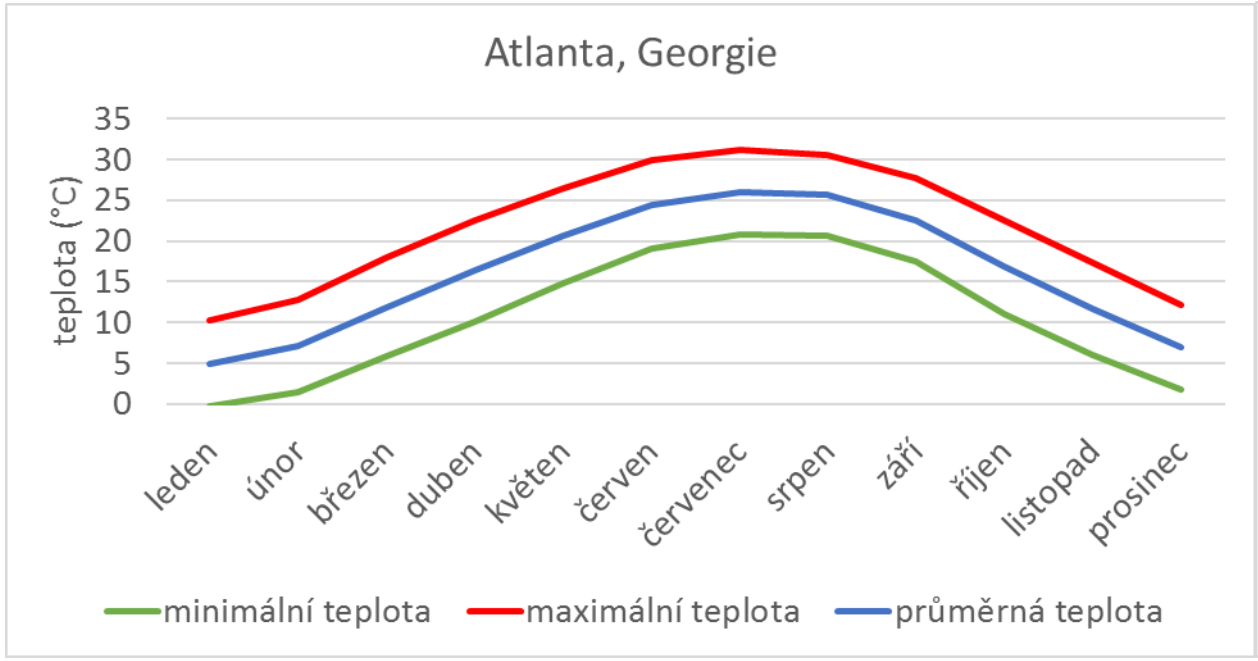
- Patří mezi ně: (variační) rozpětí, mezikvartilové (interkvartilové) rozpětí, rozptyl, směrodatná odchylka a variační koeficient.

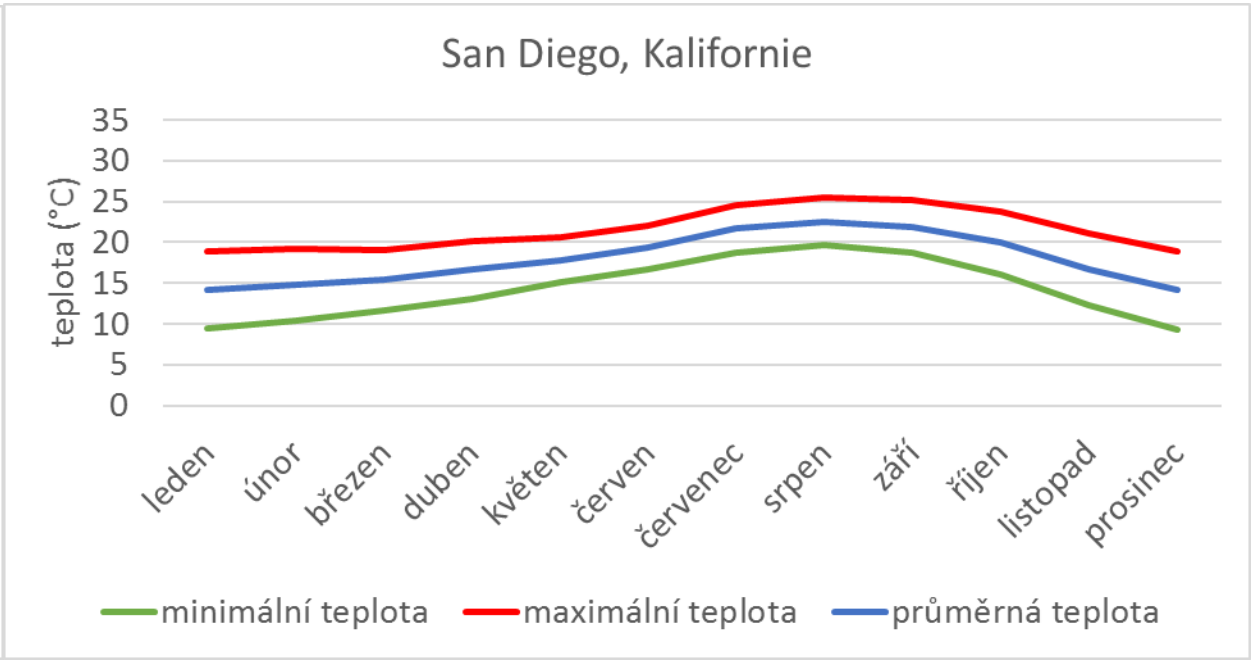
K čemu potřebujeme míry variability?



