



Katedra
aplikované matematiky



Statistika s Excelem aneb Máme data. A co dál?

Martina Litschmannová

Katedra aplikované matematiky, FEI, VŠB-TU Ostrava

ŠKOMAM 2016

Jak získat data?

Primární zdroje dat

- Vlastní měření (fyzika, biologie, chemie, ...)
- Vlastní dotazníkové šetření, resp. agenturní výzkum na míru

Sekundární zdroje dat

- Interní zdroje (údaje obchodního charakteru, finanční povahy, technického rázu)
- Externí zdroje
 - Klasické statistické přehledy (ministerstva, ČSÚ, ...)
 - Agenturní výzkumy (panelová šetření – panel domácností, prodejen, televizního diváka, ...)
 - Databáze

Jak získat data?

Výběr vysoké školy

Vážené studentky, vážení studenti,
chtěli bychom zjistit jaké jsou vaše představy o dalším studiu na VŠ a jak vám při volbě vysoké školy pomáhají webové stránky vysokých škol a jejich studijních oborů.

Prosíme vás proto o vyplnění následujícího anonymního dotazníku. Jak se takovýto dotazník analyzuje si ukážeme v rámci ŠKOMAMu na přednášce Statistika s Excelem.

Za vyplnění dotazníku předem děkuji,
Martina Litschmannová

PS: Jako odměnu pro všechny respondenty dotazníkového šetření uvádím v předstihu otázku pro ŠKOMAM CUP:

Pokud v dotazníku uvedete, že chcete studovat na VŠ, budete dotazováni i na vaše očekávání ohledně nástupního platu po ukončení studia. Odhadněte, jaký bude průměrný očekávaný nástupní plat těch, kteří na otázku odpoví.

***Povinné pole**

Chcete po maturitě studovat na VŠ? *

- ☐ Ano
- ☐ Ne
- ☐ Nevím

<https://www.google.cz/intl/cs/forms/about/>

Máme data

https://docs.google.com/spreadsheets/d/1RgB2THxImbdclqthC-YgN55q

Průzkum (Odpovědi)

Soubor Upravit Zobrazit Vložit Formát Data Nástroje Formulář Doplnky Návod Poslední úprava provedena p

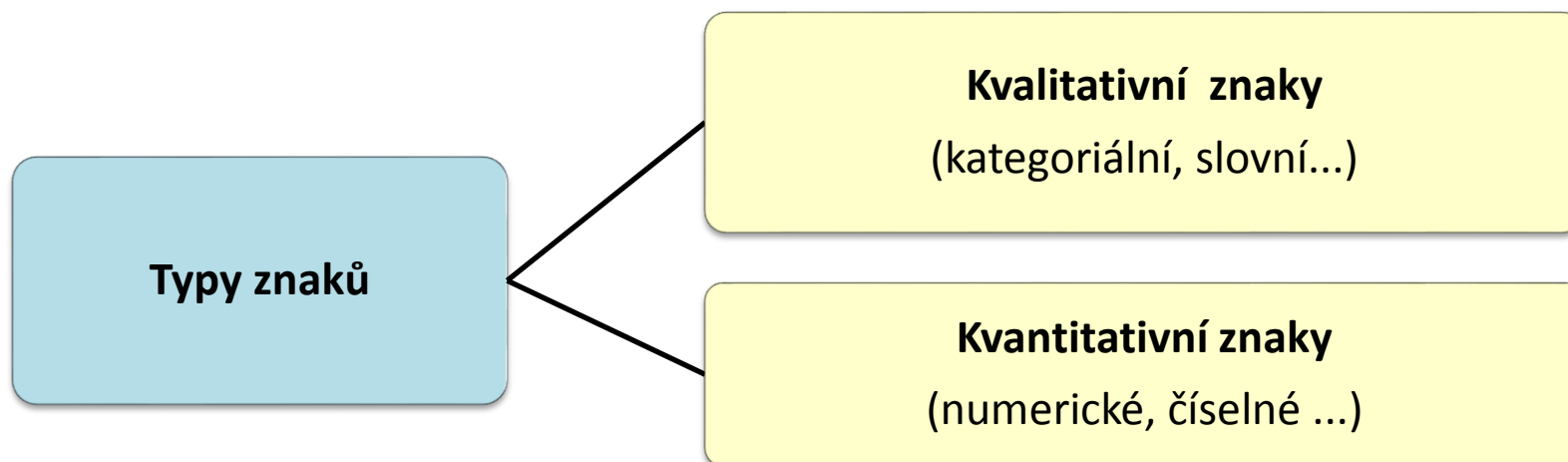
Komentáře Sdílet

Časová značka

	A	B	C	D	E	F	
1	Časová značka	Chcete po maturitě studo	Chcete studovat nějaký n	Kde sídlí VŠ, kterou chce	Jakou roli při vašem rozho	Jakou roli při vašem rozho	Jakou rol
11	3.2.2016 9:11:00	Ano	Ne	Praha	5	5	
12	3.2.2016 9:11:53	Ano	Zatím jsem si studijní obc	Ostrava	2	1	
13	3.2.2016 9:49:39	Ano	Ano	Praha	2	2	
14	3.2.2016 9:49:59	Ano	Zatím jsem si studijní obc	Praha	3	3	
15	3.2.2016 9:50:18	Ano	Ne	Praha	5	1	
16	3.2.2016 9:52:09	Ano	Zatím jsem si studijní obc	Zatím jsem si konkrétní v	5	4	
17	3.2.2016 9:52:44	Ano	Ano	Zatím jsem si konkrétní v	4	4	
18	3.2.2016 9:53:04	Ano	Ne	V zahraničí	1	2	
19	3.2.2016 9:54:57	Ano	Ne	Ostrava	4	4	
20	3.2.2016 9:54:57	Ano	Zatím jsem si studijní obc	Zatím jsem si konkrétní v	4	2	
21	3.2.2016 9:55:22	Ne					
22	3.2.2016 9:56:03	Ne					
23	3.2.2016 9:56:16	Nevím					
24	3.2.2016 9:58:02	Ano	Zatím jsem si studijní obc	Zatím jsem si konkrétní v	2	2	
25	3.2.2016 10:03:40	Ano	Ano	Praha	3	2	

A co dál?

Typy statistických znaků (proměnných)



Explorační analýza aneb popisná statistika

EDA pro kvalitativní (slovní) znaky

- Tabulky četností – tady není co zkazit?
- Jak vizualizovat kvalitativní znaky?

EDA pro kvantitativní (číselné) znaky

- Ořidný průměr
- Medián a spol.
- Jak identifikovat odlehlá pozorování?
- Proč potřebujeme míry variability?
- Jak vizualizovat kvantitativní znaky?

Jak jednoduše analyzovat kvalitativní znak?

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	$p_2=n_2/n$
x_k	n_k	$p_k=n_k/n$
Celkem:	$n_1+n_2+...+n_k=n$	1

+ Modus (název nejčetnější varianty)

Číselné charakteristiky

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

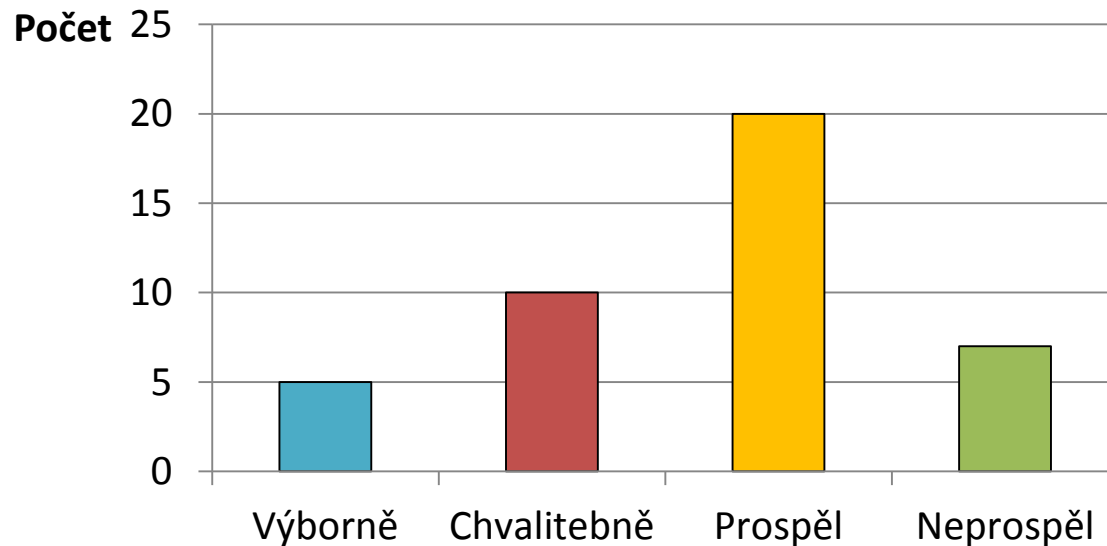
Relativní četnosti uvádějme vždy pouze jako doplněk absolutních četností, nikoliv samostatně!



Jak kvalitativní znak vizualizovat?

