

# IDENTIFIKACE BIMODALITY V DATECH

JIRÍ MILITKÝ,

*Katedra textilních materiálů, Technická universita v Liberci, Hálkova 6  
461 17 Liberec, e- mail: jiri.miliky@vslib.cz*

MILAN MELOUN,

*Katedra analytické chemie, Universita Pardubice, Pardubice*

**Motto:** *Je normální předpokládat normální data ?*

## **Abstrakt:**

*Jsou uvedeny základní postupy průzkumové analýzy dat zaměřené na identifikaci bimodality v datech. Je pojednáno o možnosti ověření bimodality pomocí parametrických modelů složených ze dvou normálních rozdělení. Tyto techniky jsou demonstrovány na reálných datech z měření chlupatosti přízí. Je zmíněn program v jazyce MATLAB umožňující realizaci graficky orientovaných metod a odhad parametrů směsi dvou normálních rozdělení.*

## **1. Úvod**

Většina praktických úloh z oblasti analytické chemie vede na zpracování jednorozměrných výběrů. Stejně typy úloh se vyskytují prakticky ve všech oborech a patří mezi základní metrologické problémy. Obvykle se postupuje tak, že se na základě měření ( $x_1, \dots, x_N$ ) stanoví vhodný odhad střední hodnoty  $\mu$  a rozptylu  $\sigma^2$ . Data z laboratoří a zkušeben mají často některé specifické zvláštnosti:

- I.** Obsahují občas extrémně velké hodnoty, které však nejsou důsledkem chyb měření (to je typické pro stopové analýzy).
- II.** Mohou být cenzurována zdola s ohledem na limitu detekce přístrojů.
- III.** Jsou vždy kladná a výrazně zešikmená k vyšším hodnotám
- IV.** Rozsahy zpracovávaných dat jsou buď malé nebo extrémně velké
- V.** Jsou často prostorově nebo časově závislá. Zejména časová závislost je častá tam, kde se odebírají vzorky z téže soustavy.
- VI.** Rozdělení dat je jen zřídka normálnímu jak se běžně předpokládá ve standardní statistické analýze.
- VII.** Z nejrůznějších důvodů může rozdělení dat obsahovat více lokálních maxim (polymodalita).

Tyto zvláštnosti pak omezují použití různých technik založených na průzkumové analýze a identifikaci vybočujících měření. Také robustní techniky mohou selhat, zejména pro situace ad **VI** a **VII**.

Z hlediska korektního použití statistických metod je proto žádoucí mít možnost zkoumat statistické zvláštnosti dat (průzkumová analýza), ověřovat základní předpoklady o datech a hodnotit kvalitu výsledků s ohledem na základní schéma [1]

### **"data - model - statistická metoda"**

Toto schéma se považuje za základ interaktivní tvorby statistických modelů všeho druhu.

V této práci je pojednáno o možnostech především geometrické analýzy dat s ohledem na možnou bimodalitu. Je pojednáno také o způsobech testování bimodality a konstrukci parametrických modelů složených ze dvou normálních rozdělení. Tyto

techniky jsou demonstrovány na reálných datech z měření chlupatosti přízi na přístroji Uster. Jedná se o výběr velikosti 14 000 pro kompaktní přízi C 30tex [2].

## 2. Základní pojmy

Zdánlivě nejjednodušším úkolem je stanovení odhadu vhodných parametrů charakterizujících jednorozměrné výběry, tj. výsledky měření  $x_j$   $i = 1, \dots, N$ . Účelem je běžně odhad parametru polohy a stanovení jeho neurčitosti (resp. přesnosti). Vychází se přitom z aditivního modelu měření [1].

$$x = \mu + \varepsilon \quad (1)$$

V této rovnici označuje:

- $\mu$  skutečnou hodnotu měřené veličiny (střední hodnota)
- $\varepsilon$  náhodnou chybu měření resp. „šumovou“ složku

Pro standardní statistickou analýzu dat se běžně vychází z těchto předpokladů:

- střední hodnota chyb měření je nulová, t.j.  $E(\varepsilon) = 0$ ,
- rozptyl chyb měření je konstantní, t.j.  $D(\varepsilon) = \sigma^2$
- chyby jsou vzájemně nezávislé .t.j.  $E(\varepsilon_i * \varepsilon_j) = 0$
- chyby mají normální rozdělení t.j.  $\varepsilon \approx N(0, \sigma^2)$ , což zahrnuje předpoklad, že data mají unimodální rozdělení.

