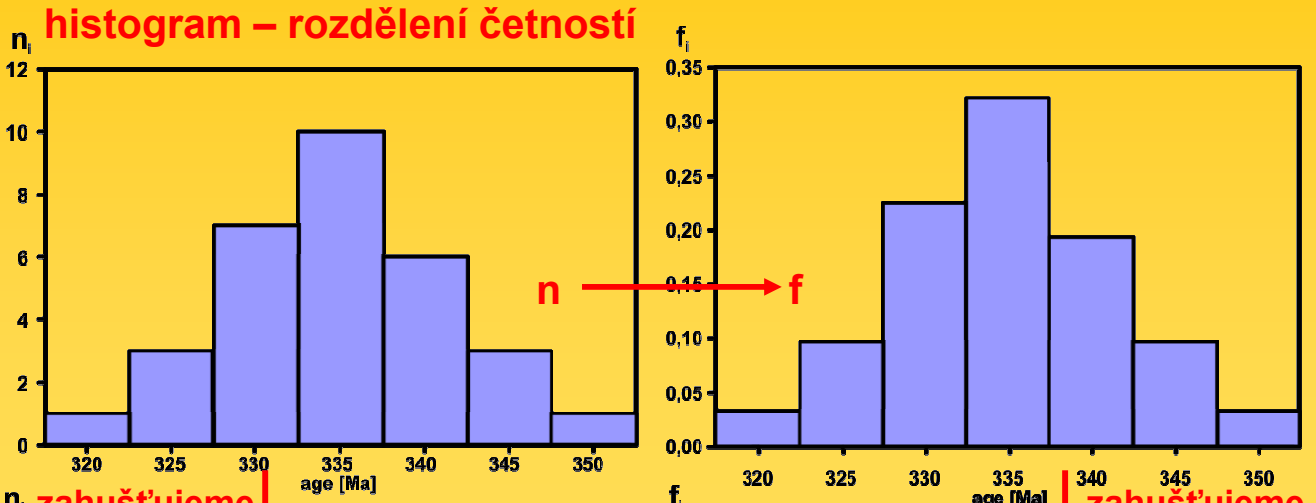
Základy zpracování geologických dat

Od histogramu (rozdělení četností) k rozdělení pravděpodobnosti a základní charakteristiky polohy a variability

