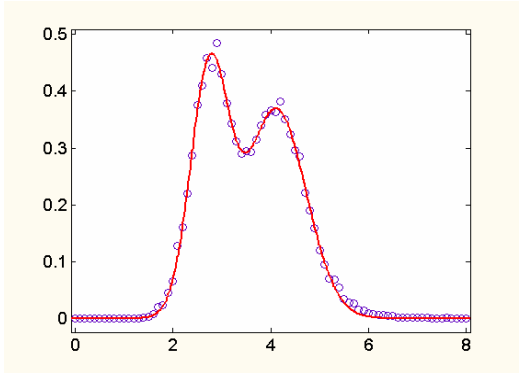
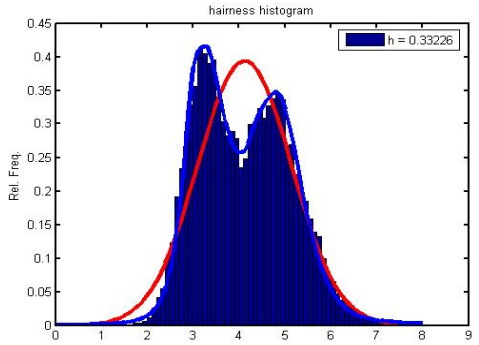




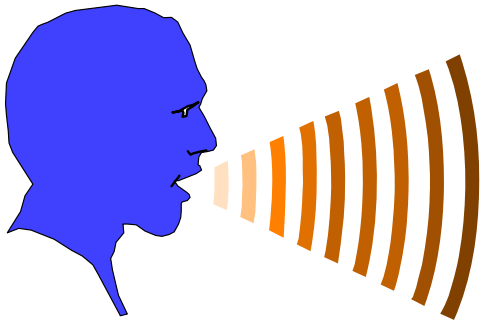
IDENTIFIKACE BIMODALITY V DATECH

Jiří Militky

Technická universita v Liberci
e- mail: jiri.miliky@vslib.cz

Milan Meloun

Universita Pardubice,
Pardubice

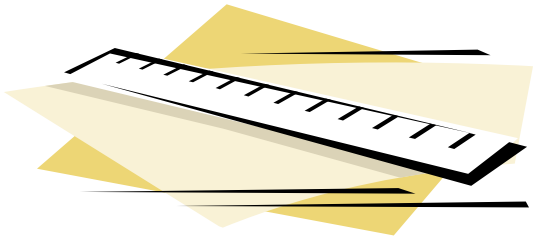


Motto: *Je normální předpokládat normální data ?*

Zvláštnosti dat

- Obsahují občas extrémně velké hodnoty, které však nejsou důsledkem chyb měření.
- Mohou být cenzurována zdola s ohledem na limitu detekce přístrojů.
- Jsou vždy kladná a výrazně zešikmená k vyšším hodnotám
- Rozsahy zpracovávaných dat jsou buď malé nebo extrémně veliké
- Jsou často prostorově nebo časově závislá.
- Rozdělení dat je jen zřídka.
- Z nejrůznějších důvodů může rozdělení dat obsahovat více lokálních maxim (polymodalita).

$$C_K = \int_{-\infty}^{\infty} (x - M_1)^K f(x)dx \qquad \hat{C}_K = \frac{1}{N}\sum_{i=1}^{N} (x_i - \hat{M}_1)^K$$



Základní model

$$M_K = \int_{-\infty}^{\infty} x^K f(x)dx \qquad \hat{M}_K = \frac{1}{N}\sum_{i=1}^{N} x_i^K$$

Aditivní model měření $x = \mu + \varepsilon$

μ skutečná hodnota měřené veličiny (střední hodnota)

ε náhodná chyba měření resp. „šumová“ složka

Předpoklady o chybách:

1. střední hodnota chyb měření je nulová, t.j. $E(\varepsilon) = 0$
2. rozptyl chyb měření je konstantní, t.j. $D(\varepsilon) = \sigma^2$
3. chyby jsou vzájemně nezávislé .t.j. $E(\varepsilon_i * \varepsilon_j) = 0$
4. chyby mají normální rozdělení t.j. $\varepsilon \approx N(0, \sigma^2)$

Za předpokladu nezávislosti měření platí,že známé momentové odhady aritmetický průměr a výběrový rozptyl jsou odhady střední hodnoty a rozptylu

