RUTINNÍ ANALÝZA DAT V ANALYTICKÉ LABORATOŘI

Jiří Militký

KTM, Technická universita v Liberci, 461 17 LIBEREC, Česká Republika

Milan Meloun,

KACH, Universita Pardubice,530 099 PARDUBICE Česká Republika

Souhrn

Příspěvek je zaměřen na problematiku standardního zpracování rutinních výsledků měření. Jsou uvedeny základní statistické pojmy a jejich interpretace. Jsou navrženy jednoduché charakteristiky pro vyjádření parametru polohy pro případ, že data nemají normální rozdělení. Jsou uvedeny možnosti konstrukce intervalů spolehlivosti pro tyto charakteristiky. Cílem je ukázat, že i rutinní zpracování dat přináší řadu problémů, které je nutné vhodnými postupy odstranit nebo alespoň omezit jejich vliv na výsledek analýzy.

1.ÚVOD

Zpracování dat v analytické praxi využívá kombinace poznatků klasické *analytické chemie, matematické statistiky a informatiky* na jedné straně a speciálních postupů. *chemometrie* na straně druhé. Důležitou součástí analýzy dat jsou metody k získávání relevantních informací z experimentů a pozorování.

Stále větší počet výkonných osobních počítačů třídy PC podporuje na pracovištích trend decentralizace a interaktivnosti při zpracování experimentálních dat a interpretaci výsledků. To klade větší nároky na pracovníky, kteří již těžko obhájí jednoduché postupy vyhodnocování dat, založené mnohdy na zjednodušených nebo i nesprávných předpokladech. Nabídka a možnosti počítačově orientovaného statistického zpracování dat nutí experimentátora k hlubší analýze, což vede většinou i k radikální změně pohledu na rutinně prováděnou výzkumnou práci.

Existuje celé spektrum méně či více dokonalých a komplexních programů a programových systémů pro statistické vyhodnocování dat. Jiné jsou budovány jako univerzálně použitelné, i když zaměřené na specifické oblasti (chemometrie, biometrie, ekonometrie, medicínská statistika, obchodní statistika, statistika pro sociology, psychology, atd.).

Úlohy vyhodnocení experimentálních dat v analytické praxi mají některé společné rysy:

(a) rozsahy zpracovávaných dat nejsou obyčejně velké,
(b) v datech se vyskytují výrazné nelinearity, neaditivity a vzájemné vazby, které je třeba identifikovat a popsat,
(c) rozdělení dat jen zřídka odpovídá normálnímu běžně předpokládanému ve standardní statistické analýze,
(d) v datech se vyskytují vybočující měření a různé heterogenity,
(e) statistické modely se často tvoří na základě předběžných informací z dat (datově orientované přístupy),
(f) parametry statistických modelů mají mnohdy definovaný fyzikální význam, a musí proto vyhovovat velikostí, znaménkem nebo vzájemným poměrem,
(g) existuje jistá neurčitost při výběru modelu, popisujícího chování dat.

Z hlediska 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

"data - model - statistická metoda"

Pak je možné využívat i alternativních postupů statistické analýzy včetně robustních a adaptivních metod.

V této přednášce jsou uvedeny jen techniky používané pro rutinní zpracování jednorozměrných výběrů. Vychází se z N-tice výsledků experimentů, t.j.dat $\{x_i\}$ i = 1,...N. Celý postup je demonstrován na jednoduchém příkladu.

Příklad

Tento příklad ilustruje také vliv transformace dat na jejich statistické chování. Nechť původní data

$$w_i = i \qquad i = 1.....19$$

pocházejí z rovnoměrného rozdělení ($w_{min} = 1$, $w_{max} = 19$)

V průběhu získávání dat došlo k jejich reciproké transformací, takže jsou k dispozici "naměřené" hodnoty

$$x_i = w_i^{-1} = \frac{1}{i} \qquad i = 1...19$$

K této transformaci může v praxi dojít v řadě případů. Např. je-li w_i frekvence, je x_i úměrné vlnové délce, pokud je w_i povrchový odpor, je x_i povrchová vodivost atd.

V dalším bude demonstrováno jak reciproká transformace ovlivnila výběr $\{x_i\}$ i = 1,..19. Tento příklad přes svoji zdánlivou jednoduchost demonstruje jaké obtíže lze i při rutinním zpracování dat očekávat.

S ohledem na rozsah příspěvku jsou vynechány ***techniky ověřování základních předpokladů o datech***, které vlastní rutinní analýze předcházejí. Jejich přehled je uveden v knize [1]

2. ZÁKLADNÍ POJMY

Standardně se předpokládá, že výběr $\{x_i\}$ i = 1,.. N tvoří realizace spojité náhodné veličiny jejíž chování je úplně popsáno např. **distribuční funkcí** F(x).