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

K čemu potřebujeme míry variability?





Atlanta, Georgie
prům. teplota 16°C

San Diego, Kalifornie
prům. teplota 17°C

Výběrový rozptyl

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

Na co si dát pozor?

Rozměr rozptylu je druhou mocninou rozměru proměnné.

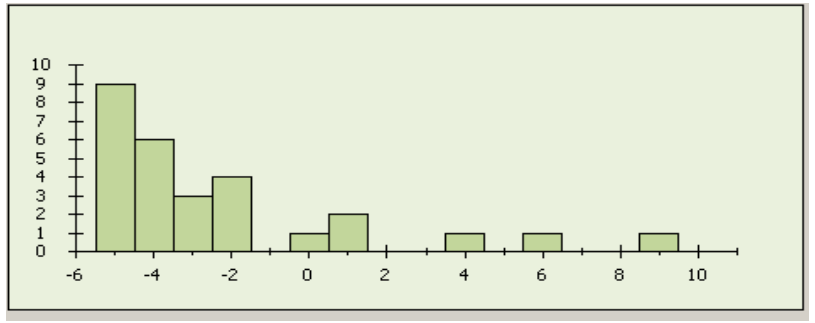
Výběrová směrodatná odchylka

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

Jakou představu o variabilitě dat nám dává sm. odchylka?

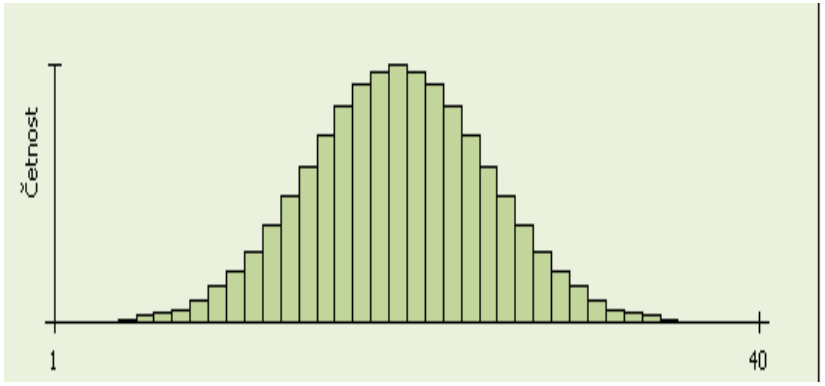
Čebyševova nerovnost: $\forall k > 0: P(\mu - k\sigma < X < \mu + k\sigma) > 1 - \frac{1}{k^2}$

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



Empirické pravidlo 3 sigma

k	$P(\mu - k\sigma < X < \mu + k\sigma)$
1	0,682
2	0,954
3	0,998



Variační koeficient

(Směrodatná odchylka v procentech aritmetického průměru. Používá se většinou pro proměnné nabývající nezáporných hodnot.)

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

- Čím nižší var. koeficient, tím homogennější soubor.
- $V > 50\ \%$ značí silně rozptýlený soubor.

Proč potřebujeme bezrozměrnou míru variability?

Umožňuje srovnání variability proměnných, které mají různé jednotky.

Interkvartilové rozpětí

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

Užití: např. při identifikaci odlehlých pozorování

Odlehlá pozorování

- ty hodnoty proměnné, které se mimořádně liší od ostatních hodnot a tím ovlivňují např. vypovídací hodnotu průměru.

Jak postupovat v případě, že v datech identifikujeme odlehlá pozorování?

- V případě, že odlehlost pozorování je způsobena:
  - hrubými chybami, překlepy, prokazatelným selháním lidí či techniky ...
  - důsledky poruch, chybného měření, technologických chyb ...

tzn., známe-li příčinu odlehlosti a předpokládáme-li, že již nenastane, jsme oprávněni tato pozorování vyloučit z dalšího zpracování.

