# Základy zpracování geologických dat

R. Čopjaková

- Předmět je zaměřen na získání teoretických základů statistické analýzy numerických dat v geologických vědách a její praktické provádění pomocí programu Microsoft Excel

# Zpracování geologických dat

- **Úvod.** Sběr dat. Analýza a výběr dat. Vlastní zpracování dat, grafická prezentace.
- **Popis jednorozměrných statistických souborů.** Náhodný výběr, Uspořádání dat zákl. souboru - rozdělení četností. Četnost absolutní, relativní, kumulativní.
- **Základní typy rozdělení četností** – rozdělení četností u geologických jevů.
- **Základní statistické charakteristiky.** Míry polohy - aritmetický průměr, medián, kvantily, modus; Míry variability - rozptyl, směrodatná odchylka, variační rozpětí; bodové a intervalové odhady.
- **Testování statistických hypotéz** - Základní pojmy a postup testování. Základní parametrické a neparametrické testy.
- **Vzájemné vztahy veličin -** Regresní analýza a korelační analýza.

# Doporučená literatura


- Brázdil, Rudolf - Kolář, Miroslav - Prošek, Pavel. *Statistické metody v geografii*. Brno : Masarykova univerzita Brno, 1993. 177 s.
- Brázdil, Rudolf. *Statistické metody v geografii :cvičení*. 3. vyd. Brno : Vydavatelství Masarykovy univerzity, 1995. 177 s.
- Sattran, Vladimír - Soukup, Blahomil. *Použití matematických metod v geologii*. Vyd. 1. Praha : Ústřední ústav geologický v Academii, 1973. 153 s.
- *Biostatistika*. Edited by Karel Zvára. 1. vyd. Praha : Univerzita Karlova-Vydavatelství Karolinum, 2001. 210 s.
- Hanousek, Jan - Charamza, Pavel. *Moderní metody zpracování dat :matematická statistika pro každého*. 1. vyd. Praha : Grada, 1992. 210 s.

- Při statistickém zkoumání nás zajímají hromadné jevy a procesy, u kterých zkoumáme zákonitosti, které se projevují u velkého počtu prvků.
  - Petrologie - celohorninové analýzy, mineralogie - analýzy minerálů
  - Geochemie, hydrologie – kontaminace půd, vod atd.
  - Pórovitost, hustota hornin
  - Měření geologickým kompasem
  - Měření morfologických parametrů na schránkách organismů
- Prvky zkoumání nazýváme statistické jednotky.
- Pozorovaním nebo měřením hodnot zkoumaného znaku (veličiny) na několika statistických jednotkách získáme datový soubor.

- **Statistický soubor jednorozměrný**, jestliže sledujeme jeden znak – stanovení stáří, pórovitost
  nebo **vícerozměrný**, pokud sledujeme více znaků - celohorninové analýzy, chemické analýzy minerálů

- statistické znaky:
  - kvantitativní, popsané číselnou hodnotou (průtok, stáří, hustota);
    kvantitativní pořadové – např. stupeň vybělení horniny
  - kvalitativní, popsané vlastnostmi (barva)

**Statistický soubor: z pohledu úplnosti**

- **základní soubor** je soubor všech statistických jednotek
- **výběrový soubor** je vybraná část ze základního souboru

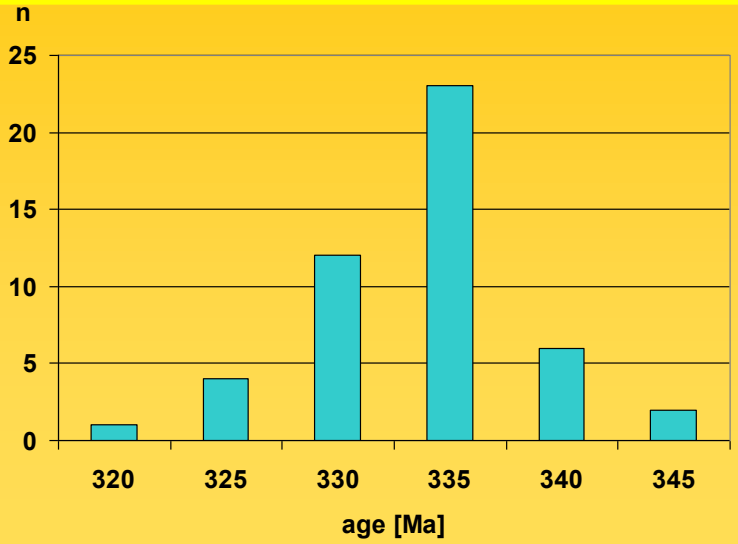
- **Základní soubor** není vždy k dispozici (např. změřit všechny objekty je časově nebo finančně neúnosné nebo nemožné).
- Data pak zobrazují jen část objektů (**výběrový soubor**), avšak my chceme získat obraz o parametrech celého základního souboru. Z výběrového souboru samozřejmě nemůžeme určit přesné parametry základního souboru, ale pouze jejich **odhady**.