Platí, že F(x) odpovídá pravděpodobnosti, s jakou nabývá náhodná veličina hodnot menších než x. Tedy

$$F(x) = P(\xi \le x) \qquad (1)$$

Je zřejmé, že:

a) F(x) je neklesající funkcí svého argumentu (pravděpodobnost je nezáporná),

b) $F(x_{min}) = 0$, $F(x_{max}) = 1$, kde [x_{min} -obyčejně -∞, x_{max}-obyčejně ∞] je definiční interval náhodné veličiny ζ .

Derivací distribuční funkce je **hustota pravděpodobnosti** (frekvenční funkce) f(x) = dF(x) / dx. Tato funkce musí splňovat podmínky:

a) f(x) je v celém definičním intervalu nezáporná,
b) f(x) je normalizovaná, tj.
c)

$$\int_{-\infty}^{\infty} f(x)dx = 1 \qquad (2)$$

Inverzní funkcí k funkci distribuční je **funkce kvantilová** $Q(P) = F^{-1}(x)$. Její konkrétní hodnota, tj. 100P%ní kvantil x_P je takové číslo, pro které je pravděpodobnost, že náhodní veličina bude menší rovna právě P. Tedy

$$P(\xi \le x_P) \qquad (3)$$

Jako **charakteristika polohy** náhodné veličiny s rozdělením f(x) se obyčejně používá **střední hodnota** E(x), pro kterou platí

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx \qquad (4)$$

Pro symetrická unimodální rozdělení je E(x) rovno **módu** M_o, tedy maximu na funkci f(x).
Robustní charakteristikou polohy je medián $\tilde{x}_{0.5} = M$, pro který platí

$$P(\xi \le M) = 0.5 \;\text{ t.j. }\; M = Q(0.5) \qquad (5)$$

Pro symetrická unimodální rozdělení je $E(x) = M_o = M$.

U rozdělení zešikmených k vyšším hodnotám je $E(x) > M > M_o$ a u rozdělení zešikmených k nižším hodnotám je $M_o > M > E(x)$.

Pro charakterizaci **variability** (koncentrace kolem střední hodnoty) se používá druhý centrální moment D(x) označovaný jako rozptyl

$$D(x) = \int_{-\infty}^{\infty} (x - E(x))^2 f(x)dx = E[x - E(x)]^2 \qquad (6)$$

Odmocnina z rozptylu $\sigma = \sqrt{D(x)}$ se nazývá směrodatná odchylka.

Pro charakterizaci **šikmosti** rozdělení se používá třetí normovaný centrální moment

$$g_1 = C_3(x) / \sqrt{D^3(x)} \qquad (7)$$

a pro **špičatost** čtvrtý normovaný centrální moment)

$$g_2 = C_4 / D^2(x) \qquad (8)$$

Zde obecně n-tý centrální moment je definován vztahem

$$C_n(x) = E[x - E(x)]^n \qquad (9)$$

Pro symetrická rozdělení je $g_1 = 0$. Kladné g_1 ukazuje na rozdělení zešikmené k vyšším hodnotám a záporné g_1 na rozdělení zešikmené k nižším hodnotám. Špičatost (exces) g_2 je pro normální rozdělení roven 3.

V některých případech se tvar rozdělení f(x) charakterizuje také **entropickým koeficientem** E, pro který platí

$$E = 0.5 * \exp(H)/\sigma \qquad (10)$$

kde H je Shannonova entropie

$$H = \int_{-\infty}^{\infty} f(x)\ln(f(x))dx \qquad (11)$$

(Standardně se předpokládá, že H se používá pro vyjádření rozdělení chyb měření, pak je. Střední hodnota E(x) = 0.). Maximální E = 2.066 má normální rozdělení. Obyčejně platí, že 1.11 <E < 2.055

Je zřejmé, že pro známé rozdělení náhodné veličiny definované např. hustotou pravděpodobnosti f(x) lze přímo z definičních výrazů určit přímo všechny základní statistické charakteristiky.

Při zpracování dat $\{x_i\}$ i = 1,.. N se obyčejně uvažuje **aditivní model měření** definovaný vztahem

$$x_i = \mu + \varepsilon_i \qquad (12)$$

kde ε_i jsou chyby měření a μ je skutečná hodnota měřené veličiny.

Zde pak f(x) odpovídá tvarem rozdělení chyb f( ) a liší se jen nenulovou střední hodnotou E(x) = μ oproti E(x) = 0.

Je tedy účelné zkoumat rozdělení f( ) způsobené především měřicími přístroji. Uveďme základní třídy rozdělení, které jsou časté při analýze experimentálních dat [2,3].

A. Třída lichoběžníkových rozdělení

Vzniká při konvoluci dvou rovnoměrných rozdělení se stejnou střední hodnotou a šířkami rozdělení (x_{max} - x_{min}) rovnými A a B. Podle poměru B/A (uvažuje se, že A je vždy větší nebo rovno B), pak vycházejí různá rozdělení lišící se délkou konců. Přehled situací, ve kterých dochází ke vzniku rovnoměrně rozdělených chyb je uveden v práci [3].