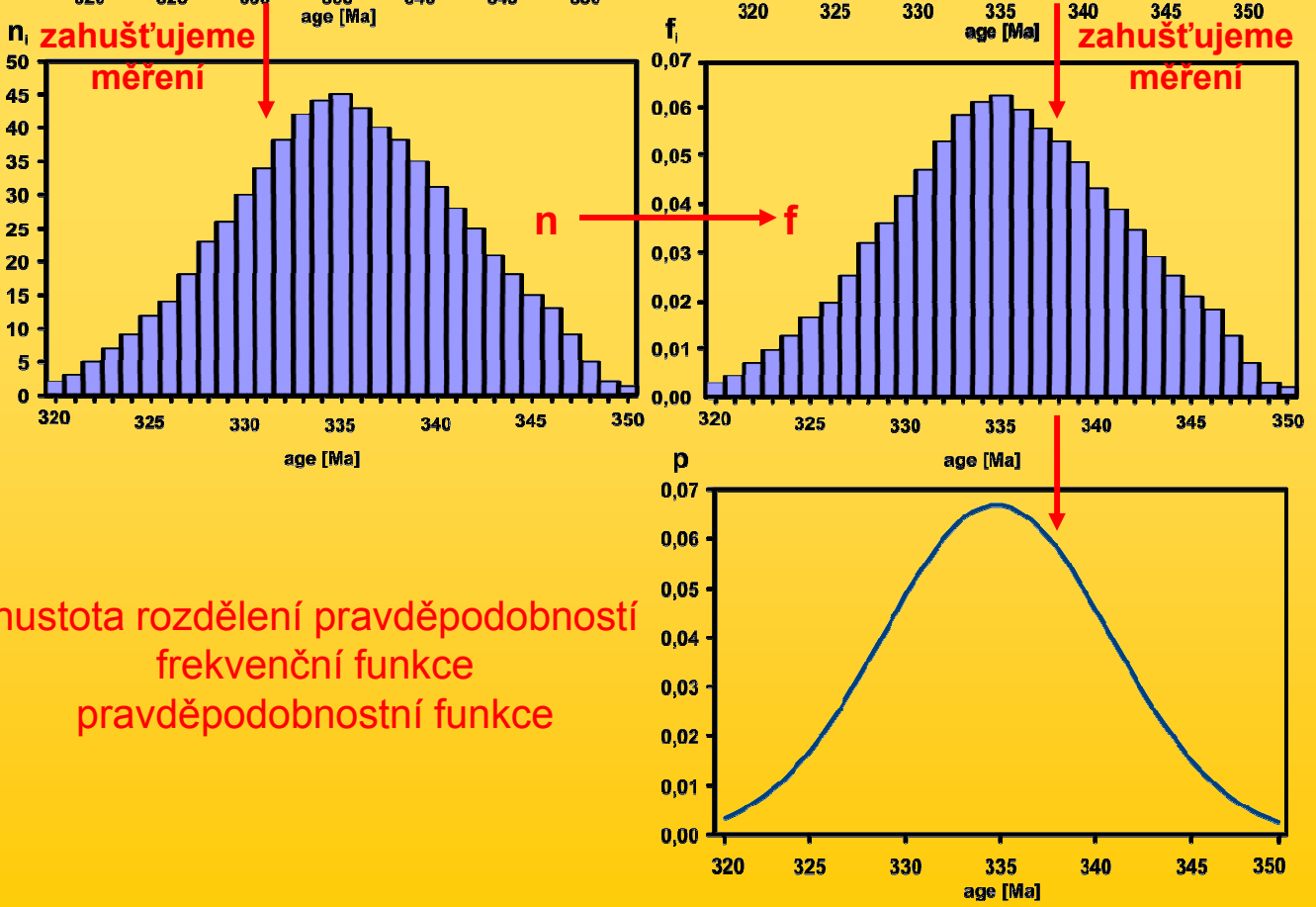
R. Čopjaková

Od histogramu (rozdělení četností) k rozdělení pravděpodobnosti

střed int	n_i	f_i
320	1	0,03
325	3	0,10
330	7	0,23
335	10	0,32
340	6	0,19
345	3	0,10
350	1	0,03
suma	31	1



střed int	n_i	f_i
320	2	0,00
321	3	0,00
322	5	0,01
323	7	0,01
324	9	0,01
325	12	0,02
326	14	0,02
327	18	0,03
328	23	0,03
329	26	0,04
330	30	0,04
331	34	0,05
332	38	0,06
333	42	0,06
334	44	0,07
335	45	0,07
336	43	0,06
337	40	0,06
338	38	0,06
339	35	0,05
340	31	0,05
341	28	0,04
342	25	0,04
343	21	0,03
344	18	0,03
345	15	0,02
346	13	0,02
347	9	0,01
348	5	0,01
349	2	0,00
350	1	0,00
suma	676	1