V případě , kdy  $E(\varepsilon) = k$  , se tato konstanta vlastně přičte k  $\mu$  a resultují systematicky vychýlené odhady střední hodnoty, protože vyjde, že  $E(x) = \mu + k$  .

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

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

Obecně platí, že odhady  $x_A$  a  $s^2$  jsou korelované s korelačním koeficientem

$$\rho(x_A, s^2) = \frac{g_1(x)}{\sqrt{n}}$$

kde  $g_1(x)$  je šikmost. Pro pozitivně zešikmená rozdělení jsou tedy tyto odhady pozitivně korelované.

Při nesplnění některých předpokladů o rozdělení a chování „šumové“ složky může dojít buď ke zkreslení nebo až ke znehodnocení statistické analýzy. Na druhou stranu je malé porušení předpokladů často nevýznamné. Protože prakticky nelze předem stanovit co to je malé porušení předpokladů je třeba je ověřit a případně omezit jejich vliv.

Pokud je rozdělení dat bimodální (obecně polymodální) nelze prakticky model (1) použít, protože parametr  $\mu$  již necharakterizuje dobře polohu (resp. jde pouze o formální parametr). Standardně je bimodalita způsobena tím, že data pocházejí ze směsi rozdělení. Pak je třeba pro každou složku směsi rozdělení počítat charakteristiku polohy,

rozptýlení atd. zvlášť. Bimodalita může být také způsobena měřicím zařízením resp. principem měření. V každém případě je její indikace klíčová, protože mění zásadním způsobem představu o statistickém chování dat a může vést až k potřebě nových měření resp. dělení dat podle externích kritérií (typ materiálů, geometrie objektů atd.).

### 3. Vizualizace rozdělení dat

Pro grafickou vizualizaci rozdělení dat existuje celá řada jednoduchých technik pro odhad tvaru rozdělení jako jsou: histogram, kvantilový graf, Q-Q graf a jádrový odhad hustoty pravděpodobnosti. [1].

Nejjednodušší je odhad kvantilové funkce, který je zejména pro malé výběry založen na pořádkových statistikách

$$x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(i)} \dots \leq x_{(N)}$$

Pořádkové statistiky  $x_{(i)}$  jsou vzestupně seříděné hodnoty prvků výběru. Necht' je  $F_e(x)$  distribuční funkce, ze které prvky výběru  $x_i$  pocházejí. Je známo, že transformovaná náhodná proměnná

$$z_{(i)} = F_e(x_{(i)}) \quad (2)$$

má nezávisle na rozdělení dat  $F_e$  Beta rozdělení  $Be[i, N-i+1]$ . Odpovídající střední hodnota je dána vztahem

$$E(z_{(i)}) = \frac{i}{N+1}$$

kde  $E(.)$  je operátor střední hodnoty (matematického očekávání). Lze ukázat, že prvky  $V_{ij}$  odpovídající kovarianční matice  $V$  pro všechny dvojice  $z_{(i)}, z_{(j)}$   $i, j = 1, \dots, N$  jsou funkcí pouze  $i, j$  a  $N$ . Použitím zpětné transformace  $E[z_{(i)}]$  lze dospět k výrazu

$$E(x_{(i)}) = F_e^{-1}(z_{(i)}) = Q_e(P_i)$$

V rov je  $Q_e(P_i)$  kvantilová funkce a

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

je pořadová pravděpodobnost. Vlastnosti kvantilové funkce a její výhody pro konstrukci výběrové distribuční funkce jsou uvedeny v knize [1].

Kvantilová funkce je obecně inverzní k funkci distribuční. Pokud je  $F(x) = P$  spojitá distribuční funkce a  $f(x)$  je odpovídající hustota pravděpodobnosti je  $Q(P) = x$  kvantilová funkce. Lze snadno ověřit, že  $F(Q(P)) = P$  pro všechna  $0 \leq P \leq 1$  a  $f(Q(P)) \cdot q(P) = 1$ . Veličina  $q(P) = dQ(P)/dP$  se označuje jako kvantilově hustotní funkce  $f(Q(P)) = 1/q(P)$  je hustotně kvantilová funkce.

