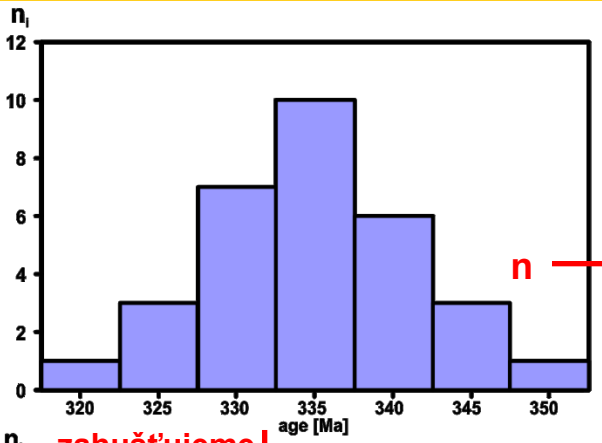
Základy zpracování geologických dat

Rozdělení pravděpodobnosti

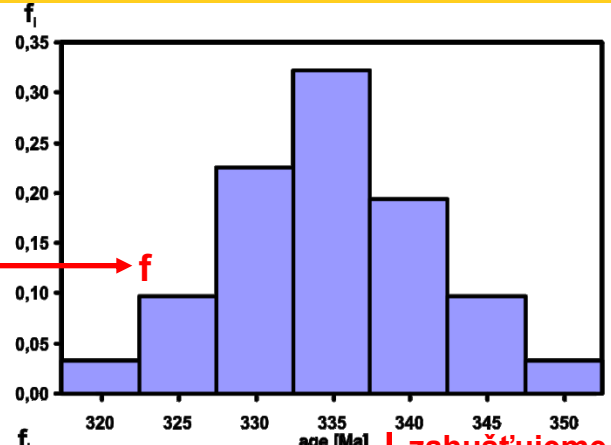
R. Čopjaková

Od četnosti k pravděpodobnosti

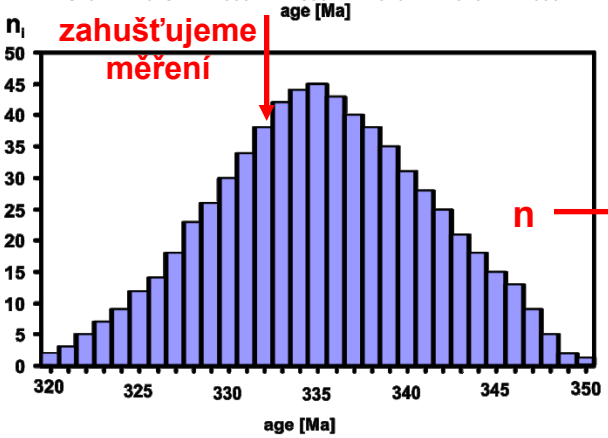
střed int	n_i	f_i
320	1	0,03
325	3	0,10
330	7	0,23
335	10	0,32
340	6	0,19
345	3	0,10
350	1	0,03
suma	31	1



