

Statistické metody v ochraně kulturního dědictví

Lubomír Prokeš

1

Data a práce s nimi

- 1) sběr a zpracování dat (tvorba databáze)
- 2) analýza dat (výběr a použití vhodné metody)
- 3) prezentace výsledků (špatná prezentace dat může vést k chybným závěrům)
- 4) metaanalýza (srovnávání výsledků z různých publikací)

Pozor !!!

Nepoučený uživatel může často založit zásadní rozhodnutí na základě

1) volby nesprávné metody statistické analýzy, která poskytne nesmyslné výsledky

2) nesprávné interpretace správných výsledků

Statistika

= nauka o tom, jak získat
informace z numerických dat.

- 1) **Získávání dat.** Zahrnuje metody pro sběr dat, jež zodpoví předem danou otázku. Základní přístupy k výběru měřených objektů, návrhu experimentů (experimental design) a validaci instrumentů pro získávání dat.
- 2) **Analýza dat.** Zahrnuje organizaci dat a jejich popis užitím grafů a numerických souhrnů (popisná statistika, průzkumová analýza dat (EDA)).
- 3) **Statistické usuzování (inference).** Usiluje o získání závěrů o širším univerzu jevů na základě analýzy dat, včetně zhodnocení spolehlivosti těchto závěrů, k čemuž využívá pravděpodobnostní pojmy (statistická inference, statistická indukce).

Statistický software

WYSIWYG:

MS Excel,
STATISTICA,
SPSS, NCSS
Kyplot, **PAST**, aj.

ne WYSIWYG:

MATLAB,
S+, R,
SciPy

Typy dat

Kvalitativní (nominální):

Ize sledovat jen identitu (=) a odlišnost (\neq).

- ***alternativní (dichotomické)*** – znak má pouze dvě varianty (ano / ne).
- ***množné (polytomické)*** – znaky s větším počtem variant.

Typy dat

- ***Kvantitativní znaky***

- 1) ***pořadové (ordinální) znaky.***

jejich varianty jsou uspořádané podle intenzity sledovaného znaku.

- ***porovnávací:*** není předem daná pořadová stupnice, varianty se třídí podle míry zastoupení (intenzity) sledovaného znaku

- ***zařazovací:*** předem se vymezí pořadí variant, tj. zadá se jejich „stupnice“.

Typy dat

- ***Kvantitativní znaky***
- ***2) číselné (kardinální) znaky.***
Měřitelné znaky, jejichž varianty lze vyjádřit číselnou hodnotou.
- ***intervalové.-*** nemají smysluplnou nulu
- ***podílové (poměrové).*** - mají smysluplnou nulu

Typy dat

Kvantitativní data

Diskrétní: nabývají konečně mnoha hodnot (např. četnosti)

Spojitá: nabývají hodnot všech reálných čísel v daném intervalu (např. rozměry)

- Typ dat je nutno respektovat při výběru metod analýzy dat !!

Transformace dat

poměrové → intervalové → pořadové → nominální.

- „Dummy variables“:
- *Heavisidova funkce*:

$$\Theta(x) = \begin{array}{ll} 1 & \text{když } x > 0 \\ 0 & \text{když } x \leq 0 \end{array}$$

Transformace dat

- *Absolutní četnost* (n_i) = počet případů, v nichž se určitá hodnota x_i vyskytne ve statistickém souboru.
- *Relativní četnost* (f_i) = podíl případů z celkového rozsahu souboru, v nichž se hodnota x_i vyskytne ve statistickém souboru.

$$f_i = \frac{n_i}{n}$$

Transformace dat

- *Třídní (skupinové, intervalové) četnosti* = kvantitativní znaky rozdělíme na intervaly a všechna pozorování z téhož intervalu nahradíme jedinou hodnotou, nejčastěji průměrem z nejnižší a nejvyšší hodnoty v dané třídě.

Počet tříd má vliv na přesnost výpočtu ukazatelů a pracnost výpočtů. Čím je počet tříd menší, tím je délka intervalů větší a tím jsou výpočty méně přesné.

Transformace dat

- **Transformace do pořadí:** převádí hodnoty x_i podle velikosti do intervalu $i = 1$ až n . Stejným hodnotám přiřazujeme průměrné pořadí, které této skupince hodnot odpovídá.

Popisná statistika I.

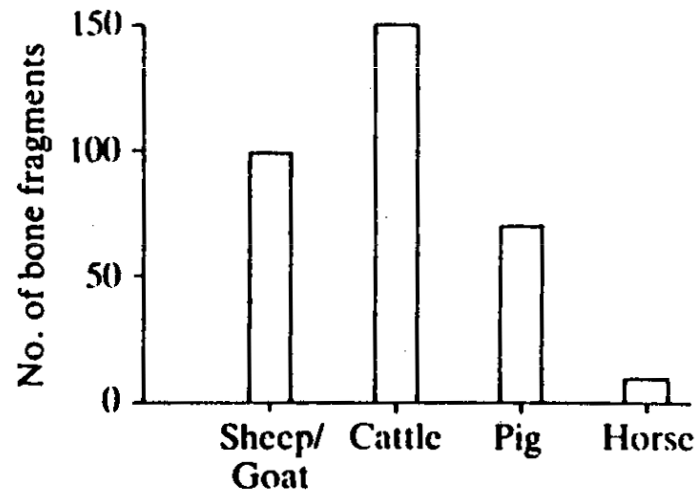
Popisná statistika

- 1) grafické metody
- 2) tabulky
- 3) číselné parametry

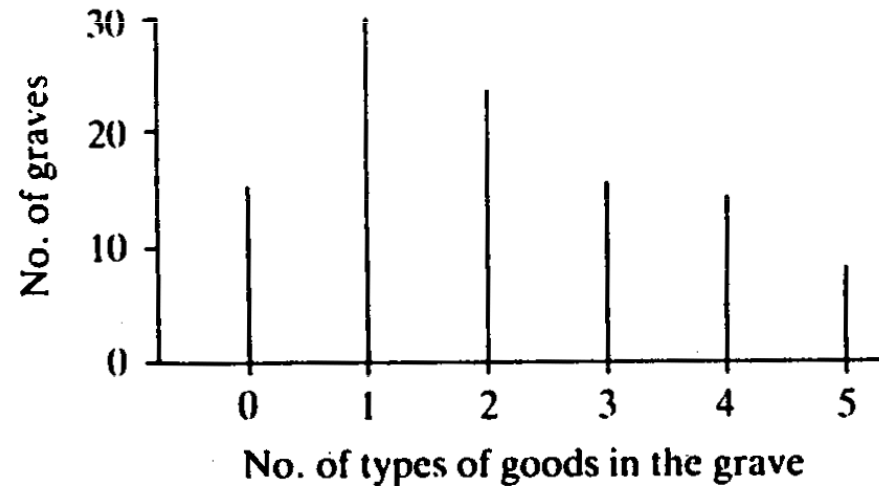
Sloupcový graf (bar chart)

Modus (\hat{x})

nejčastěji se vyskytující hodnota v souboru.

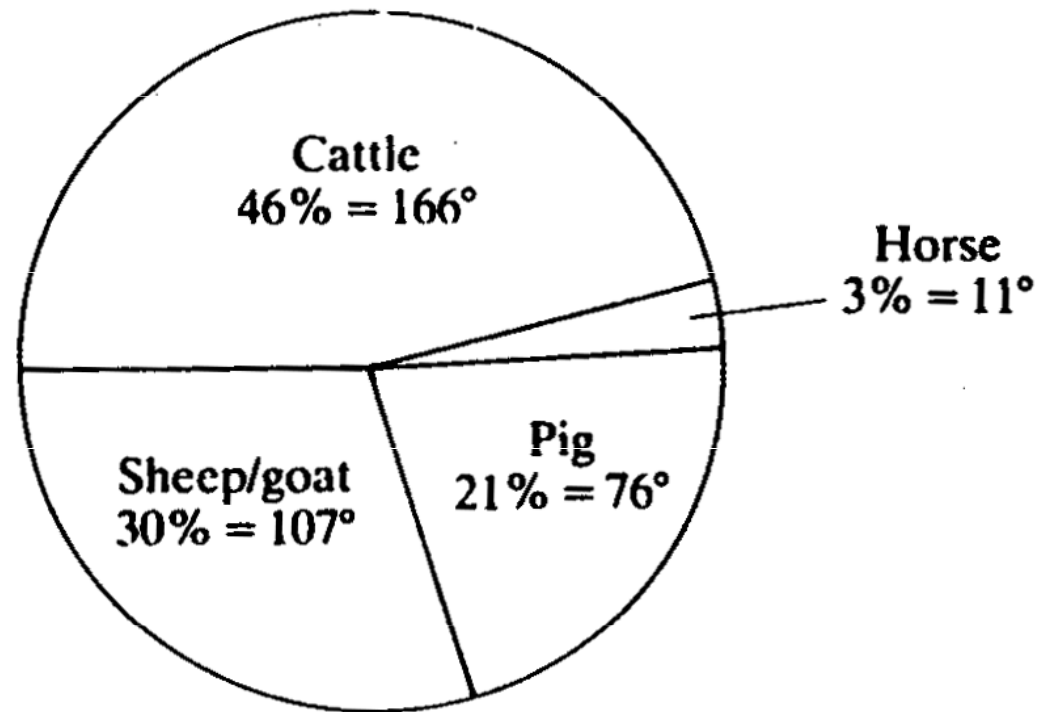


Bar chart of the number of bone fragments of different domestic animal species from a hypothetical British iron age site.