Z rov. (3) je patrné, že pořádkové statistiky  $x_{(i)}$  jsou hrubé odhady kvantilové funkce  $Q_e(P_i)$  v místech  $P_i$ . Pro odhad kvantitu  $x_P = Q_e(P)$  v místě  $P$  pro které platí, že  $i/(n+1) < P < (i+1)/(n+1)$  se používá po částech lineární interpolace

$$x_{(P)} = (N+1) \left( \frac{PN + P - i}{N+1} \right) (x_{(i+1)} - x_{(i)}) + x_{(i)} \quad (3)$$

Rozptyl tohoto odhadu  $D(x_P)$  se počítá ze vztahu

$$D(x_P) = \frac{P(1-P)}{N f^2(x_P)} \quad (4)$$

Symbol  $f_e(x_P)$  označuje hustotu pravděpodobnosti odpovídající distribuční funkci  $F_e$ . Asymptotické rozdělení kvantitu  $x_P$  je normální se střední hodnotou  $Q(P)$  a rozptylem definovaným rov. (4).

Interpolace definovaná rov. (3) se dá použít pro odhad speciálních výběrových kvantilů  $x_{P_i}$  resp.  $x_{1-P_i}$  pro  $P_i = 2^{-i}$  ( $i = 1, \dots, N$ ). Tyto kvantily se označují jako písmenové hodnoty

[1]. Všechny písmenové hodnoty kromě mediánu ( $i = I$ ) jsou v párech. Např. lze určit dolní kvartil  $x_{0.25}$  ( $P_i = 0.25$ ) a horní kvartil  $x_{0.75}$  ( $P_i = 0.75$ ) atd.

Možnosti přesnějšího odhadu  $P_i$  jsou uvedeny v práci [1]. Pro charakterizaci polohy se pak používá medián  $x_{0.5}$  a pro charakterizaci variability se používá difference mezi horním  $x_{0.75}$  ( $P_i = 0.75$ ) a dolním  $x_{0.25}$  ( $P_i = 0.25$ ) kvantilem. Kvantilová odchylka je pak dvojnásobek této difference.

$$DQ = 2 * (x_{0.75} - x_{0.25}) \quad (5)$$

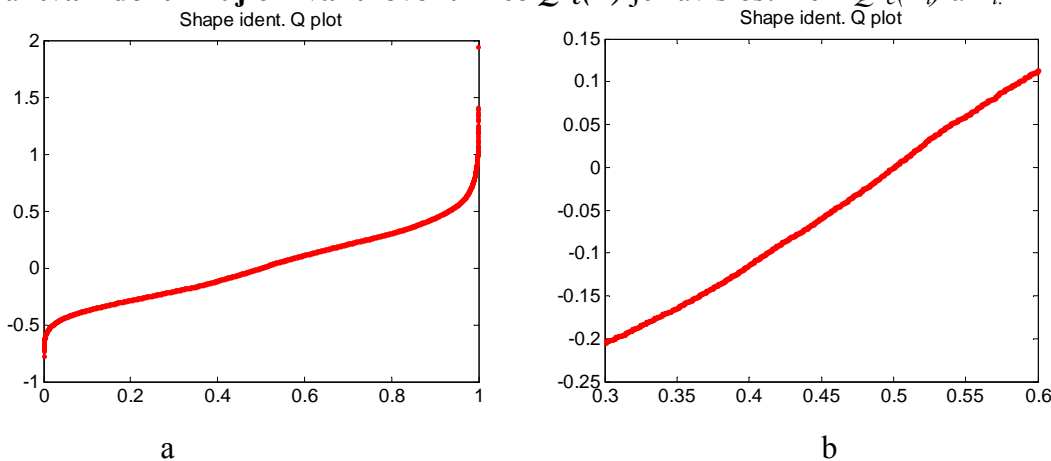
$DQ$  je vlastně numerická derivace  $Q(P)$  v místě  $P = 0.5$ , která přibližně aproximuje hustotně kvantilovou odchylku

$$DfQ = 1 / f(Q(0.5)) = 1 / f(x_{0.5}) = q(0.5)$$

Použitím těchto odhadů polohy a rozptýlení je možné standardizovat výběrovou hustotu pravděpodobnosti, tak aby bylo  $DQ = 1$  a medián  $x_{0.5} = 0$ . Standardizovaná kvantilová funkce se označuje jako tvar identifikující kvantilová funkce  $QI(P)$ . Výběrová funkce  $QI_e(P)$  je určena vztahem [1]

$$QI_e(P) = \frac{(x_{(i)} - x_{0.5})}{2 * (x_{0.75} - x_{0.25})} \quad (6)$$

Graf tvar identifikující kvantilové funkce  $QI_e(P)$  je závislost mezi  $QI_e(P_i)$  a  $P_i$ .



