CHEMOMETRIE I

Statistické zpracování jednorozměrných dat

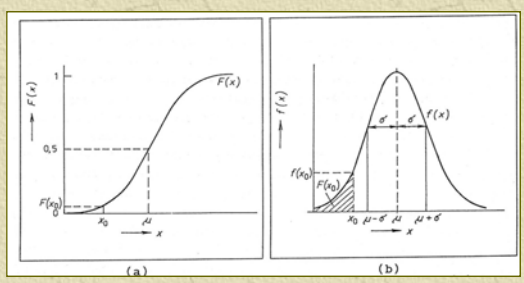
Doporučená literatura

- Meloun M., Militký J.: Statistické zpracování experimentálních dat. Plus, Praha 1994. + novější vydání.
- Meloun M., Militký J.: Kompendium statistického zpracování dat. Academia, Praha 2002.
- Podklady na webových stránkách katedry.
- http://meloun.upce.cz

Rozdělení pravděpodobnosti

- diskrétní rozdělení
- spojitá rozdělení

Normální rozdělení



- Distribuční funkce
- Graf hustoty pravděpodobnosti

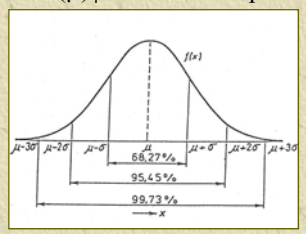
$$p(x) = \frac{1}{\sigma \cdot \sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

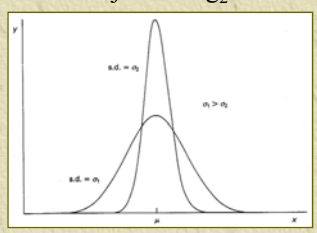
- Normování – transformace z $N(\mu, \sigma^2)$ na $N(0, 1)$:

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

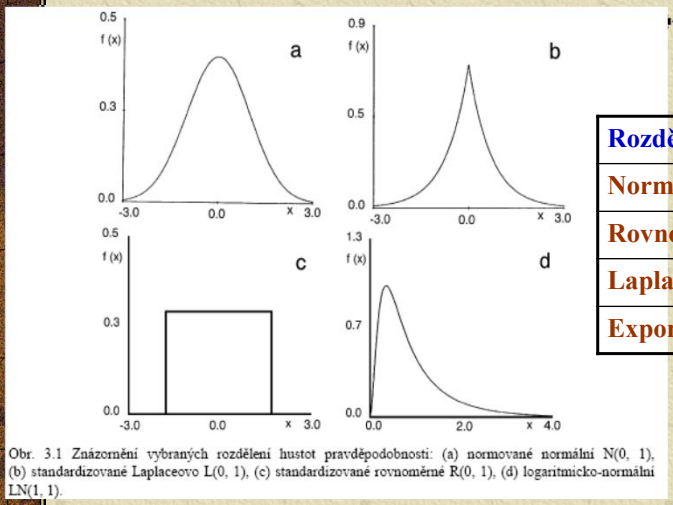
Normální rozdělení

- Momenty = číselné charakteristiky rozdělení podávající informace o vlastnostech rozdělení:
- Obecný moment r-tého stupně $M_r = \int x^r p(x)\,dx$
- Centrální moment r-tého stupně $M(\mu)_r = \int (x-\mu)^r p(x)\,dx$
  - M_1 – střední hodnota – charakterizuje polohu
  - $M(\mu)_2$ – rozptyl – charakterizuje přesnost
  - $M(\mu)_3$ – koeficient šikmosti – charakterizuje tvar – g_1
  - $M(\mu)_4$ – koeficient špičatosti – charakterizuje tvar – g_2





„Experimentální“ rozdělení

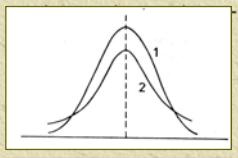


Obr. 3.1 Znázornění vybraných rozdělení hustot pravděpodobnosti: (a) normované normální N(0, 1), (b) standardizované Laplaceovo L(0, 1), (c) standardizované rovnoměrné R(0, 1), (d) logaritmicko-normální LN(1, 1).

Rozdělení	g_1	g_2
Normální	0	3
Rovnoměrné	0	1,8
Laplaceovo	0	6
Exponenciální	2	9

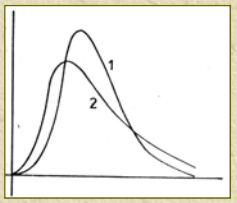
- Exponenciální rozdělení (=jednostranné Laplaceovo)

„Teoretická“ rozdělení



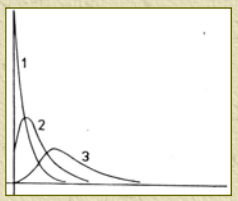
Studentovo rozdělení
1: $\nu = \infty$; 2: $\nu = 1$

$$t = \frac{|x - \mu|}{\sigma}\sqrt{n}$$



Fisherovo-Snedecorovo rozdělení
1: $\nu_1 = 10$, $\nu_2 = 50$; 2: $\nu_1 = 10$, $\nu_2 = 4$