B. Třída arkussínových rozdělení

Arkussínové rozdělení vzniká všude tam, kde se v konstantních intervalech stanovují hodnoty periodického (sinusoidového) signálu (např. střídavé napětí). Je definováno v intervalu (-A, A) a pro jeho hustotu pravděpodobnosti platí

$$f(x) = \frac{1}{\pi\sqrt{A^2 - x^2}} \qquad (13)$$

Další rozdělení z této třídy jsou kompozicí dvou různých arkussinových rozdělení (s parametry A_1, A_2, což odpovídá vzorkování signálu složeném ze dvou periodických složek lišících se frekvencí). Podle poměru A_1/A_2 pak vycházejí různá rozdělení tohoto systému.

Pokud je signál, ze kterého se odebírají vzorky složen ze dvou složek výrazně odlišných frekvencí, vzniká bimodální rozdělení. To může být např. případ, kdy se na deterministickém přístroji měří materiál. jehož vlastnosti se periodicky mění (měření tloušťky).

C. Třída exponenciálních rozdělení

Tato skupina symetrických rozdělení je velmi oblíbená jak v teoretických úvahách, tak při praktických aplikacích i proto, že jako speciální případy zahrnuje rozdělení rovnoměrné, normální a Laplaceovo. Vyskytuje se také při měření nejčastěji. Hustota pravděpodobnosti této třídy se dá vyjádřit ve tvaru

$$f(x) = \frac{\alpha}{2\lambda\Gamma(1/\alpha)} \exp\left(-\left|\frac{x-\mu}{\lambda\sigma}\right|^{\alpha}\right) \qquad (14)$$

zde μ je střední hodnota, σ je směrodatná odchylka, $\Gamma(x)$ je gamma funkce a

$$\lambda = \sqrt{\frac{\Gamma(1/\alpha)}{\Gamma(3/\alpha)}}$$

Parametr α je tvarový faktor, určující jak tvar v okolí střední hodnoty, tak i délku konců. Pro $\alpha < 1$ jde o rozdělení blízké tvarem k **Cauchymu**, pro $\alpha = 1$ jde o **Laplaceovo**, pro $\alpha = 2$ o **normální**, pro $\alpha > 2$ o přibližně **lichoběžníkové** a pro $\alpha \to \infty$ jde o **rovnoměrné** rozdělení.

Pro tuto třídu rozdělení je možno špičatost g_2 vyjádřit ve tvaru

$$g_2 = \Gamma(1/\alpha)\Gamma(5/\alpha)/[\Gamma(3/\alpha)]^2 \qquad (15)$$

a entropický koeficient ve tvaru

$$E = 1/\alpha * \exp(1/\alpha)\sqrt{\Gamma(1/\alpha)/\Gamma(3/\alpha)} \qquad (16)$$

Např. pro $\alpha = 0.25$ je $g_2 = 458$, $E = 0.085$, pro $\alpha = 1$ je $g_2 = 6$, $E = 1.92$ a pro $\alpha = 2$ je $g_2 = 3$, $E = 2.066$.

Pozor, často se stává, že se tato třída rozdělení považuje za univerzální pro popis experimentálních chyb ε, což není zdaleka pravda.

Konvolucí rovnoměrného rozdělení (se směrodatnou odchylkou σ_R) a exponenciálního rozdělení (se směrodatnou odchylkou σ_E) vzniká třída **zploštělých rozdělení**. Ta je plošší v oblasti střední hodnoty a přitom má dlouhé konce. Tvar této třídy rozdělení určuje parametr $C = \sigma_R / \sigma_E$. Platí pro ně, že při stejné hodnotě špičatosti g_2 jako u exponenciální třídy rozdělení mají výrazně nižší entropický koeficient E.

D. Třída Studentova rozdělení

Studentovo rozdělení patří mezi výběrová rozdělení (charakterizuje rozdělení střední hodnoty výběru z normálního rozdělení). Pro případ použití tohoto rozdělení při popisu chyb měření se používá v nejjednodušším (normovaném) tvaru, kdy je hustota pravděpodobnosti dána vztahem

$$f(x) = \frac{\Gamma((\alpha+1)/2)}{\sqrt{2\pi}\,\Gamma(\alpha/2)(1+x^2/\alpha)^{\frac{\alpha+1}{\alpha}}} \tag{17}$$