Obr. 1  $QI_e(P)$  pro přízi C 30 tex (a.. celý rozsah, b.. detail)

**Tvar identifikující kvantilová funkce  $QI_e(P)$  pro přízi C 30 tex je zobrazen na obr. 1.** Sigmoidální tvar v centrální oblasti je typický pro bimodální rozdělení (je patrný pouze v detailu). V horní části grafu jsou patrné vybočující body.

Hodnoty  $|QI_e(P)| \geq 1$  jsou identifikací vybočujících měření (pro normální rozdělení) nebo indikací rozdělení s dlouhými konci.

Hodnoty  $QI_e(P)$  se mohou snadno použít pro specifikaci šikmosti a délky konců výběrových rozdělení. Veličina  $SQ = QI_e(0.25) + QI_e(0.75)$  se používá jako míra šikmosti (pro symetrická rozdělení je rovna nule).

Míry délky konců jsou  $QI_e(0.05)$  a  $QI_e(0.95)$ . Platí, že:

pro rozdělení s krátkými konci je  $QI_e(0.95) < 0.5$ ,

pro rozdělení s dlouhými konci je  $QI_e(0.95) > 1$

pro rozdělení s středními konci je  $0.5 < QI_e(0.95) < 1$ .

**Empirický kvantilový graf  $Q(P)$  je závislost  $x_{(i)}$  na  $P_i$ .** Na tomto grafu lze snadno identifikovat statistické zvláštnosti dat jako symetrie, lokální koncentrace a přibližná normalita.

Detailní rozbor  $QP$  je popsána v knize [1]. Pro zlepšení interpretace je výhodné do tohoto grafu přidat graf funkce normálního rozdělení

$$Q_N(P) = \mu + \sigma * u_p$$

kde  $u_p$  jsou kvantily standardizovaného normálního rozdělení  $N(0, 1)$ . Parametry  $\mu$  a  $\sigma$  jsou odhady polohy a měřítka.

Standardně je možné zobrazit dva typy normálních kvantilových funkcí. První je založena na momentových odhadech tj. výběrovém průměru  $x_M$  a výběrové směrodatné odchylce  $s$ . Druhá používá robustní kvantilový odhad polohy tj. medián  $x_{0.5}$  a směrodatnou odchylku počítanou z kvartilů

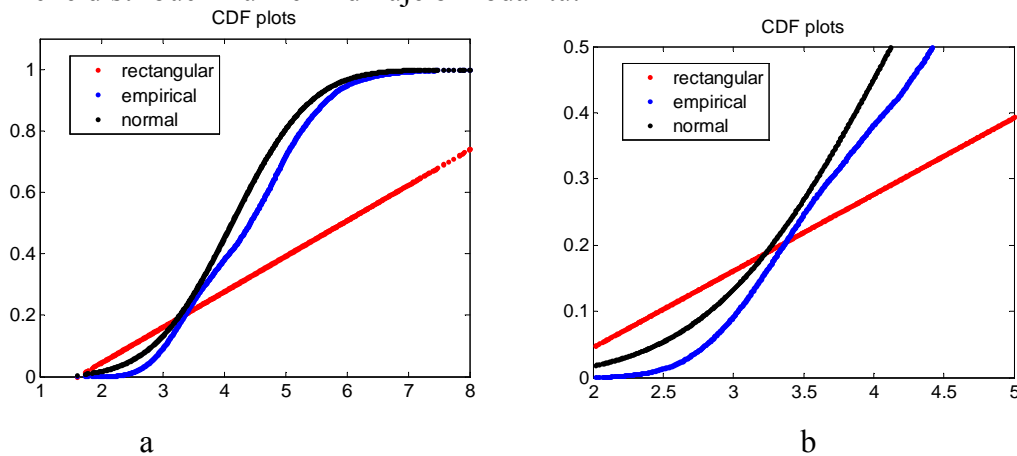
$$s_M = \frac{x_{0.75} - x_{0.25}}{1.349}$$

Tato varianta  $QP$  umožňuje lépe porovnat odchylky od normality. Pokud lze a priori předpokládat, že data pocházejí z nějakého známého rozdělení lze místo kvantilové funkce normálního rozdělení použít kvantilovou funkci tohoto rozdělení.