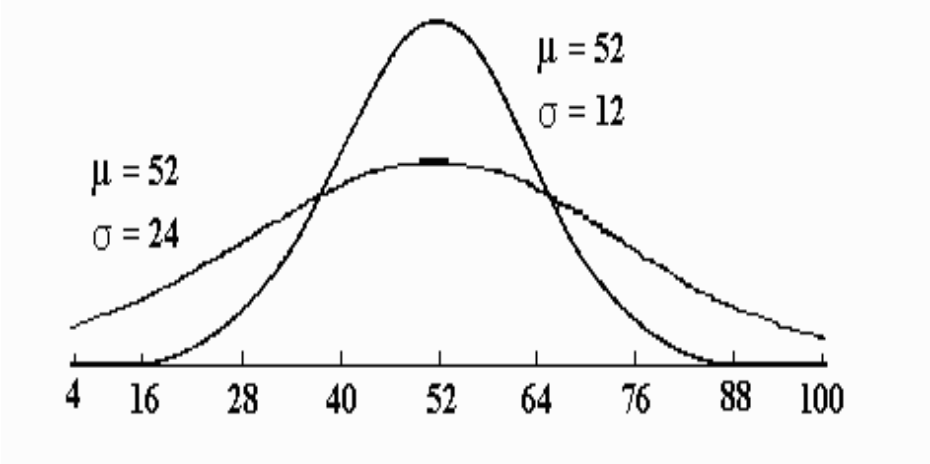
Normální rozdělení N(μ, σ^2): $$f(x) = \frac{1}{\sqrt{2 * \pi}} e^{(x-\mu)^2 / 2\sigma^2}$$

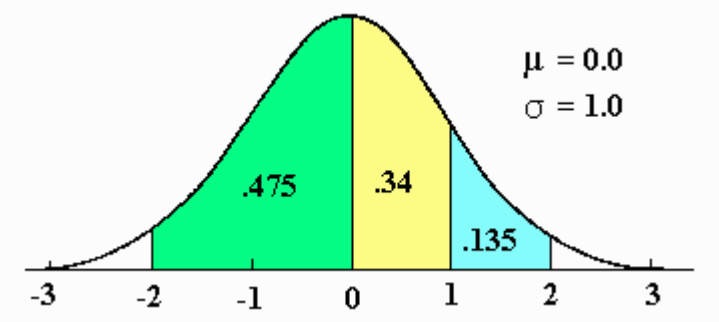
Normální rozdělení

Nezávislá měření ne nutně normální

Parametr μ odhad $\bar{x}$ rozptyl $D(\bar{x}) = \frac{\sigma^2}{N}$

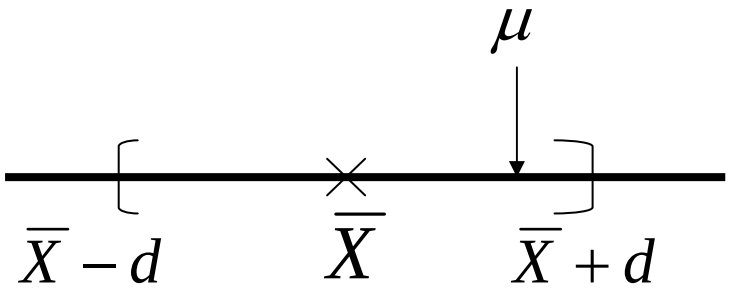
Parametr σ^2 odhad s^2 rozptyl $D(s^2) = \frac{2 * \sigma^4}{N}$





$$Z = \frac{X - \mu}{\sigma}$$

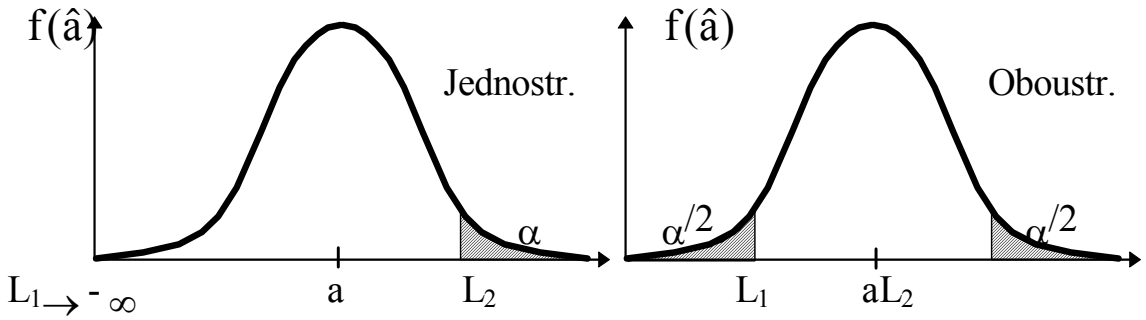
Intervalové odhady



- "IS": interval obsahující se zadanou pravděpodobností (1-α) parametr a. $P(L_1 \le a \le L_2) = 1 - \alpha$

(1 - α) koeficient konfidence, statistická jistota (0.99, 0.95)

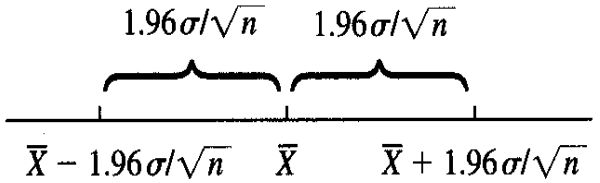
α hladina významnosti (α = 0.01, 0.05)



větší N → užší IS
větší σ^2 → širší IS
větší α → užší IS

Platí pouze pro normální rozdělení !