Bar chart of the number of graves containing different numbers of grave-good types for a hypothetical central European bronze age cemetery.

Koláčový graf (pie chart)

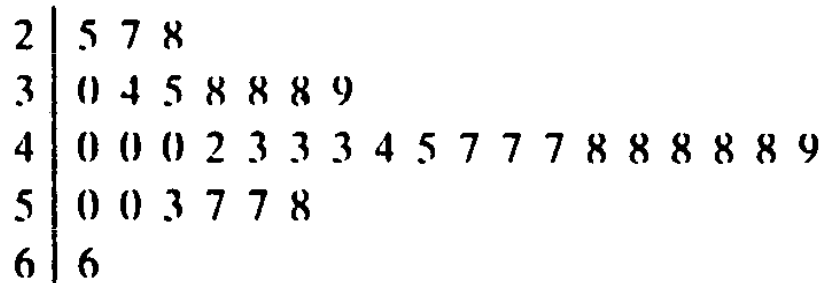


Pie chart of the relative proportions of bone fragments of different domestic species

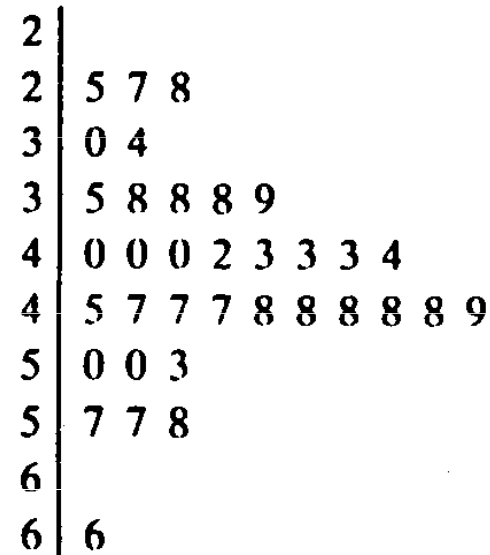
Čárkovací metoda

Interval	Postupný zápis četností
961 – 965	
966 – 970	###
971 – 975	###
976 – 980	### ###
981 – 985	### ### ###
986 – 990	### ### ### ### ### ###
991 – 995	### ### ### ### ### ### ### ### ### ###
996 – 1 000	### ### ### ### ### ### ### ### ### ###
1 001 – 1 005	### ### ### ### ### ### ###
1 006 – 1 010	### ### ###
1 011 – 1 015	###
1 016 – 1 020	###
1 021 – 1 025	;
1 026 – 1 030	
1 031 – 1 035	

Stem and leaf plot



Stem-and-leaf diagram of the
Mount Pleasant post-hole diameters



Stem-and-leaf diagram of the
Mount Pleasant post-hole diameters
with stem intervals 5 units wide
instead of 10.

Variační rozpětí:

$$R = x_{\max} - x_{\min}$$

Histogram a frekvenční polygon

$$0,05R < k < 0,12R$$

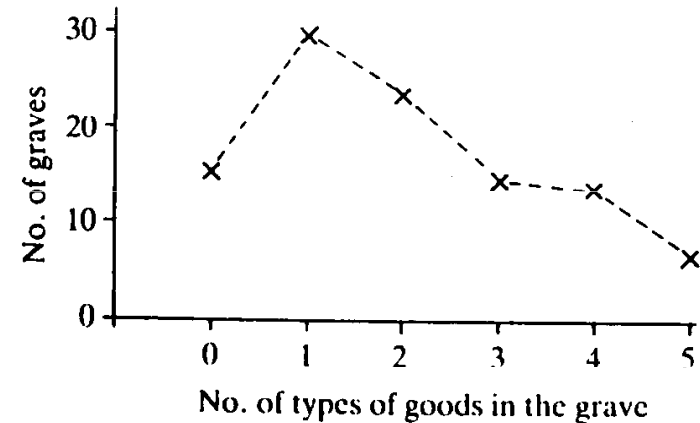
$$k \approx 1 + \log_2(2n) = 1 + 3,3 \log n$$

(Sturgesovo pravidlo)

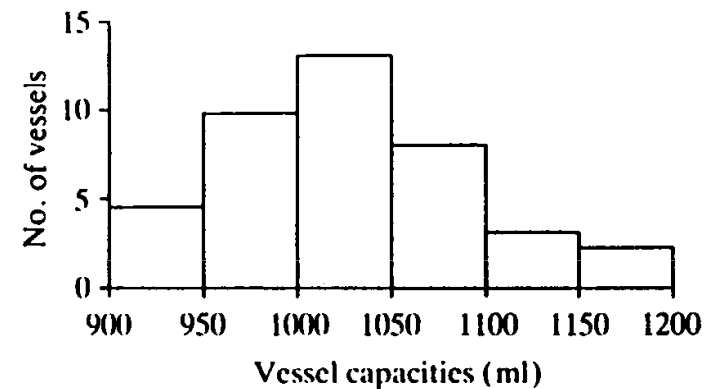
$$k \approx 5 \log n$$

$$k = \text{int}(2,46(n - 1)^{0,4})$$

modus



Frequency polygon



Bar chart of the distribution of vessel capacities for a group of 40 bell beakers.

Kvantily a percentily

Rozdělují soubor na danou procentuální část.

Nejvýznamnější kvantily:

Medián: 2. kvartil (50% percentil)

Q_I : Dolní kvartil (1. kvartil, 25% percentil)

Q_{III} : Horní kvartil (3. kvartil, 75% percentil)

- *Medián* (\tilde{x}) rozděljuje uspořádané (podle velikosti) zjištěné hodnoty na dvě stejně početné části.

Pro výpočet mediánu a ostatních kvantilů platí:

Je-li n liché

- $\tilde{x} = x_k$ kde $k = (n + 1)/2$

Je-li n sudé

- $\tilde{x} = \frac{x_k + x_{k+1}}{2}$ kde $k = n/2$

Výhodou mediánu je, že bezprostředně nezávisí na extrémních hodnotách.

Od mediánu se odvozují i některé parametry rozptýlení:

Mediánová odchylka

$$MD = \frac{\sum_i (x_i - \tilde{x})}{n}$$

Absolutní mediánová odchylka

$$MAD = \text{med}(X_i - \text{med})$$

Interkvartilové rozpětí

$$Q = Q_{III} - Q_I$$

Kvartilový koeficient šikmosti

$$KS = \frac{Q_{III} + Q_I - 2\tilde{x}}{Q_{III} - Q_I}$$

Pearsonův koeficient šikmosti

$$SK = \frac{3(\bar{x} - \tilde{x})}{s}$$

Momentové charakteristiky

Aritmetický průměr (\bar{x})

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

Geometrický průměr

$$\bar{x}_g = \sqrt[n]{\prod_{i=1}^n x_i}$$

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

Momentové charakteristiky

Rozptyl

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

resp.

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

Kladná druhá odmocnina z rozptylu se nazývá *směrodatná odchylka*.