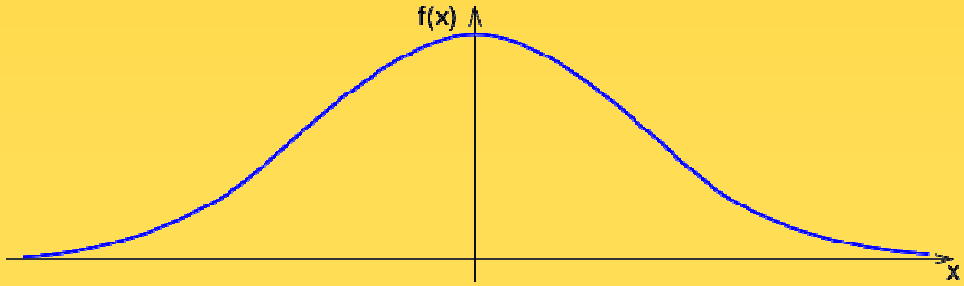


hustota rozdělení pravděpodobností
frekvenční funkce
pravděpodobnostní funkce

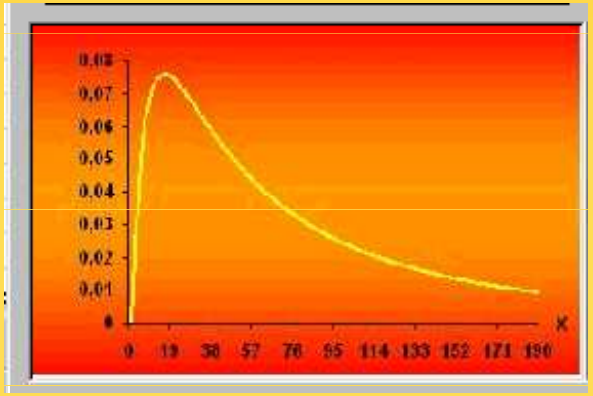
Rozdělení pravděpodobnosti

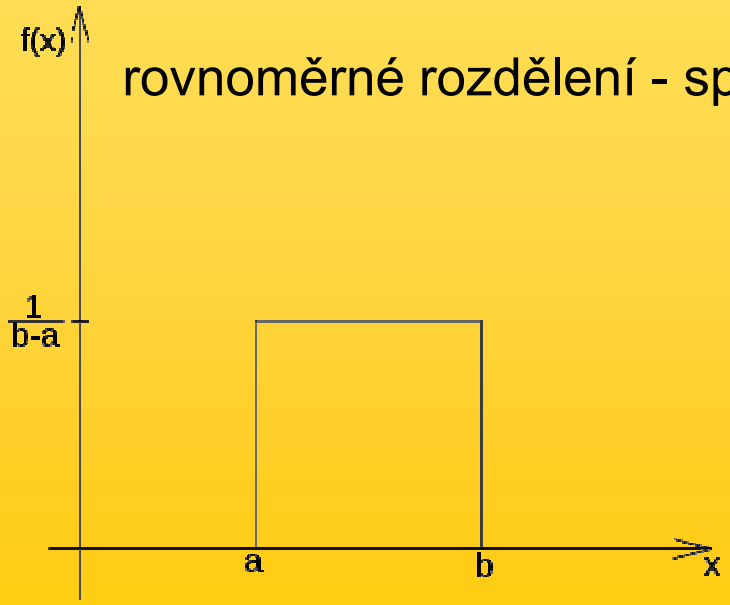
- Spojité náhodné veličiny
- Diskrétní náhodné veličiny

normální rozdělení - spojité



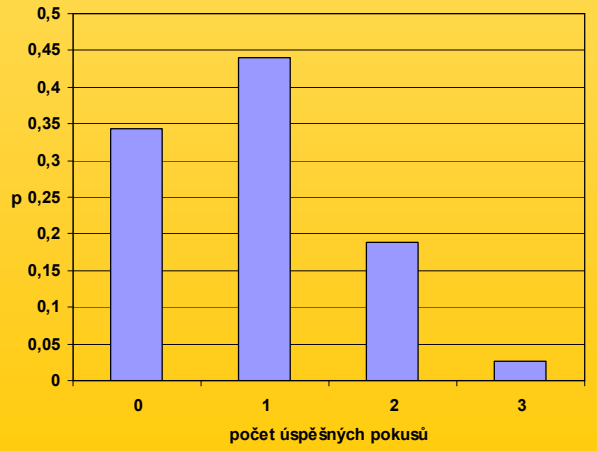
lognormální rozdělení - spojité





rovnoměrné rozdělení - spojité

binomické rozdělení - nespojité



Numerická charakteristika souborů dat

- Charakteristiky (míry) polohy
  - charakterizují střední hodnotu souboru dat (např. aritmetický průměr)
- Charakteristiky (míry) variability
  - charakterizují rozptýlenost, variabilitu dat, (např. variační rozpětí)

- míry parametrické
  - počítané ze všech hodnot souboru
  - závislé na typu rozdělení pravděpodobností
- míry neparametrické
  - počítané jen z některých hodnot souboru
  - univerzálně použitelné pro různé typu rozdělení pravděpodobností

Charakteristiky (míry) polohy

Střední hodnoty

- aritmetický průměr - (parametrická míra) normální rozdělení pravděpodobností

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

- geometrický průměr - (parametrická míra) lognormální rozdělení pravděpodobností
  - n-tá odmocnina součinu hodnot souboru

$$G(x_1, x_2, \ldots, x_n) = \sqrt[n]{x_1 \cdot x_2 \cdots x_n} = \left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}}$$

- medián - (neparametrická míra) je hodnota, jež dělí soubor dat seřazených podle velikosti na dvě stejně početné poloviny.