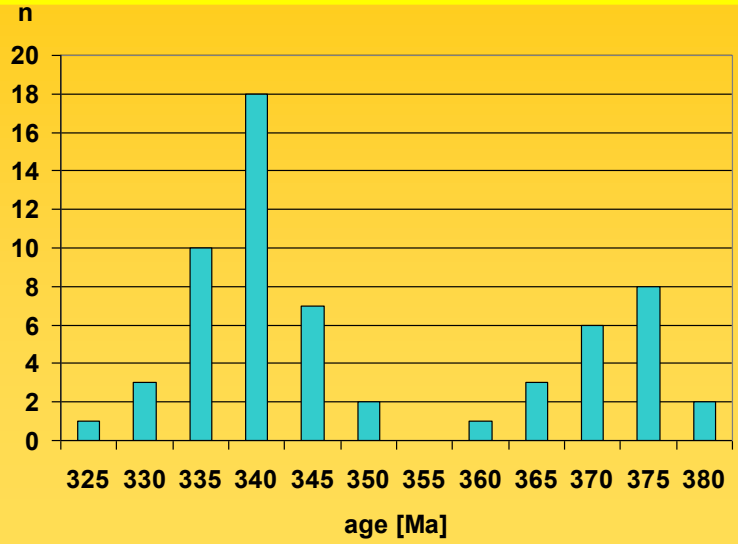
- rozsah základního (výběrového) souboru je počet jednotek v souboru; n = počet statistických jednotek

# Zpracování kvantitativních dat

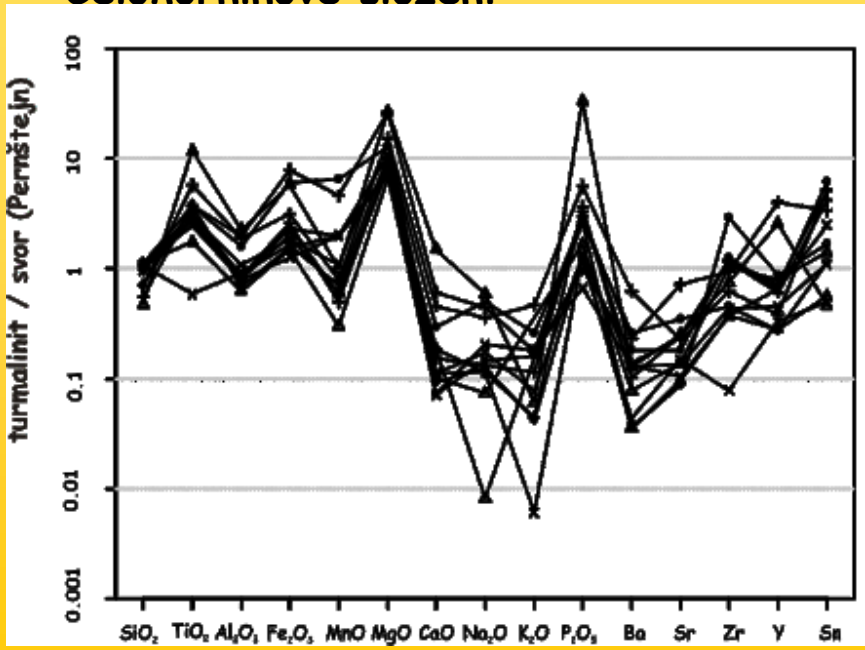
- Grafické zpracování – správné čtení a interpretace
  - Funkce – lineární, logaritmické, exponenciální
- Početní – míry polohy a variability – např. aritmetický průměr, směrodatná odchylka, minimum, maximum,
- Vzájemné vztahy a závislosti

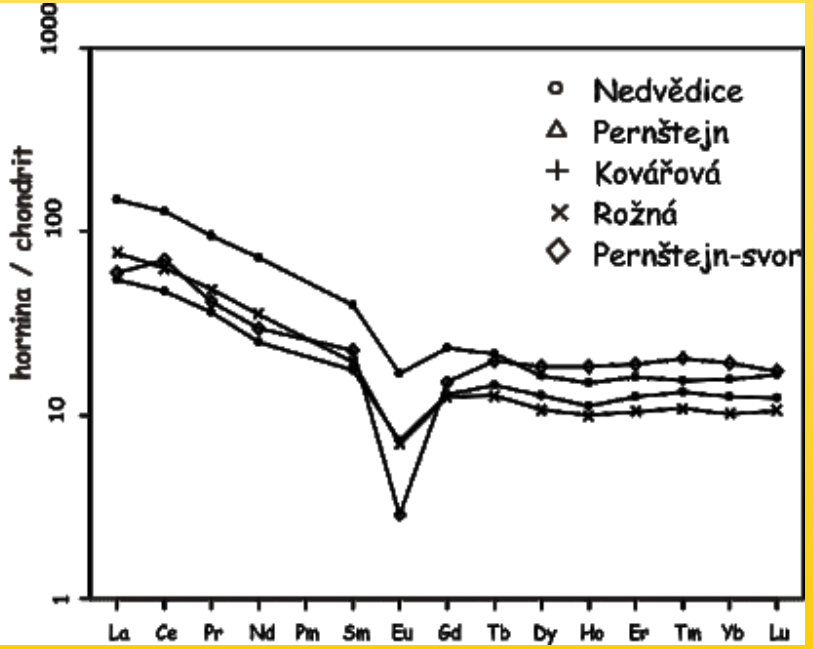
# Histogram – stáří metamorfovaných hornin