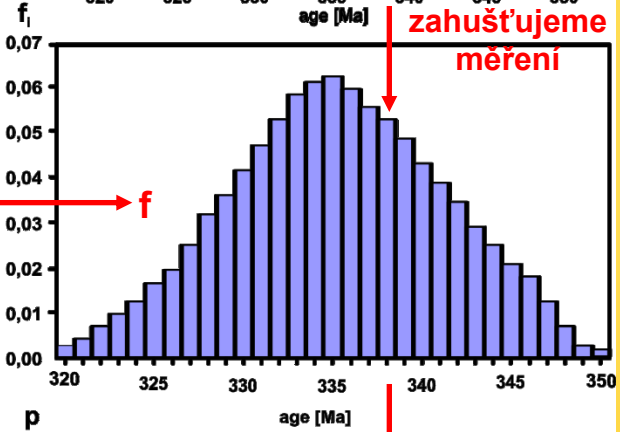
n → f

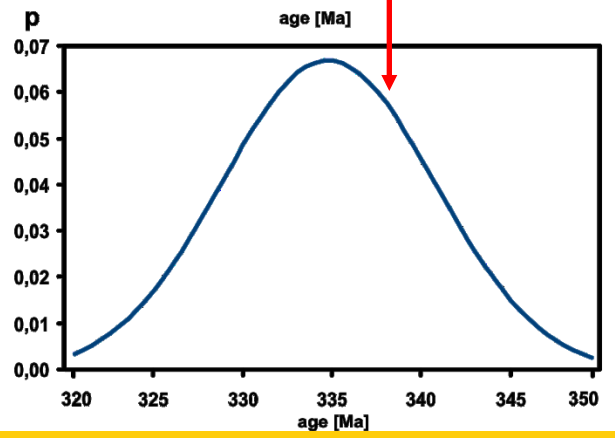


střed int	n_i	f_i
320	2	0,00
321	3	0,00
322	5	0,01
323	7	0,01
324	9	0,01
325	12	0,02
326	14	0,02
327	18	0,03
328	23	0,03
329	26	0,04
330	30	0,04
331	34	0,05
332	38	0,06
333	42	0,06
334	44	0,07
335	45	0,07
336	43	0,06
337	40	0,06
338	38	0,06
339	35	0,05
340	31	0,05
341	28	0,04
342	25	0,04
343	21	0,03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