Variační koeficient

$$s_r = \frac{s}{\bar{x}}$$

Momentové charakteristiky

- $$m_k = \frac{(x_i - \bar{x})^k}{n}$$

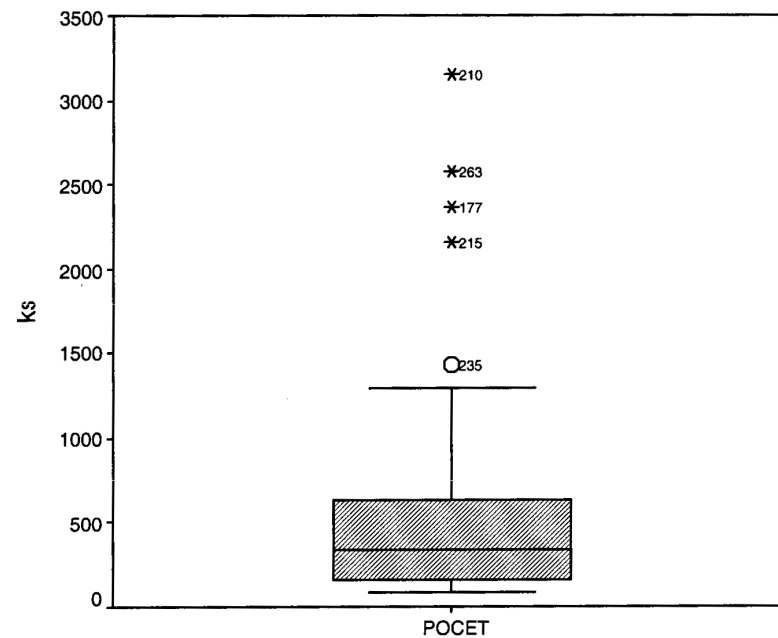
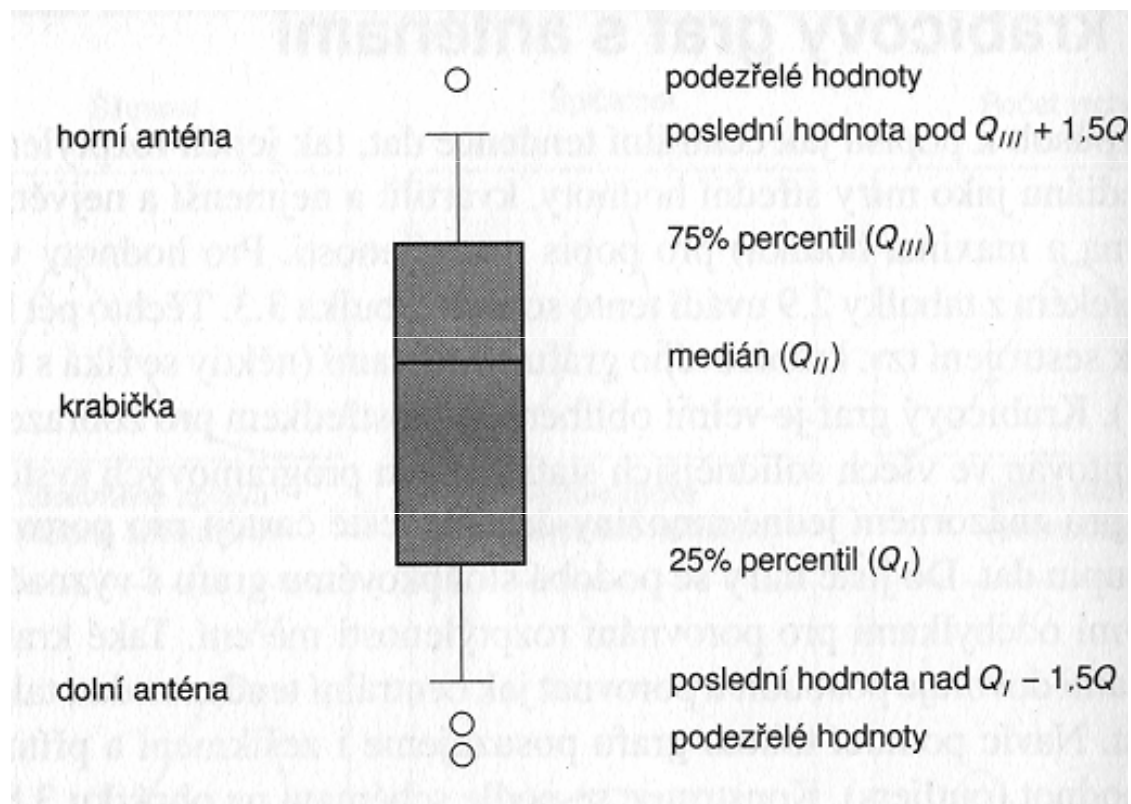
- Šikmost: měří asymetrii dat

$$S_1 = \frac{m_3}{m_2^{3/2}}$$

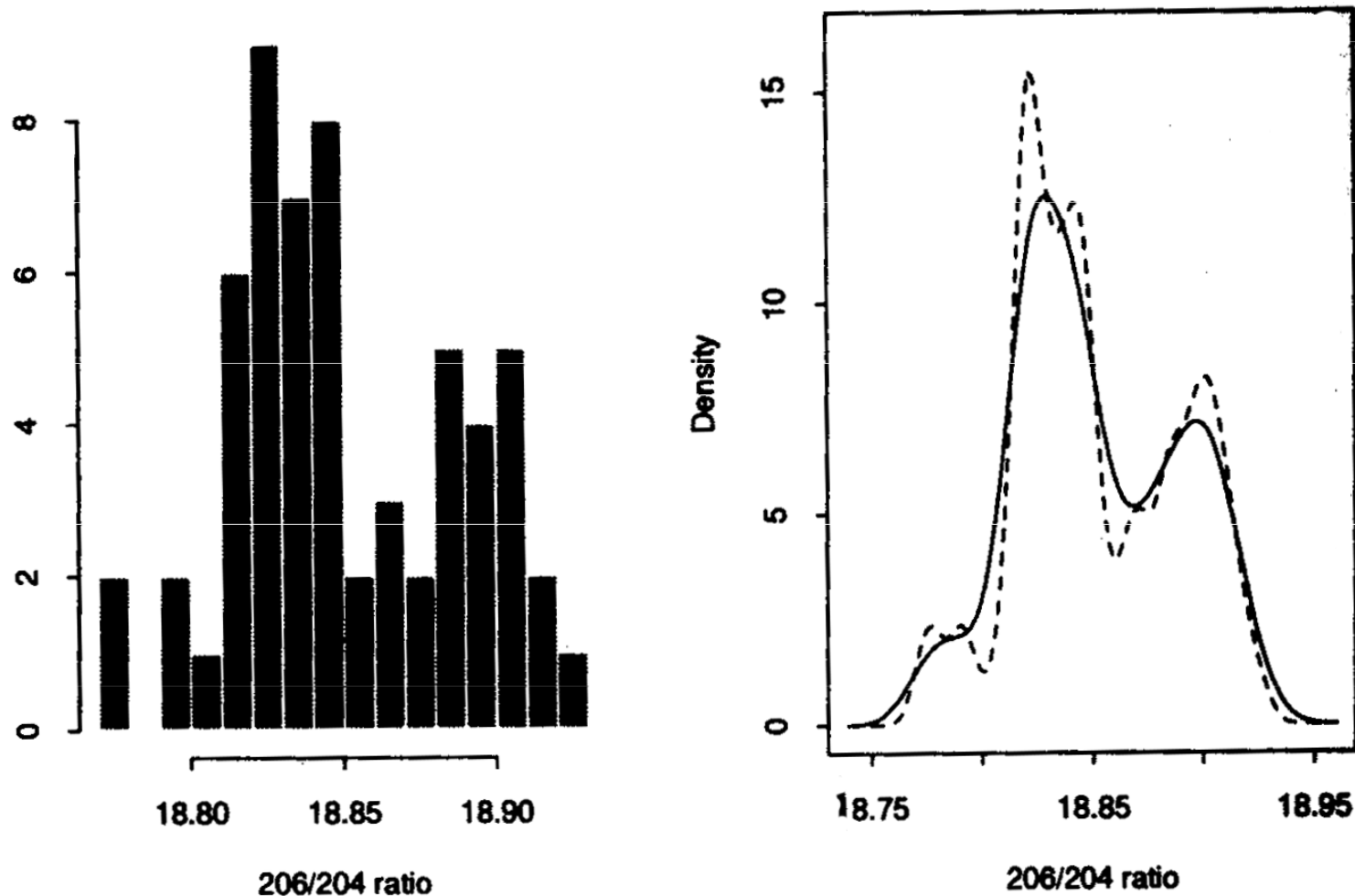
- Špičatost:

$$S_2 = \frac{m_4}{m_2^2} - 3$$

Box and whisker plot



Jádrové odhady (KDE)



A histogram (left panel) and KDEs (right panel) for the Lavrion $^{206}\text{Pb}/^{204}\text{Pb}$ lead isotope ratio data from Stos-Gale et al. (1996). The solid KDE uses a smoothing parameter, h , determined by the method of Sheather and Jones (1991); the dashed KDE uses a subjectively determined h .