Zde tvarový faktor α odpovídá stupňům volnosti (α = N - 1). Pro $\alpha > 4$ je možno určit směrodatnou odchylku σ dle vztahu

$$\sigma = \sqrt{\alpha/(\alpha-2)} \tag{18}$$

špičatost podle vzorce

$$g_2 = (3\alpha-2)/(\alpha-4) \tag{19}$$

a entropický koeficient z výrazu

$$E = \frac{\Gamma(\alpha/2)\sqrt{\pi(\alpha-2)}}{2\Gamma((\alpha+1)/2)}\exp[\beta(\alpha)(\alpha+1)] \tag{20}$$

kde $\beta(1) = \ln(2)$, $\beta(2) = 1 - \ln(2)$ a dále

$$\beta(\alpha) = \left[1 - \frac{1}{2} + \frac{1}{3} + \frac{(-1)^{\alpha-1}}{\alpha-1} - \ln(2)\right] * (-1)^{\alpha} \tag{21}$$

Zajímavou zvláštností této třídy rozdělení je to, že pro $\alpha = 4$ je $g_2 = \infty$, ale entropický koeficient E = 1.903. Navíc pro $\alpha = 6$ je špičatost rovna $g_2 = 6$ jako u Laplaceova rozdělení, ale entropický koeficient E = 2.0053 je značně vyšší. Pro $\alpha \to \infty$ přechází toto rozdělení na normální rozdělení. Tato třída rozdělení je příkladem, že špičatost g_2 nevystihuje jednoznačně tvar symetrických unimodálních rozdělení.
Různé typy dalších analytických rozdělení (i zešikmených) jsou podrobně diskutovány v práci [2].

3. STANDARDNÍ ANALÝZA DAT

Při rutinním zpracování experimentálních dat se obyčejně provádí:

A. výpočet popisných charakteristik (odhad parametrů polohy, rozptýlení a tvaru),
B. určení konfidenčních intervalů,
C. testování významnosti parametrů.

Pokud nemá být statistická analýza pouhým numerickým počítáním bez hlubšího smyslu, je pochopitelně třeba, aby byly ověřeny všechny předpoklady, které vedly k návrhu daného postupu analýzy.

Při zpracování výsledků rutinních měření se běžně předpokládá aditivní model měření. O datech $\{x_i\}$ i = 1, ...N se apriorně soudí, že jde o nezávislé stejně rozdělené veličiny, pocházející z normálního rozdělení N (μ, σ^2). Tyto předpoklady jsou základem prakticky všech klasických metod analýzy experimentálních dat. Jejich ověření je podrobně popsáno v knize[1].

Poznámka

Je třeba mít na paměti, že malé porušení předpokladu normality nemusí být katastrofické s ohledem na výsledek statistické analýzy. Na druhé straně je však špatné, když odhady i testy závisejí na spíše jiných faktorech než je chování většiny dat (na velikosti výběru, uspořádání výsledků nesledovaných proměnných atd.).

Pokud data nesplňují předpoklad normality, je v řadě případů možné zlepšit jejich rozdělení vhodnou **transformací**. Je také možné konstruovat **empirické pravděpodobnostní modely**, které umožňují zpracování dat podle zvláštností jejich chování a nikoliv apriorních předpokladů. Tyto postupy řešení jsou uveden např. v [1,,3].

A. Popisné charakteristiky

Při klasické popisné analýze dat se počítá aritmetický průměr, rozptyl, šikmost a špičatost podle dále uvedených vztahů,. Standardně se předpokládá, že:

- jednotlivé prvky výběru x_i jsou vzájemně nezávislé
- výběr je homogenní, tj. všechna x_i pocházejí ze stejného rozdělení f(x)
- rozdělení f(x) je gaussovské - normální (což je třeba zejména pro další fáze zpracování).

Pro určení výběrových charakteristik se využívá **momentů** (analogie definičních vztahů). Datově orientované techniky předpokládají obyčejně pouze nezávislost prvků výběru a používají **kvantilové** charakteristiky.

Výběrovým odhadem střední hodnoty E(x) je aritmetický průměr $\bar{x}$ s rozptylem $D(\bar{x})$ definovaným známými vztahy

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad D(\bar{x}) = \frac{\sigma^2}{N} \tag{22}$$

kde σ je směrodatná odchylka rozdělení f(x). Odhaduje se jako odmocnina z výběrového rozptylu s^2, pro který platí

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2 \qquad D(s^2) = 2\sigma^4/(N-1) \tag{23}$$

Označení $\bar{x}$ a s^2 se používá s ohledem na jejich běžný zápis, i když jde o odhady, které se standardně označují stříškou.