Grafické znázornění

A) Sloupcový graf (bar chart)

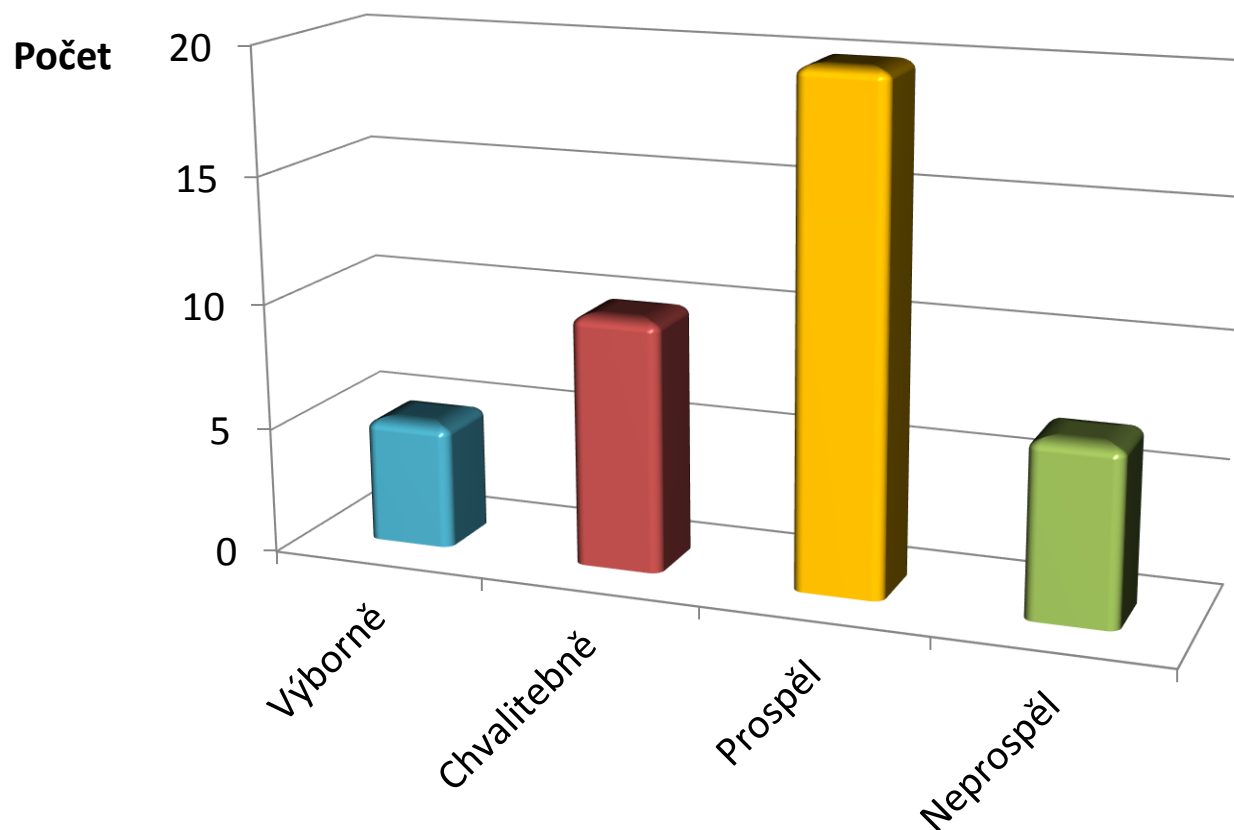


„...můžete vytvořit sloupcový graf a dodat mu zcela nový a přitažlivý vzhled“

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

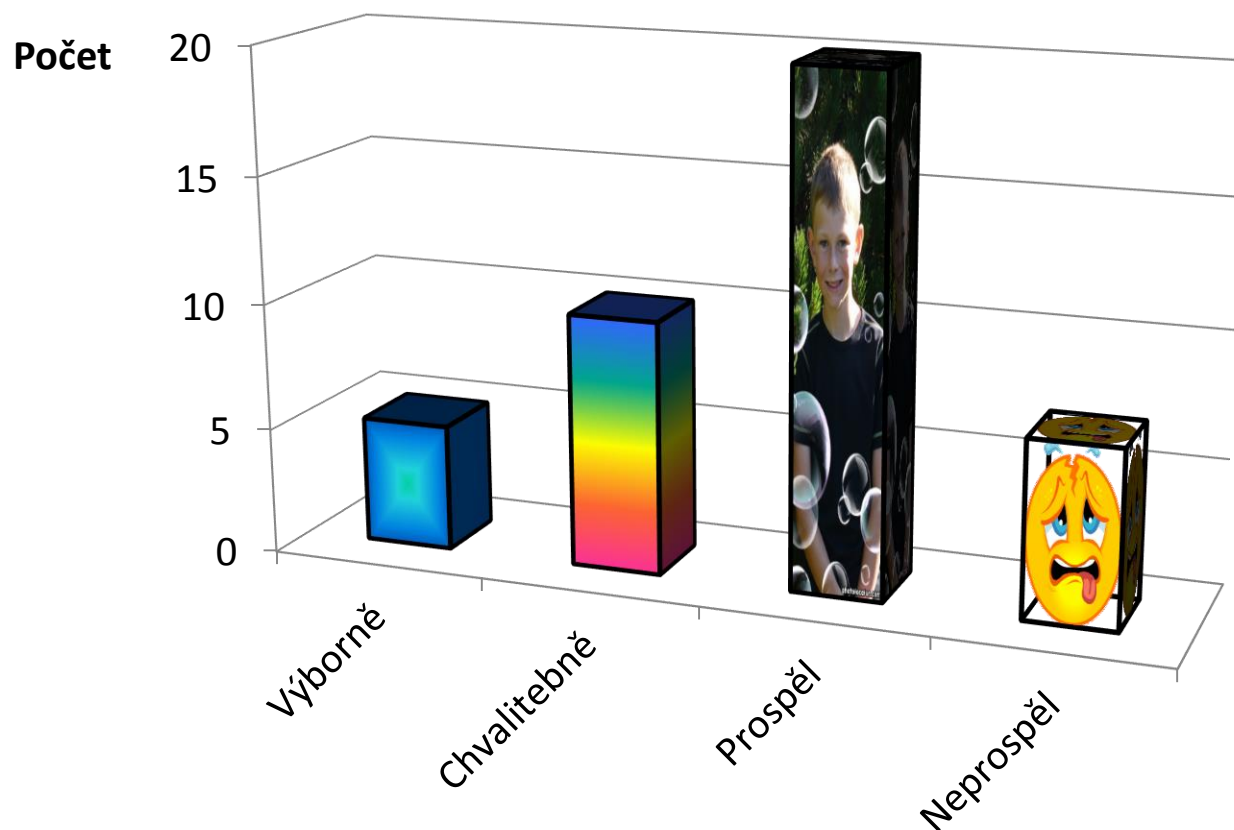
Grafické znázornění

A) Sloupcový graf (bar chart)



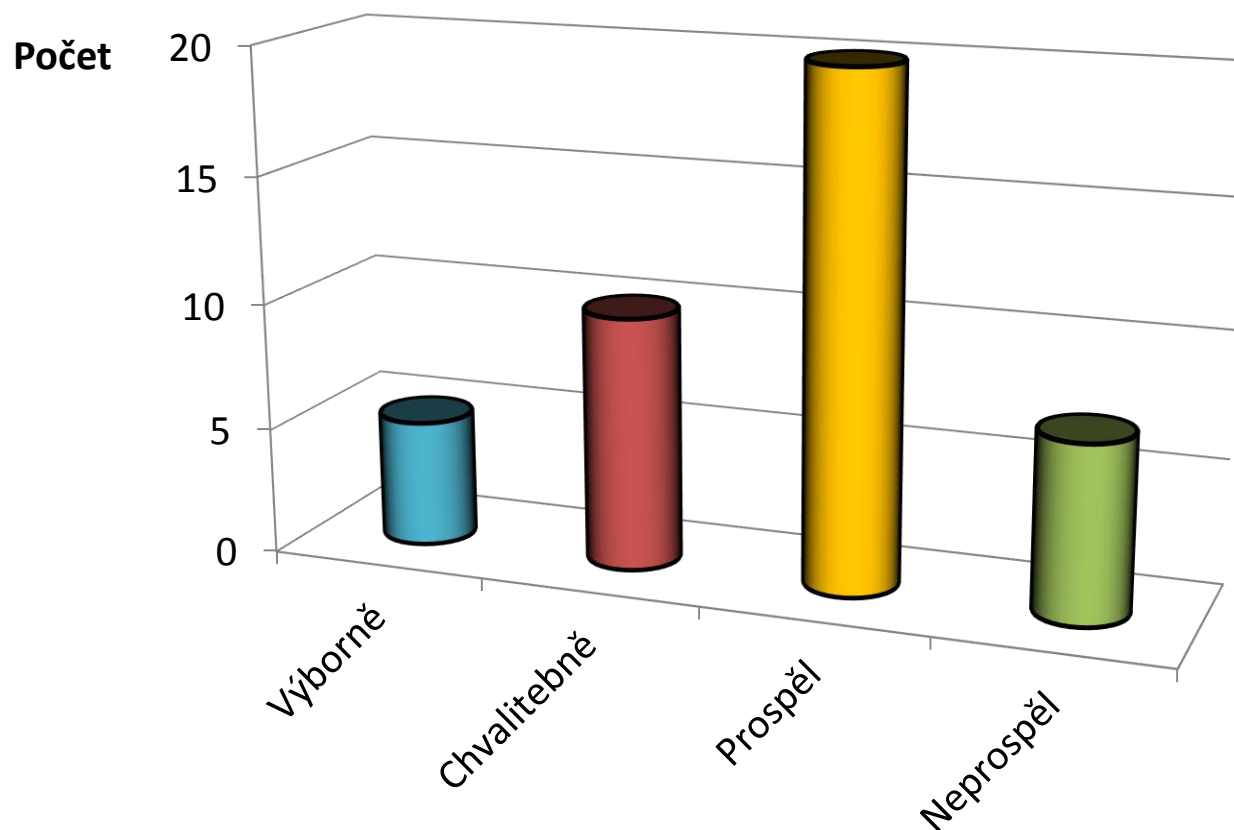
Grafické znázornění

A) Sloupcový graf (bar chart)