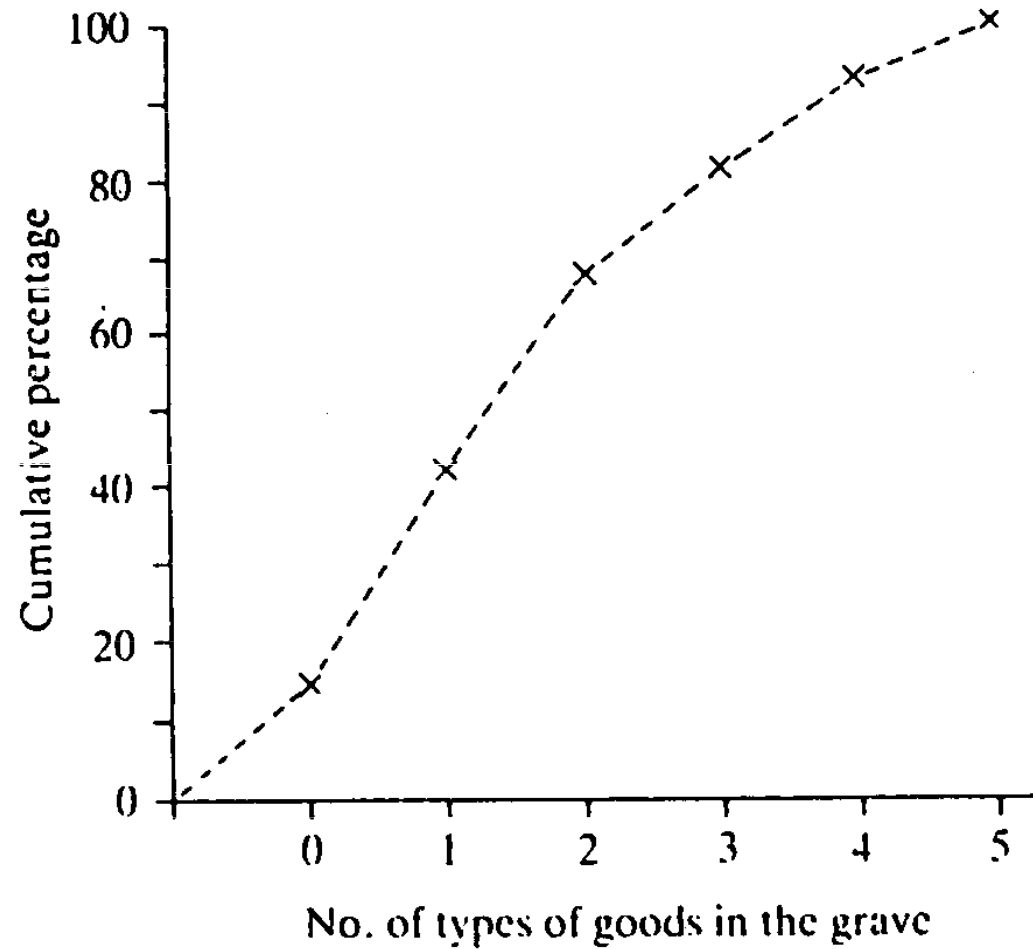
Jádrové odhady (KDE)

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^n K\left[\frac{x - x_i}{h}\right]$$

kde $K(x)$ je funkce symetrická kolem nuly, šířka pásu h určuje stupeň vyhlazení:

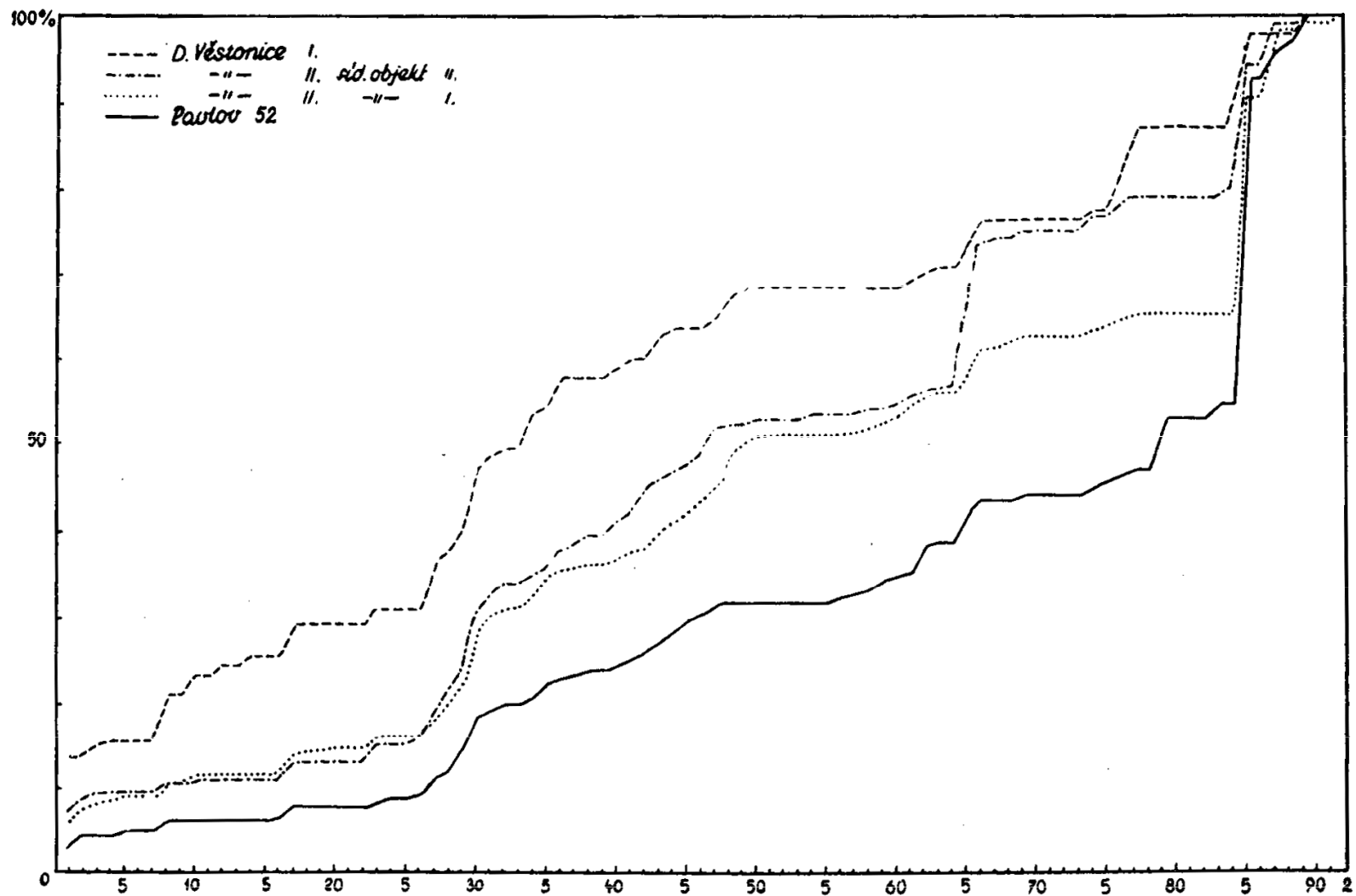
$$h_{\text{opt}} = 2,34\sigma n^{-0,2}$$

Kumulativní graf



Cumulative curve of the data on numbers of grave-good types

Kumulativní graf



Hromadný graf kamenných industrií z jednotlivých sídlištních celků paleolitických stanic pod Pavlovskými vrchy.

- Při posuzování grafů je třeba sledovat:
 - 1) zhuštění dat (místo či místa s největší četností)
 - 2) shluky dat
 - 3) mezery v datech (intervaly bez hodnot)
 - 4) odlehlé hodnoty (přítomnost údajů odlišných od zbytku dat)
 - 5) tvar rozdělení (např. z histogramu)

Základní soubor a výběr

- ***Základní populace*** (základní soubor) je množina všech teoreticky možných objektů (jedinců) v uvažované situaci. V mnoha případech má pouze hypotetický význam.
- ***Výběr (vzorek)*** je podmnožinou základní populace (velmi často totiž nelze podrobit výzkumu celou základní populaci). Počet prvků (objektů) n ve výběru se nazývá rozsah výběru.