Konstrukce IS



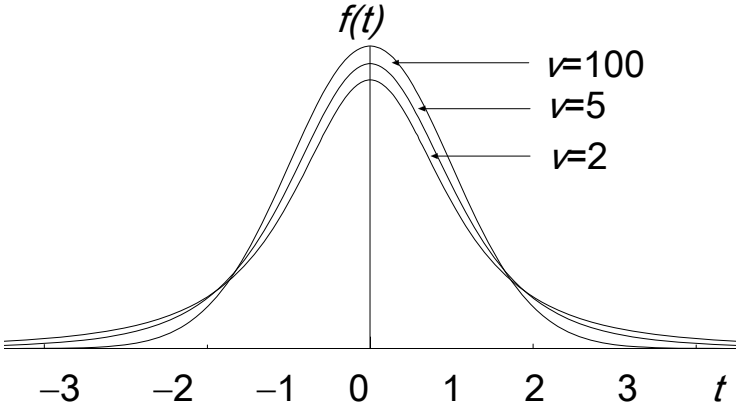
data x_i ... $N(\mu, \sigma^2)$

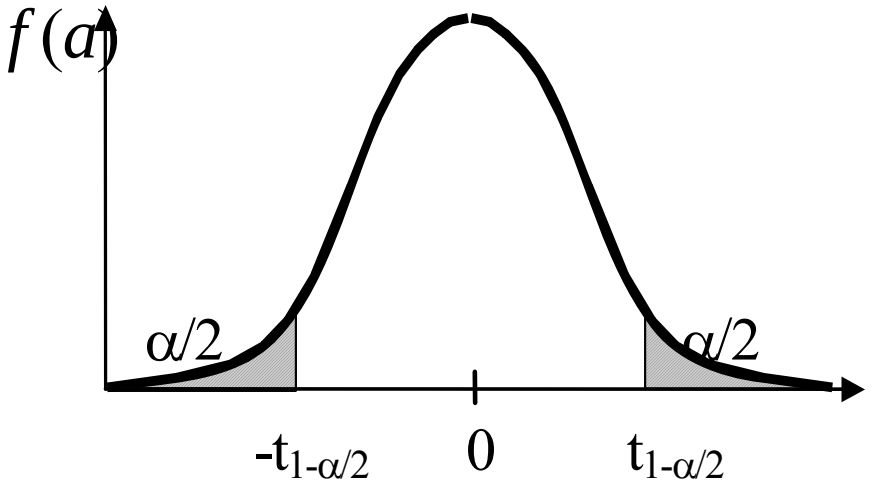
$t = (\bar{x} - \mu) / s.\sqrt{N}$ Studentovo rozdělení, d.f. = N - 1

$\chi^2 = (N-1).s^2 / \sigma^2$ Chí-kvadrát rozdělení, d.f. = N - 1

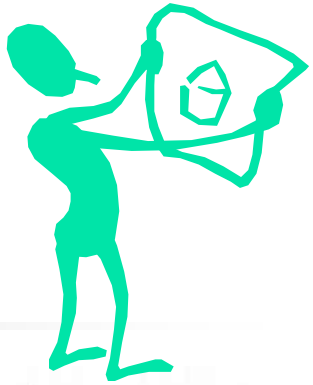
$$P(-t_{1-\alpha/2} \le (\bar{x} - \mu) / s.\sqrt{N} \le t_{1-\alpha/2}) = 1 - \alpha$$

$$\bar{x} - t_{1-\alpha/2}.s/\sqrt{N} \le \mu \le \bar{x} + t_{1-\alpha/2}.s/\sqrt{N}$$





Interpretace IS



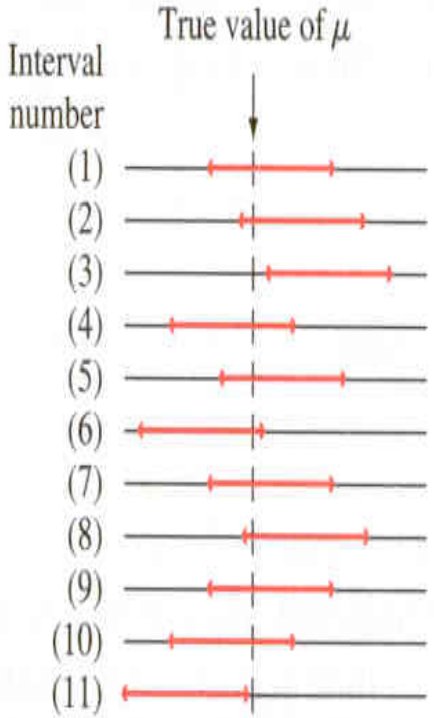
95% interval spolehlivosti.

