Statistika

aneb známe tři druhy lži:

- úmyslná
- neúmyslná
- statistika

Statistika je metoda, jak vyjádřit nejistá data s přesností na setinu procenta.

Statistika

- Shromažďování a třídění dat
- Analýza za účelem formulování obecných závěrů a rozhodování

Popisná statistika

Metody zjišťování a sumarizace informací

Inferenční statistika

Měření spolehlivosti závěrů o populaci založených na informacích získaných z výběru

náhodný výběr

výběrový soubor $\subset$ **základní soubor**

den	teplota
1.4.2008	11
2.4.2008	10
3.4.2008	10
4.4.2008	9
5.4.2008	8
6.4.2008	7
7.4.2008	8
8.4.2008	9
9.4.2008	4
10.4.2008	9
11.4.2008	8
12.4.2008	7
13.4.2008	8
14.4.2008	9
15.4.2008	12
16.4.2008	13
17.4.2008	15
18.4.2008	11
19.4.2008	12
20.4.2008	10
21.4.2008	9
22.4.2008	8
23.4.2008	9
24.4.2008	11
25.4.2008	10
26.4.2008	9
27.4.2008	6

Elementární zpracování dat

- **cílem je zjednodušit nějaká data tak, abychom se v nich lépe vyznali**
- **důsledkem je ztráta informací!**

průměrná teplota: 9.2 °C
minimum: 4 °C
maximum: 15 °C
rozsah: 11 °C
modus: 9 °C
medián: 9 °C
rozptyl: 5.1 °C
směrodatná odchylka: 2.3 °C

STATISTICKÝ ZNAK

Dopravní nehody celá Česká republika

Kraj	Počet nehod	Následky				Příčiny					Pod vlivem alkoholu
		Mrtví	Těžce ranění	Lehce ranění	Škoda (v Kč)	Nepřiměřená rychlost	Nedání přednosti v jízdě	Nesprávné předjíždění	Nesprávný způsob jízdy	Jiné	
Praha	82	0	0	3	3 208 000	2	17	2	59	2	0
Středočeský	61	0	3	7	2 363 000	3	10	2	26	20	0
Jihočeský	20	0	0	1	534 000	0	2	1	16	1	0
Plzeňský	12	0	0	1	448 000	1	2	0	5	4	0
Karlovarský	9	0	2	3	326 000	2	1	0	3	3	2
Ústecký	19	0	1	3	662 000	1	3	0	6	9	0
Liberecký	17	0	0	5	817 000	4	2	0	6	5	0
Královéhradecký	15	0	1	1	419 000	0	0	1	5	9	1
Pardubický	17	0	1	2	1 063 000	1	2	0	7		
Vysočina	16	0	0	3	929 000	3	3				
Jihomoravský	31	0	0	2	1 467 000						
Olomoucký	17	0	2								
Moravskoslezský	4										
Zlínský	1										
Česká republika	370								165	107	9

je společnou vlastností prvků statistického souboru, jejíž proměnnost je předmětem zkoumání
Např. věk, rodinný stav, bydliště, zaměstnání nebo u věcí účel, stáří
těžce zranění, vliv alkoholu, vzniklá škoda, známky…

Hodnota statistického znaku

1. **Kvalitativní**
   - Nominální – kvalitativní znaky
     Jednotlivé hodnoty jsou neporovnatelné, data nelze seřadit, neexistuje nic jako „velikost“
   - Ordinární – lze je seřadit a přiřadit jim číslo, tj. určit pořadí
     Neumožňují posoudit vzdálenost
2. **Kvantitativní**
   - Diskrétní (je možné je spočítat)
   - Spojité (údaje, které lze měřit -přečíst na stupnici)

Určete typ dat:

věk člověka, výška, počet podlaží, teplota, pohlaví, zaměstnání, účel budovy, barva očí, dosažené vzdělání, počet studentů, známka ze zkoušky,

MĚŘÍCÍ STUPNICE (ŠKÁLY)