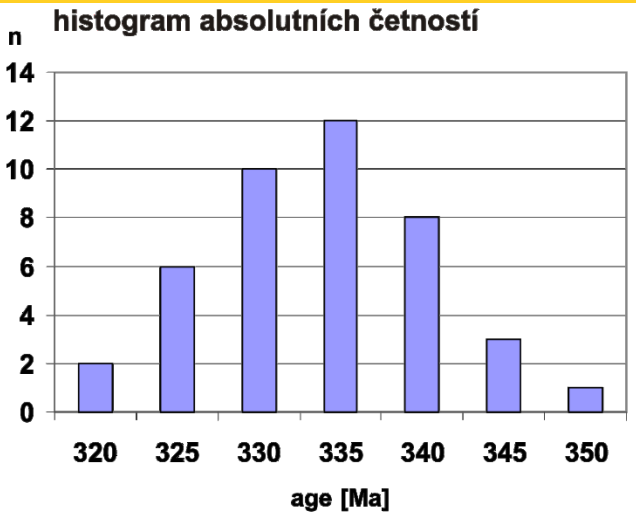
n → f

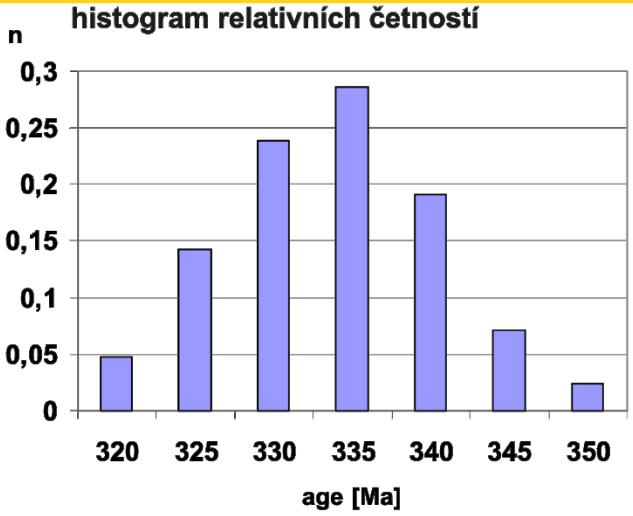




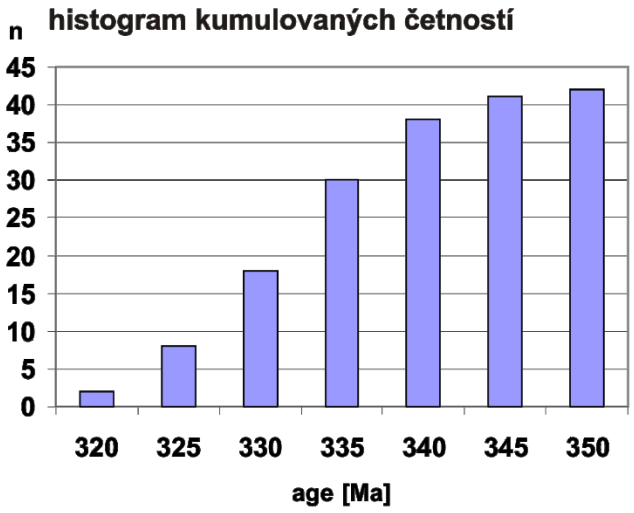
Hustota rozdělení pravděpodobností
frekvenční funkce
pravděpodobnostní funkce

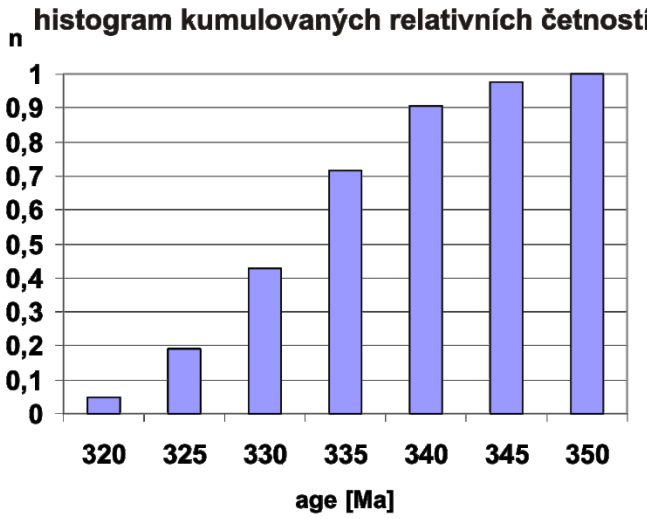
Vztah mezi frekvenční a distribuční funkcí





f(x)





F(x)

pro distribuční fci diskrétní náhodné veličiny platí: $F(x) = P(X \leq x)$ a tedy $F(x) = \sum_{x_i \leq x} p(x_i)$

Co je to distribuční funkce

(cumulative probability)?

$$F(x) = P(X \leq x)$$

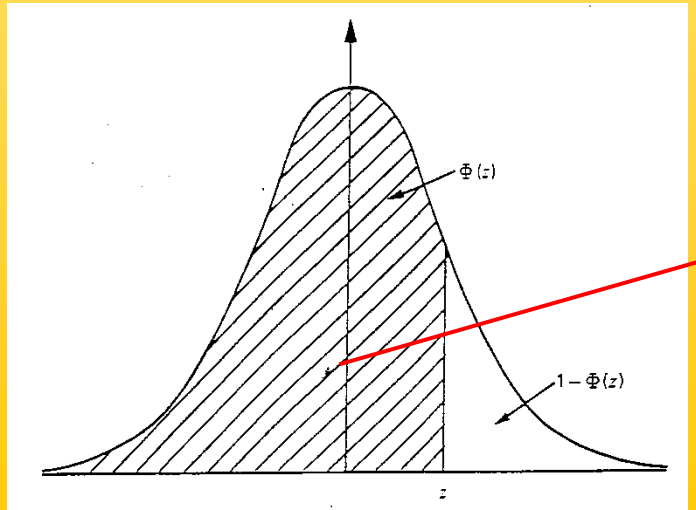
Distribuční funkce v bodě x je rovna pravděpodobnosti jevu, že náhodná veličina X nepřevýší hodnotu x.