- správná interpretace "95% confidence" se týká četnosti jevu A
- Jev A:

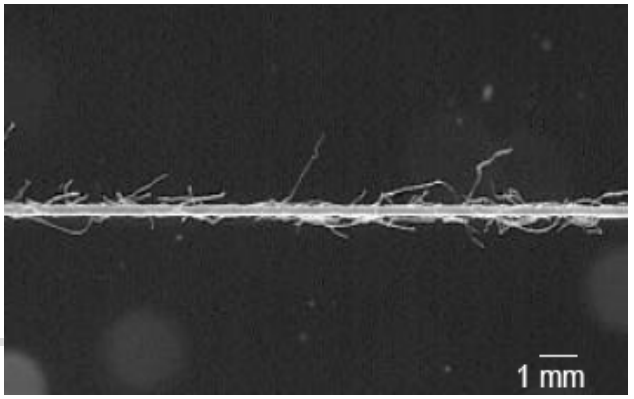
$$\bar{X} - 1.96 \cdot \sigma / \sqrt{n} < \mu < \bar{X} + 1.96 \cdot \sigma / \sqrt{n}$$

- P(A) = 0.95

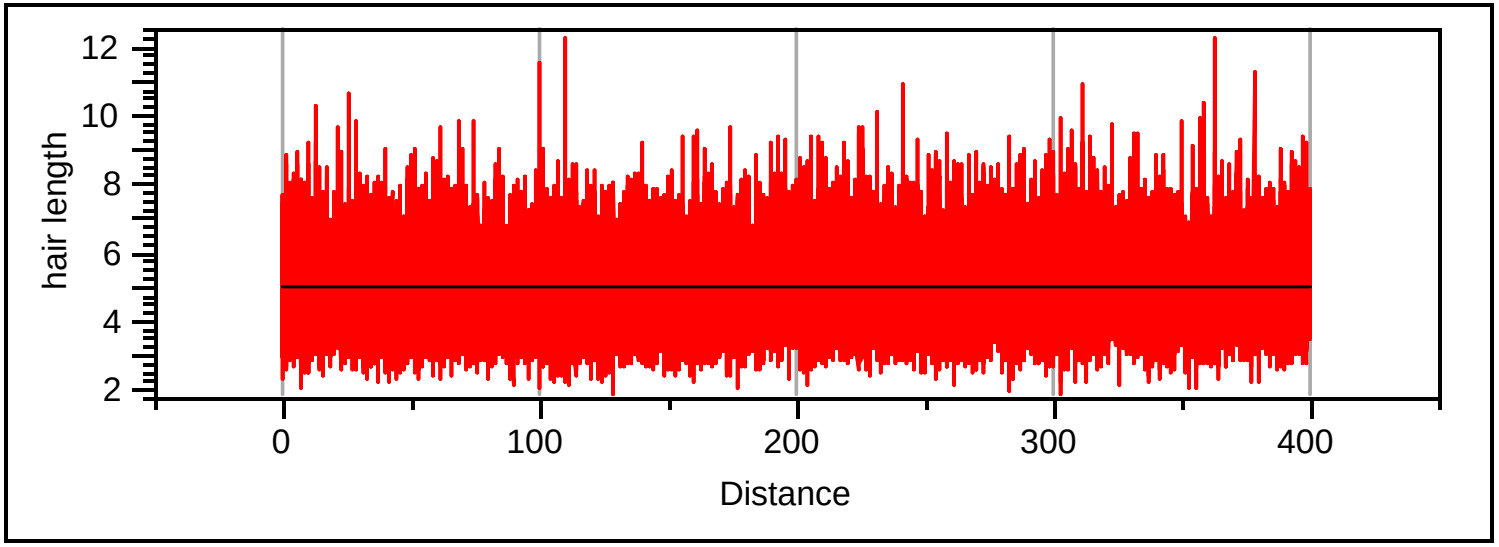
→ 95% všech intervalů spolehlivosti obsahuje μ.

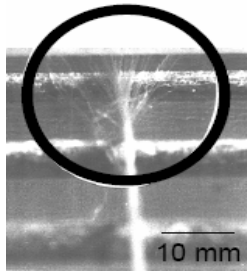


Testovací data

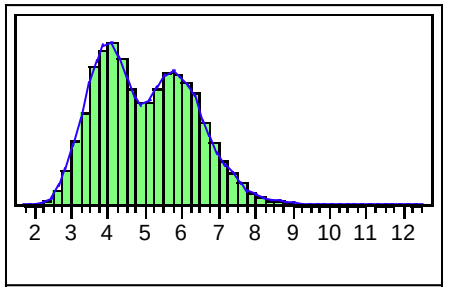


- Reálná data z měření chlupatosti přízí na přístroji Uster. Jedná se o výběr velikosti 14 000 pro kompaktní přízi C 30tex