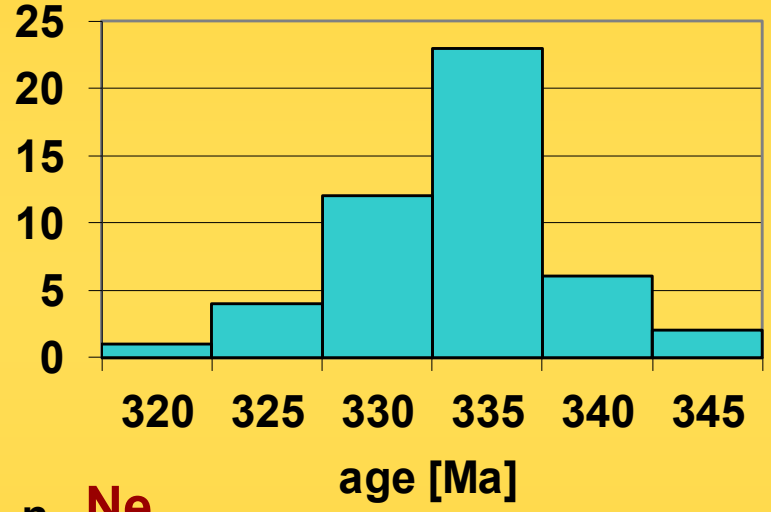
- modus - (neparametrická míra) - nejčetnější hodnota souboru – např. u bimodálních rozdělení četností

Charakteristiky (míry) polohy

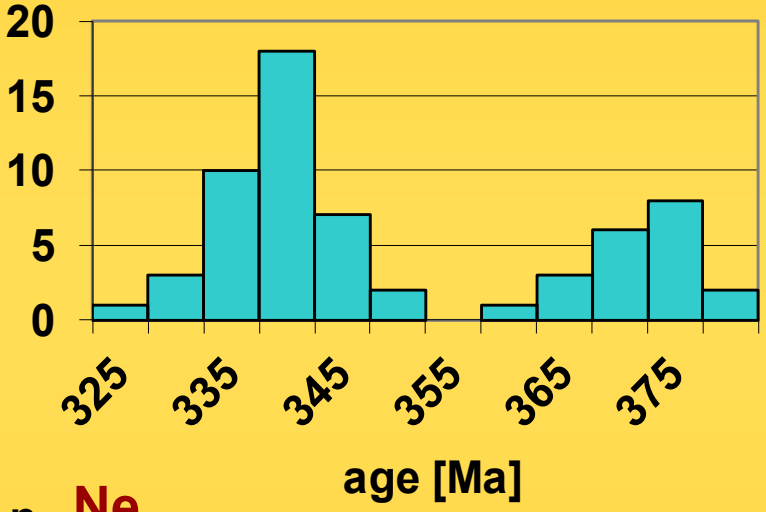
Nejznámější a nejčastěji používanou charakteristkou polohy je aritmetický průměr hodnot souboru.

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

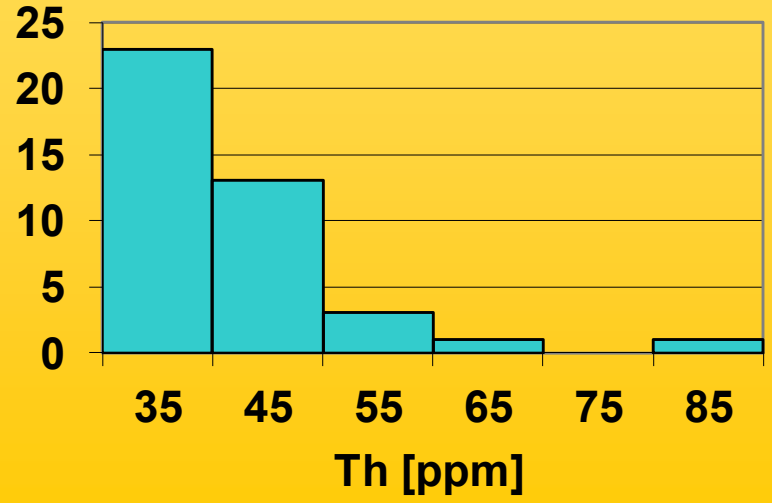
Ano



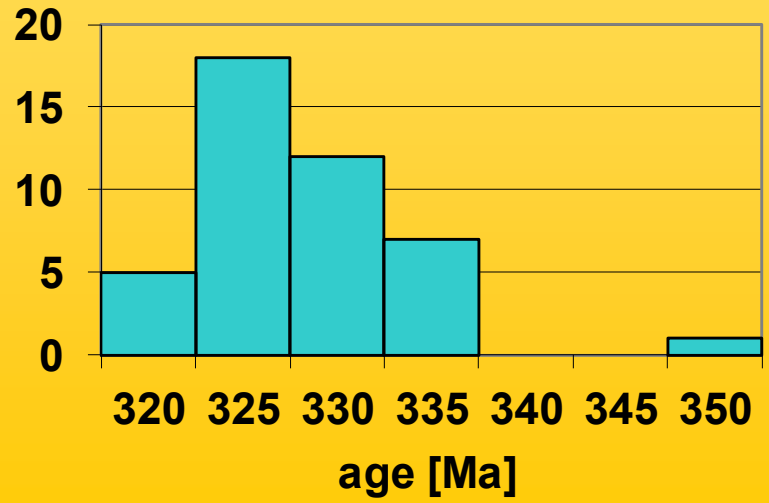
Ne



Ne



Ne



Charakteristiky (míry) polohy

- **Medián – (neparametrická míra polohy)**
- **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.