Poznámka

Je méně známé, že směrodatná odchylka s je vychýleným odhadem veličiny σ. Platí, že E(s) < σ. Pro nevychýlený odhad lze odvodit

$$\hat{\sigma} = K_u * s \approx \sqrt{\frac{s^2(N-1)}{N-1.45}} \tag{24}$$

kde

$$K_u = \Gamma(\frac{N-1}{2})\sqrt{\frac{N-1}{2}} \Big/ \Gamma(\frac{N}{2}) \approx \sqrt{\frac{N-1}{N}}\left(1 - \frac{3}{4N} - \frac{37}{32N^2}\ldots\right) \tag{25}$$

Platí, že

$$D(\hat{\sigma}) = \sigma^2 \Big/ (2N) \tag{26}$$

a pro D(s) je možno psát

$$D(s) \approx \frac{\sigma^2}{2(N-1)} \tag{27}$$

V technické praxi se někdy používá také odvozená charakteristika variační koeficient δ = E(x)/D(x).Jeho výběrový odhad V a odpovídající rozptyl D(V) lze vyjádřit ve tvaru

$$V = \frac{\bar{x}}{s} \qquad D(V) \approx \delta\left[\frac{N + (2N-1)\delta^2}{2N(n-1)}\right] \propto \frac{\delta^2}{2N}(1 + 2\delta^2) \tag{28}$$

Pro odhad šikmosti g_1 se používá výběrová šikmost $\hat{g}_1$ s rozptylem $D(\hat{g}_1)$, kde

$$\hat{g}_1 = \frac{\sqrt{N}\sum_{i=1}^{N}(x_i - \bar{x})^3}{\left[\sum_{i=1}^{N}(x_i - \bar{x})^2\right]^{3/2}} \qquad D(\hat{g}_1) \approx \frac{6N(N-1)}{(N-2)(N+1)(N+3)} \tag{29}$$

a pro odhad špičatosti g_2 platí

$$\hat{g}_2 = \frac{N\sum_{i=1}^{N}(x_i - \bar{x})^4}{\left[\sum_{i=1}^{N}(x_i - \bar{x})^2\right]^2} \qquad D(\hat{g}_2) \approx \frac{24N(N-1)^2}{(N-3)(N-2)(N+5)(N+3)} \tag{30}$$

Zde odhady $D(\hat{g}_1)$ a $D(\hat{g}_2)$ jsou pouze asymptotické (platí pro velké N).

Poznámka

U malých výběrů jsou odhady $\hat{g}_1$ *a* $\hat{g}_2$ *často velmi zkreslené, protože jsou značně citlivé na vybočující pozorování, resp. malé heterogenity výběru.*

Dosavadní úvahy vycházely z předpokladu normality. Pokud je rozdělení, ze kterého pochází daný výběr symetrické ($\hat{g}_1$ = 0), ale s délkou konců $\hat{g}_2 \neq 3$, lze určit rozptyl směrodatné odchylky podle přibližného vztahu

$$D(s) \approx \frac{\sigma^2 (g_2 - 1)}{4N} \tag{31}$$

Poznámka

Pro určení entropického koeficientu E je třeba provést seskupení dat do histogramu šíře h s celkem M sloupci, kde v i-tém je n_i *hodnot. Pak je*

$$\hat{E} = \frac{hN}{2s} 10^{-\frac{1}{N}\sum_{i=1}^{M} n_i \log n_i} \tag{32}$$

Pro symetrická rozdělení je vhodná volba

$$M = \frac{(g_2^2 + 1.5)}{6N^{0.4}} \tag{33}$$

Pro rozptyl parametru $\hat{E}$ *byl u symetrických rozdělení nalezen aproximativní výraz*

$$D(\hat{E}) = \frac{0.81}{g_2^2 E^3 N} \tag{34}$$

Problém je v tom, že v závislosti na tvaru rozdělení je nutné mít dostatečný počet dat pro zajištění požadované přesnosti odhadů. Pro třídu symetrických rozdělení platí[2]:

- pokud je g_2 v intervalu 1.56 - 2.78, lze nalézt výše uvedené odhady s, a $\hat{E}$ dostatečně přesně (střední kvadratická chyba odhadu δ = 5%) již u menších výběrů (N≈ 50 - 200)
- při g_2 > 25 nepostačuje pro přesné určení s ani výběr rozsahu 1000
- při g_2 < 1.56 lze určit s přesně i z menších výběrů N≈20, ale E je stanoven velmi nepřesně (je třeba řádově stovek hodnot).

Jedním z nejdůležitějších je odhad střední hodnoty (středu symetrie u symetrických rozdělení). Aritmetický průměr $\bar{x}$ je efektivní (má minimální rozptyl) pouze pro normální rozdělení. Pro případ Laplaceova rozdělení je efektivní výběrový **medián** $\tilde{x}_{0.5}$ pro jehož rozptyl platí

$$D(\tilde{x}_{0.5}) = \frac{1}{4Nf^2(M)} \tag{35}$$

kde $f^2(M)$ je hodnota hustoty pravděpodobnosti v místě teoretického mediánu M. Pro případ Laplaceova rozdělení je $f(M) = 1/(\sigma\sqrt{2})$, a tedy $D(\tilde{x}_{0.5}) = \sigma^2/(2N)$. Pro Laplaceovo rozdělení je poměr

$$E_M = D(\bar{x}) \Big/ D(\tilde{x}_{0.5}) = 2 \tag{36}$$

tj. medián je 2x efektivnější než aritmetický průměr. Pro třídu exponenciálních rozdělení s parametrem α lze obecně určit E_M ze vztahu