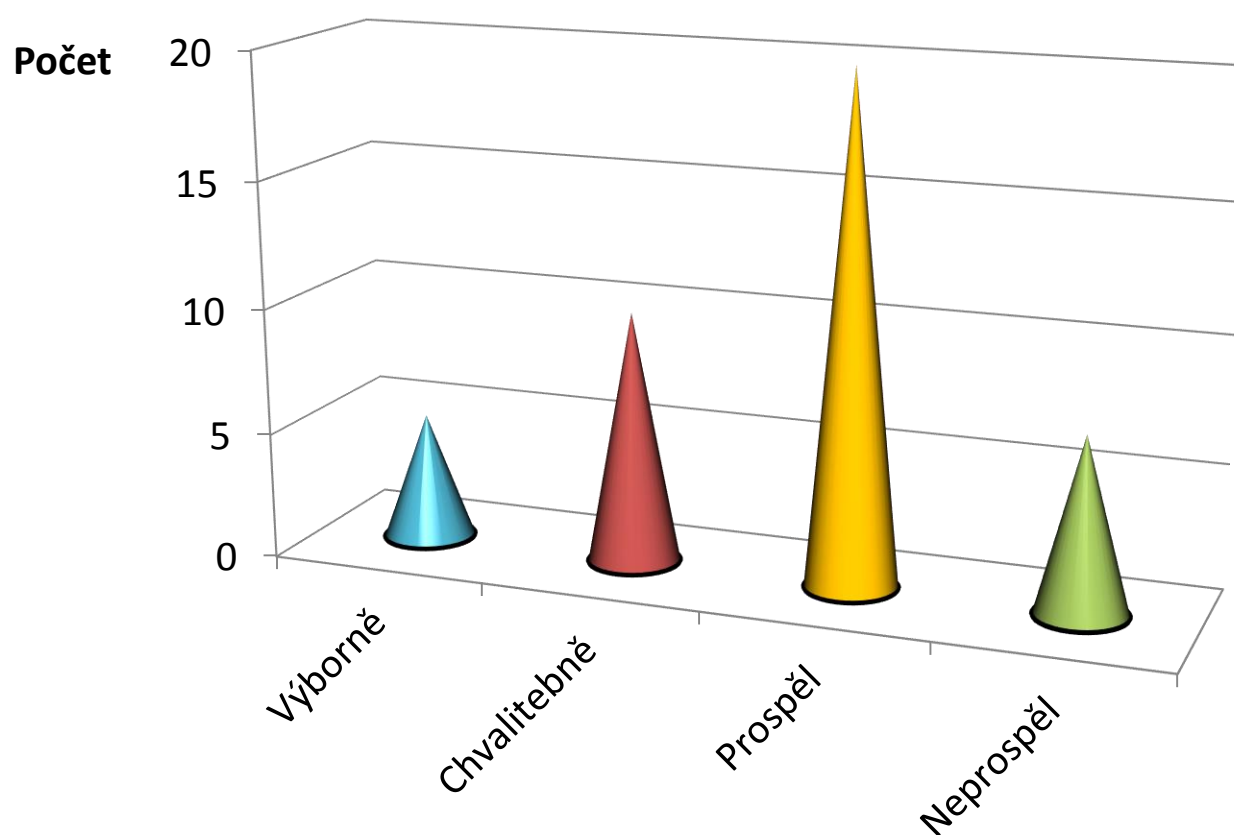
Grafické znázornění

A) Sloupcový graf (bar chart)



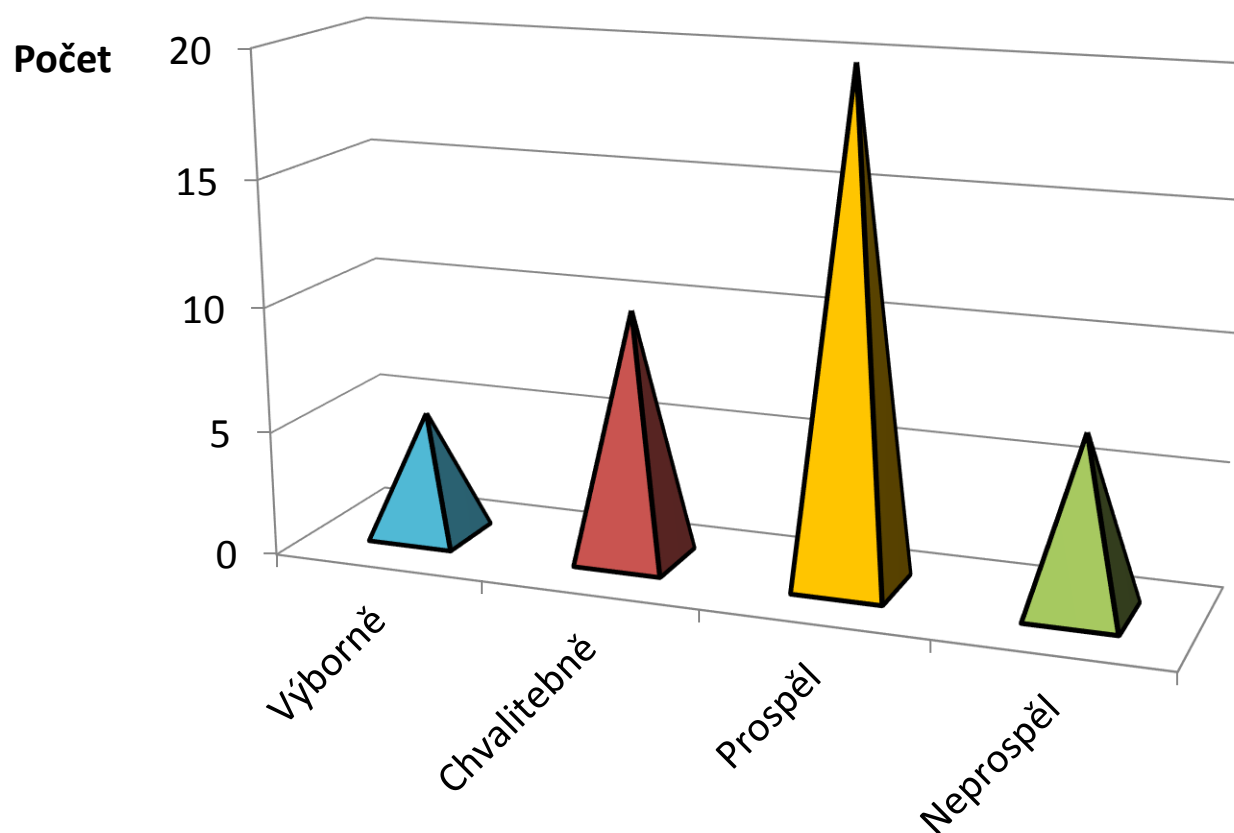
Grafické znázornění

A) Sloupcový graf (bar chart)



Grafické znázornění

A) Sloupcový graf (bar chart)



Grafické znázornění

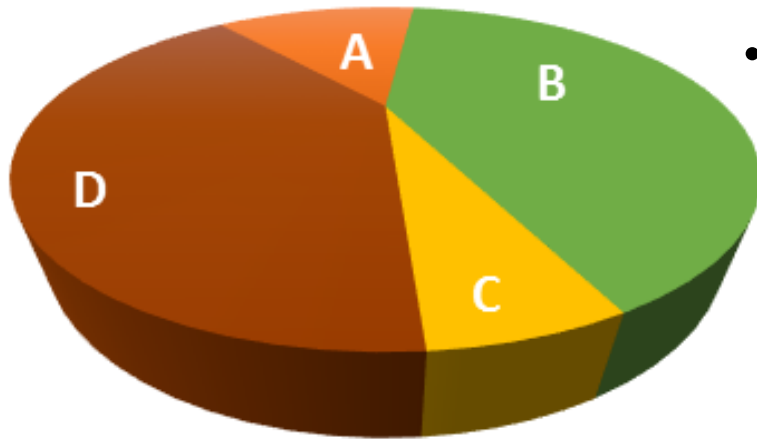
A) Sloupcový graf (bar chart)

Na co si dát pozor?