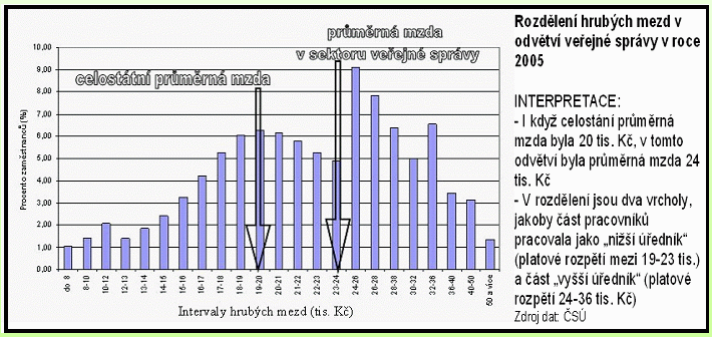
	Příklad	Matematické operace	Statistické charakteristiky
NOMINÁLNÍ	Číselné označení: barev, psychologického typu, pohlaví atd.	=, ≠	Modus, absolutní a relativní četnost
ORDINÁLNÍ	Školní známky, stupnice tvrdosti, služební pořadí, Richterova stupnice, …	=, ≠, <, >	+ medián, kvantily a kvantilové odchylky, procentily
INTERVALOVÁ	Teplota ve °C, Farenheita, letopočet, IQ, …	Navíc: intervaly, **nula zvolená** „+“ „-“ „*“	+ arim. průměr, směrodatná odchylka, šikmost, špičatost
POMĚROVÁ	Teplota ve °Kelvina, věk, váha, výška, velikost úhlu, čas, …	Navíc: **nula absolutní** „/“	+ koeficient variability, geom. průměr

MĚŘÍCÍ STUPNICE (ŠKÁLY)

	Testy významnosti	Míry závislosti	Statistické metody
NOMINÁLNÍ	χ^2 – test, McNemar test, Cochran test, …	Kontingenční tabulka a čtyřpolní tabulka	Některé neparametrické
ORDINÁLNÍ	Znaménkový test, Mann-Whitney U-test, Friedmanova pořadová analýza variance, aj.	+ pořadová korelace	Všechny neparametrické
INTERVALOVÁ	Parametrické metody: F-test t-test pro závislé či nezávislé soubory)	+ Pearsonova součinová korelace	Všechny neparametrické a parametrické
POMĚROVÁ	Parametrické metody: F-test t-test pro závislé či nezávislé soubory)	+ Pearsonova součinová korelace	Všechny neparametrické a parametrické

Třídění – rozdělení dat do tříd

1. Tříd musí být „tak akorát“
   cílem je potlačit kolísání četností, ale nesmíme setřít charakteristické rysy
2. Třídy musí být disjunktní
3. Stejná šířka intervalu



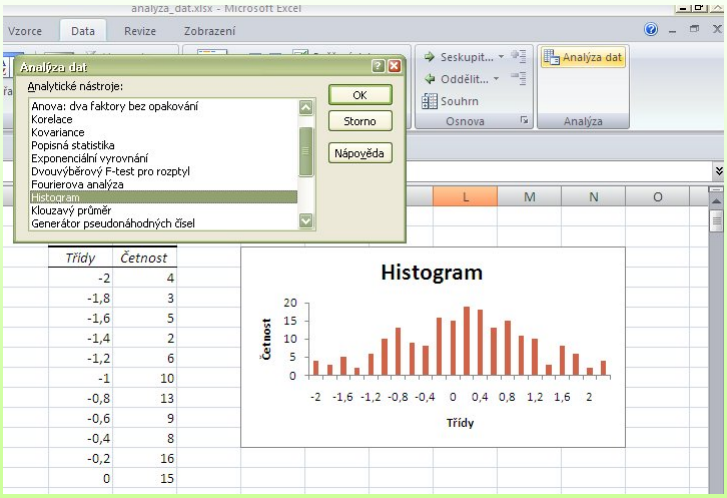
Rozdělení hrubých mezd v odvětví veřejné správy v roce 2005

INTERPRETACE:
- I když celostátní průměrná mzda byla 20 tis. Kč, v tomto odvětví byla průměrná mzda 24 tis. Kč
- V rozdělení jsou dva vrcholy, jakoby část pracovníků pracovala jako „nižší úředník“ (platové rozpětí mezi 19-23 tis.) a část „vyšší úředník“ (platové rozpětí 24-36 tis. Kč)