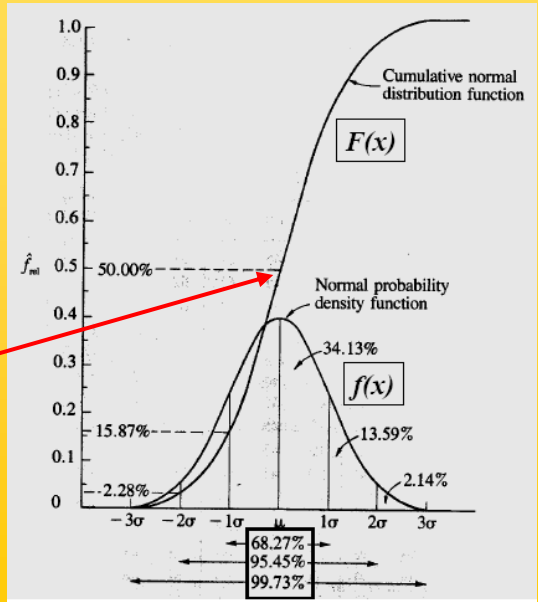
Diskrétní n.v.

$$F(x) = \sum_{x_i \leq x} P(x_i)$$

Spojitá n.v.

$$F(x) = \int_{-\infty}^{x} f(x)d(x)$$



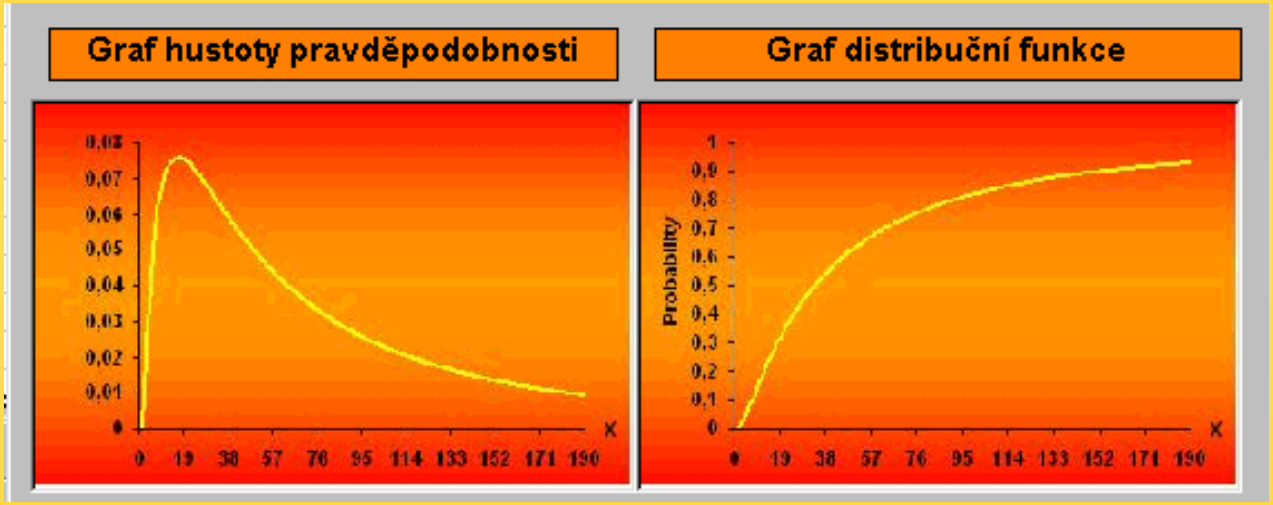


Lognormální rozdělení (logaritmicko-normální rozdělení)

Frekvenční funkce lognormálního rozdělení LN (μ, σ^2):

$$y = f(x \mid \mu, \sigma) = \frac{1}{x\sigma\sqrt{2\pi}} e^{\frac{-(\ln x - \mu)^2}{2\sigma^2}}$$

μ,σ^2 jsou parametry lognormálního rozdělení



Má-li náhodná veličina X rozdělení LN (μ, σ^2), má potom náhodná veličina Y = ln X rozdělení N (μ, σ^2). Má-li veličina Y rozdělení N (μ, σ^2), potom veličina $X = e^Y$ má rozdělení LN (μ, σ^2).