Jednoduchým odhadem **distribuční funkce** je lokální součet pořádkových statistik.

$$cdf(x) = \frac{\sum_{i=1}^i x_{(i)}}{N} \quad , \quad \text{for } x_{(i)} \leq x \leq x_{(i+1)} \quad (7)$$

Tento odhad je pro přízi C 30 tex ukázán na obr. 2. Na stejném obr. jsou ukázány průběhy distribučních funkcí normálního a rovnoměrného rozdělení. Typický „skok“ na empirické distribuční funkci indikuje bimodalitu.



Obr. 2 Distribuční funkce pro přízi C 30 tex (a.. celý rozsah, b.. detail)

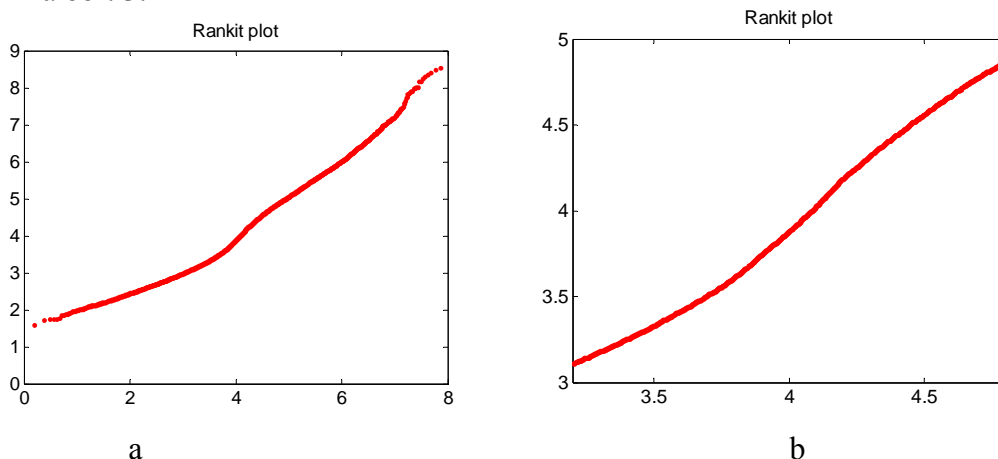
Pro porovnání empirického rozdělení výběru s vybraným teoretickým rozdělením je možné jednoduše použít varianty **Q-Q grafu**. Klasický **Q-Q** graf je založen na porovnání empirické kvantilové funkce  $Q_e(P_i) \cong x_{(i)}$  s vybranou teoretickou kvantilovou funkcí  $Q_T(P_i)$ . Pokud je teoretická distribuční funkce typu  $F_T((x-T)/S)$  je výhodné použít standardizovanou kvantilovou funkci  $Q_{TS}(P_i)$  (viz. [1]) nebo tvar identifikující kvantilovou funkci  $Q_{IT}(P_i)$ . Pokud je výběrové rozdělení shodné s teoretickým vyjde na **Q-Q** grafu přímka

$$x_{(i)} = T + S Q_{TS}(P_i) \quad (8)$$

Parametr  $T$  charakterizuje polohu a parametr  $S$  rozptýlení. Pro některá tří parametrická rozdělení je parametr tvaru použit jako parameter grafů.

Nevýhodou **Q-Q** grafu je častá nelinearita způsobená závislostí mezi pořádkovými statistikami a nekonstantním rozptylem.

Q-Q graf pro případ normálního teoretického rozdělení (rankitový graf) je pro přízi C 30 tex na obr. 3.



Obr. 3 Q-Q Rankitový graf pro přízi C 30 tex (a.. celý rozsah, b.. detail)

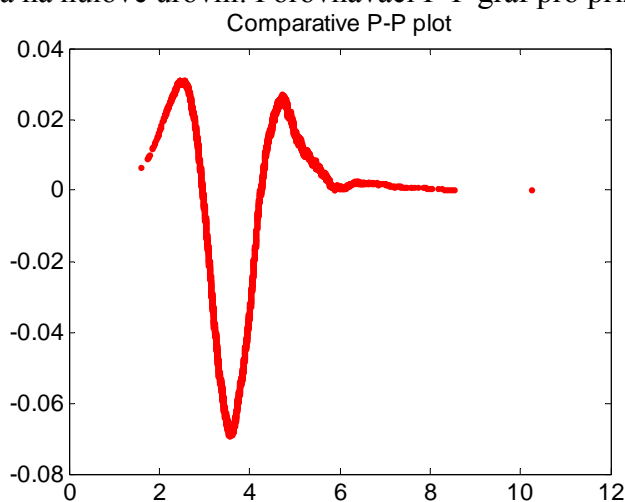
Ohyb v centrální oblasti zde opět indikuje bimodalitu.