Vizualizace rozdělení dat



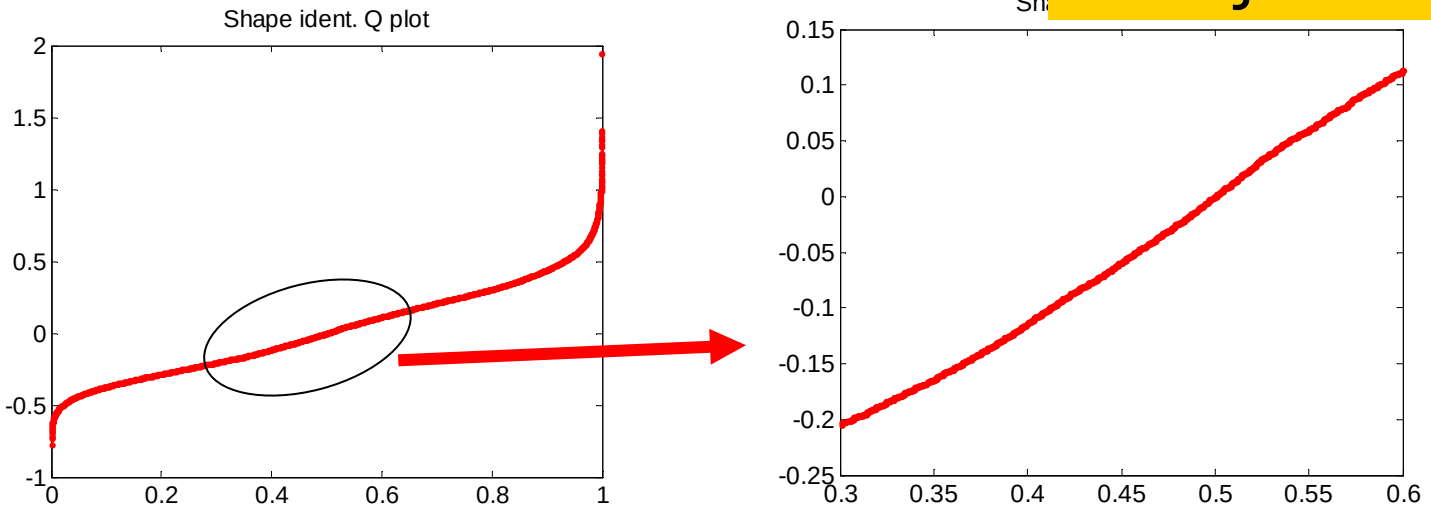
Tvar identifikující kvantilová funkce *QI(P)*

$$QI_e(P_i) = \frac{(x_{(i)} - x_{0.5})}{2 * (x_{0.75} - x_{0.25})}$$

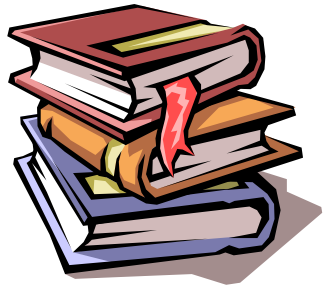
$$P_i = \frac{i}{N+1}$$

Graf **tvar identifikující kvantilové funkce:** závislost mezi $QI_e(P_i)$ a P_i.