$$E_M = \alpha^2\Gamma(3/\alpha) \Big/ \Gamma^3(1/\alpha) \tag{37}$$

Z této rovnice plyne, že pro rozdělení $\alpha > 2$ (dlouhé konce) je medián efektivnější než aritmetický průměr. Pro rozdělení s plochými vrcholy se doporučuje použití **kvartilové polosumy** definované vztahem

$$P_F = (\tilde{x}_{0.75} - \tilde{x}_{0.25})/2 \tag{38}$$

kde $\tilde{x}_{0.75}$ resp. $\tilde{x}_{0.25}$ je horní, resp. dolní kvartil.

V případě ohraničených rozdělení (arkussínové a lichoběžníkové třídy) je efektivní tzv. **polosuma**

$$\hat{x}_P = (x_{max} - x_{min})/2$$

kde x_{max} je maximální a x_{min} je minimální prvek výběru Polosuma $\hat{x}_P$ je efektivnější než $\bar{x}$ pro $g_2 > 2.2$.

Poznámka

Rozptyl odhadu polosumy je pro normální rozdělení roven

$$D(\hat{x}_P) = (\pi^2\sigma^2) \Big/ (24\ln(N))$$

pro rovnoměrné rozdělení

$$D(\hat{x}_P) = (6\sigma^2) \Big/ [(N-1)(N-2)]$$

a pro arkussínové rozdělení

$$D(\hat{x}_P) = (5\pi^4\sigma^2) \Big/ N^4$$

Z tohoto porovnání je patrné, že aritmetický průměr není zdaleka vždy efektivním odhadem. Samostatným problémem jsou vybočující měření, které zkreslují zejména $\bar{x}$ a $\hat{x}_P$.

Existuje řada technik, jak odhadovat polohu v případě nebezpečí, že ve výběru jsou vybočující měření (tzv. robustní techniky) [1,3]. Mezi nejjednodušší patří tzv. α-uřezaný průměr $\bar{x}(\alpha)$, který je definován vztahem

$$\bar{x}(\alpha) = \frac{1}{n - 2M} \sum_{i=M+1}^{N-M} x_{(i)} \qquad (39)$$

kde M = int(αN/100) je celá část výrazu α N/100 a $x_{(i)}$ jsou pořádkové statistiky (vzestupně setříděné prvky výběru).

V práci [2] bylo pro symetrické rozdělení s možnými vybočujícími hodnotami doporučeno volit jako odhad středu symetrie (centrální hodnoty) medián $\tilde{\tilde{x}}_C$ z odhadů $\bar{x}$, $\tilde{x}_{0.5}$, $\hat{x}_P$, P_F, $\bar{x}(0.25)$, tedy

$$\tilde{\tilde{x}}_C = \text{med}\{\bar{x},\ \tilde{x}_{0.5},\ \hat{x}_P,\ P_F,\ \bar{x}(0.25)\}$$

kde med{.} označuje medián z prvků v závorce.Dá se ukázat, že pro rozdělení od **Cauchyho** až k **normálnímu** je $\tilde{\tilde{x}}_C$ efektivnější než $\tilde{x}_{0.5}$, $\bar{x}(0.25)$ a $\bar{x}$

Poznámka

Pro odhad rozptylu odhadu $\tilde{\tilde{x}}_C$ je možno použít interkvantilové délky

$$k_{0.9} = (\tilde{x}_{0.95} - \tilde{x}_{0.05})/2 \qquad (40)$$

Pak

$$D(\tilde{\tilde{x}}_C) = k_{0.9}^2 / (2.72N) \qquad (41)$$

délku $k_{0.9}$ je možno určit spolehlivě od rozsahu výběru N > 40.

B. Konfidenční intervaly

V předchozím byly uvedeny postupy získávání bodových odhadů. Ty mají ze statistického hlediska pouze velmi malou cenu, protože nic neříkají o tom, v jakých mezích se skutečné parametry souboru pohybují. K tomu slouží **konfidenční** intervaly (intervaly spolehlivosti). Interval spolehlivosti parametru A je vyjádřen ve tvaru $c_1 < A < c_2$, kde hraniční hodnoty c_1, c_2 se volí tak, aby platilo

$$\mathbf{P\ (c_1 \le A \le c_2) = 1 - \alpha}$$

kde (1 - α) je konfidenční koeficient (statistická jistota).

Obyčejně se volí (1 - α) = 0.99, resp. 0.95. Parametr α je hladina významnosti.

Platí obecně:

- čím je vyšší statistická jistota, tím je konfidenční interval širší,
- čím je odhad kvalitnější (má menší rozptyl), tím je konfidenční interval užší,
- čím je rozsah výběru vyšší, tím je konfidenční interval užší.