**P-P (resp. Probability-Probability) graf** umožňuje přímé porovnání distribučních funkcí. V P-P grafu je empirická distribuční funkce (odhadovaná jako pořadová pravděpodobnost  $P_i$ ) vynesena proti teoretické distribuční funkci  $F_T(x_{(i)})$ . Pokud tvoří tato závislost přímku (s nulovým úsekem a jednotkovou směrnici) je empirické rozdělení shodné s teoretickým. Nevýhodou tohoto grafu je nutnost úplné specifikace teoretické distribuční funkce včetně jejích parametrů. To obvykle vyžaduje odhad parametrů zadaného rozdělení např. metodou maximální věrohodnosti.

Parzen [3] navrhl použít pro tyto účely tzv. **porovnávající distribuční funkci** definovanou výrazem

$$CDD = F_T(Q_e(P_i)) \quad (9)$$

Funkce CDD je přibližně rovna teoretické distribuční funkci v místě  $x_{(i)}$ . Závislost CDD na  $P_i$  je tedy přibližně rovna P-P grafu. Pro indikaci bimodalit je výhodnější použití tzv. **porovnávacího P-P grafu**. V tomto grafu je CDD nahrazeno rozdílem  $CDD - P_i$ . V případě unimodálního normálního rozdělení je odpovídající porovnávající P-P graf horizontální přímkou na nulové úrovni. Porovnávací P-P graf pro přízi C 30 je na obr.4.



Obr. 4 Porovnávací P-P graf pro přízi C 30

Je patrné, že data mají normální rozdělení. Kombinace  $Q-Q$  a  $P-P$  grafů je tedy výhodná pro indikaci bimodality.

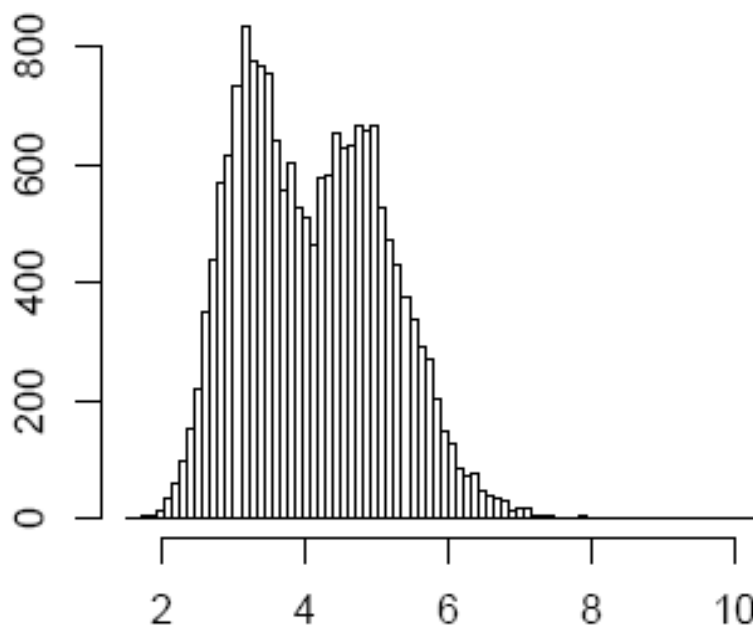
Jednoduchý odhad hustoty pravděpodobnosti (PDF) je **histogram** s konstantní šířkou sloupců -tříd (počet tříd je  $M$ ). Histogram je po částech konstantní odhad hustoty pravděpodobnosti. Výška sloupce v  $j$  – té třídě ohraničené hodnotami  $(t_{j-1}, t_j)$  je určena ze vztahu

$$f_H(x) = \frac{C_N(t_{j-1}, t_j)}{N h_j} \quad (10)$$

Zde funkce  $C_N(a, b)$  označuje počet hodnot výběru v intervalu  $\langle a, b \rangle$  a  $h_j = t_j - t_{j-1}$  je délka  $j$ -té třídy (intervalu). Při konstrukci histogramu je třeba určit hraniční hodnoty  $\{t_j\}$   $j=1, \dots, M$ , počet intervalů  $M$  a jejich délky  $h_j$  s ohledem na jeho kvalitu. Pro přibližně normální data je konstantní délka tříd určena vztahem

$$h = 3.49 * (\min(s, Dq / 2) / 1.34) / n^{1/3} \quad (11)$$

Histogram pro přízi C 30 tex a  $h$  určené z rov. (11) je na obr. 5.