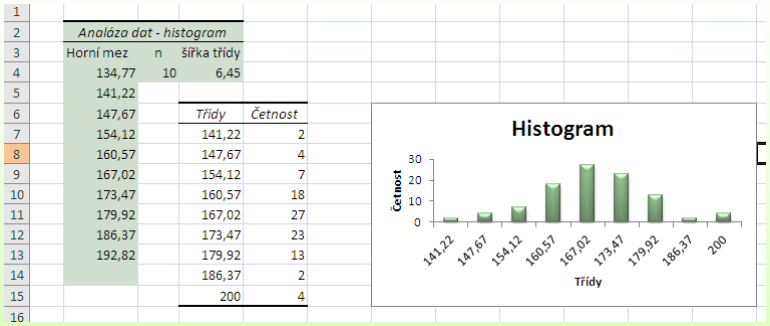
Zdroj dat: ČSÚ

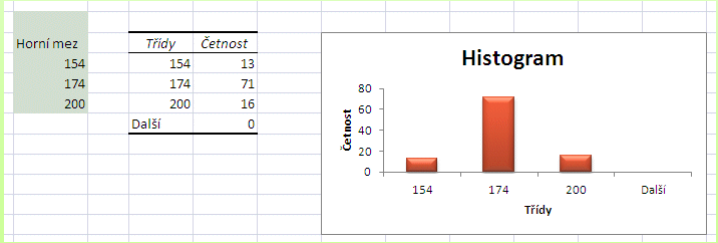
Histogram

- histogramu v Excelu – doplněk analýza dat



Histogram



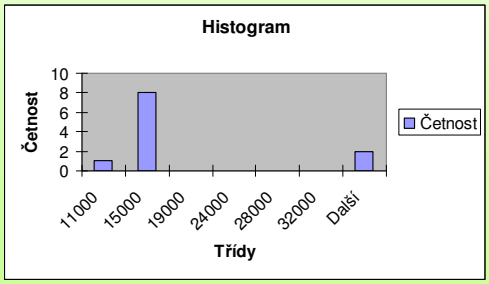


Základní popisné statistiky

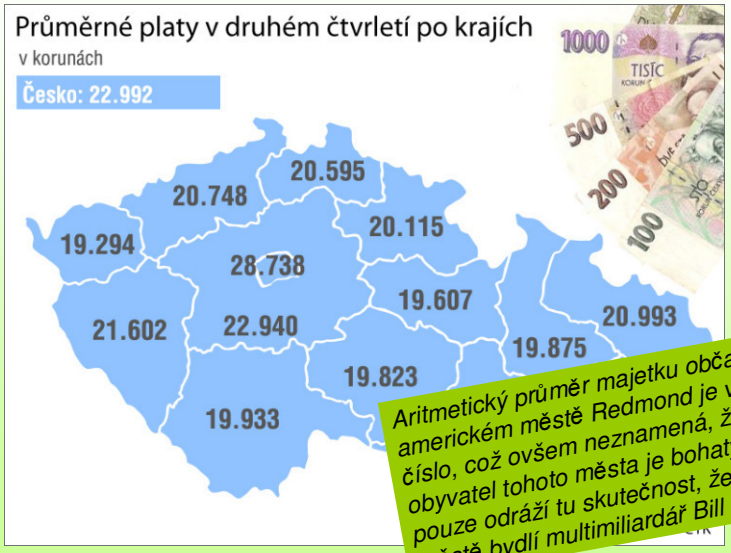
	A	B	C	D
1	134,77			
2	166,60		*Analáza dat - popisná statistika*	
3	156,34		Stř. hodnota	164,73
4	173,73		Chyba stř. hodnot	1,10
5	167,15		Medián	165,23
6	164,50		Modus	#N/A
7	161,15		Směr. odchylka	10,99
8	177,59		Rozptyl výběru	120,76
9	174,26		Špičatost	1,00
10	171,64		Šikmost	0,03
11	155,62		Minimum	134,77
12	175,76		Maximum	199,27
13	170,55		Součet	16472,85
14	165,34		Počet	100
15	159,87			

Základní popisné statistiky

- pokud mám platy v podniku:
- 14 520; 11 350; 12 645; 14 520; 13 562; 14 520; 32 458; 38 452; 10 235; 11 548;
- „průměrný plat“ = 16 824
- medián = 13 562



Základní popisné statistiky



Aritmetický průměr majetku občanů v americkém městě Redmond je velice vysoké číslo, což ovšem neznamená, že typický obyvatel tohoto města je bohatý. Tento fakt pouze odráží tu skutečnost, že v daném městě bydlí multimiliardář Bill Gates."