- V ostatních případech je nutno zvážit, zda se vyloučením odlehlých pozorování nepřipravíme o důležité informace o jevech vyskytujících se s nízkou četností.

Identifikace odlehlých pozorování

Metoda vnitřních hradeb

$$\left[\left(x_i < x_{0,25} - 1{,}5IQR\right) \vee \left(x_i > x_{0,75} + 1{,}5IQR\right)\right] \Rightarrow x_i \ \ je\ odlehlým\ \ pozorováním$$

Dolní mez vnitřních hradeb

Horní mez vnitřních hradeb

Identifikace extrémních pozorování

Metoda vnějších hradeb

$$\left[\left(x_i < x_{0,25} - 3IQR\right) \vee \left(x_i > x_{0,75} + 3IQR\right)\right] \Rightarrow x_i \;\; je\ extrémním\ \ pozorováním$$

Dolní mez vnějších hradeb

Horní mez vnějších hradeb

V předložených datech identifikujte odlehlá pozorování:

$MN_{0,25} = \mathbf{6,8}$

$MN_{0,5}$**=7,3**

$MN_{0,75} = 8,7$

MN (%)
4,9
6,8
6,8
6,8
6,8
7,8
7,8
8,7
9,7
15,7

$IQR = MN_{0,75} - MN_{0,25} = 1,9$

$1,5 \cdot IQR = 2,85$

<u>Vnitřní hradby:</u>

Dolní mez: $6,8 - 2,85 = \mathbf{3,95}$ Horní mez: $8,7 + 2,85 = \mathbf{11,55}$

V předložených datech identifikujte odlehlá pozorování:

$MN_{0,25} = \mathbf{6,8}$

$MN_{0,5}$**=7,3**

$MN_{0,75} = 8,7$

MN (%)
4,9
6,8
6,8
6,8
6,8
7,8
7,8
8,7
9,7
15,7

$IQR = MN_{0,75} - MN_{0,25} = 1,9$

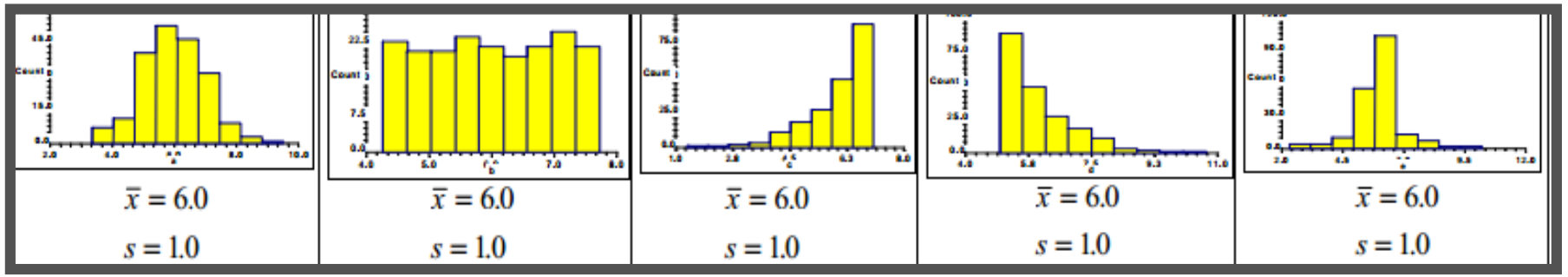
$1,5 \cdot IQR = 2,85$

<u>Vnitřní hradby:</u>

Dolní mez: $6,8 - 2,85 = \mathbf{3,95}$ Horní mez: $8,7 + 2,85 = \mathbf{11,55}$

Míry šikmosti a špičatosti

Jsou míry polohy a míry variability dostatečné pro posouzení rozdělení sledovaných veličin?



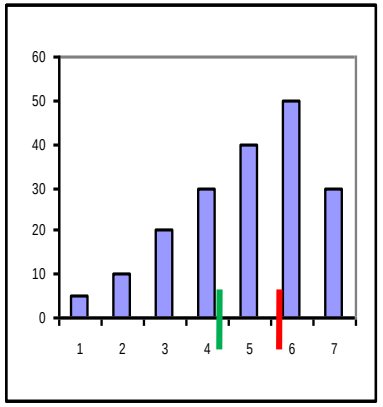
Zdroj: TVRDÍK, J.: Základy matematické statistiky, Ostravská univerzita, 2008

Všech pět ukázek má stejné charakteristiky polohy i variability (průměry i směrodatné odchylky jsou shodné). Přesto na první pohled vidíme, že tvary rozdělení dat jsou různé.