Poznámka

Proto je snahou požívat efektivní odhady, u nichž je dosaženo dostatečně nízkého rozptylu již pro malé rozsahy výběru (např. u Laplaceova rozdělení postačuje pro dosažení stejného rozptylu při použití mediánu polovina měření než při použití aritmetického průměru.

Při tvorbě konfidenčních intervalů je obecně nutné znát rozdělení daného odhadu, resp. jeho funkce. Tak za předpokladu, že výběr pochází z normálních rozdělení $N(\mu,\sigma^2)$, lze stanovit, že náhodná veličina

$$t = [(x - \mu)\sqrt{N}]/s \tag{42}$$

má Studentovo rozdělení s f = N - 1 stupni volnosti a náhodná veličina

$$\chi^2 = \frac{(N-1)s^2}{\sigma^2} \tag{43}$$

má χ^2 rozdělení s f = N - 1 stupni volnosti. Pro 100 (1 - α) %ní konfidenční interval střední hodnoty μ normálního rozdělení pak vyjde

$$\bar{x} - t_{1-\alpha/2}\frac{s}{\sqrt{N}} \le \mu \le \bar{x} + t_{1-\alpha/2}\frac{s}{\sqrt{N}} \tag{44}$$

kde $t_{1-\alpha/2}$ je kvantil Studentova rozdělení (pro N > 30 ho lze nahradit kvantilem $u_{1-\alpha/2}$ normovaného normálního rozdělení). Pro větší rozsahy výběru (uvádí se od N > 40) lze u normálního rozdělení použít i pro ostatní parametry (rozptyl, špičatost, variační koeficient) aproximace, že náhodná veličina

$$u = \frac{\text{odhad} - \text{teor. hodnota}}{\sqrt{\text{rozptyl odhadu}}}. \tag{45}$$

má normované normální rozdělení a určit konfidenční interval typu

$$\text{odhad} \pm u_{1-\alpha/2}\sqrt{\text{rozptyl odhadu}} \tag{46}$$

Pro jiná než normální rozdělení lze využít vlastností interkvantilové odchylky a definovat **90% ní** interval spolehlivosti

$$\text{odhad} \pm 1.65\sqrt{\text{rozptyl odhadu}} \tag{47}$$

Při analýze jednorozměrných výběrů by vždy měly být určeny parametr polohy, rozptýlení, špičatosti a odpovídající 90%ní intervaly spolehlivosti. Pro nesymetrické rozdělení je třeba nejdříve provést symetrizační transformaci nebo určit typ rozdělení dat.

C. Testy významnosti

Testy významnosti úzce souvisí s konfidenčními intervaly. Detailně jsou popsány v práci [1]. Opět je lze realizovat na základě znalostí funkcí odhadů majících rozdělení t a χ^2 pro normální rozdělení dat. V případě, že rozdělení dat normální není, je nutné buď vhodným způsobem určit rozdělení testovací statistiky nebo použít různých simulačních postupů. Platí, že pro symetrická rozdělení lze klasických t-testů použít i při různé délce konců rozdělení, ze kterého data pocházejí. Podstatně komplikovanější jsou případy, kdy je rozdělení dat zešikmené.

Příklad 1 - řešení

Stanovme výběrové charakteristiky pro veličiny w a transformované veličiny x.

A. P*ůvodní veličiny (w)*

Přímým dosazením do odpovídajících vztahů vyjde

$\bar{x} = 10$, $s^2 = 31.667$, $\hat{g}_1 = 0$, $\hat{g}_2 = 1.793$.

Také $\tilde{x}_{0.5} = \hat{x}_P = P_F = \bar{x}(0.25) = 10$, a tedy $\tilde{x}_C = 10$. Protože rozdělení dat není zřejmě normální ($\hat{g}_2$ indikuje skutečně rektangulární rozdělení), použijeme pro vyjádření konfidenčních intervalů rov. (47). Z rov. (31) určíme

D(s) = $s^2(g_2-1) / 4N = 31.667*(1.793 - 1)/(4*19) =$ **0.33**.

Pak 90%ní interval pro směrodatnou odchylku je 5.627 ± 0.95. Pokud bychom zanedbali fakt, že rozdělení dat není normální, vyšlo by **D(s)** = **0.833** a 90%ní konfidenční interval by vyšel nesprávně vyšší 5.627 ± 1.51.

Pro stanovení 90%ního intervalu střední hodnoty $\tilde{x}_C = 10$ určíme nejdříve rozptyl z rov.(41). Vzhledem k malému rozsahu výběru použijeme vztah $k_{0,,9} = 1.65*\sigma = 1.65*5.627 = 9.285$. Pak $D(\tilde{x}_C) = 9.285^2 /(2.72*19) = 1.668$. Tedy 90%ní konfidenční interval je 10 ± 2.752. Vzhledem k nemožnosti výpočtu zpřesněného intervalu $k_{0.9}$ z kvantilů je tento interval stejný jako klasický, určený z rozptylu $D(\bar{x}) = \sigma^2 / N$.