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

Grafické znázornění

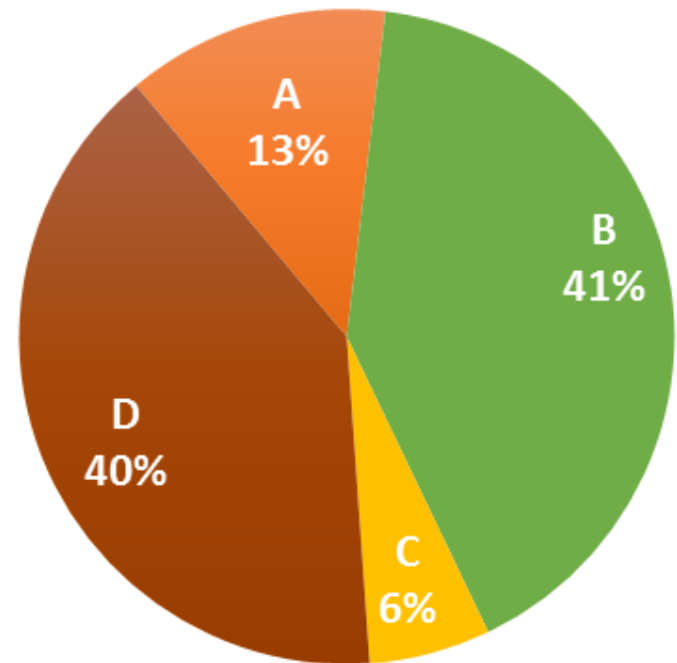
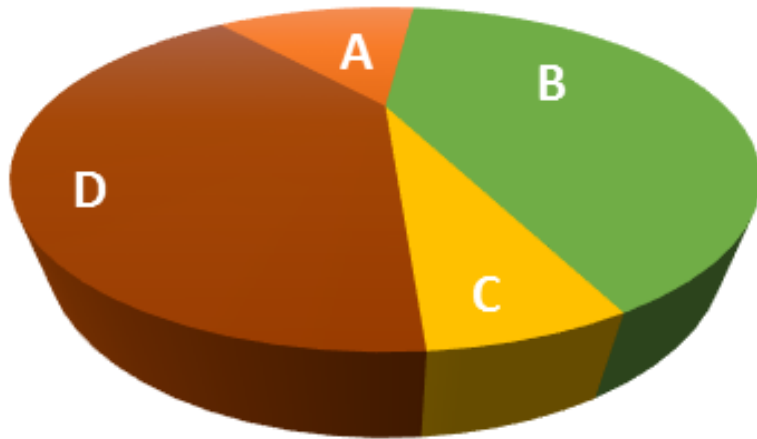
B) Výsečový graf – koláčový graf (pie chart)



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

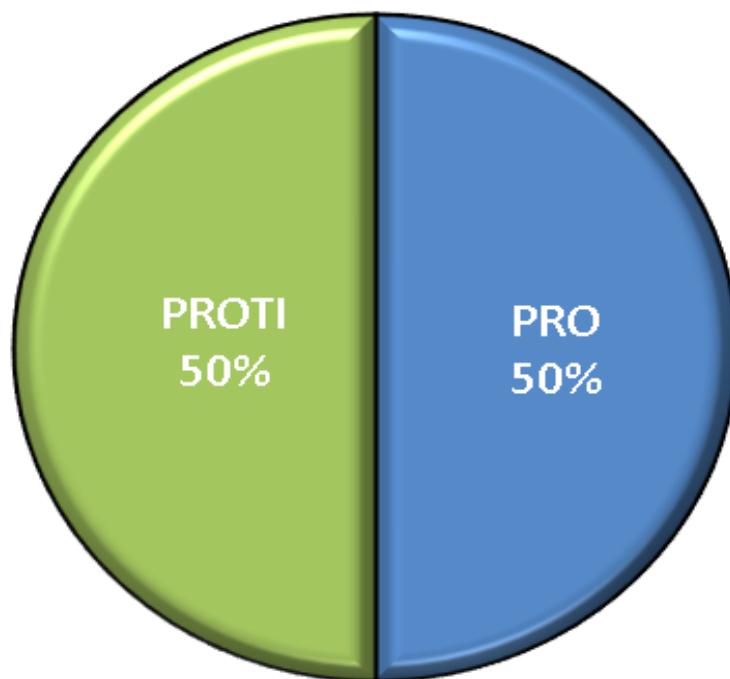
Grafické znázornění

B) Výsečový graf – koláčový graf (pie chart)



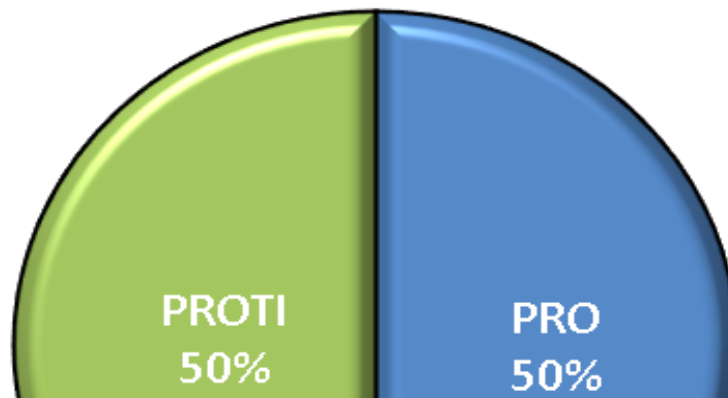
Anketa

Jste pro navýšení hodinové dotace Matematiky na SŠ?



Anketa

Jste pro navýšení hodinové dotace Matematiky na SŠ?



TAKHLE NE!!!

Grafické znázornění

B) Výsečový graf – koláčový graf (pie chart)

Na co si dát pozor?

Grafické znázornění

B) Výsečový graf – koláčový graf (pie chart)

Na co si dát pozor?

- Otáčení 3D výsečových grafů
- Neuvádění absolutních četností, resp. celkového počtu respondentů v „blízkosti“ grafu

A nyní prakticky...

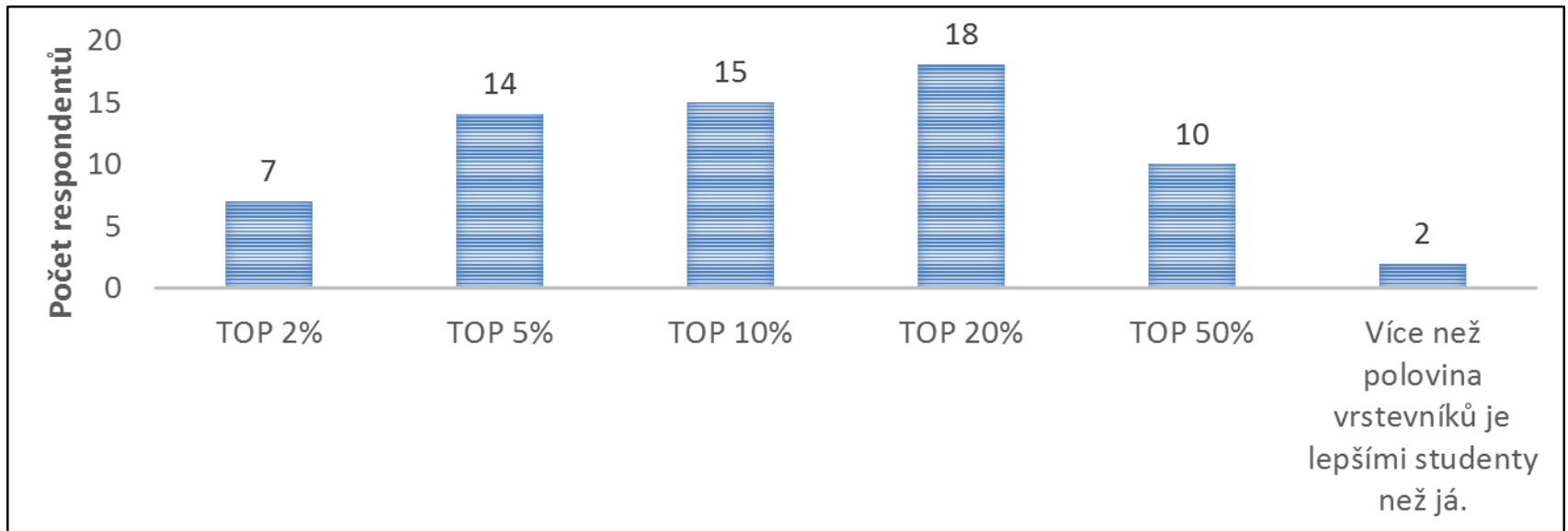
ID	Sebehodnocení
3.2.16 9:11	TOP 20%
3.2.16 9:11	TOP 2%
3.2.16 9:49	TOP 20%
3.2.16 9:49	Více než polovina vrstevníků je lepšími studenty než já.
3.2.16 9:50	TOP 10%

Jak analyzovat kvalitativní znak v Excelu?