=>

Soubor dat s lognormálním rozdělením pravděpodobností lze snadno transformovat na soubor dat s normálním rozdělením pravděpodobností

např. $y_i = \ln x_i$ => $x_i = e^{y_i}$

$y_i = \log x_i$ => $x_i = 10^{y_i}$

Lognormální rozdělení mají např.

Vhodné pro jednostranně ohraničená data – např. fyzikální veličiny (teplota, tlak, hmotnost, objem, …)

- **zrnitost některých sedimentů**
- **mocnost sedimentárních hornin**
- **propustnost sedimentárních hornin**
- **koncentrace stopových prvků v horninách**
- **pórovitost magmatických hornin**
- **průtok vody v řekách**

Lognormální rozdělení (logaritmicko-normální rozdělení)

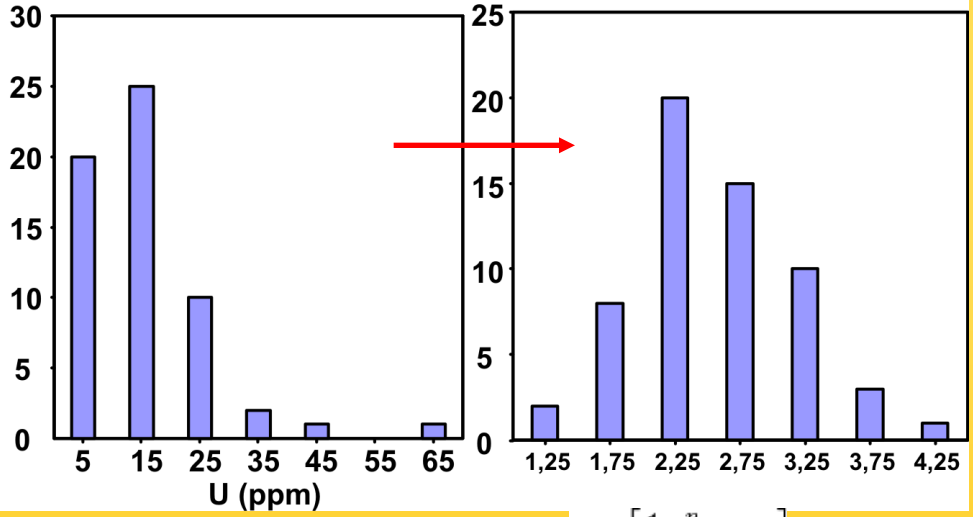
X → Y

U (ppm)	ln(x)
3	1,099
4	1,386
4	1,386
5	1,609
5	1,609
6	1,792
6	1,792
7	1,946
7	1,946
26	3,258
27	3,296
27	3,296
28	3,332
29	3,367
34	3,526
36	3,584
47	3,850
63	4,143
aritmetický průměr	2,530 $\bar{y}$
NE	12,559 $\bar{x}=e^{\bar{y}}$

střed int	U (ppm)
5	20
15	25
25	10
35	2
45	1
55	0
65	1
n	59

→

střed int	ln (U)
1,25	2
1,75	8
2,25	20
2,75	15
3,25	10
3,75	3
4,25	1
n	59



Střední hodnota - při použití transformace ln(x) spočtu jako $\bar{x} = e^{\bar{y}}$; tedy $\exp\left[\frac{1}{n}\sum_{i=1}^{n}\ln x_i\right]$

tzv. **geometrický průměr** - další možný dopočet dle:

$$\bar{x}_G = \sqrt[n]{x_1.x_2.\ \ldots\ .x_n} = \sqrt[n]{\prod_{x=1}^{n} x_i}$$

lze chápat jako zobecněný průměr s transformací f(x) = ln x

Míra variability - směrodatnou odchylku NEUŽÍVAT - (výpočet analogicky jako pro střední hodnotu)

Binomické rozdělení

- **Binomické rozdělení Bi (n,p)** popisuje četnost výskytu náhodného jevu v ***n*** nezávislých pokusech, v nichž má jev stále stejnou známou pravěpodobnost **p**.