Sigmoidální tvar v centrální oblasti indikuje bimodalitu

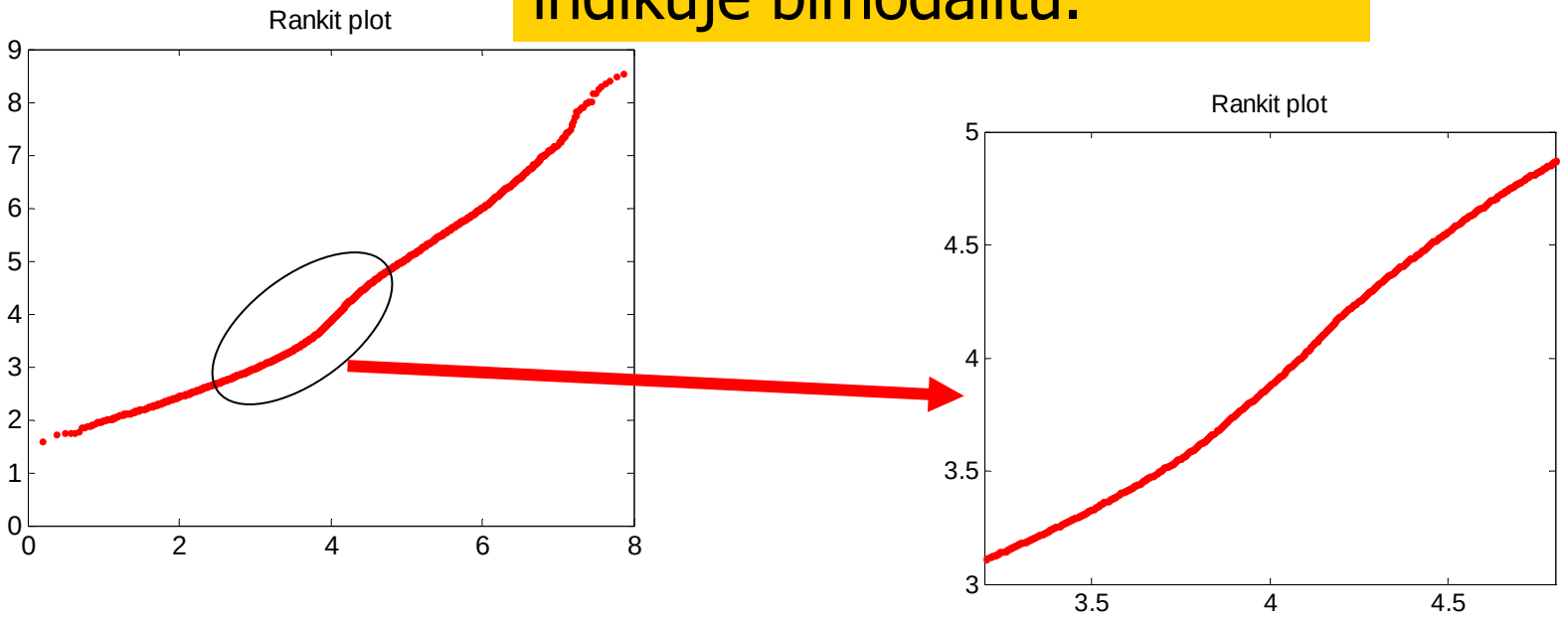


Rankitový graf



***Q-Q* graf** je založen na porovnání empirické kvantilové funkce $Q_e(P_i) \cong x(i)$ s vybranou teoretickou kvantilovou funkcí *QT(Pi)*. Pro normální kvantilovou funkci jde o rankitový graf

Ohyb v centrální oblasti zde indikuje bimodalitu.



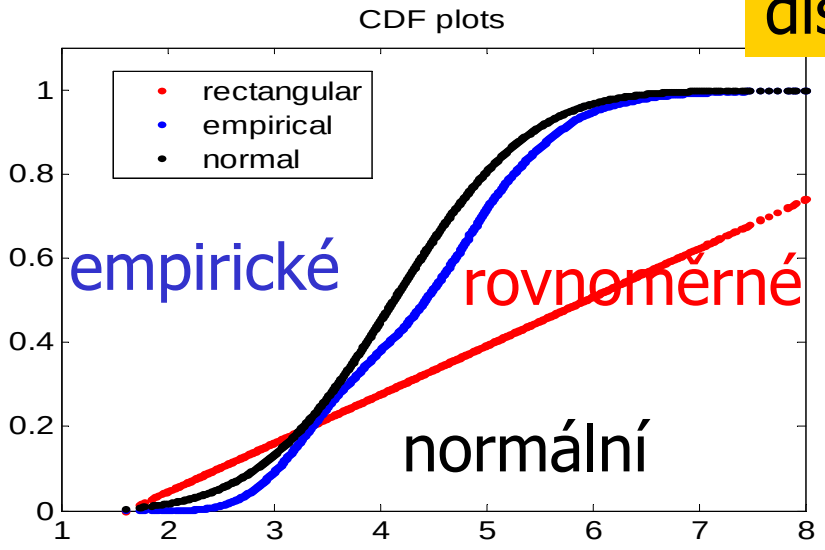
Pořádkové statistiky $x_{(1)} \le x_{(2)} \le .. \le x_{(i)} .. \le x_{(N)}$

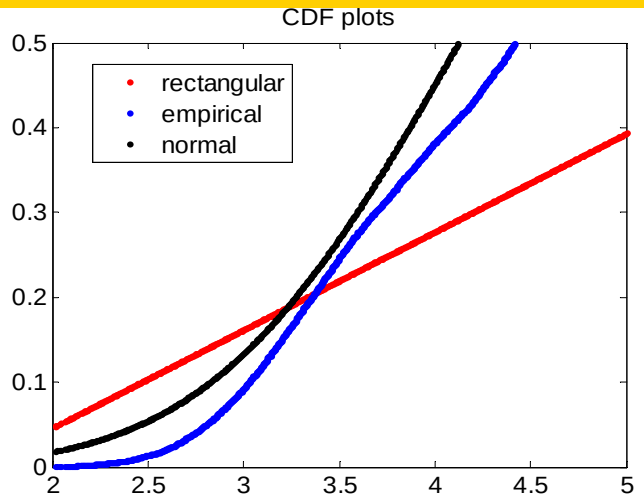
Distribuční funkce

Odhad distribuční funkce: **lokální součet pořádkových statistik**

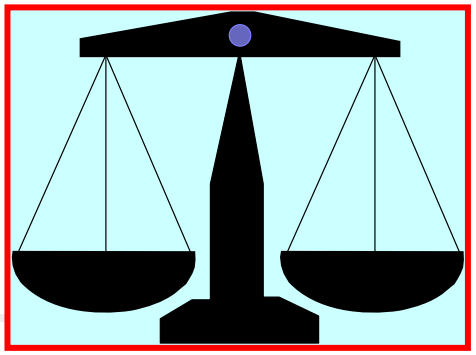
$$cdf(x) = \frac{\sum_{i=1}^{i} x_{(i)}}{\sum_{i=1}^{N} x_{(i)}} \quad , \quad \text{for } x_{(i)} \le x \le x_{(i+1)}$$