- **Populační parametr** dané proměnné je číselná hodnota, která tuto proměnnou charakterizuje v základní populaci (např. aritmetický průměr). Má nějakou fixní číselnou hodnotu, kterou v praxi zpravidla neznáme (pokud neprovedeme úplné šetření); odhadujeme ji na základě výběrových statistik.
- **Výběrová statistika** charakterizuje vzorek, získaný výběrem ze základní populace (výběrové šetření); má číselnou hodnotu, jež charakterizuje výběr (např. výběrový průměr). Co je parametr pro populaci, to je výběrová statistika pro výběr.

Distribuční funkce

Pro distribuční funkci platí: je neklesající, spojitá zleva, $0 \leq F(x) \leq 1$ pro všechna reálná $-\infty < x < \infty$

$$\lim_{x \rightarrow -\infty} F(x) = 0$$

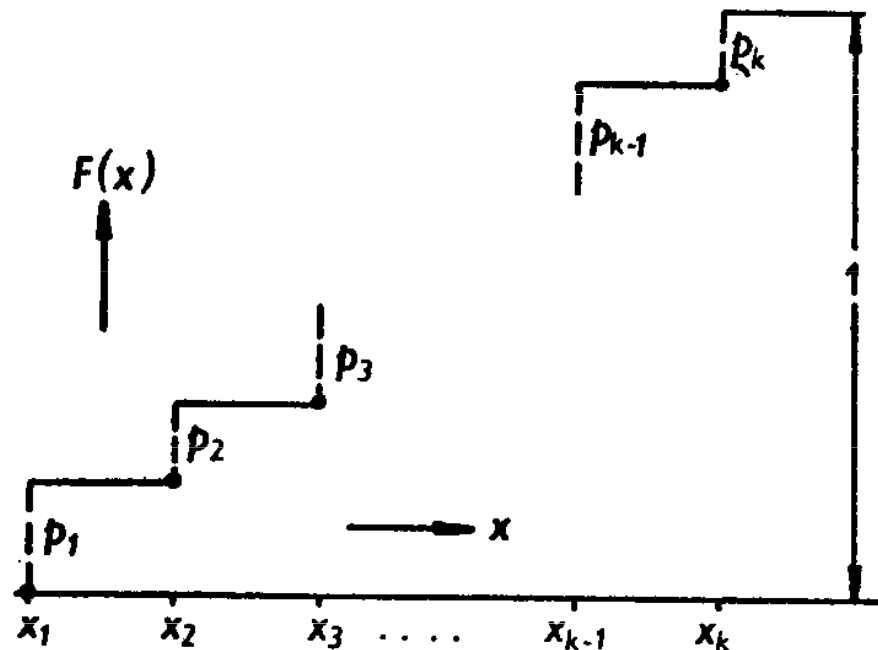
$$\lim_{x \rightarrow \infty} F(x) = 1$$

a $P(a \leq X < b) = F(b) - F(a)$ pro libovolná $a < b$.

Distribuční funkce

- Distribuční funkce *diskrétní náhodné veličiny* je schodovitá funkce s body skoku x_1, x_2, \dots, x_k .

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

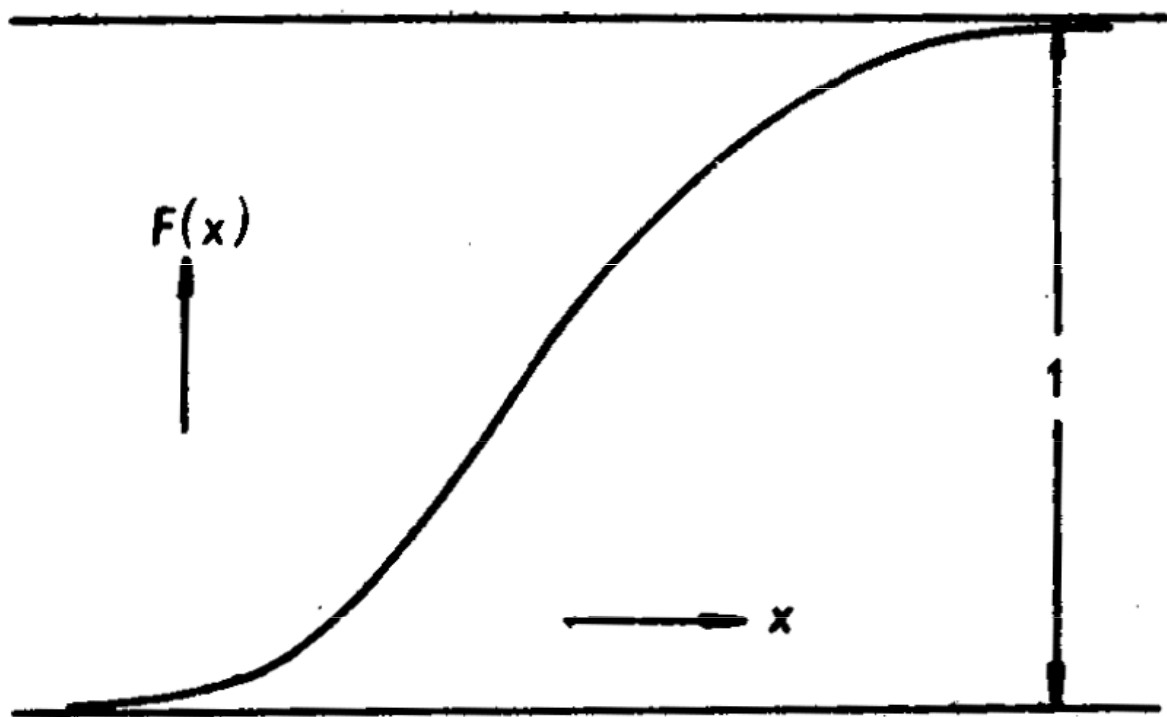


Distribuční funkce diskrétní náhodné veličiny.

Distribuční funkce

- Pro *spojitou náhodnou veličinu* má distribuční funkce tvar

-



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

kde $f(x)$ je hustota pravděpodobnosti distribuční funkce.

Distribuční funkce spojité náhodné veličiny.

Charakteristiky náhodné veličiny

umožňují shrnutí informace o náhodné veličině do několika číselných hodnot.

Momentová metoda

- k-tý obecný moment: $m_{ok} = E(X^k)$
- k-tý centrální moment: $m_{ck} = E\{[X - E(X)]^k\}$

Metoda maximální věrohodnosti

- mnohem složitější výpočty

Parametr polohy (střední hodnota)

- diskrétní:

$$E(X) = \sum_{x_j} x_j p_j$$

- spojitě:

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx$$

Parametr polohy (střední hodnota)

- $E(kx) = kE(x)$

kde k je konstanta.

- $E(x_1 + x_2 + \dots + x_n) = E(x_1) + E(x_2) + \dots + E(x_n)$

- $E(x_1 \cdot x_2 \cdot \dots \cdot x_n) = E(x_1) \cdot E(x_2) \cdot \dots \cdot E(x_n)$

- $E(k_1x_1 + k_2x_2 + \dots + k_nx_n) = \sum_{i=1}^n k_i E(x_i)$

kde k_1, k_2, \dots, k_n jsou konstanty.