$$F = \frac{\sigma_1^2}{\sigma_2^2}$$



χ^2 rozdělení (chí kvadrát)
1: $\nu = 2$; 2: $\nu = 4$; 3: $\nu = 10$

$$\chi^2 = \sum_i \left(\frac{x_i - \bar{x}}{\sigma}\right)^2$$

Bodové odhady polohy rozptýlení a tvaru

Vlastnosti bodových odhadů

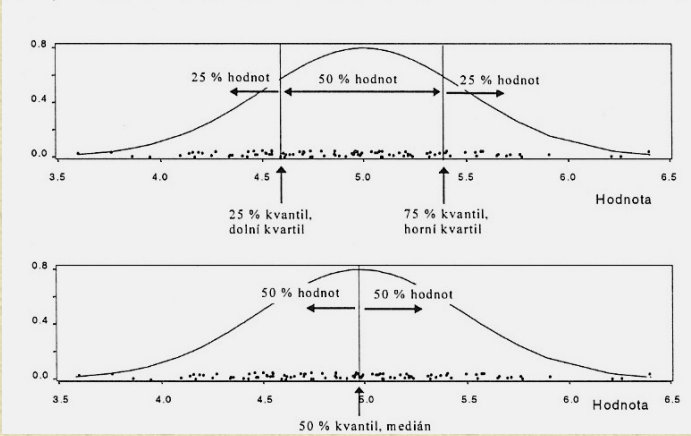
- KONZISTENTNOST odhadu – s rostoucí četností n se zmenšuje rozdíl mezi odhadem (= průměrem) a skutečnou hodnotou μ.
- NESTRANOST odhadu – pro n blížící se k ∞ $\bar{x} = \mu$
- VYDATNOST (efektivnost) odhadu – rozptyl odhadu okolo skutečné hodnoty μ se s rostoucí četností n zmenšuje.
- ROBUSTNOST odhadu – necitlivost na odchylky od předpokládaného rozdělení.

- Míry polohy mohou být založeny na kvantilech (kvantilové charakteristiky) nebo na momentech (momentové charakteristiky).

Kvantilové charakteristiky polohy

- KVANTILY – hodnoty znaku, které dělí data na určitý počet skupin o stejném počtu prvků.
- Medián $\tilde{x}_{0,5}$ kvantil, který rozděluje data na 2 části o 50 % rozsahu souboru.
  - pro n liché – prostřední člen uspořádaného výběru $x_{(k)}$
  - pro n sudé: $\tilde{x}_{0,5} = \frac{x_{(k)} + x_{(k+1)}}{2}$ kde $k = \frac{n+1}{2}$
- Kvartily – rozdělují uspořádanou řadu hodnot na 4 skupiny se stejnou četností. Prostřední kvartil = medián. Dolní a horní kvartil: $\tilde{x}_{0,25}$ $\tilde{x}_{0,75}$
- Decily – rozdělují uspořádanou řadu hodnot na 10 skupin o stejně velké četnosti: $\tilde{x}_{0,1}, \tilde{x}_{0,2}$ L

Kvantily



Momentové charakteristiky polohy

- Aritmetický průměr $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- Vážený aritmetický průměr

$$\bar{x}_w = \frac{\sum w_i \cdot x_i}{\sum w_i}, \text{ kde } w_i \text{ je statistická váha}$$

- Geometrický průměr $\bar{x}_g = \sqrt[n]{x_1 \cdot x_2 \, L \, x_n}$
- Kvadratický průměr $\bar{x}_{sq} = \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n}}$

Momentové charakteristiky polohy

- Další 2 průměry se někdy řadí mezi robustní charakteristiky.
- Uřezaný průměr

$\bar{x}(\upsilon)$ 10% uřezání $\bar{x}(10)$

$$\bar{x}(\upsilon)=\frac{1}{n-2M}\sum_{i=M+1}^{n-M} x_i \qquad M=\text{int}\left(\frac{\upsilon\cdot n}{100}\right)$$

- Winsorizovaný průměr

$$\bar{x}_w(\upsilon)=\frac{1}{n}\left[(M+1)\cdot(X_{(M+1)}+X_{(n-M)})+\sum_{i=M+2}^{n-M-1} x_i\right]$$

Winsorizace = nahrazení odlehlých výsledků sousedními výsledky uspořádaného souboru, které již nejsou odlehlé; nezmenšuje se četnost souboru a zachovává se charakter dat.

Bodové odhady míry rozptýlení

- Rozpětí $R = x_{max} - x_{min}$
- Interkvartilové rozpětí $R = \tilde{x}_{0,75} - \tilde{x}_{0,25}$
- Rozptyl σ^2 / směrodatná odchylka σ

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

- Výběrový odhad rozptylu s^2 / výběrový odhad směrodatné odchylky s

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

- Relativní směrodatná odchylka - RSD

$$RSD=\frac{s}{\bar{x}}\cdot 100 \quad [\%]$$

Bodové odhady tvaru rozdělení

- Koeficient šikmosti g_1 – číslo, které charakterizuje nesouměrnost rozdělení, dává informace o tvaru rozdělení co do zešikmení resp. Nesouměrnosti.
- Koeficient špičatosti g_2 – číslo, charakterizující zkoncentrování (protažení) prvků souboru v blízkosti určité hodnoty znaku.

$$g_1 = \frac{\sqrt{n}\cdot\sum(x_i-\bar{x})^3}{\left[\sum(x_i-\bar{x})^3\right]^{3/2}} \qquad g_2 = \frac{n\cdot\sum(x_i-\bar{x})^4}{\left[\sum(x_i-\bar{x})^2\right]^2}$$