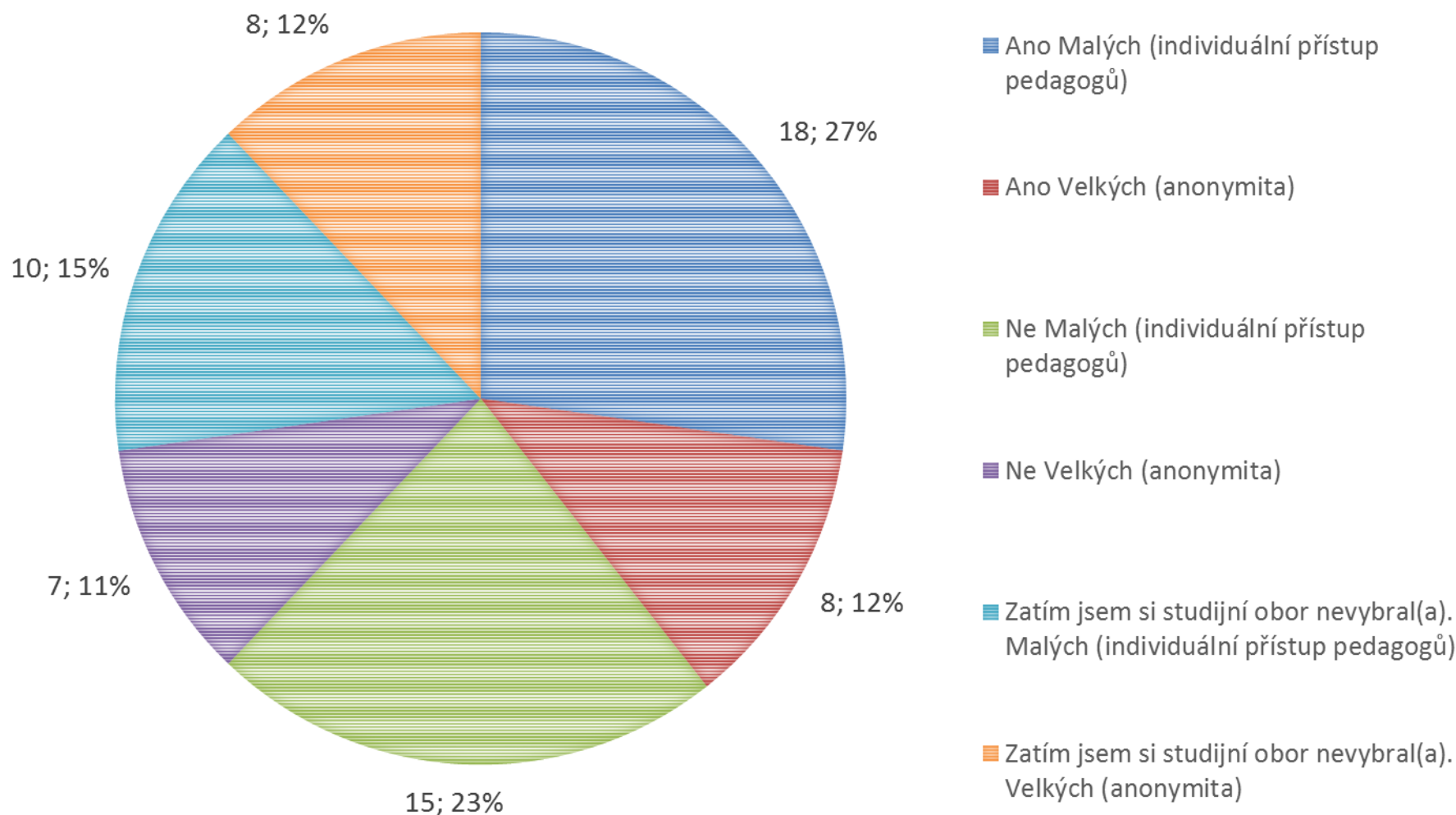
Vložení / Grafy / Kontingenční graf / Kontingenční graf a kontingenční tabulka

A nyní prakticky...

Kategorie studentů	Četnost	Relativní četnost
TOP 2%	7	11%
TOP 5%	14	21%
TOP 10%	15	23%
TOP 20%	18	27%
TOP 50%	10	15%
Více než polovina vrstevníků je lepšími studenty než já.	2	3%
Celkový součet	66	100%



Preference studijních oborů dle jejich velikostí v závislosti na tom, zda studenti již mají představu, jaký studijní obor chtějí studovat



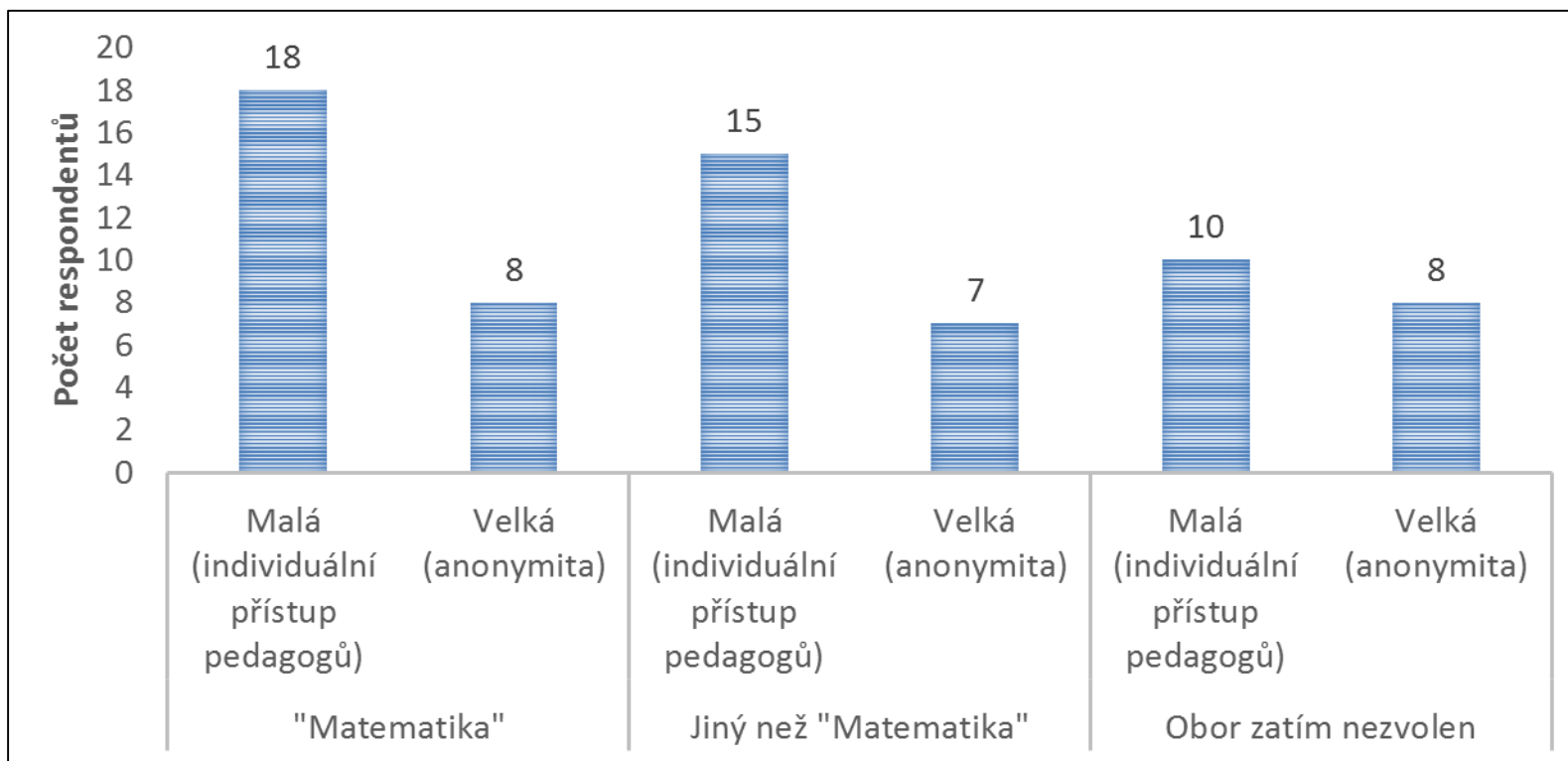
Grafické znázornění

B) Výsečový graf – koláčový graf (pie chart)

Na co si dát pozor?

- Otáčení 3D výsečových grafů
- Neuvádění absolutních četností, resp. celkového počtu respondentů v „blízkosti“ grafu
- Ne vždy je graf přehlednější než tabulka

Preference studijních oborů dle jejich velikostí v závislosti na tom, zda studenti již mají představu, jaký studijní obor chtějí studovat



Preferovaná velikost studijní skupiny	Zvolený studijní obor			Celkový součet
	"Matematika"	Jiný než "Matematika"	Obor zatím nezvolen	
Malá (individuální přístup pedagogů)	18	15	10	43
Velká (anonymita)	8	7	8	23
Celkový součet	26	22	18	66

Explorační analýza aneb popisná statistika

EDA pro kvalitativní (slovní) znaky

- Tabulky četností – tady není co zkazit?
- Jak vizualizovat kvalitativní znaky?

EDA pro kvantitativní (číselné) znaky

- Ořídny průměr
- Medián a spol.
- Jak identifikovat odlehlá pozorování?
- Proč potřebujeme míry variability?
- Jak vizualizovat kvantitativní znaky?

Jak jednoduše analyzovat kvantitativní znak?

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 ledniče,
hovoří o příjemné průměrné teplotě.