Parametr disperze (rozptyl)

- diskrétní:

$$D^2(X) = \sum_{x_j} [X - E(X)]^2 p_j$$

- spojitě:

$$D^2(X) = \int_{-\infty}^{\infty} [X - E(X)]^2 f(x) dx$$

Parametr disperze (rozptyl)

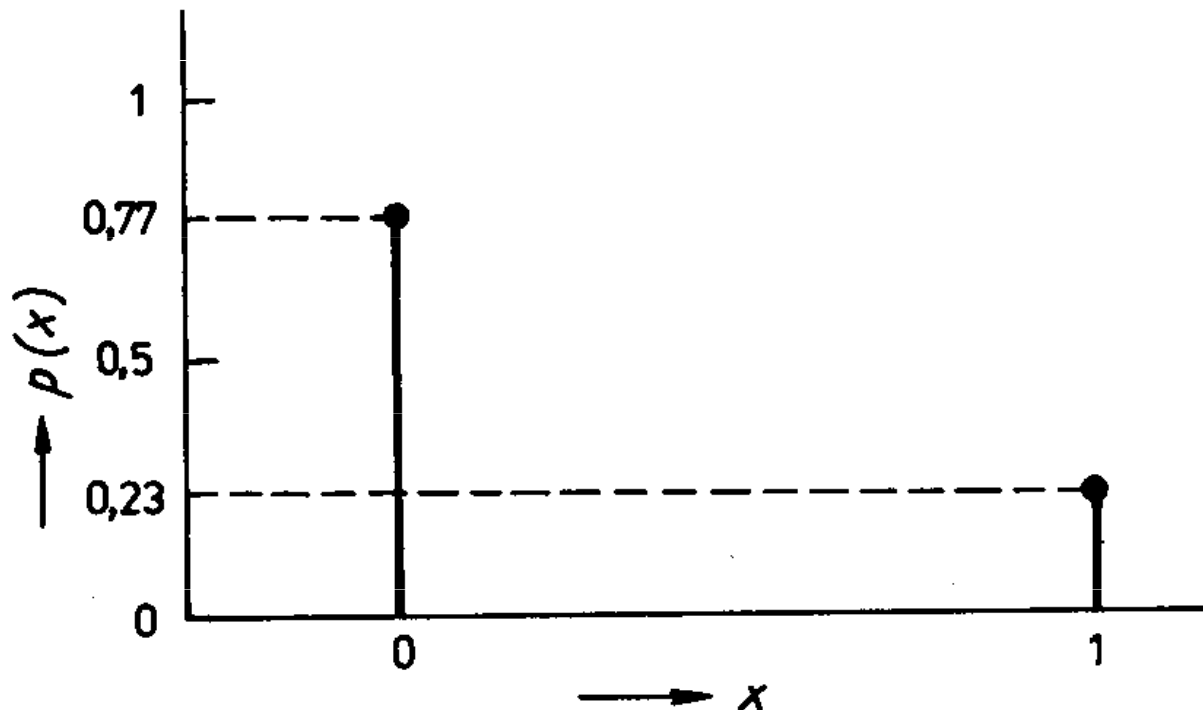
- $D^2(kx) = k^2D^2(x)$ kde k je konstanta.
- $D^2(x_1 + x_2 + \dots + x_n) = D^2(x_1) + D^2(x_2) + \dots$
• $+ D^2(x_n)$
- $D^2(k_1x_1 + k_2x_2 + \dots + k_nx_n) = \sum_{i=1}^n k_i^2 D^2(x_i)$
kde k_1, k_2, \dots, k_n jsou konstanty.
- $D^2(x_1 - x_2) = D^2(x_1) + D^2(x_2)$

Alternativní rozdělení

- veličina může nabývat hodnot 0 nebo 1 (přítomnost či nepřítomnost určitého znaku).

$$p(x) = 1 - p \quad \text{pro } x = 0$$

$$p(x) = p \quad \text{pro } x = 1$$



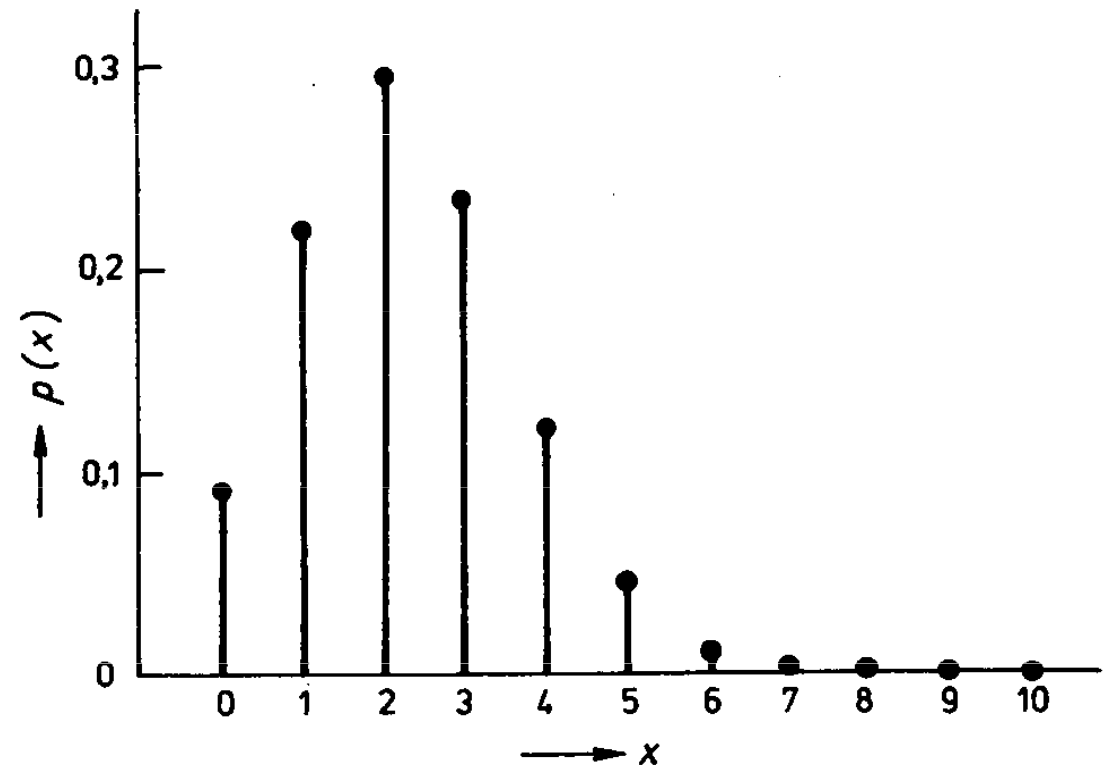
Alternativní rozdělení

- $F(x) =$
 - 0 pro $x \leq 0$
 - p pro $0 < x \leq 1$
 - 1 pro $x > 1$
- střední hodnota: $E(X) = p$
- rozptyl: $D^2(X) = p(1 - p)$

Binomické rozdělení

- náhodná veličina nabývá pouze hodnot 0, 1, 2, ..., n (= počet kladných výsledků z n nezávislých pokusů).

$$p_k = \binom{n}{k} p^k (1-p)^{n-k}$$



Binomické rozdělení

- $F(x) = 0$ pro $x < 0$
- $F(x) = \sum_{i=0}^x \binom{n}{i} p^i (1-p)^{n-i}$ pro $0 \leq x \leq n$
- $F(x) = 1$ pro $x > n$

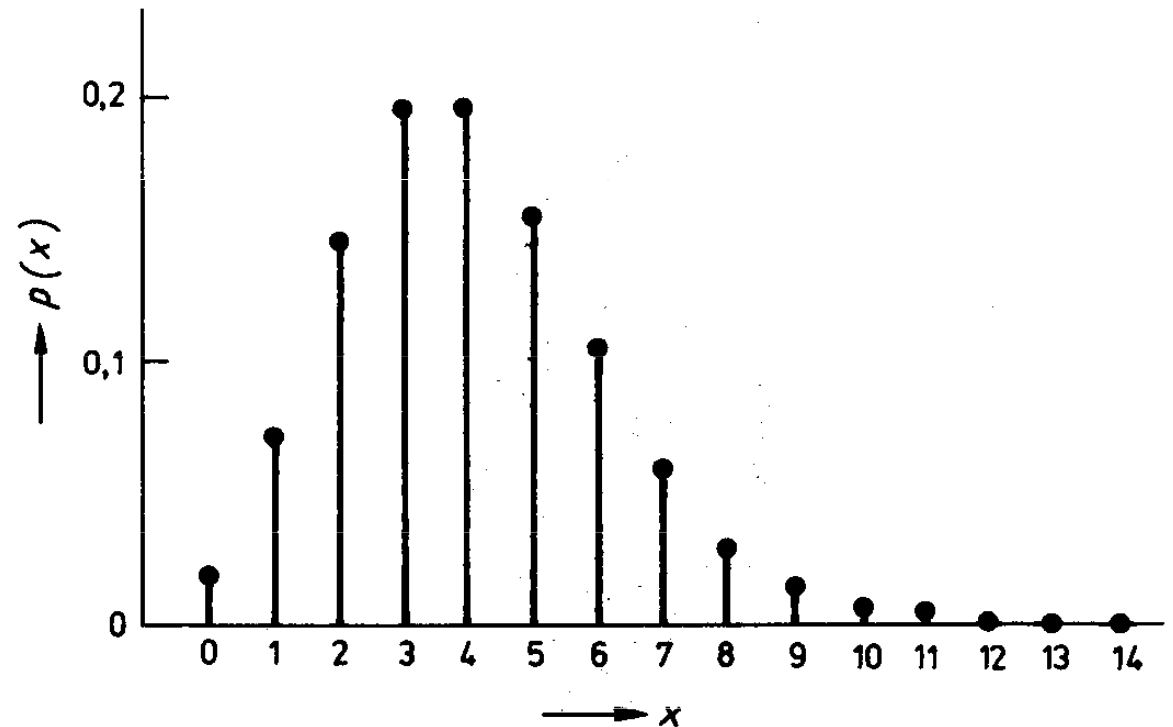
Střední hodnota: $E(X) = np$

Rozptyl: $D^2(X) = np(1-p)$

Poissonovo rozdělení

je limitou binomického rozdělení, je to „rozdělení vzácných jevů“.