- Diskrétní náhodná veličina *X* s binomickým rozdělením může nabývat celočíselných hodnot od nuly po *n*. Pravděpodobnost, že jev nastane právě *x*-krát z *n* pokusů při pravděpodbnostijevu *p*, je určena rozdělením

$$P[X = x] = \binom{n}{x} p^x (1-p)^{n-x}$$

počet pokusů n
pravděpodobnost úspěchu p
pravděpodobnost neúspěchu 1-p = q
počet úspěšných pokusů x

- Pro $n \to \infty$ a malé pravděpodobnosti $p \to 0$ přechází binomické rozdělení v rozdělení Poissonovo.
- Pro *p* blízké 0,5 lze binomické rozdělení již od *n* v řádu několika desítek velmi dobře aproximovat normálním rozdělením.

- střední hodnota: E(x) = np

- rozptyl: $\sigma^2(X) = np(1-p) = npq$
- a směrodatná odchylka je odmocninou z rozptylu:

Jaká je pravděpodobnost, že při 5 vrzích kostkou padne právě 2× číslo 1?

pak n = 5; pravděpodobnost úspěchu p = 1/6;

pravděpodobnost neúspěchu 1-p = q = 5/6

$$P[X = x] = \binom{n}{x} p^x (1-p)^{n-x}$$

Kombinační číslo - počet kombinací *x*-té třídy z *n* prvků bez opakování

počet K(*x*, *n*) všech *x*-členných kombinací z *n* prvků je: $\frac{n!}{x!(n-x)!}$

$$p_2 = \binom{5}{2} \left(\frac{1}{6}\right)^2 \left(1 - \frac{1}{6}\right)^{(5-2)} \approx 0,16 = 16\%$$

Binomické rozdělení

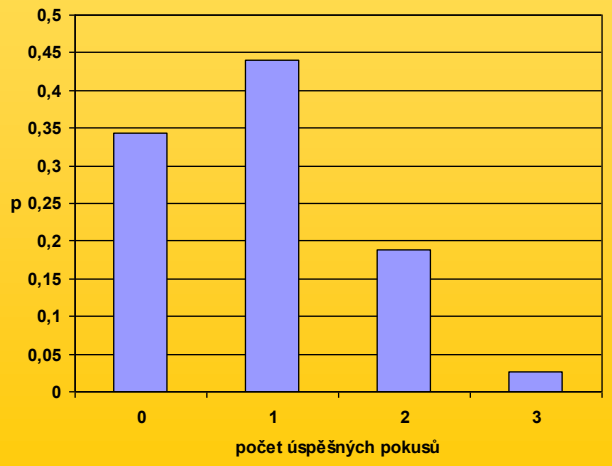
Ropná společnost provede 3 vrty, pravděpodobnost, že narazí na ropu je 0,3. Spočti hustotu pravděpodobnosti pro binomické rozdělení a stanov pravděpodobnost, že společnost minimálně dvěma vrty narazí na ropu.

	x	P(x)
žádný úspěšný vrt	0	$\binom{3}{0}(0,3)^0\,(0,7)^3 = 0,343$
jeden úspěšný vrt	1	$\binom{3}{1}(0,3)^1\,(0,7)^2 = 0,441$
dva úspěšné vrty	2	$\binom{3}{2}(0,3)^2\,(0,7)^1 = 0,189$
tři úspěšné vrty	3	$\binom{3}{3}(0,3)^3\,(0,7)^0 = 0,027$