Výběrová šikmost (standardizovaná)

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

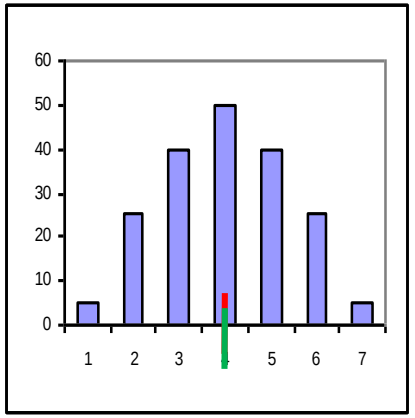
Negativně zešikmená data



$a < -2$

$\bar{x} < \tilde{x}_{0,5} < \hat{x}$

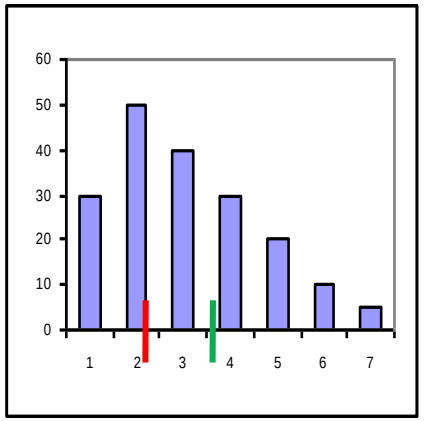
Symetrická data



$a \in \langle -2; 2 \rangle$

$\bar{x} = \tilde{x}_{0,5} = \hat{x}$

Pozitivně zešikmená data



$a > 2$

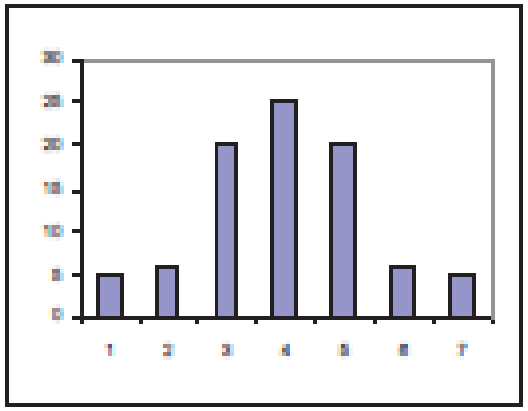
$\bar{x} > \tilde{x}_{0,5} > \hat{x}$

empirické pravidlo

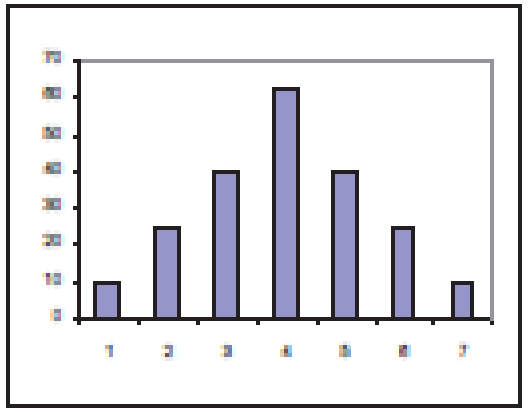
Výběrová špičatost (standardizovaná)

míra koncentrace kolem průměru

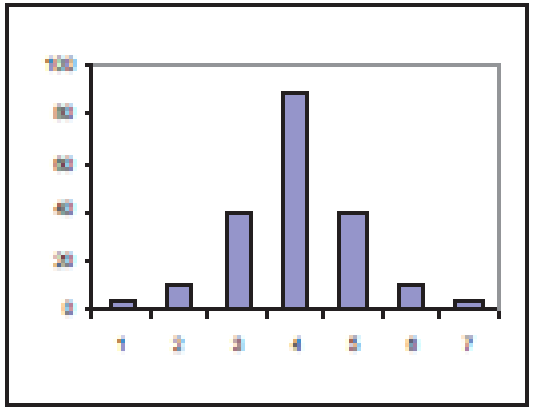
$$b = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \cdot \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{s^4} - 3\frac{(n-1)^2}{(n-2)(n-3)}$$



$b < -2$
rozdělení plošší než normální r.