Obr. 5 Histogram s konstantní délkou tříd ( $h = 0.134$ ) pro přízi C 30 tex

Pro případ nekonstantní délky tříd je možné použít jednoduchou dvoustupňovou metodu. V první fázi se určí počet tříd ze vztahu

$$M = \text{int}[2.46 (N - 1)^{0.4}] \quad (12)$$

Zde  $\text{int}[x]$  celá část čísla  $x$ . Optimální počet tříd určený z rov. (12) je pro přízi C 30 tex roven  $M = 125$ .

Ve druhé fázi se počítají jednotlivé délky tříd na základě principu stejné pravděpodobnosti ve všech třídách. K tomuto účelu je vhodné použít empirickou kvantilovou funkci  $Q(P_i)$  využívající pořádkových statistik  $x_{(i)}$

Prakticky to znamená, že osa  $P$  se rozdělí na identické intervaly o velikosti  $1/M$ . Pro tyto intervaly se určí odpovídající odhady kvantilů  $t_j = x_{(j/M)}$  s využitím rov. (4).

Praktické zkušenosti ukazují, že tento postup je výhodný i pro silně zešíkmená rozdělení [1].

Přirozeným zobecněním histogramu je **jádrový odhad hustoty pravděpodobnosti**  $f(x)$  (hladká funkce, závislá na parametru vyhlazení  $h$ ). Tento odhad má pro případ konstantního parametru vyhlazení  $h$  tvar

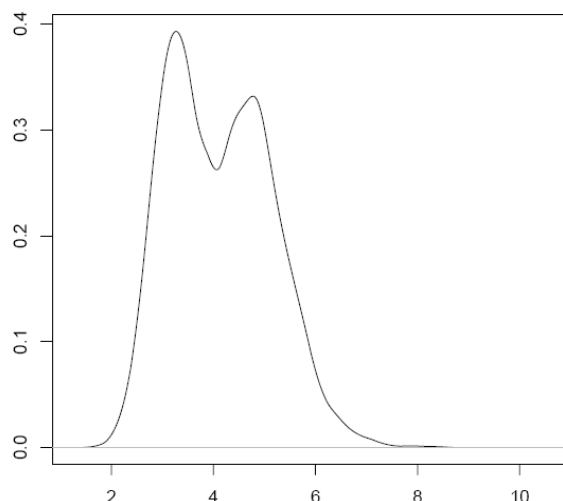
$$\hat{f}(x) = \frac{1}{N} \sum_{i=1}^N K \left[ \frac{x - x_i}{h} \right] \quad (13)$$

Výběr jádrové funkce  $K[x]$  a výpočet parametru vyhlazení je uveden v knize [1]. Jednoduchá bikvadratická jádrová funkce  $K[x]$  má tvar

$$K(x) = 0.9375 * (1 - x^2)^2 \quad \text{for } -1 \leq x \leq 1 \quad (14)$$

Parametr  $h$  se dá určit z rov (11).

Jádrový odhad hustoty definovaný rov. (13) a jádrovou funkcí vyjádřenou rov. (14) pro přízi C 30 tex je na obr. 6.



Obr. 6 Jádrový odhad hustoty ( $h = 0.128$ ) pro přízi C 30 tex

Výše uvedené postupy jsou součástí programu BIMODAL v jazyku MATLAB.

#### 4. Parametrický model bimodality

Pokud lze předpokládat, že data jsou výběrem buď z jednoho nebo ze směsi dvou normálních rozdělení je možné porovnat unimodální model

$$f_u(x_i) = A0 * \exp \left( -\frac{(x_i - B1)^2}{2 * C1^2} \right) \quad (15)$$

a bimodální model (směs dvou normálních rozdělení)

$$f_B(x_i) = A1 * \exp \left( -\frac{(x_i - B1)^2}{2 * C1^2} \right) + A2 * \exp \left( -\frac{(x_i - B2)^2}{2 * C2^2} \right) \quad (16)$$

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.