Typický „skok" na empirické distribuční funkci indikuje bimodalitu





Porovnávací PP graf

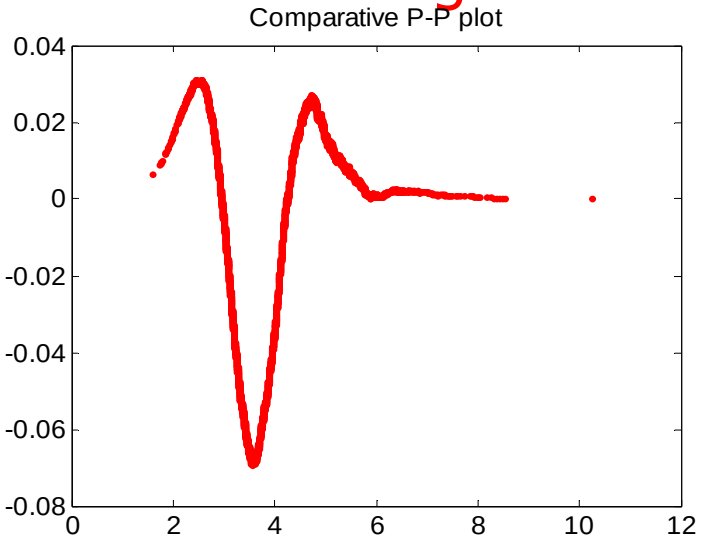


Porovnávající distribuční funkce

$$CDD = F_T(Q_e(P_i))$$

Pro indikaci bimodality se jako $F_T(.)$ volí distribuční funkce normálního rozdělení.

Porovnávací P-P graf : závislost $CDD - P_i$ vs. P_i.



V případě unimodálního normálního rozdělení je odpovídající porovnávající *P-P* graf horizontální přímka na nulové úrovni. Bimodální rozdělení:typické píky

Histogram

- Histogram je po částech konstantní odhad hustoty pravděpodobnosti. Výška sloupce v j – té třídě ohraničené hodnotami (t_{j-1}, t_j) je určena ze vztahu

$$f_H(x) = \frac{C_N(t_{j-1}, t_j)}{N\ h_j}$$

délka j-té třídy (intervalu).

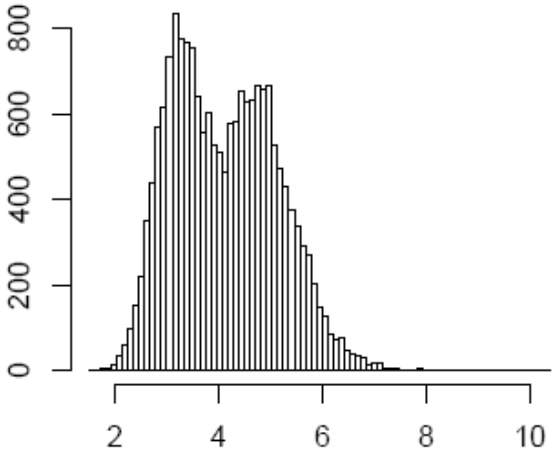
$$h_j = t_j - t_{j-1}$$

Zde funkce $C_N(a, b)$ označuje počet hodnot výběru v intervalu <a, b>

Pro přibližně normální data je délka tříd určena vztahem

$$h = 3.49 * (\min(s, Dq/2)/1.34)/n^{1/3}$$

$$Dq = 2 * (x_{0.75} - x_{0.25})$$



Jádrový odhad hustoty

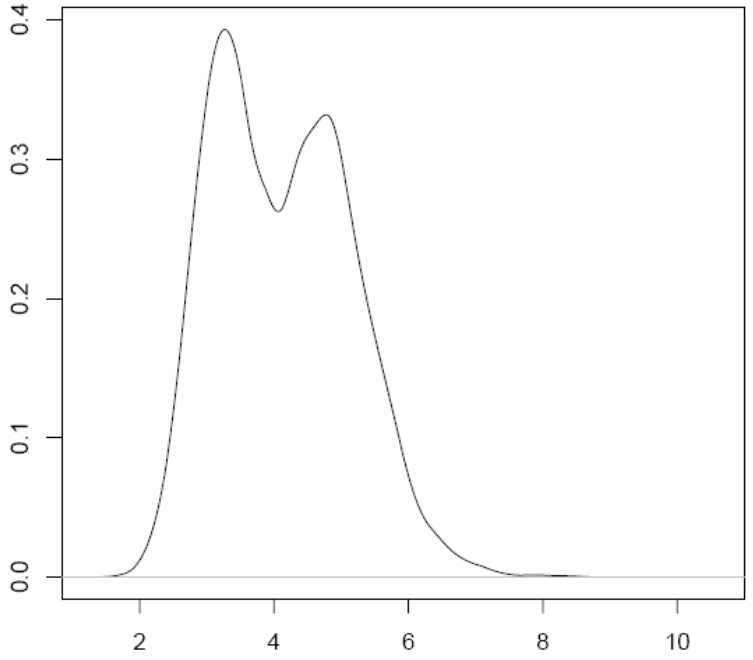
Jádrový odhad hustoty pravděpodobnosti *f(x)* (hladká funkce, závislá na parametru vyhlazení *h*). Tento odhad má pro případ konstantního parametru vyhlazení *h* tvar

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

Jednoduchá bikvadratická jádrová funkce *K[x]* má tvar

$$K(x) = 0.9375 * (1 - x^2)^2$$

$$\text{for } -1 \le x \le 1$$



Parametrický model bimodality

Unimodální model $$f_u(x_i) = A0 * \exp\left(-\frac{(x_i - B1)^2}{2 * C1^2}\right)$$

Bimodální model (směs dvou normálních rozdělení)

$$f_B(x_i) = A1 * \exp\left(-\frac{(x_i - B1)^2}{2 * C1^2}\right) + A2 * \exp\left(-\frac{(x_i - B2)^2}{2 * C2^2}\right)$$

Zde A1, A2 jsou podíly prvního normálního rozdělení (index 1) a druhého normálního rozdělení (index 2). Parametry B1 a B2 jsou střední hodnoty jednotlivých rozdělení a parametry C1, C2 jsou směrodatné odchylky.