Autor neznámý

Aritmetický průměr

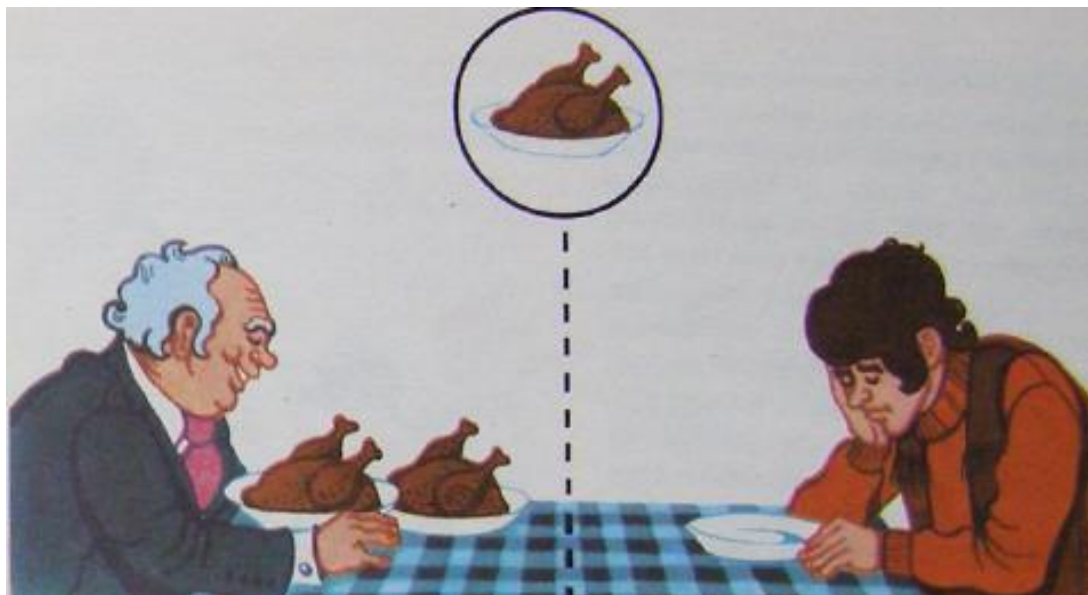
$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

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

Ošidnost průměru

ŠKOMAM CUP

Pokud v dotazníku uvedete, že chcete studovat na VŠ, budete dotazováni i na vaše očekávání ohledně nástupního platu po ukončení studia. Odhadněte, jaký bude průměrný očekávaný nástupní plat těch, kteří na otázku odpoví.

Aritmetický průměr

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

Na co si dát pozor?

- Průměr není rezistentní vůči odlehlým pozorováním!
- Vážený průměr
- Harmonický průměr (úlohy o společné práci, průměrná rychlost, ...)
- Geometrický průměr (tempa růstu)
- 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** – $\tilde{x}_{0,1}; \tilde{x}_{0,2}; \dots; \tilde{x}_{0,9}$

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

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

Kde se s kvantily setkáme v praxi?

- Vyhodnocení Národních srovnávacích zkoušek, ...

Percentil

Skóre se převede na **percentil** - číslo udávající „kolik procent účastníků jste svým výsledkem v daném testu předstihli“. Percentil 90 tedy znamená, že jste byli úspěšnější (nebo stejně úspěšní) než 90 % všech účastníků.

Zdroj: <https://scio.cz/nsz/vyhodnoceni.asp>

Odlehlá 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

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

**Dolní mez
vnitřních hradeb**

**Horní mez
vnitřních hradeb**

$$IQR = x_{0,75} - x_{0,25}$$

(interkvartilové rozpětí)

Identifikace odlehlých pozorování

Metoda vnějších hradeb

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

Dolní mez
vnějších hradeb

Horní mez
vnějších hradeb

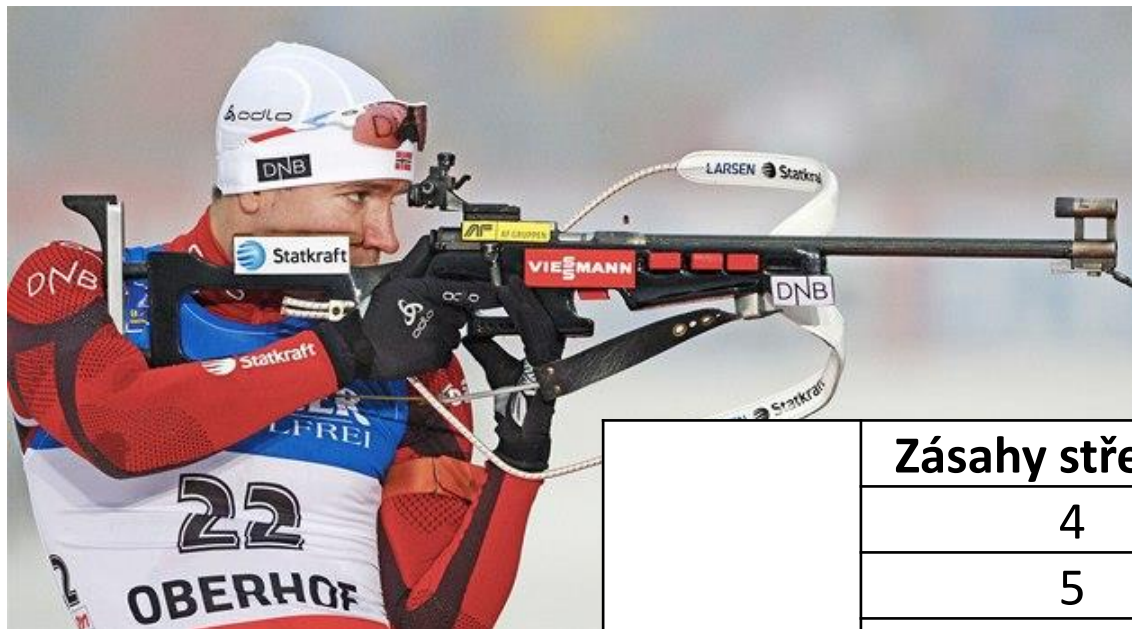
$$IQR = x_{0,75} - x_{0,25}$$

(interkvartilové rozpětí)

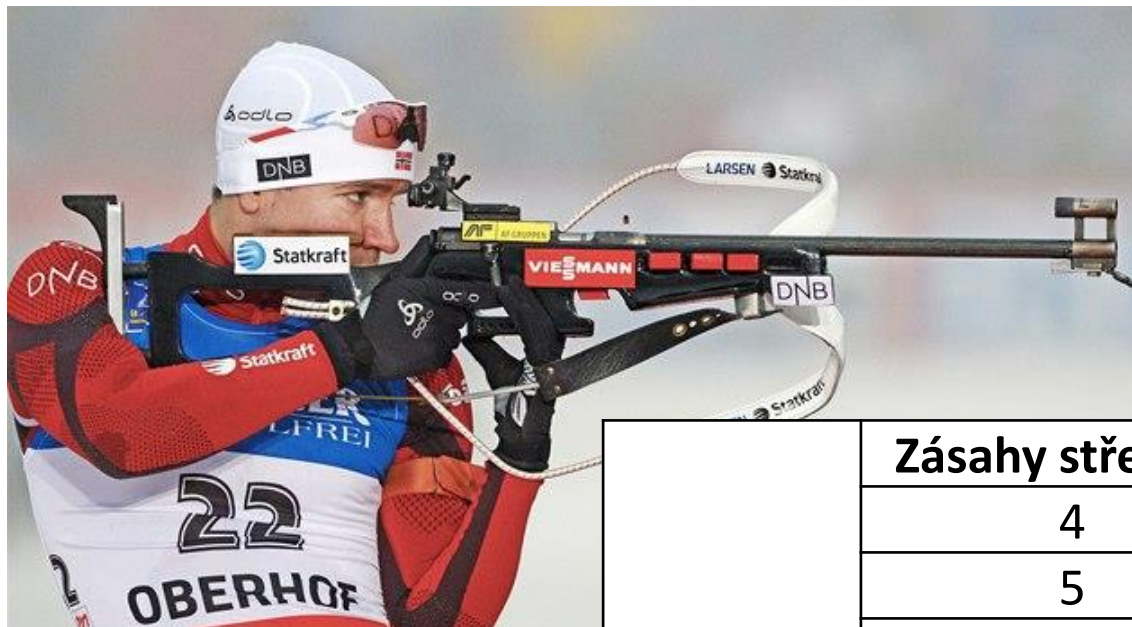
Míry variability

- Charakteristiky hodnotící rozptýlenost hodnot statistického souboru kolem nějaké míry polohy.
- Patří mezi ně: (variační) rozpětí, mezikvartilové (interkvartilové) rozpětí, rozptyl, směrodatná odchylka a variační koeficient.

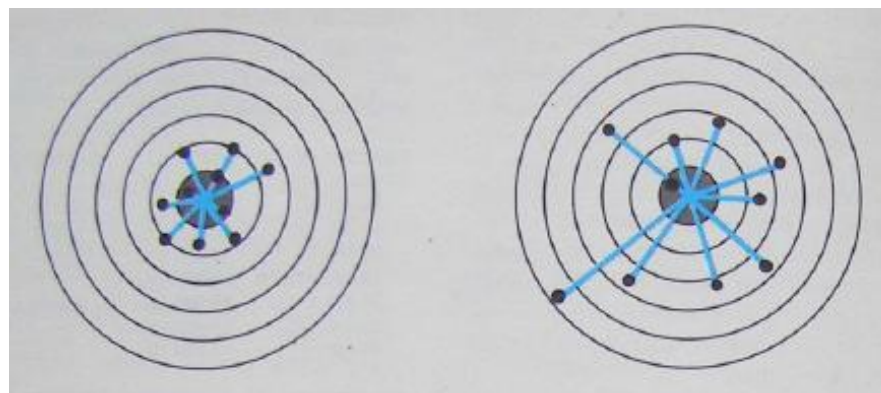
K čemu potřebujeme
míry variability?