Důležité je, že směs dvou normálních rozdělení může být unimodální nebo bimodální. Pro případ unimodality musí platit, že



$$|B1 - B2| < 2 * \min(C1, C2) \quad (17)$$

Pro jedno normální rozdělení jsou parametry  $A1 =$  aritmetický průměr,  $B1 =$  výběrová směrodatná odchylka. 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. Residuum  $r_i$  pro  $i$  tou hodnotu je rozdíl mezi empirickou hustotou (viz rov.(13)) a modelovou hustotou (viz rov.(16))

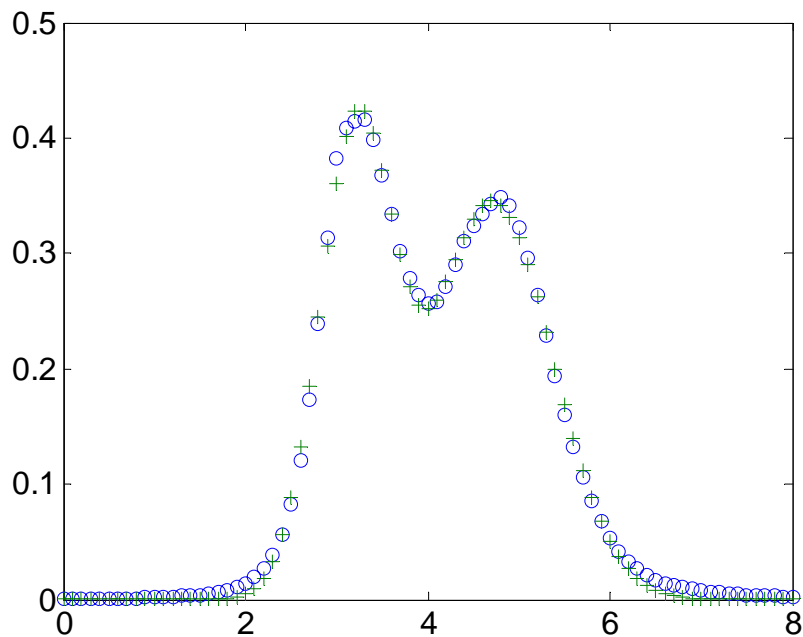
$$r_i = \hat{f}(x_i) - f_B(x_i)$$

Reziduální součet čtverců je pak roven

$$S = \sum_{i=1}^N r_i^2 \quad (18)$$

kde  $N$  je počet tříd. Předpoklady vedoucí k použití metody nejmenších čtverců (tj. minimalizace  $S$ ) jsou uvedeny v knize [1]. V programu BIMODAL v jazyku MATLAB jsou použity algoritmy minimalizace nejmenších čtverců na bázi Levenberg-Marquardtovy metody a jejího vylepšení (Trust-region)

Popis těchto algoritmů je uveden v knize [1]. Aproximace empirické hustoty pravděpodobnosti směsí dvou normálních rozdělení je pro přízi C 30 tex zobrazena na obr. 7.



Obr. 7 Aproximace empirické hustoty pravděpodobnosti (kroužky) směsí dvou normálních rozdělení (křížky) pro přízi C 30 tex

Pro porovnání normálního rozdělení a směsí dvou normálních rozdělení (unimodálního a bimodálního modelu) je výhodné použít věrohodnostního poměru

$$LR = 2 * \ln \left( \frac{L_B}{L_U} \right) \quad (19)$$

Věrohodnostní funkce  $L_U$  má tvar

$$L_U = \prod_{i=1}^N \frac{1}{\sqrt{2 * \pi * C1^2}} \exp\left(-\frac{(x_i - B1)^2}{2 * C1^2}\right) \quad (20)$$

a pro věrohodnostní funkci  $L_B$  platí

$$L_B = \prod_{i=1}^N A1 * \exp\left(-\frac{(x_i - B1)^2}{2 * C1^2}\right) + A2 * \exp\left(-\frac{(x_i - B2)^2}{2 * C2^2}\right) \quad (21)$$

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 potřeba směsi dvou rozdělení (bimodalita) je ověřena.

#### 4. Závěr

Je patrné, že 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 [2].

**Poděkování:** Tato práce vznikla v rámci řešení projektu MŠMT. – č. 1M4674788501 „Centrum Textil II“

#### 5. Literatura

- [1] Meloun M., Militký J.: *Zpracování experimentálních dat*, Academia Praha 2004
- [2] Militký J., Ibrahim S.: “*Yarn Hairiness Complex Characterization*”, Proc. Annual Fiber Soc. Conf., St Galen, May 2005.
- [3] Parzen E.: *Statistical methods mining and non parametric quantile domain data analysis*, Proc Ninth int. conf. on quantitative methods for environmental science, July 1988, Melbourne