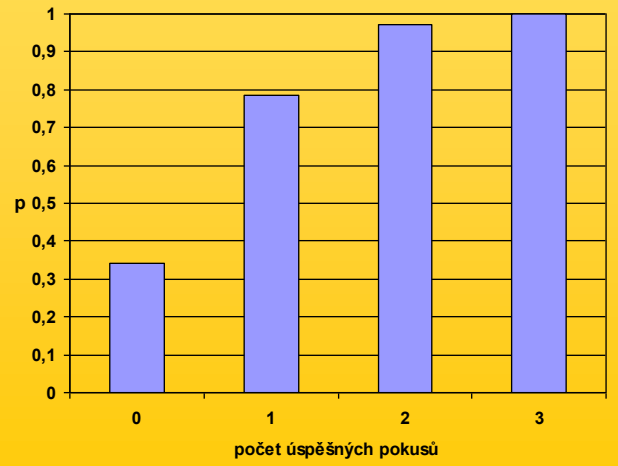
střední hodnota
E(X) = n.p = 3 . 0,3 = 0,900
rozptyl
npq = 3.0,3.0,7 = 0,63

minimálně dva úspěšné vrty
z f(x): 0,189+0,027 = 0,216
z F(x): 1- F(1) = 1-0,784 = 0,216

frekvenční kunkce binomického rozdělení



distribuční funkce binomického rozdělení



$$F(x) = \sum_{x_i \le x} p(x_i)$$

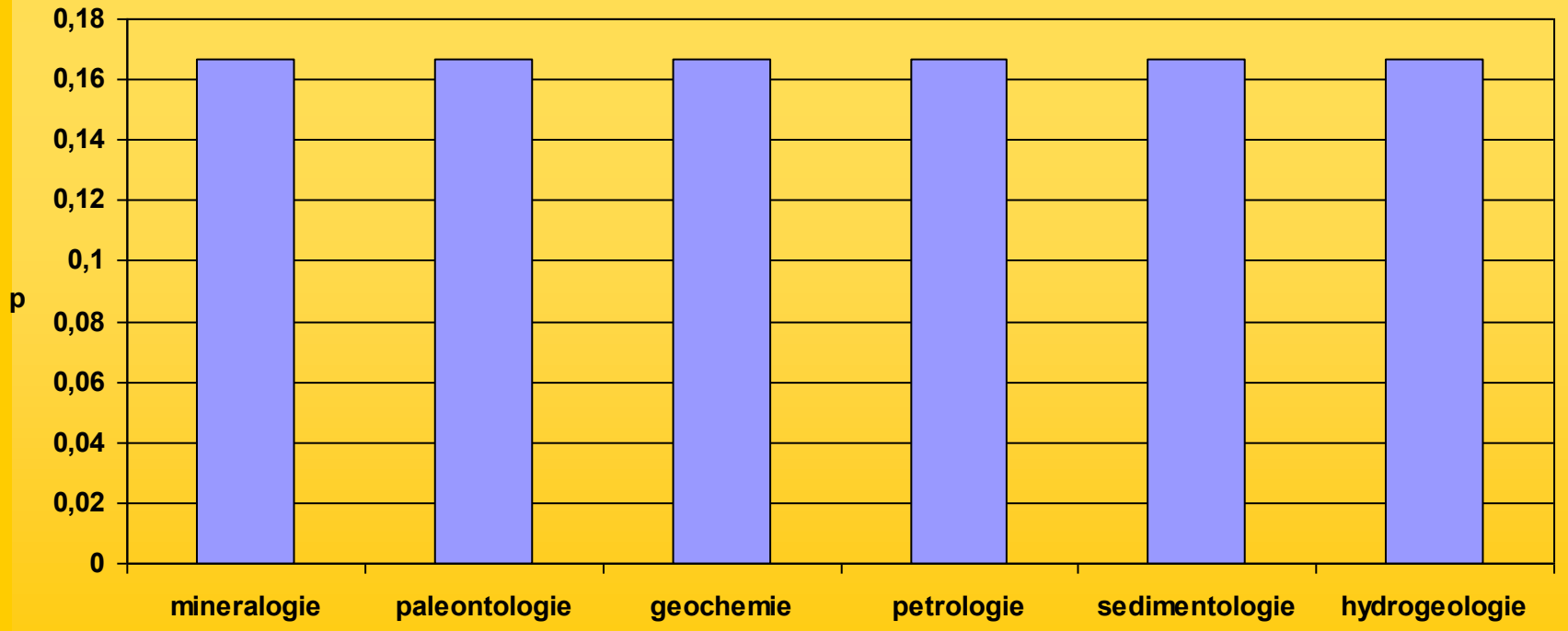
Rovnoměrné rozdělení

Rovnoměrné (diskrétní) rozdělení - jev může nabývat jednoho z *k*-stavů, všechny stavy mají stejnou pravděpodobnost