	Zásahy střelce A	Zásahy střelce B
	4	1
	5	5
	6	9
Průměr	?	?



	Zásahy střelce A	Zásahy střelce B
	4	1
	5	5
	6	9
Průměr	5	5



Zdroj: Swoboda Helmut, Moderní statistika, 1977

Výběrový rozptyl

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

Na co si dát pozor?

Rozměr rozptylu charakteristiky je **druhou mocninou rozměru proměnné**.

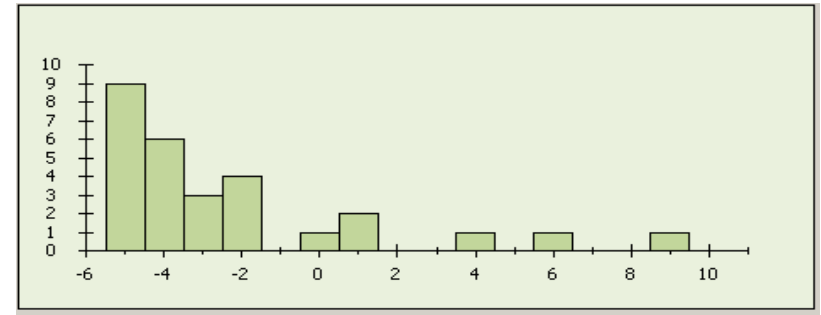
Výběrová směrodatná odchylka

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

Jakou představu o variabilitě dat nám dává sm. odchylka?

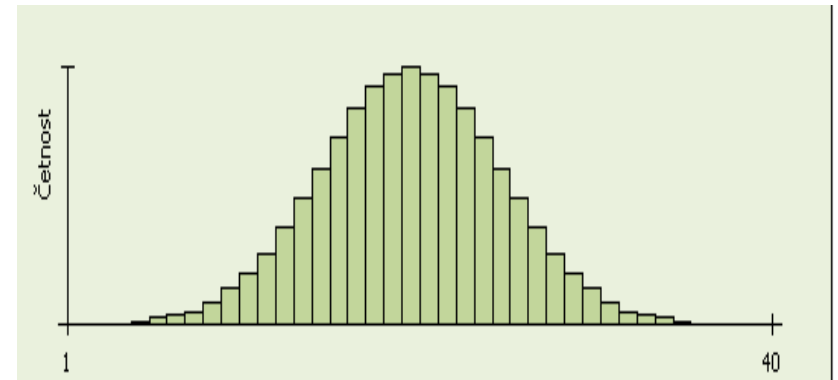
Čebyševova nerovnost: $\forall k > 0: P(\mu - k\sigma < X < \mu + k\sigma) > 1 - \frac{1}{k^2}$

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



Pravidlo 3 sigma

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



μ – populační průměr, σ – populační směrodatná odchylka

Variační koeficient

(Směrodatná odchylka v procentech aritmetického průměru)

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

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

Proč potřebujeme bezrozměrnou míru variability?

Umožňuje srovnání variability proměnných, které mají různé jednotky.

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

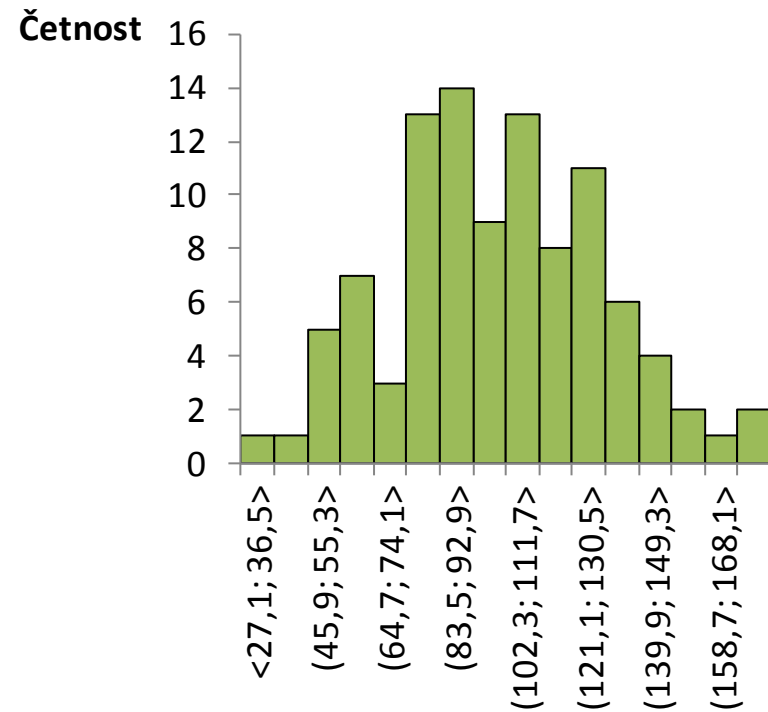
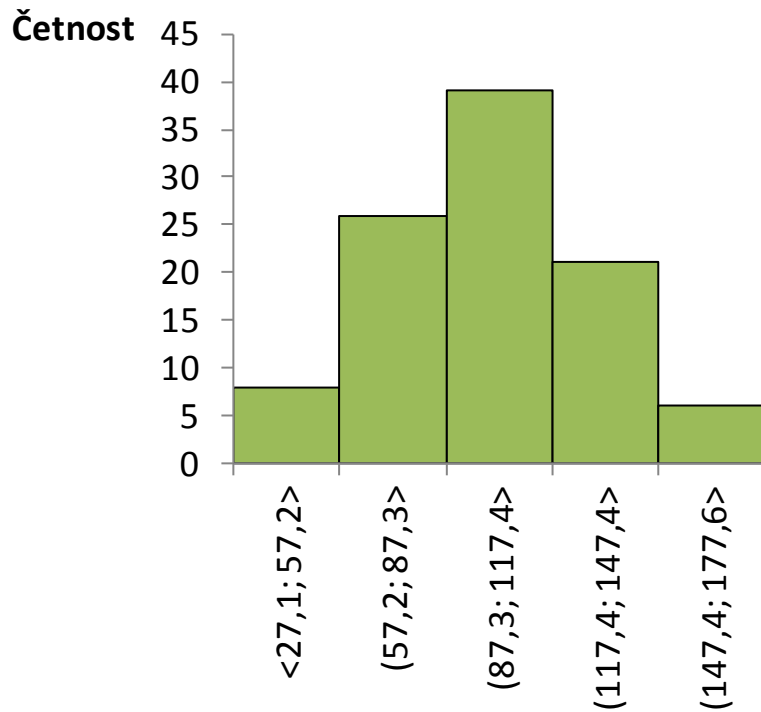
	Délka (m)	Váha (kg)	Teplota (°C)
Průměr	2,26	127,6	14 567
Medián	2,675	117,8	13 700
Směrodatná odchylka	0,78	23,3	1 200 (před zaokrouhlením 1235)

Správný zápis číselných charakteristik

	Délka (m)	Váha (kg)	Teplota (°C)
Průměr	2,26	128	14 600
Medián	2,68	118	13 700
Směrodatná odchylka	0,78	24	1 300