$$p_k = \frac{e^{-\lambda} \lambda^k}{k!}$$



Poissonovo rozdělení

- $F(x) = 0$ pro $x < 0$

- $F(x) = \sum_{i=0}^x \frac{e^{-\lambda} \lambda^i}{i!}$ pro $x \geq 0$

Střední hodnota a rozptyl:

$$E(X) = D^2(X) = \lambda$$

Rovnoměrné rozdělení

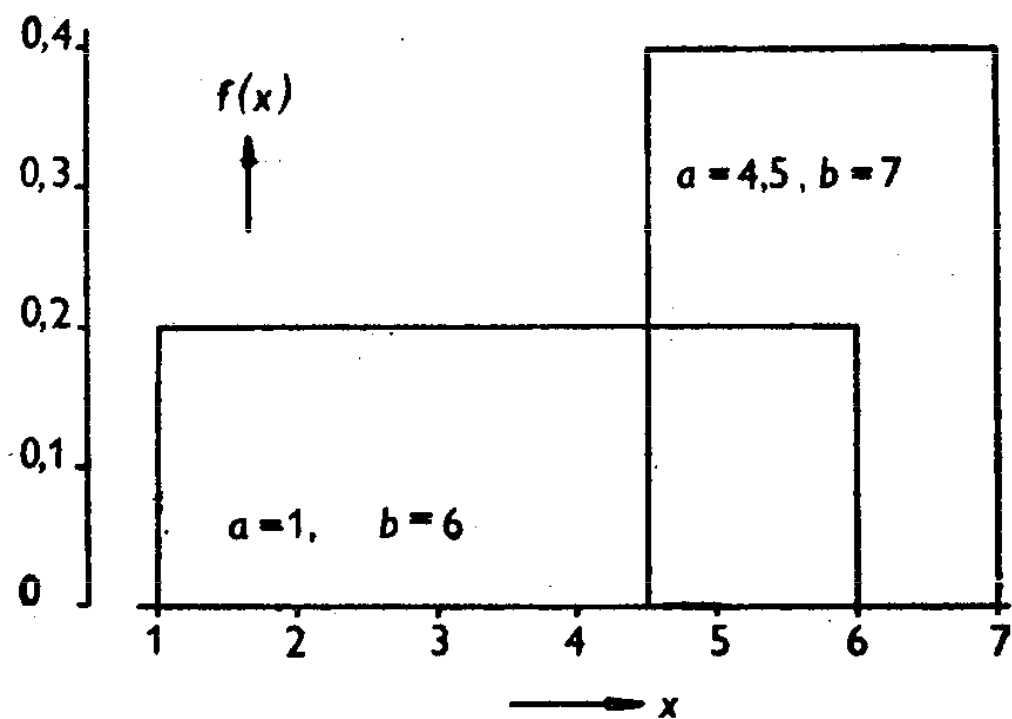
Hustota pravděpodobnosti v intervalu (a, b) má tvar:

$$f(x) = \frac{1}{b-a}$$

$$x \in (a, b)$$

$$f(x) = 0$$

ostatní



Rovnoměrné rozdělení

Distribuční funkce je

- $F(x) = 0,$ pro $x < a$
- $F(x) = \frac{x - a}{b - a}$ pro $a \leq x \leq b$
- $F(x) = 1,$ pro $x \geq b$
- Střední hodnota: $E(X) = \frac{a + b}{2}$
- Rozptyl: $D^2(X) = \frac{(b - a)^2}{12}$

Normální (Gaussovo) rozdělení

- Hustota pravděpodobnosti

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-[(x-\mu)/\sigma]^2 / 2}$$

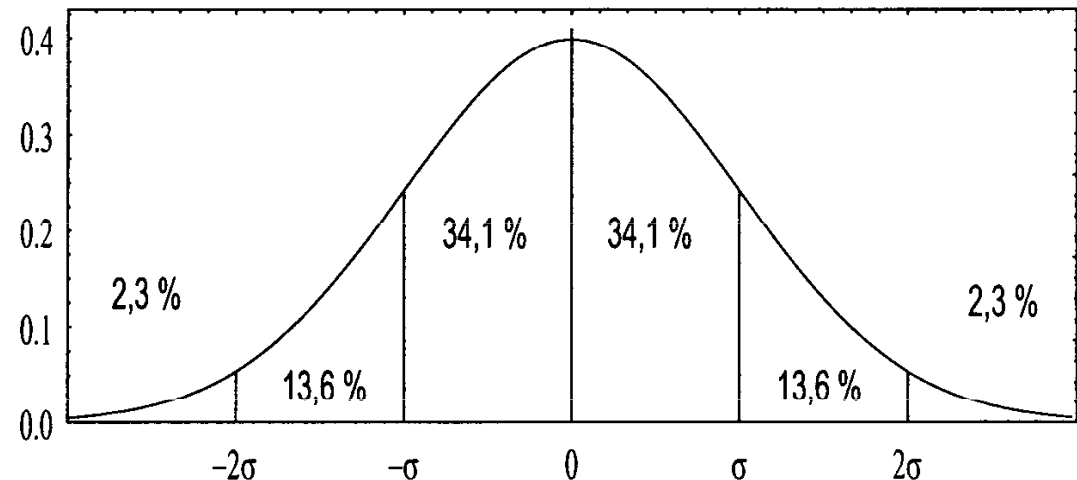
- Distribuční funkce

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^x e^{-[(y-\mu)/\sigma]^2 / 2} dy$$

Střední hodnota: $E(x) = \mu$,

Rozptyl: $D^2(x) = \sigma^2$

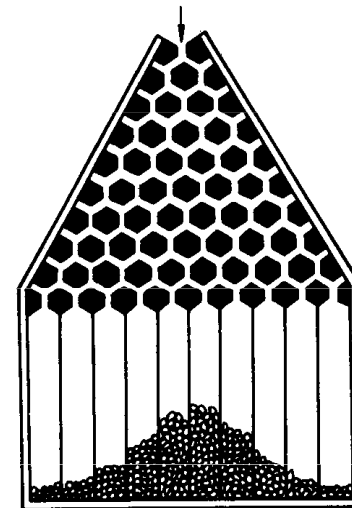
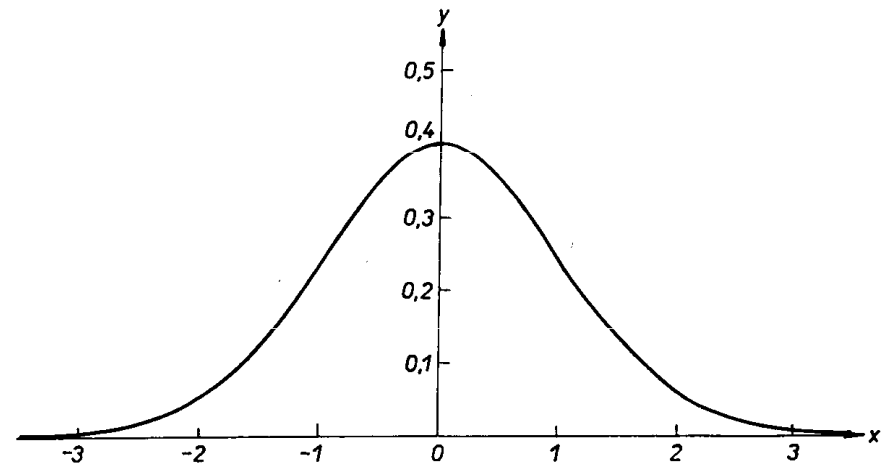
Hustota normovaného normálního rozdělení



Normální rozdělení

Centrální limitní věta:

průměr „velmi velkého“
náhodného výběru je
náhodnou veličinou s
přibližně normálním
rozdělením, i když má
základní soubor rozdělení jiné
než normální.