Pro případ unimodality musí platit, že $|B1 - B2| < 2 * \min(C1, C2)$

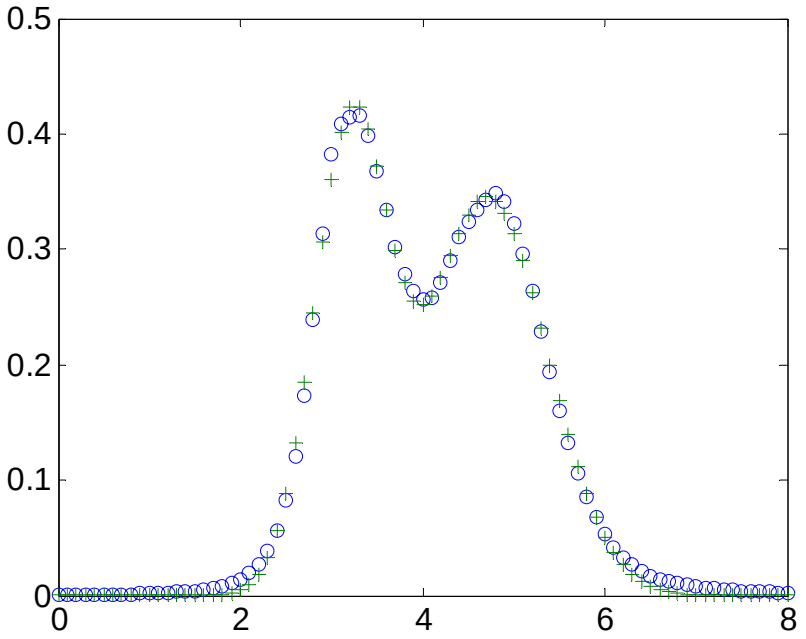
Odhad parametrů

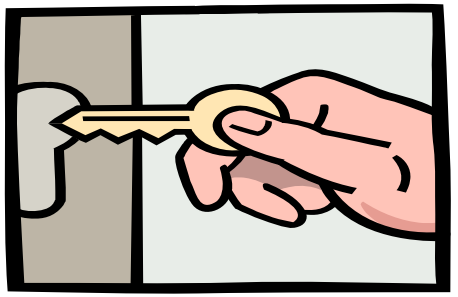
Pro odhad parametrů směsi dvou normálních rozdělení (A1, A2, B1, B2, C1 and C2) je možno použít nelineární metody nejmenších čtverců, kde data jsou získána z histogramu

$$r_i = \hat{f}(x_i) - f_B(x_i)$$

$$S = \sum_{i=1}^{N} r_i^2$$

V programu BIMODAL v jazyku MATLAB jsou použity algoritmy minimalizace nejmenších čtverců na bázi Levenberg-Marquardtovy metody a jejího vylepšení (Trust-region)





Test bimodality

Věrohodnostní poměr $$LR = 2 * \ln\left(\frac{L_B}{L_U}\right)$$

Funkce L_U $$L_U = \prod_{i=1}^{N} \frac{1}{\sqrt{2 * \pi * C1^2}} \exp\left(-\frac{(x_i - B1)^2}{2 * C1^2}\right)$$

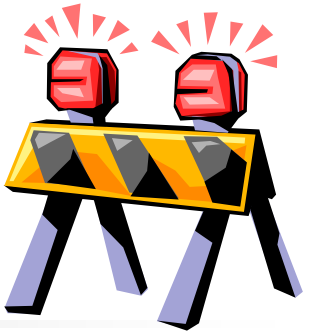
Funkce L_B $$L_B = \prod_{i=1}^{N} A1 * \exp\left(-\frac{(x_i - B1)^2}{2 * C1^2}\right) + A2 * \exp\left(-\frac{(x_i - B2)^2}{2 * C2^2}\right)$$

Statistika *LR* má přibližně $\chi^2(4)$ rozdělení, tj. pro $LR \leq 9$

je možné akceptovat jednoduché unimodální rozdělení.

Pro přízi C 30 tex je LR = 244.3 a použití směsi dvou rozdělení (bimodalita) je ověřeno.

Závěr



Bimodalitu lze indikovat jak pomocí grafických pomůcek průzkumové analýzy dat tak i pomocí parametrických modelů.

Existuje ještě celá řada formálních testů bimodality, které však nejsou tak informativní (neumožňují komplexní posouzení statistických zvláštností dat).

Samostatným problémem je následná analýza objasňující příčiny vzniku bimodality a je jí případná eliminace.

Pro případ chlupatosti příze je možno bimodalitu objasnit na základě modelu dvou typů chlupatosti