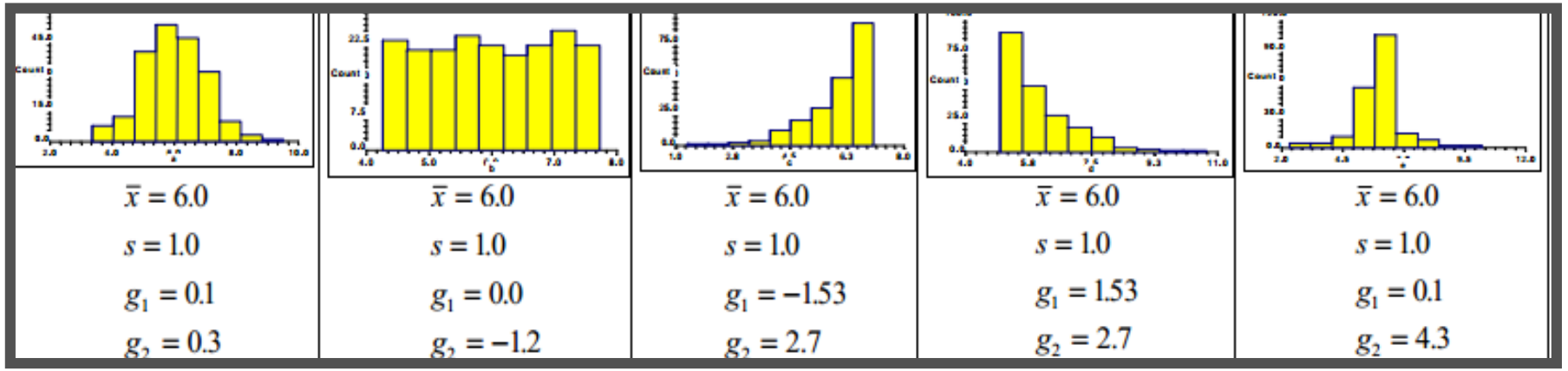


$b \in \langle -2; 2 \rangle$
špičatost odpovídá normálnímu r.



$b > 2$
rozdělení špičatější než normální r.

Jsou míry polohy a míry variability dostatečné pro posouzení rozdělení sledovaných veličin?



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

Všech pět ukázek má stejné charakteristiky polohy i variability (průměry i směrodatné odchylky jsou shodné). Přesto na první pohled vidíme, že tvary rozdělení dat jsou různé. K číselnému vyjádření těchto rozdílů nám slouží další charakteristiky - šikmost (g_1, angl. skewness) a špičatost (g_2, angl. kurtosis).

Přesnost číselných charakteristik

Směrodatnou odchylku jakožto míru nejistoty měření zaokrouhlujeme **nahoru** na jednu, maximálně dvě platné cifry a míry polohy (průměr, kvantily…) zaokrouhlujeme tak, aby nejnižší zapsaný řád odpovídal nejnižšímu zapsanému řádu směrodatné odchylky.

Chybný zápis číselných charakteristik

	Délka (m)	Váha (kg)	Teplota (^{0}C)
Průměr	2,26	127,6	14 567
Medián	2,675	117,8	13 700
Směrodatná odchylka	0,78	23,7	1 200 (před zaokrouhlením 1235)
Proč je zápis chybný?			

Chybný zápis číselných charakteristik

	Délka (m)	Váha (kg)	Teplota (^{0}C)
Průměr	2,26	127,6	14 567
Medián	2,675	117,8	13 700
Směrodatná odchylka	0,78	23,7	1 200 (před zaokrouhlením 1235)
Proč je zápis chybný?	*Různý počet des. míst.*		

Chybný zápis číselných charakteristik

	Délka (m)	Váha (kg)	Teplota (^{0}C)
Průměr	2,26	127,6	14 567
Medián	2,675	117,8	13 700
Směrodatná odchylka	0,78	23,7	1 200 (před zaokrouhlením 1235)
Proč je zápis chybný?	*Různý počet des. míst.*	*3 platné cifry u směrodatné odchylky.*	

Chybný zápis číselných charakteristik

	Délka (m)	Váha (kg)	Teplota (^{0}C)
Průměr	2,26	127,6	14 567
Medián	2,675	117,8	13 700
Směrodatná odchylka	0,78	23,7	1 200 (před zaokrouhlením 1235)
Proč je zápis chybný?	*Různý počet des. míst.*	*3 platné cifry u směrodatné odchylky.*	*Nejnižší zapsaný řád průměru (jednotky) neodpovídá nejnižšímu zapsanému řádu směrodatné odchylky (stovky)+ směr. odch. není zaokrouhlena nahoru.*

Oprava

	Délka (m)	Váha (kg)	Teplota (^{0}C)
Průměr	2,26	127,6	14 567
Medián	2,68	117,8	13 700
Směrodatná odchylka	0,78	23,7	1 200 (před zaokrouhlením 1235)
Proč je zápis chybný?		*3 platné cifry u směrodatné odchylky.*	*Nejnižší zapsaný řád průměru (jednotky) neodpovídá nejnižšímu zapsanému řádu směrodatné odchylky (stovky)+ směr. odch. není zaokrouhlena nahoru.*