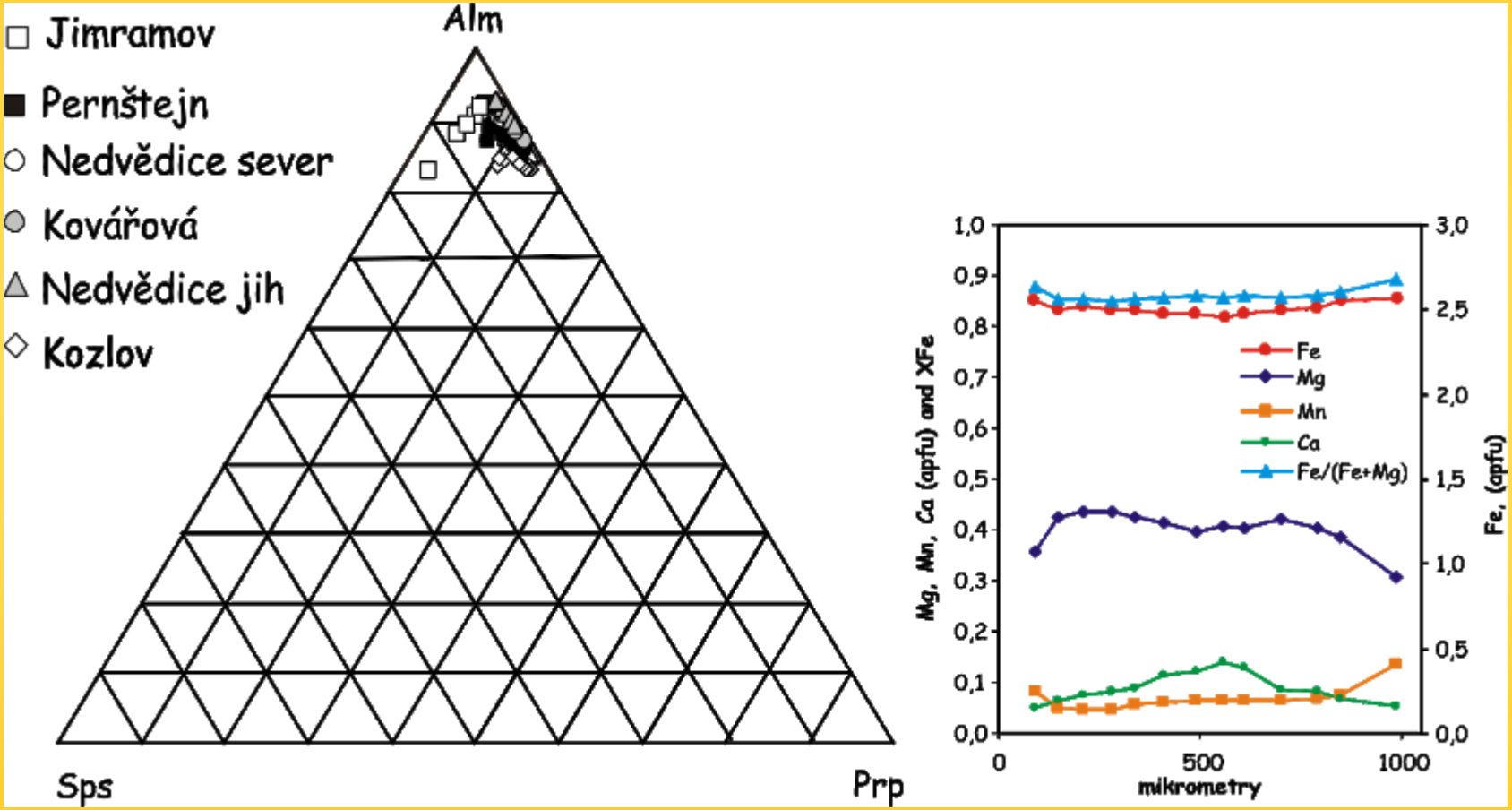




## Celohorninové složení

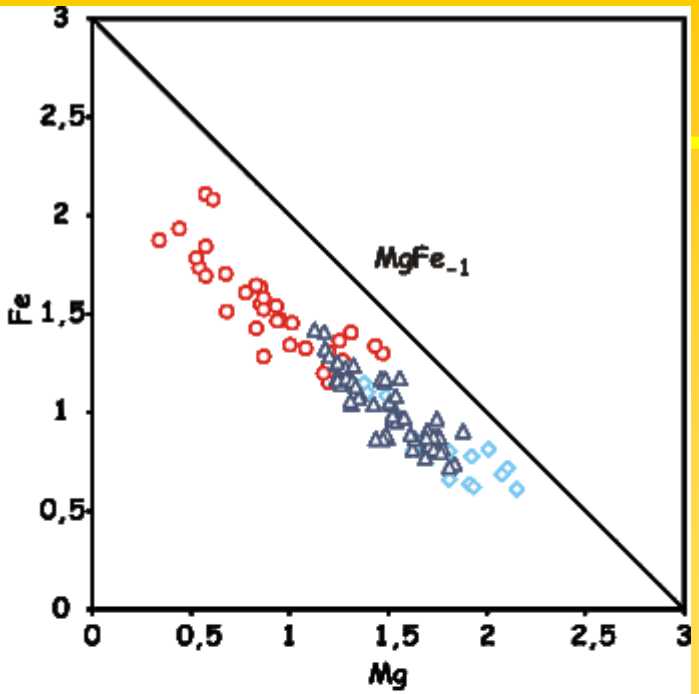


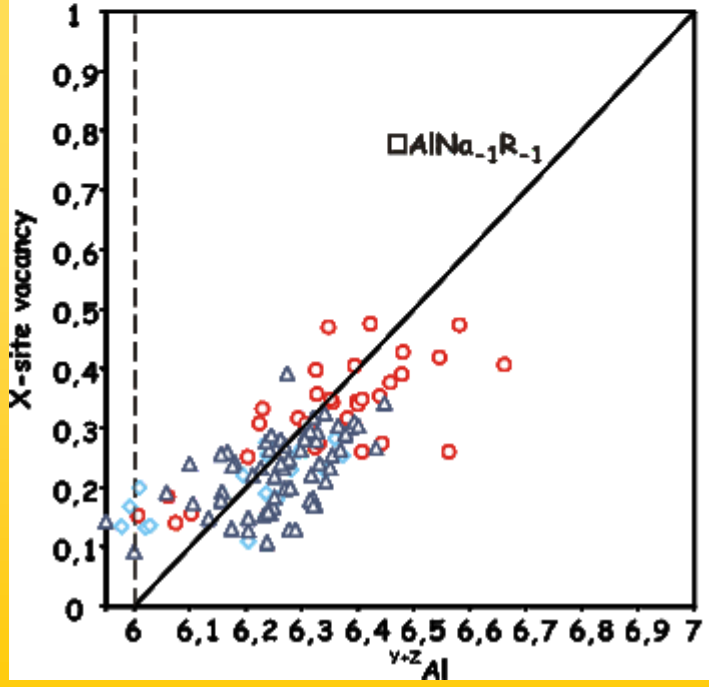



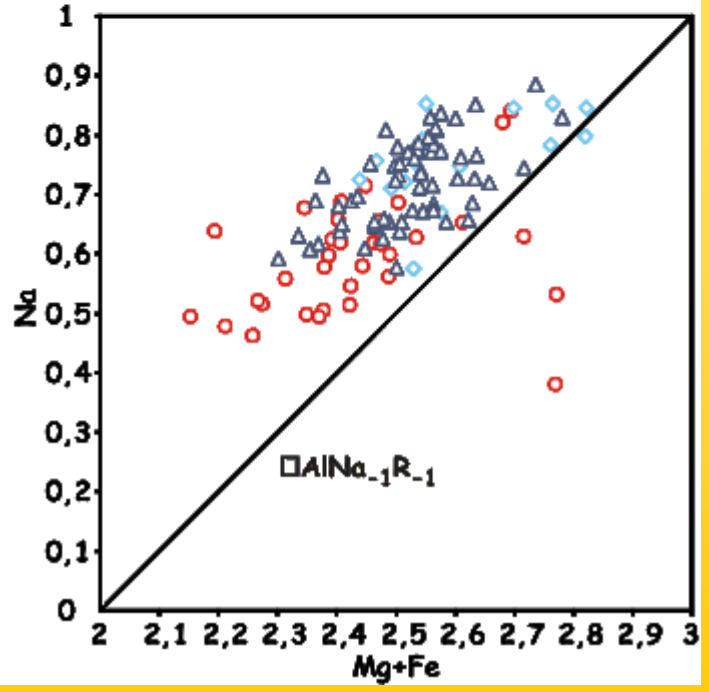
# Chemické složení a zonálnost granátu



Su


MgFe_-1
Fe
Mg
0
0,5
1
1,5
2
2,5
3


□AlNa_-1R_-1
X-site vacancy
y+z Al
0
0,1
0,2
0,3
0,4
0,5
0,6
0,7
0,8
0,9
1
6
6,1
6,2
6,3
6,4
6,5
6,6
6,7
6,8
6,9
7


□AlNa_-1R_-1
Na
Mg+Fe
0
0,1
0,2
0,3
0,4
0,5
0,6
0,7
0,8
0,9
1
2
2,1
2,2
2,3
2,4
2,5
2,6
2,7
2,8
2,9
3