$$P(X = x_i) = \frac{1}{k}, i = 1,2,...,k$$

např. zájem o jednotlivé studijní obory je rovnoměrný

Alternativní (Bernouliho, nula-jedničkové) rozdělení jev může nabývat jednoho ze dvou stavů (0 nebo 1)

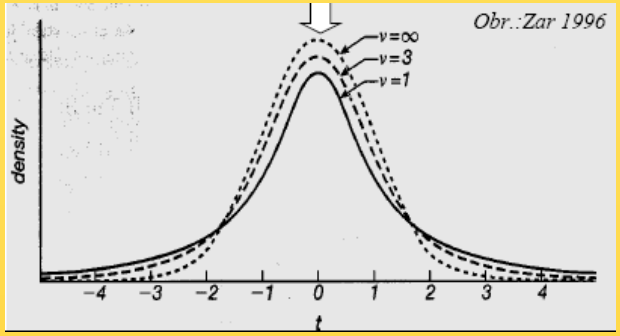


Speciální spojitá rozdělení

Studentovo rozdělení (t-rozdělení) - rozdělení odchylky průměru od střední hodnoty

Aritmetický průměr výběrového souboru s normálním rozdělením se může více či méně lišit od střední hodnoty základního souboru. Provádíme opakovaně výběry o stejném rozsahu pro k […]ritmetický průměr X a směrodatnou odchylku S. Pak veličina $t = \frac{(\overline{X} - \mu)\sqrt{n}}{S}$ má Studentovo rozdělení s n-1 stupni volnosti.

Hustota […]podobností t-rozdělení pro různé stupně volnosti



- **Studentovo rozdělení modeluje rozdělení průměrů všech možných souborů o velikosti *n***
- je podobné standardizovanému normálnímu rozdělení
- je symetrické kolem střední hodnoty μ = 0
- má pouze 1 parametr:
- stupně volnosti: $\nu = n-1$; Co to jsou stupně volnosti? ν = počet pozorování mínus počet parametrů
- Využívá se při testování statistických hypotéz – tzv. t-testy, např. testování rozdílu mezi dvěma průměry.

Speciální spojitá rozdělení

Rozdělení chí-kvadrát

Využívá se při testování statistických hypotéz

Nejčastěji při testování shody empirického rozdělení (rozdělení četností naměřeného souboru dat) s předpokládaným teoretickým rozdělením tzv. **Pearsonův test** neboli **chí-kvadrát test**

Speciální spojitá rozdělení

Fisher-Snedeckorovo rozdělení

F-rozdělení
využití při testování statistických hypotéz, při analýze rozptylu

např. test shody rozptylů dvou výběrů z normálního rozdělení.

Statistické funkce v excelu

- NORMDIST stanovení hodnoty pravěpodobnosti frekvenční nebo distribuční funkce normálního rozdělení N (μ,σ^2)
- NORMINV určí kvantil normálního rozdělení N (μ,σ^2)

pro standardizované normální rozdělení N (0, 1)

- NORMSDIST
- NORMSINV

- LOGNORMDIST stanovení hodnoty pravěpodobnosti frekvenční nebo distribuční funkce logaritmicko-normálního rozdělení
- BINOMDIST stanovení hodnoty pravěpodobnosti frekvenční nebo distribuční funkce binomického rozdělení Bi (n,p)
- TINV určí kvantil studentova rozdělení
- FINV určí kvantil Fisher-Snedeckorova rozdělení
- CHIINV určí kvantil chí-kvadrát rozdělení