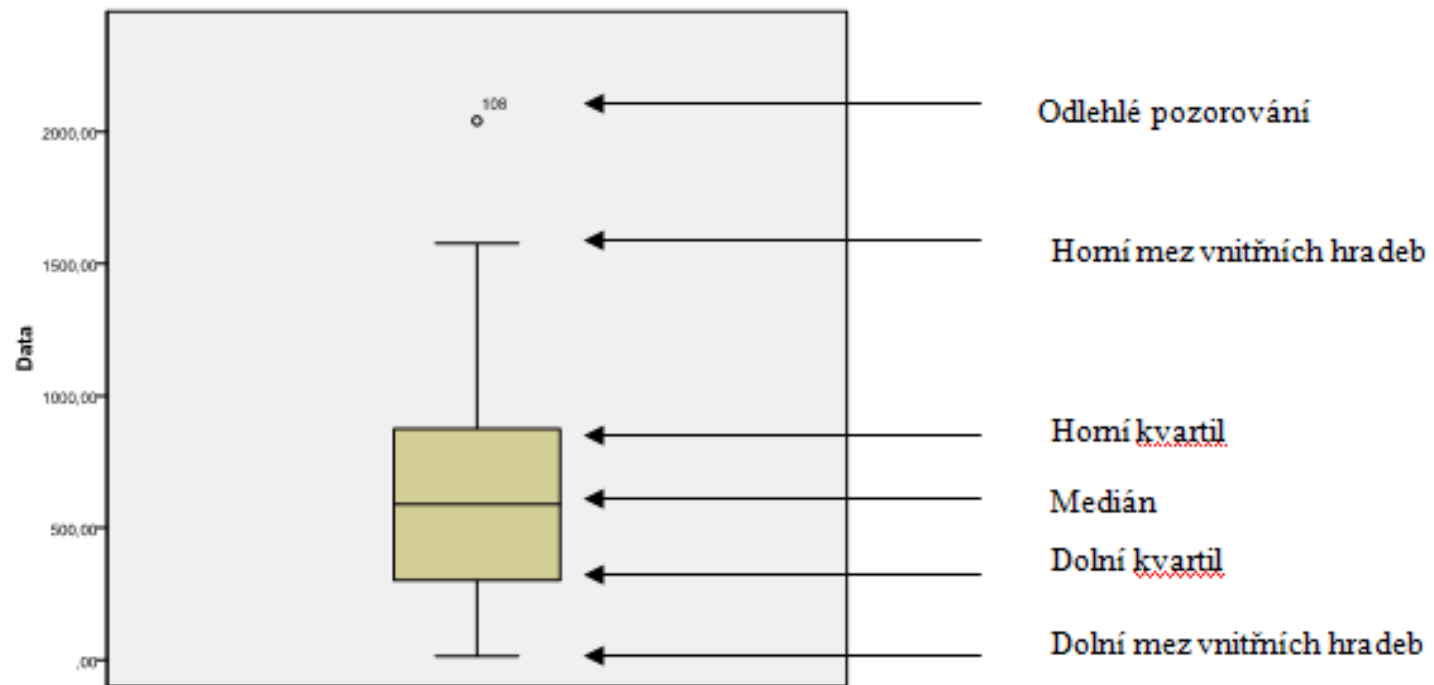
Grafické zobrazení

a) Histogram



Grafické zobrazení

b) Krabicový graf (Box plot)



A nyní prakticky...

ID	Očekávaný příjem
3.2.16 9:11	25000
3.2.16 9:11	30000
3.2.16 9:49	1,5E+12
3.2.16 9:49	24000
3.2.16 9:50	40000

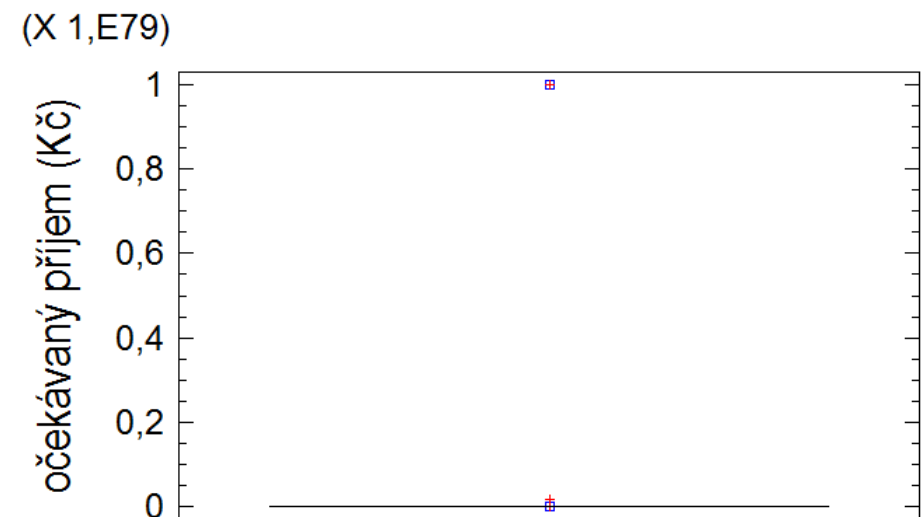
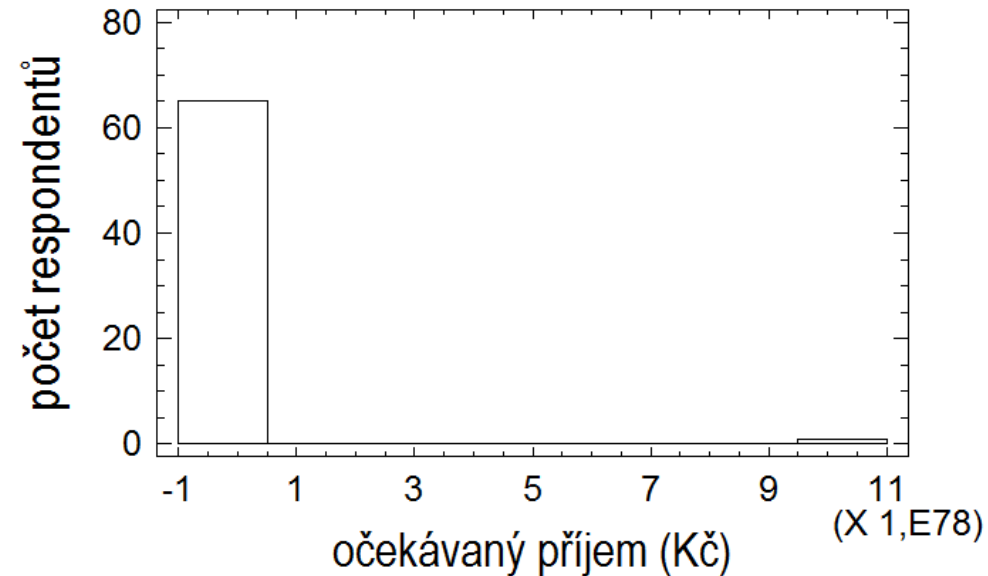
Jak analyzovat kvantitativní znak v Excelu?

viz skomam2016.xlsx

A nyní prakticky...

původní datový soubor

Očekávaný příjem (Kč)	
počet respondentů	66
počet chybějících hodnot	6
Míry polohy	
minimum	42
dolní kvartil	21250
medián	25000
průměr	1,51515E+77
horní kvartil	30000
maximum	1E+79
Míry variability	
směrodatná odchylka	1,23091E+78
variační koeficient (%)	812,4038405

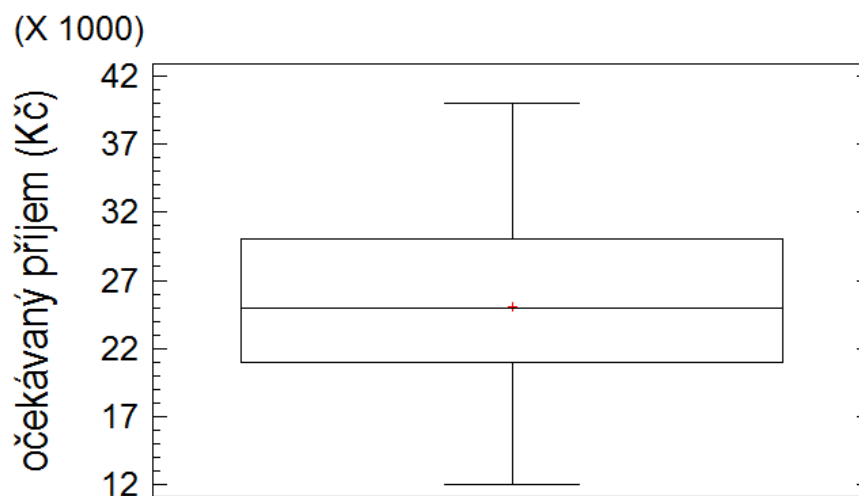
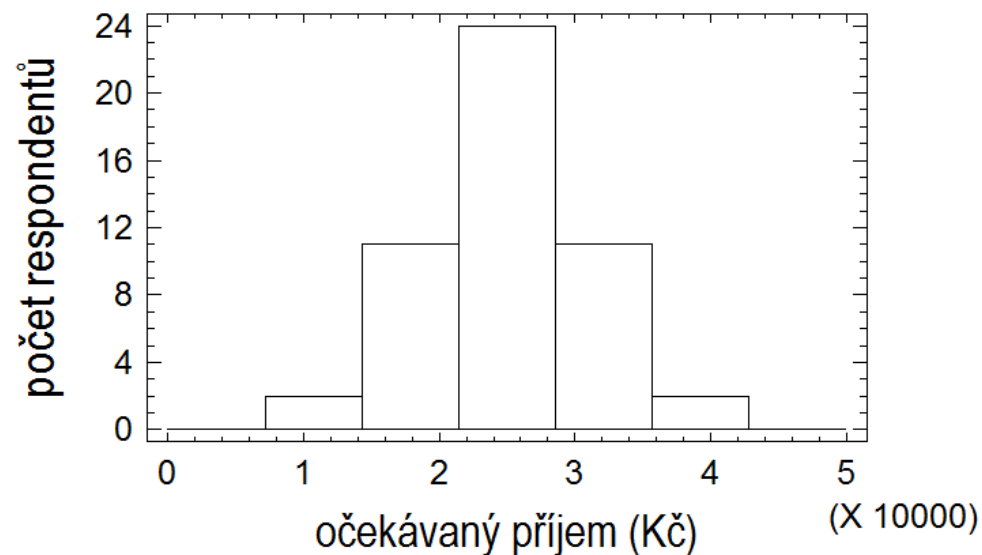


A nyní prakticky...

po odstranění odlehlých pozorování

Očekávaný příjem (Kč)

počet respondentů	50
počet chybějících hodnot	22
Míry polohy	
minimum	12000
dolní kvartil	21250
medián	25000
průměr	25024,9762
horní kvartil	29500
maximum	40000
Míry variability	
směrodatná odchylka	5371,395697
variační koeficient (%)	21,46413908



Děkuji za pozornost!