Oprava

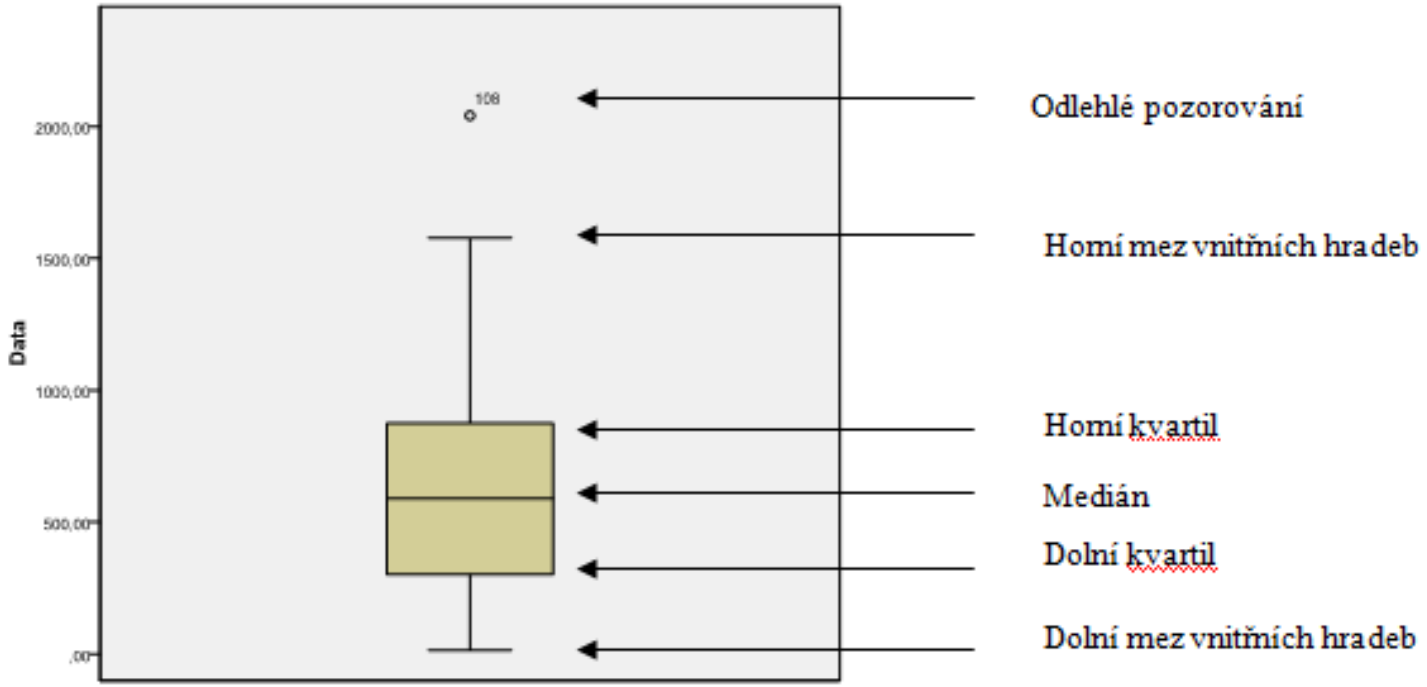
	Délka (m)	Váha (kg)	Teplota (^{0}C)
Průměr	2,26	128	14 567
Medián	2,68	118	13 700
Směrodatná odchylka	0,78	24	1 200 (před zaokrouhlením 1235)
Proč je zápis chybný?			*Nejnižší zapsaný řád průměru (jednotky) neodpovídá nejnižšímu zapsanému řádu směrodatné odchylky (stovky)+ směr. odch. není zaokrouhlena nahoru.*