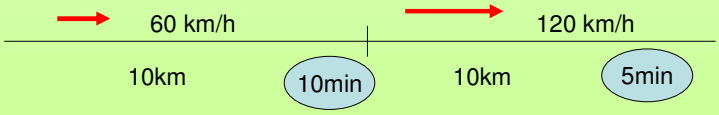
Příklad: Určete z následující tabulky průměrný počet aut, čekající na odbavení

Čas (min)	počet aut
50	0
40	1
20	2
15	3
12	4
10	5
15	6
5	7
2	8
1	9
1	10

Čas (min)	počet aut	$T_i . N_i$	
50	0	0	
40	1	40	
20	2	40	
15	3	45	
12	4	48	
10	5	50	
15	6	90	
5	7	35	
2	8	16	
1	9	9	
1	10	10	
171	55	383	2,2398

$$\bar{x} = \frac{\sum_{j=1}^{k} x_j \cdot n_j}{\sum_{j=1}^{k} n_j}$$

- Pan učitel Štoček jel s vnoučaty na výlet. První úsek své cesty dlouhý 10km jel rovnoměrnou rychlostí 60 km/h. Druhý úsek, také dlouhý 10 km jel rovnoměrnou rychlostí 120 km/h. Za jak dlouho urazil první a druhý úsek své cesty? Jakou průměrnou rychlostí se pohyboval v průběhu celé cesty?



Geometrický a harmonický průměr

Vhodnější míra polohy pro poměrné znaky. Je často používán v ekonomii a biologii (např. pro výpočet průměrného růstu).

$$\bar{x}_g = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n}$$

Jsou-li hodnoty znaku nerovnoměrně rozloženy kolem aritmetického průměru, nebo když jsou hodnoty extrémně nízké či vysoké

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

Průměr

základní soubor:

4,5,6,8,12

- průměr
  - aritmetický

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N} = \frac{4+5+6+8+12}{5} = \frac{35}{5} = 7$$

  - geometrický

$$\sqrt[N]{\prod_{i=1}^{N} X_i} = \sqrt[5]{4\times5\times6\times8\times12} = 6{,}49$$

  - harmonický

$$\frac{1}{\frac{1}{N}\sum_{i=1}^{N}\frac{1}{X_i}} = \frac{1}{\frac{1}{5}\left(\frac{1}{4}+\frac{1}{5}+\frac{1}{6}+\frac{1}{8}+\frac{1}{12}\right)} = 6{,}06$$

Charakteristiky polohy

Cíl je jedním číslem charakterizovat velikost všech číselných hodnot ve statistickém souboru.

Charakteristiky polohy nám umožňují srovnávat úroveň zkoumaného jevu u dvou nebo více souborů.

- **aritmetický průměr**
- **medián**
- **modus**.

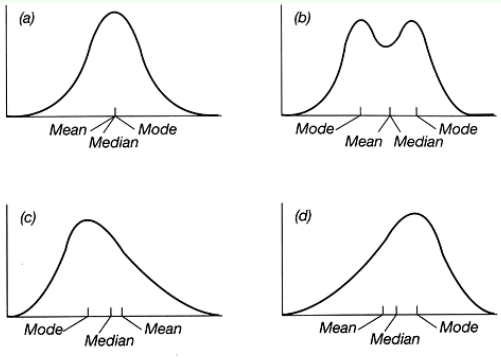


Figure 3.2 Frequency distributions showing measures of central tendency. Values of the variable are along the abscissa (horizontal axis), and the frequencies are along the ordinate (vertical axis). Distributions (a) and (b) are symmetrical, (c) is positively skewed, and (d) is negatively skewed. Distributions (a), (c), and (d) are unimodal, and distribution (b) is bimodal. In a unimodal asymmetric distribution, the median lies about one-third the distance between the mean and the mode.*