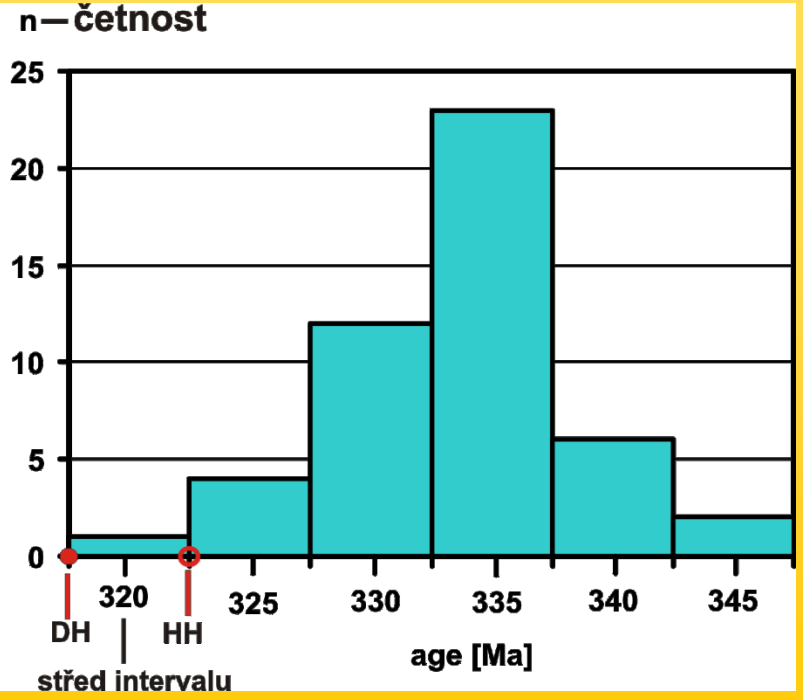
# Zpracování kvantitativních dat-jednorozměrné soubory

## Tvorba histogramu

- soubor dat: x1, x2, . . . , xn,
- soubor uspořádáme podle velikosti
- stanovení intervalů
- dolní hranice třídy
- horní hranice třídy
- střed třídy je průmer horní a dolní hranice třídy
- šírka třídy je rozdíl horní a dolní hranice trídy

Máme soubor 48 analýz - výsledky datování monazitu

| střed int | dolní hranice | horní hranice | četnost abs. |
|---|---|---|---|
| 320 | 317,5 | 322,5 | 1,00 |
| 325 | 322,5 | 327,5 | 4,00 |
| 330 | 327,5 | 332,5 | 12,00 |
| 335 | 332,5 | 337,5 | 23,00 |
| 340 | 337,5 | 342,5 | 6,00 |
| 345 | 342,5 | 347,5 | 2,00 |



# Tvorba histogramu

- najít logické hledisko pro stanovení šířky intervalu (třídy) nebo počtu intervalů
- šířka intervalů nemusí být kontantní – často zejména krajní intervaly jsou širší, případně neomezené
- počet intervalů musí být takový, aby vynikly podstatné a charakteristické rysy souboru
- jednoznačnost přiřazení statistických jednotek do určité třídy

- Pravidla pro stanovení šířky či počtu intervalů:

  Sturgesovo pravidlo $K = 1 + 3{,}3 \log n$

  $k = \sqrt{n}$

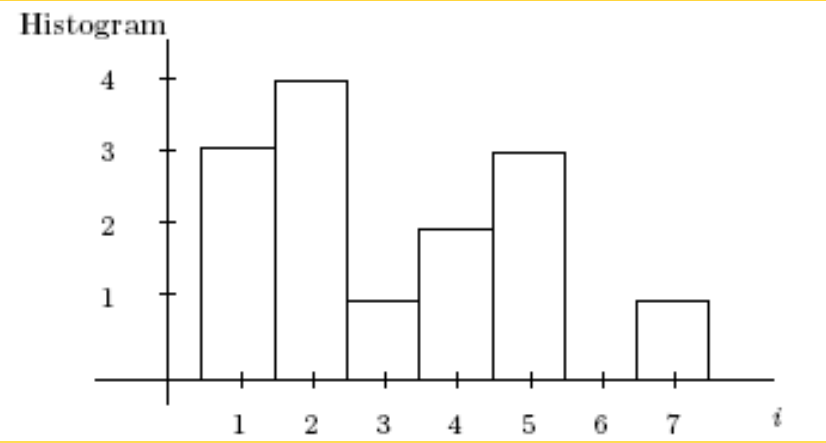
  k=celá část (5*log n)