Určeme pro tato data ještě entropický koeficient a jeho interval spolehlivosti.
Pro volbu počtu intervalů M použijeme vztahu

$$M = \frac{(\hat{g}_2^2 + 1.5)}{6N^{0.4}} \acute{} \frac{91.793 + 1.50}{6*19^{0.4}} 1.78 \approx 2$$

Rozdělíme tedy data do dvou skupin s šířkou intervalu h = 9, takže bude $n_1 = 10$ a $n_2 = 9$. Z rov.(32) pak vyjde

H = (10*.log(10) + 9*log(9)) = 0.978
E = 9.19 / (2*5.627)*10$^{-0.978}$ = 1.597

Z rov. (3.28) pak určíme příslušný rozptyl

$$D(\hat{E}) = \frac{0.81}{\hat{g}_2^2 E^3 N} = 0.81/(1.793^2 * 1.597^2 * 0.19) = 0.005199$$

Tedy konfidenční interval je 1.597 ± 0.1189. Teoretická hodnota pro rovnoměrné rozdělení je však E = 1.73, tedy těsně za hranici. Důvodem je zřejmě malý počet intervalů.
Zvolme tedy počet intervalu M = 10, takže se do osmi vejdou dva prvky a do dvou 1 prvek. Tedy h = 1.8 a $H = N^{-1} \Sigma n_i \log(n_i) = 0.253$, E = 1.295. Pro rozptyl tohoto odhadu pak vyjde $D(E) = 4.616 .10^{-3}$ a 90%ní interval je 1.695 ± 0.112. Tento interval již obsahuje teoretickou hodnotu v těsné blízkosti středu.

B. <u>Transformované veličiny (x)</u>

Přímým dosazením do odpovídajících vztahů vyjde $\bar{x} = 0.187$, $s^2 = 0.0517$, $\hat{g}_1 = 2.694$ a $\hat{g}_2 = 9.85$. Pro odhady dalších parametrů polohy platí $\tilde{x}_{0.5} = 0.1$, $\hat{x}_P = 0.526$, $P_F = 0.126$, a $\bar{x}(0.25) = 0.109$. Protože rozdělení dat není zřejmě symetrické, nelze vlastně v další analýze pokračovat. Bylo by nutné nalézt symetrizační transformaci (což je pochopitelně reciproká transformace x^{-1}).

Zajímavé je, že také řada adaptivních a robustních odhadů polohy vede k odhadům kolem mediánu 0.1, resp. dokonce nižším (Mobergova procedura, která se považuje za velmi dobrou adaptivní i robustní vede k odhadu 0.082).
Porovnáme-li tyto odhady s teoretickými výsledky nezkreslenými velikostí výběru zjistíme, že v netransformovaných datech je poměrně dobrá shoda (střední hodnoty se rovnají a rozptyly leží v 90%ním konfidenčním intervalu). U transformovaných dat se již výrazněji liší jak střední hodnoty, tak i rozptyly. Vzhledem k tomu, že jde o zešikmené rozdělení, je otázkou, co brát jako charakteristiku polohy (když nejde o střed symetrie). Většina postupů ukazuje na medián jako veličinu, která se transformací nezkresluje.

4. ZÁVĚR

Navržený způsob charakterizace centra rozdělení pomocí mediánu $\tilde{x}_C$ je poměrně jednoduchý a lze ho použít pro symetrická rozdělení. Také rozptyl tohoto odhadu se dá snadno vyčíslit. Pro rutinní analýzu dat se tak dá jednoduše využít informací, které se většinou pouze počítají ale jinak dále nevyužívají

Poděkování: *Tato práce vznikla s podporou grantu MŠMT č. VS 97084.*

4 Literatura

[1] Meloun M., Militký J.:Statistické zpracování experimentálních dat, East Publishing Praha 1998
[2] Novickij P.V.,Zograf I.A.: Oceňka pogrešnostej resultatov izmerenij, Energoatomizdat, Leningrad 1985
[3] Militký J., Militká D.: Moderní matematicko-statistické metody v hutnictví III, NHKG Ostrava 1985

Název souboru:	rutin
Adresář:	E:\Pom
Šablona:	D:\Program Files\Microsoft Office\Sablony\Normal.dot
Název:	ÚVOD
Předmět:	
Autor:	Ludmila Fridrichova
Klíčová slova:	
Komentáře:	
Datum vytvoření:	14.09.00 13:40
Číslo revize:	2
Poslední uložení:	14.09.00 13:40
Uložil:	Milan Meloun
Celková doba úprav:	1 minuta
Poslední tisk:	14.09.00 13:42
Jako poslední úplný tisk	
Počet stránek:	14
Počet slov:	3 852 (přibližně)
Počet znaků:	21 959 (přibližně)