soubor 1	0,8	1,1	1,2	1,3	1,3	1,4	1,6	1,9	2,4	medián je 1,3
soubor 2	0,8	1,1	1,2	1,3	1,3	1,4	1,6	1,9	2,4	2,5 medián je 1,35

$\tilde{x}_{0,5} = (1,3+1,4)/2$

- Kvantil - udávající hodnotu, kterou stanovená část p (z intervalu ⟨ 0 ; 1 ⟩ nebo v procentech v rozmezí 0–100 %) hodnot nepřesahuje
- **Medián jako kvantil**
- 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).

Statistické funkce v Excelu

Charakteristiky polohy

- AVERAGEA/průměr = spočte aritmetický průměr souboru
  - zadám oblast dat, z níž má průměrnou hodnotu spočítat
- GEOMEAN = spočte geometrický průměr souboru
  - zadám oblast dat, z níž má průměrnou hodnotu spočítat
- MEDIAN = stanoví medián pro soubor dat
  - zadám oblast dat, z níž má medián stanovit

stanovení daného kvantilu/percentilu

- PERCENTIL.EXC = stanoví hodnotu k-tého percentilu; k je $\in (0, 1)$
- PERCENTIL.INC = stanoví hodnotu k-tého percentilu; k je $\in \langle 0, 1 \rangle$
  - pole = oblast dat, z níž má k-tý percentil stanovit
  - k = hodnota percentilu, který chceme stanovit
    např. pro 1. kvartil k=0,25; pro 3. kvartil k=0,75; pro medián k=0,5
- MODE.SNGL = stanoví modus pro soubor dat
  - zadám oblast dat, z níž mám modus stanovit

Charakteristiky (míry) variability-rozptýlenosti

- **minimem a maximem souboru** (stat. fce v Excelu – MIN a MAX)
- **variační rozpětí** (neparametrická míra) $R = x_{max} - x_{min}$
- **desátým a devadesátým percentilem** (neparametrické míry)
  - (stat . fce v Excelu – PERCENTIL, k=0,1 a k=0,9); pokud velké soubory dat; nebo výsledky metod, kde je velká chyba stanovení, např. výsledky z LA-ICP-MS
- **mezikvartilové rozpětí** (neparametrická míra)
  - rozdíl mezi hodnotou třetího a prvního kvartilu
  - (stat . fce v Excelu – PERCENTIL - k=0,25; k=0,75 a jejich rozdíl)

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

Charakteristiky (míry) variability-rozptýlenosti

- **rozptyl** = průměrný čtverec odchylky jednotlivých hodnot souboru od aritmetického průměru; (parametrická míra) – pro soubor dat s normálním rozdělením

rozptyl (základní soubor)

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

rozptyl (výběrový soubor – tzv. odhad rozptylu)

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

- **směrodatná odchylka** = odmocnina z rozptylu (parametrická míra) – pro soubor dat s normálním rozdělením

směrodatná odchylka (základní soubor)

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

směrodatná odchylka (výběrový soubor – tzv. odhad směrodatné odchylky)

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

Statistické funkce v Excelu

Charakteristiky variability

- **VAR.P/VAR** = vypočte rozptyl základního souboru
  - zadám oblast dat, z níž má průměrnou hodnotu spočítat

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

- **VAR.S/VAR.VÝBĚR** = odhadne rozptyl základního souboru (pracujeme-li s výběrovým souborem)
  - zadám oblast dat, z níž má průměrnou hodnotu spočítat

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

- **SMODCH.P/SMODCH/STDEV** = vypočte směrodatnou odchylku základního souboru
  - zadám oblast dat, z níž má směrodatnou odchylku spočítat

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

- **SMODCH.VÝBĚR.S/SMODCH.VÝBĚR** = odhadne směrodatnou odchylku základního souboru; (pracujeme-li s výběrovým souborem)
  - zadám oblast dat, z níž má směrodatnou odchylku spočítat

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