  kde k je počet intervalů a n je rozsah souboru

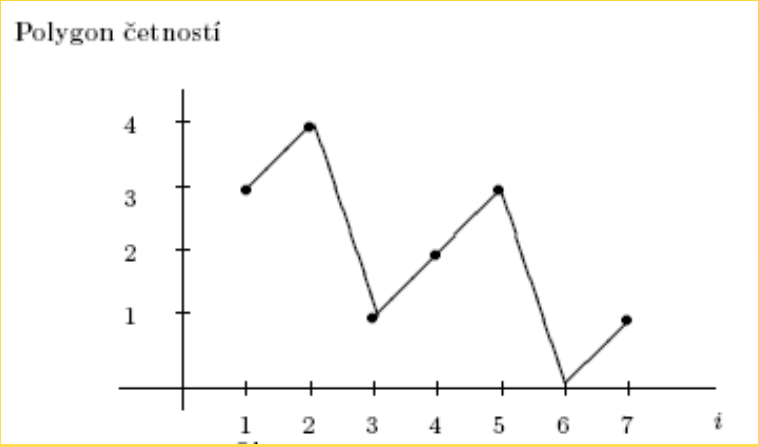
  $0{,}05R \leq h \leq 0{,}08R$

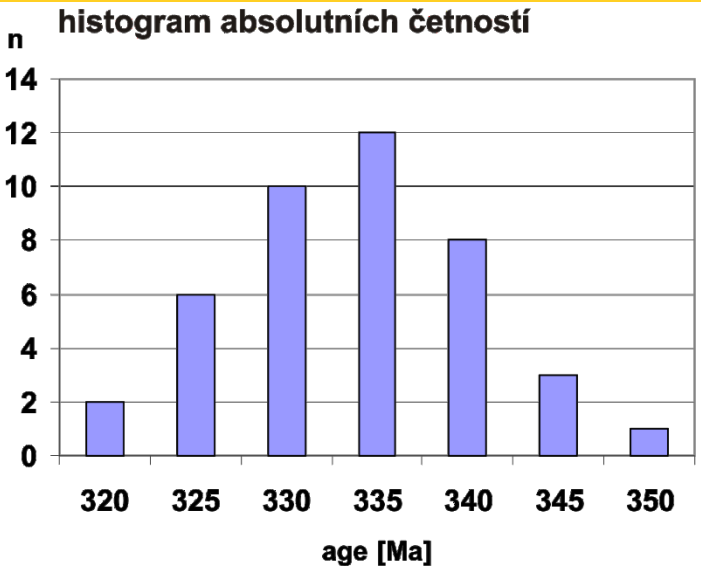
  kde h je šířka intervalu a R variační rozpětí tj. $R = X_{max} - X_{min}$