Míry rozptýlení

- **rozptyl** (variance)
  - průměrná hodnota druhé mocniny odchylky od průměru
- **směrodatná odchylka**
  - odmocnina z rozptylu
  - čím menší, tím nižší variabilita dat
- **variační koeficient** – porovnává variabilitu nestejně velkých objektů (myš a slon) – bezrozměrné číslo

$$CV = \frac{\sigma}{\overline{X}}$$

základní soubor:

4,5,6,8,12

průměr = 7

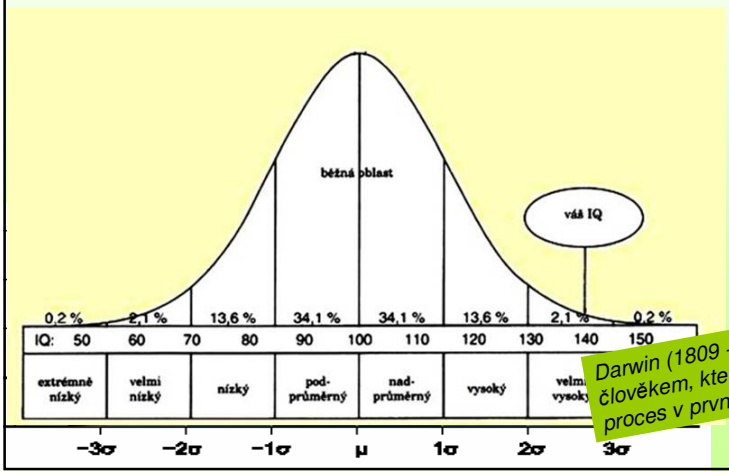
$$\sigma^2 = \frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}$$

$$\sigma^2 = \frac{(4-7)^2 + (5-7)^2 + (6-7)^2 + (8-7)^2 + (12-7)^2}{5} = \frac{9+4+1+1+25}{5} = 8$$

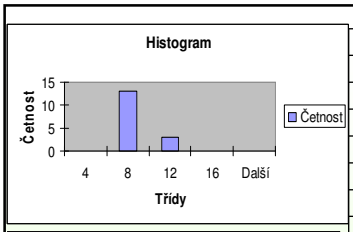
$$\sigma = \sqrt{8} = 2{,}83$$

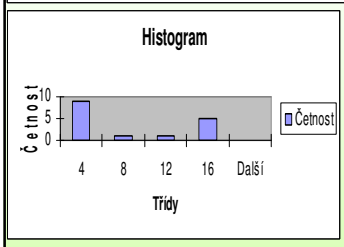
- směrodatná odchylka
  - empirické pravidlo: většina hodnot se neodlišuje od průměru o více než jednu směrodatnou odchylku a skoro všechny hodnoty jsou v pásmu do dvou směrodatných odchylek od průměru.

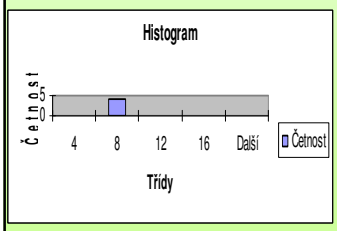
normální rozdělení:



Darwin (1809 - 1882) byl prvním člověkem, který evoluci vnímal jako proces v první řadě statistický.







	8	4	8
	7	2	7
	8	2	8
	8	4	8
	7	2	
	8	4	
	8	4	
	9	3	
	8	4	
	6	6	
	9	15	
	8	14	
	7	16	
	8	12	
	6	16	
	9	16	
průměr	**7.75**	**7.75**	**7.75**
sm. odchylka	**0.90**	**5.64**	**0.43**
modus	8	4	8
medián	8	4	8
sm. odchylka výběru	**0.93**	**5.83**	**0.50**
n	**16**	**16**	**4**

vždy uvádějte:

n, průměr,

sm. odchylku

ostatní podle potřeby