Normované normální rozdělení

Distribuční funkce normálního rozdělení závisí na μ a σ^2 . Proto se tabeluje Normované normální rozdělení, tj. normální rozdělení veličiny z (*z-skór*)

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^z e^{-y^2/2} dy$$

$$\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$

$$z = \frac{x - \mu}{\sigma}$$

Střední hodnota:

$$E(z) = 0$$

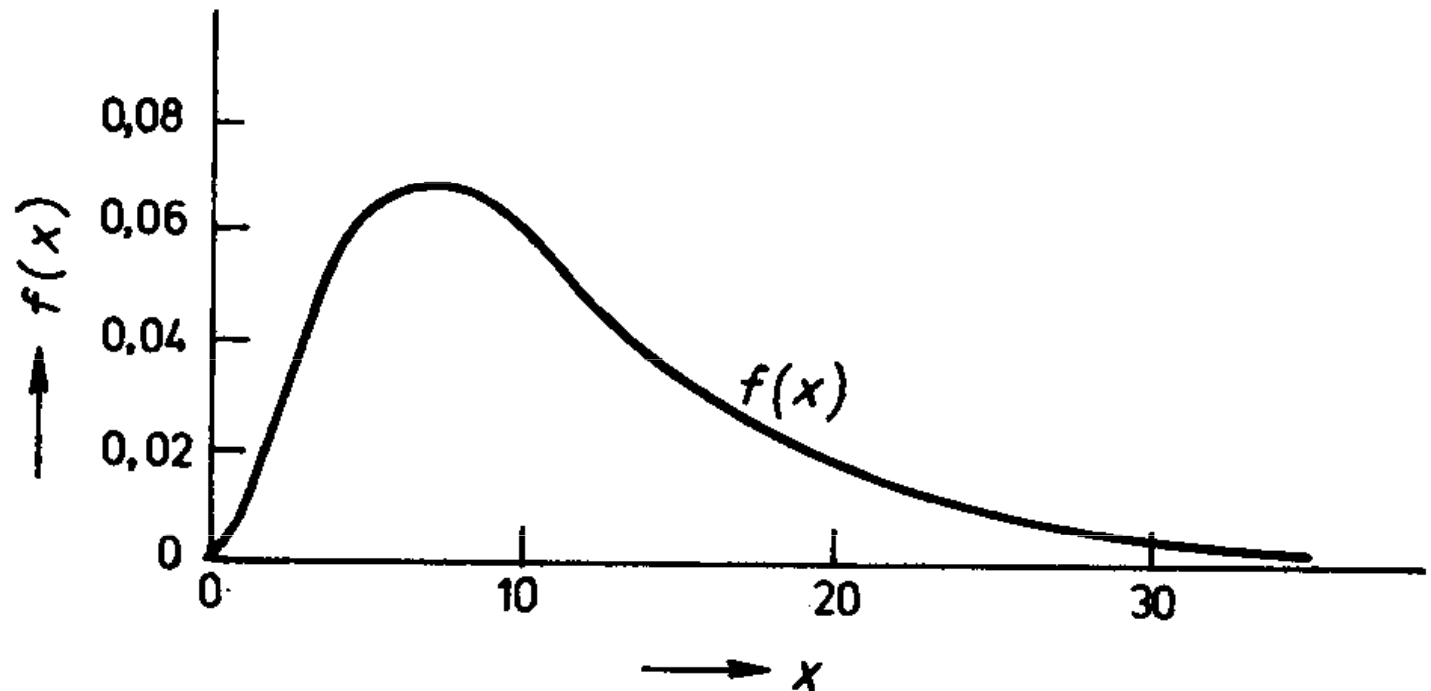
Rozptyl:

$$D^2(z) = 1$$

Logaritmicko-normální rozdělení

Hustota pravděpodobnosti

$$f(x) = \frac{0,4343}{\sigma\sqrt{2\pi}} e^{-(\log x - \mu)^2 / 2\sigma^2}$$



Logaritmicko-normální rozdělení

Distribuční funkce

$$F(x) = \frac{0,4343}{\sigma\sqrt{2\pi}} \int_0^x e^{-(\log y - \mu)^2 / 2\sigma^2} dy$$

Střední hodnota:

$$E(x) = e^{\mu/0,4343 + \sigma^2 / 2 \cdot (0,4343)^2}$$

Rozptyl:

$$D^2(x) = e^{\mu/0,4343 + \sigma^2 / 2 \cdot (0,4343)^2} \left[e^{\sigma^2 / (0,4343)^2} - 1 \right]$$

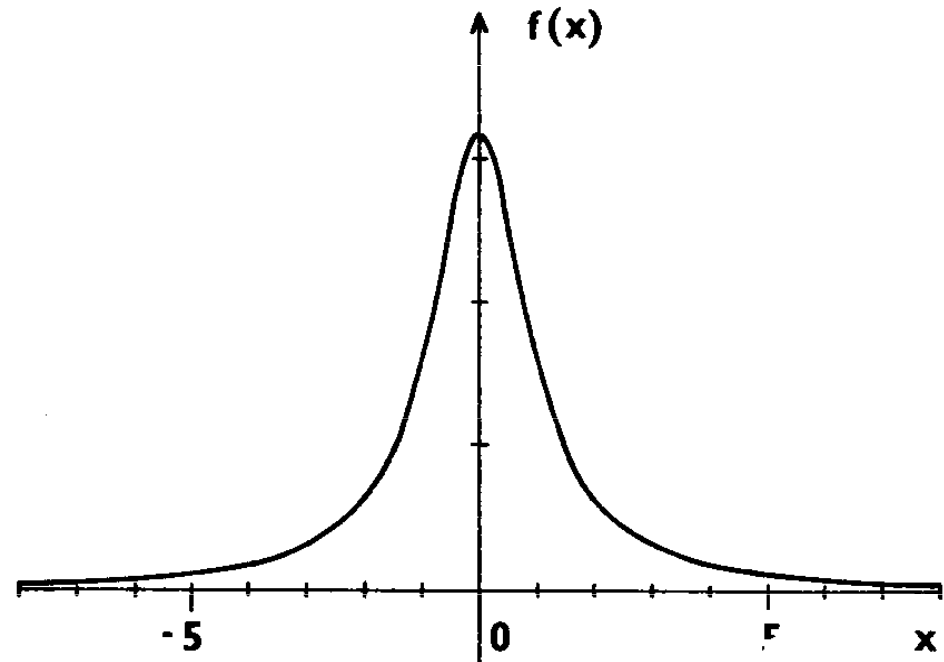
Cauchyovo rozdělení

- *Hustota pravděpodobnosti*

$$f(x) = \frac{\beta}{\pi[\beta^2 + (x - \alpha)^2]}, \quad -\infty < x < \infty$$

kde pro parametry platí

$$-\infty < \alpha < \infty, \beta > 0.$$



Cauchyovo rozdělení

Distribuční funkce

$$F(x) = \frac{1}{2} + \frac{1}{\pi} \operatorname{arctg}\left(\frac{x - \alpha}{\beta}\right) \quad , \quad -\infty < x < \infty$$

Střední hodnota: $E(x)$ není definována,

Rozptyl: $D^2(x) = \infty$.

Rozdělení na kružnici

Normální rozdělení na kružnici (von Misesovo rozdělení)

Např. úhly, hodiny během dne, dny během roku, orientace vůči světovým stranám, apod.

Jiná rozdělení spojité náhodné veličiny

- ***Smíšené rozdělení.*** Náhodná veličina je pozorována za různých podmínek a pozorované hodnoty pocházejí ze dvou nebo více různých základních souborů a to s různými pravděpodobnostmi.
- ***Cenzurované rozdělení.*** Známe pouze jednu část hodnot náhodné veličiny, hodnoty z druhé části neznáme, ale registrujeme jejich výskyt (např. hodnoty koncentrací pod mezí stanovitelnosti).
- ***Useknuté rozdělení.*** Nelze pozorovat všechny hodnoty náhodné veličiny, ale jen hodnoty z určitého intervalu.