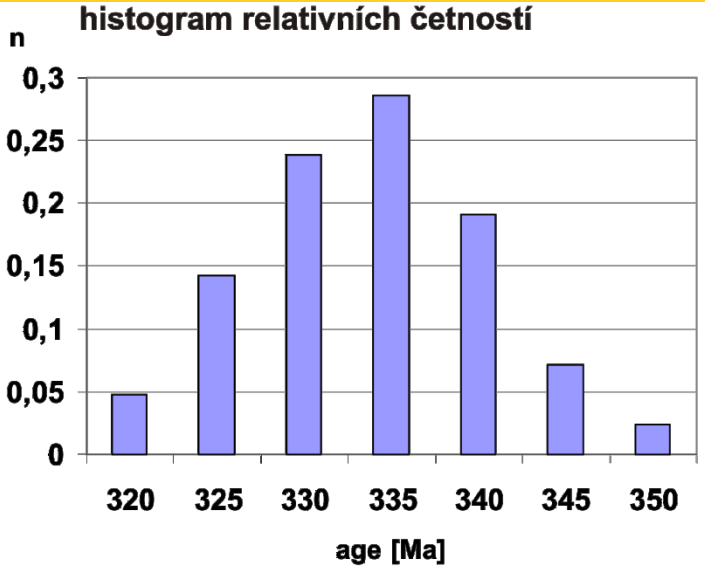
**histogram (sloupcový graf)**

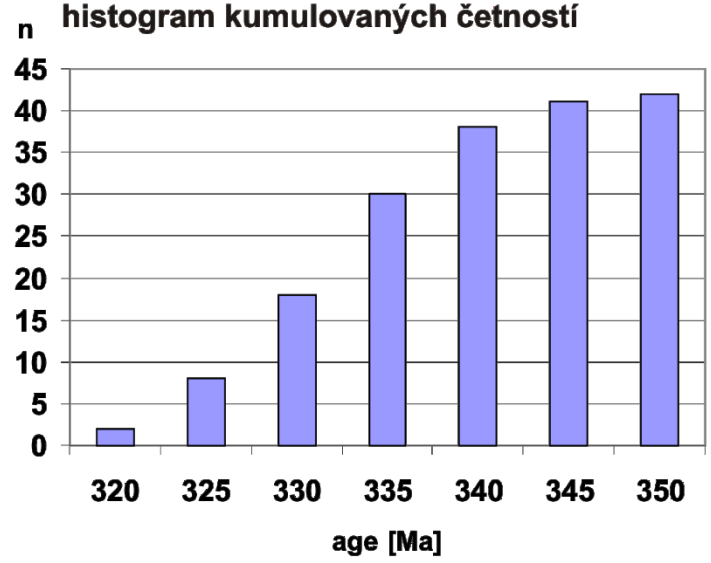


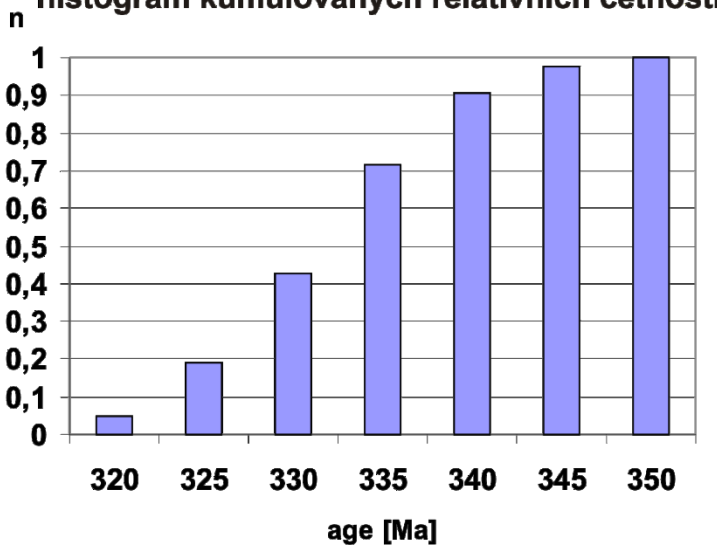

**polygon četností (spojnicový graf)**




histogram absolutních četností
n
14
12
10
8
6
4
2
0
320
325
330
335
340
345
350
age [Ma]


histogram relativních četností
n
0,3
0,25
0,2
0,15
0,1
0,05
0
320
325
330
335
340
345
350
age [Ma]

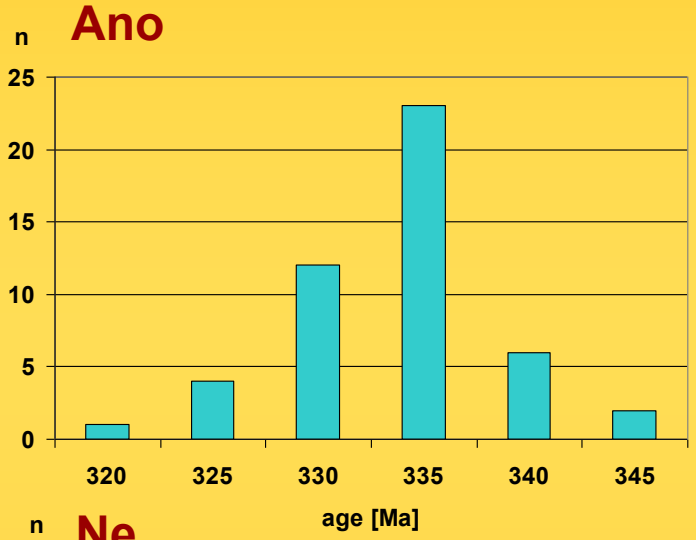

histogram kumulovaných četností
n
45
40
35
30
25
20
15
10
5
0
320
325
330
335
340
345
350
age [Ma]

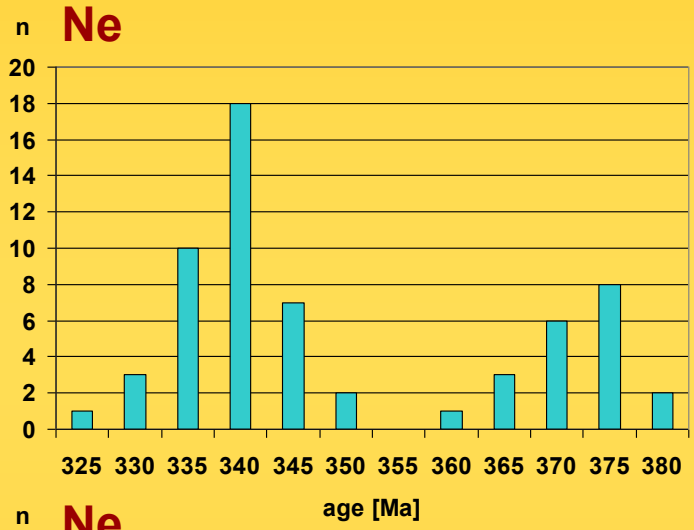

histogram kumulovaných relativních četností
n
1
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
320
325
330
335
340
345
350
age [Ma]