Správný zápis číselných charakteristik

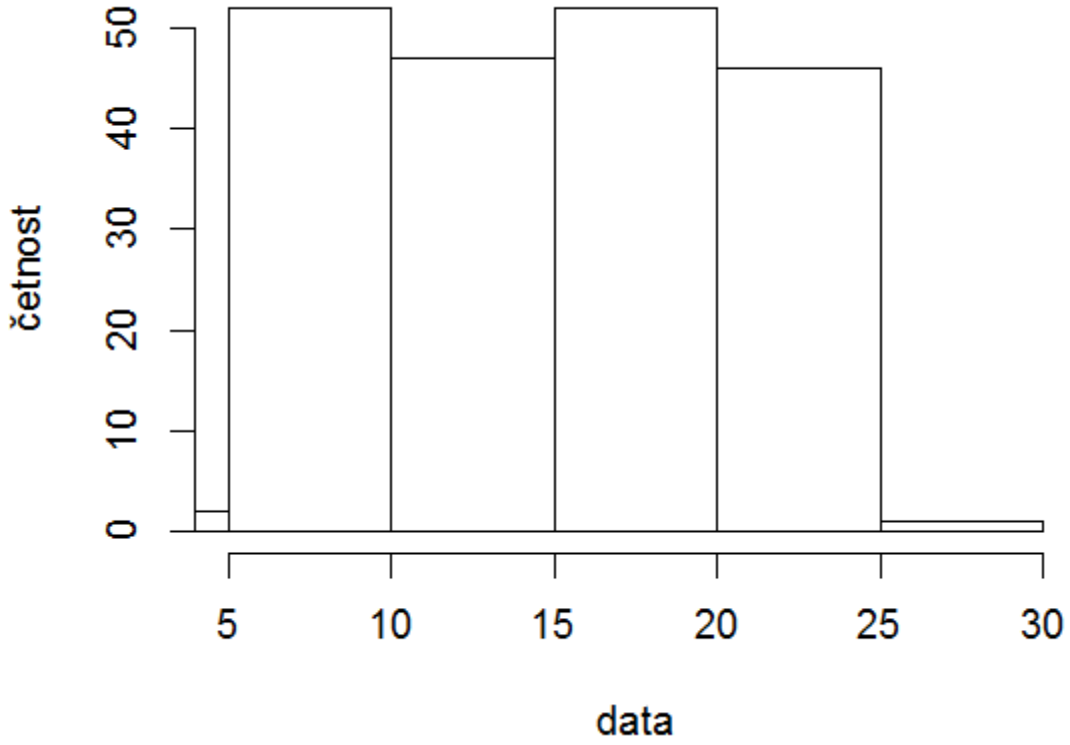
	Délka (m)	Váha (kg)	Teplota (^{0}C)
Průměr	2,26	127,6	14 600
Medián	2,675	117,8	13 700
Směrodatná odchylka	0,78	23,7	1 300

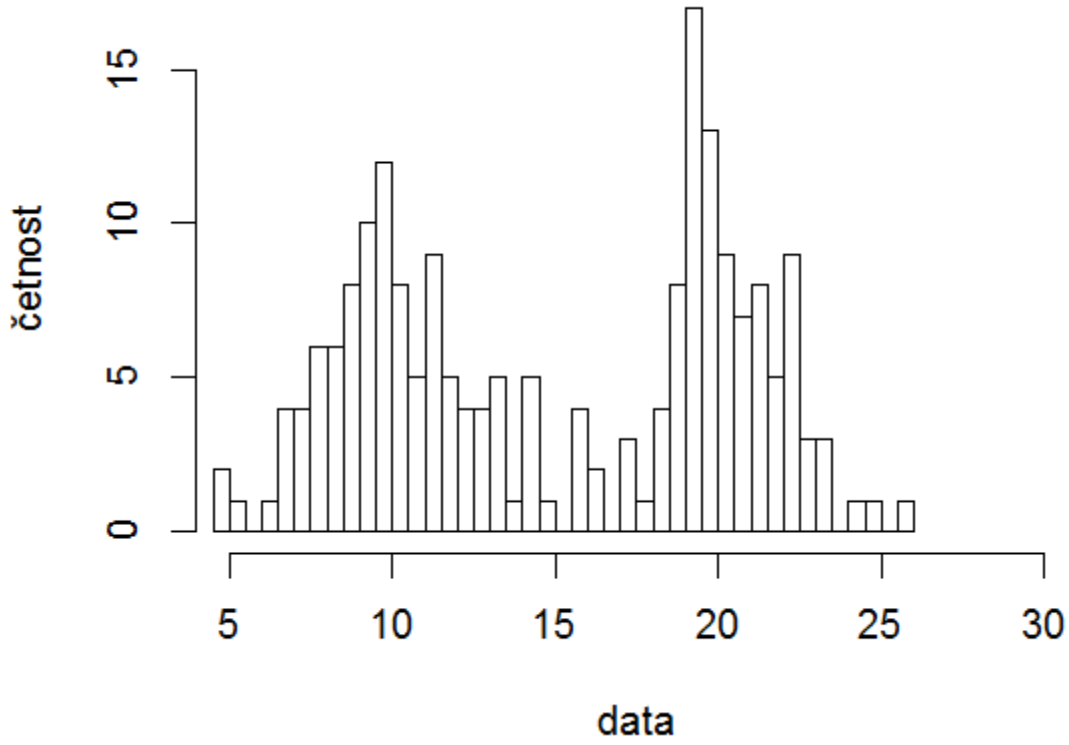
Grafické znázornění kvantitativní proměnné

Krabicový graf (Box plot)



Histogram





Pozor na zvolené členění - počet tříd!