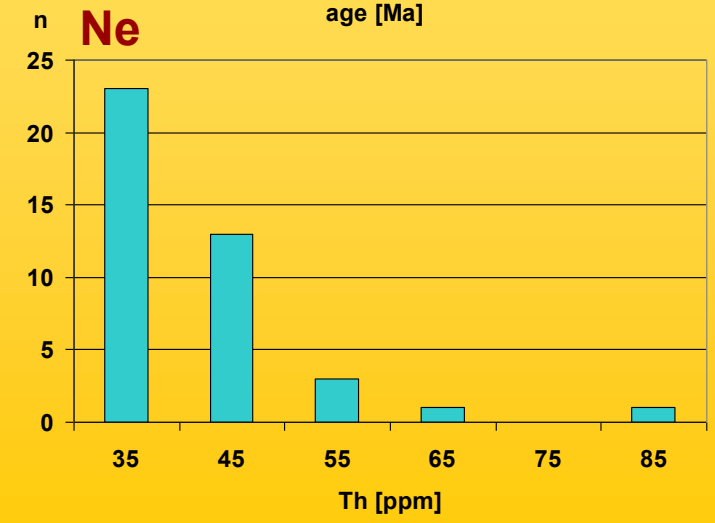
# Charakteristiky (míry) polohy

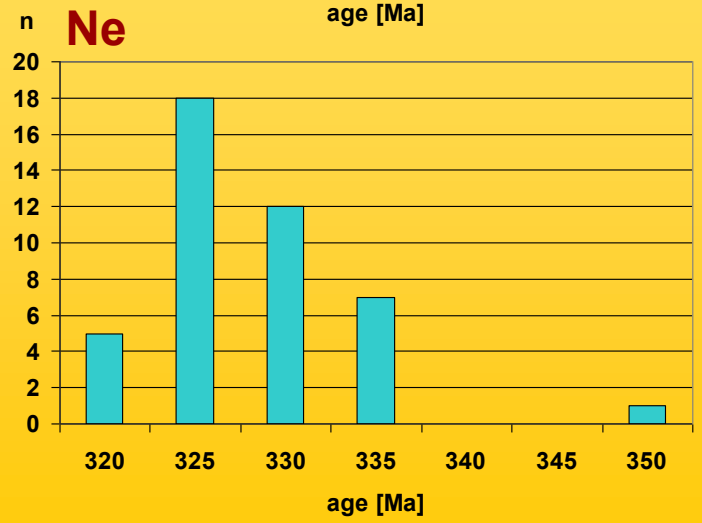
Nejznámější a nejčastěji používanou charakteristkou polohy je aritmetický průměr hodnot souboru.

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



# Charakteristiky (míry) polohy

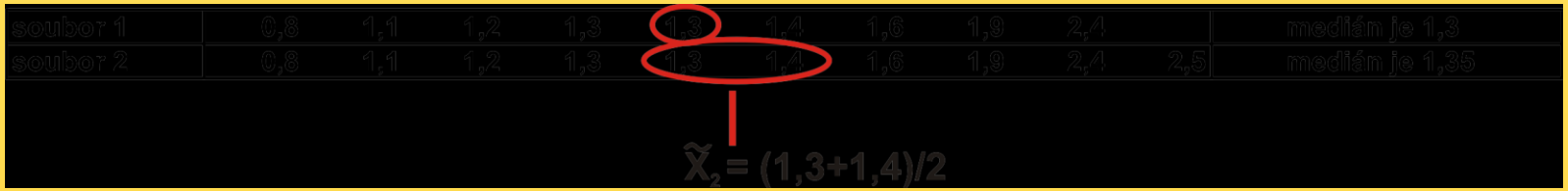
**Kvantil** - dělí soubor seřazených hodnot na několik stejně velkých částí. Kvantily tvoří inverzní funkci k funkci distribuční.

**Speciální označení kvantitů**

- **Medián -** je hodnota, jež dělí soubor dat seřazených podle velikosti na dvě stejně početné poloviny. Platí, že nejméně 50 % hodnot je menších nebo rovných a nejméně 50 % hodnot je větších nebo rovných mediánu. Pro nalezení mediánu daného souboru stačí hodnoty seřadit podle velikosti a vzít hodnotu, která se nalézá uprostřed seznamu. Pokud má soubor sudý počet prvků, obvykle se za medián označuje aritmetický průměr dvou hodnot na místech $n/2$ a $n/2+1$.

. **Výhody mediánu**

- Základní výhodou mediánu jako statistického ukazatele je fakt, že není ovlivněný extrémními hodnotami (nízkými či vysokými). Proto se často používá v případě šikmých rozdělení, u kterých aritmetický průměr dává obvykle nevhodné výsledky.




- Medián je nejpoužívanější kvantil (konkrétně kvantil dělící soubor na dvě části).
- Kromě mediánu se velmi často používají *kvartily* (soubor se dělí na čtyři části), *decily* (na deset částí) a *percentily* (na sto částí).

**Modus**

Nejčetnější hodnota souboru – užití  např. u bimodálních rozdělení četností

# Charakteristiky (míry) variability-rozptýlenosti

- variační rozpětí $R = x_{max} - x_{min}$
- mezikvartilové rozpětí $IQR = \tilde{x}_{75} - \tilde{x}_{25}$
- rozptyl - střední kvadratická odchylka od průměru
  - rozptyl (základní soubor) $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$
  - výběrový rozptyl (výběrový soubor) $S_X^2 = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \bar{X})^2$
- směrodatná odchylka – odmocnina z rozptylu; nejužívanější míra variability; vyjadřuje rozkolísanost hodnot kolem střední hodnoty
  - směrodatná odchylka (základní soubor) $\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2} = \sqrt{\left(\frac{1}{N}\sum_{i=1}^{N}x_i^2\right) - \overline{x}^2}$
  - výběrová směrodatná odchylka - pro skutečný výpočet odhadu směrodatné odchylky na empiricky zjištěné řadě čísel (výběrovém souboru) $s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2}.$