Histogram

Postup při konstrukci histogramu:

1. Seřaďte data vzestupně, tj. od nejmenší po nejvyšší hodnotu.
2. Určete minimální a maximální hodnotu v souboru (MIN(x) a MAX(x))
3. Určete variační rozpětí R, kde $R = MAX(x) - MIN(x)$.
4. Určete počet tříd histogramu, tj. počet jeho sloupců.
   Počet tříd k můžete určit buď intuitivně, nebo pomocí níže uvedených vztahů.

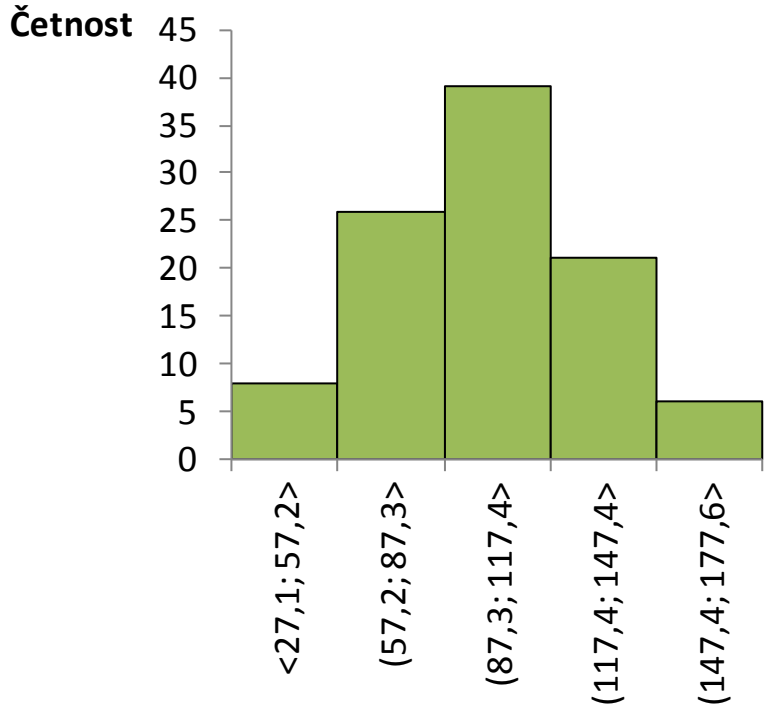
n (rozsah výběru)	k (doporučený počet tříd histogramu)
$n > 100$	$k \cong 10 \cdot \log n$
$40 < n \leq 100$	$k \cong 2\sqrt{n}$
$n \leq 40$	$k \cong 1 + 1{,}4426 \cdot \ln n$

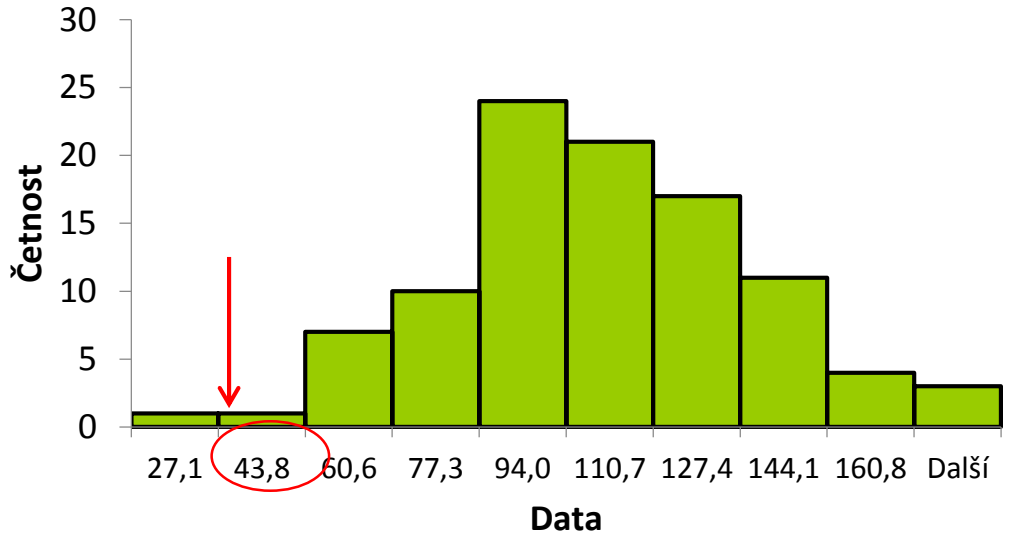
5. Vypočtěte šířku tříd h.

$$h \cong R/k$$

6. Určete meze jednotlivých tříd (viz Obr. 1.7).
7. Určete četnosti dat ve stanovených třídách a zakreslete histogram.

Histogram





Pozor na interpretaci automaticky generovaných histogramů v MS Excel!

DĚKUJEME ZA POZORNOST!