Bodový odhad parametrů polohy, rozptýlení a tvaru

A. Momentové míry polohy a rozptýlení

Míry polohy: Momentové míry polohy zahrnují různé druhy průměrů, pomocí kterých můžeme charakterizovat centrální tendenci dat. Momentové míry polohy jsou jednoduché číselné charakteristiky, které se vyčíslují ze všech prvků výběru.

Základní momentovou charakteristikou polohy je *aritmetický průměr* $\bar{x}$, který je zároveň maximálně věrohodným odhadem střední hodnoty *E(x)* normálního rozdělení. Představuje první obecný statistický moment m_1. Z *n* prvků výběru $x_1, x_2, ..., x_n$, se vypočte aritmetický průměr $\bar{x}$ dle vzorce

$$\bar{x} = \frac{x_1 + x_2 + \dots x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Tento odhad má rozptyl $D(\bar{x}) = \sigma^2/n$, kde σ^2 je rozptyl souboru, ze kterého výběr pochází. Má-li každý prvek x_i normální rozdělení s rozptylem $\sigma^2(x_i)$, lze pro odhad střední hodnoty odvodit vztah

$$\hat{x}_w = \frac{\sum_{i=1}^{n} \frac{x_i}{\sigma^2(x_i)}}{\sum_{i=1}^{n} \frac{1}{\sigma^2(x_i)}} ,$$

který se nazývá *vážený aritmetický průměr* s vahami $1/\sigma^2(x_i)$. Rozptyl tohoto odhadu má tvar $D(\bar{x}_w) = 1/\sum_{i=1}^{n} 1/\sigma^2(x_i)$. Obě rovnice jsou použitelné při znalosti rozptylů $\sigma^2(x_i)$ nebo jim odpovídajících "vah" pro jednotlivé prvky výběru.

Míry rozptýlení (variability): Momentové charakteristiky rozptýlení slouží jako odhad variability základního souboru. Míry rozptýlení, které charakterizují proměnlivost výběru v absolutní velikosti, tj. ve stejných jednotkách jako sledovaný prvek, nazýváme *mírami absolutního rozptýlení (variability)*. Když však srovnáváme rozptýlení výběrů lišících se svojí úrovní, užíváme *míry relativního rozptýlení (variability)*. Jsou to buď bezrozměrná čísla, nebo čísla vyjádřená v procentech.

Důležité jsou takové míry rozptýlení, jejichž velikost je závislá na velikostech všech prvků výběru. Mírou rozptýlení, která měří současně rozptýlení všech prvků kolem střední hodnoty se nazývá *rozptyl* σ^2. Je definován jako druhý centrální statistický moment m_2. Pro odhad rozptylu platí vztah

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

Rozptyl tohoto odhadu je $D(\hat{\sigma}^2) = 2\,\sigma^4/n$. V praktických situacích není parametr střední hodnoty μ znám a nahrazuje se aritmetickým průměrem $\mu = \bar{x}$. Takto definovaný rozptyl $\hat{\sigma}^2$ však představuje vychýlený odhad, protože $E(\hat{\sigma}^2) = K\ \sigma^2$, kde $K = (n - 1)/n$. Jako nevychýlený odhad se užívá *výběrový rozptyl*

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

Z praktického hlediska je určitou nevýhodou, že výběrový rozptyl je vyjádřen ve čtvercích užité jednotky. Proto se za míru rozptýlení volí obvykle druhá odmocnina z rozptylu, označená jako *směrodatná odchylka* $s = \sqrt{s^2}$. Její výhodou je to, že je uvedena ve stejných jednotkách jako zkoumaný výběr.

Pro charakterizaci relativního rozptýlení dat se užívá míry relativního rozptýlení, nazvané *variační koeficient* $\delta = \sigma/\mu$ nebo-li *relativní směrodatná odchylka* (často vyjádřená v procentech) se svým rozptylem

$$D(\delta) \approx \delta^2\ (n + \delta^2\ (2\ n + 1))/(2\ n\ (n - 1)).$$

Jeho odhad $\hat{\delta}$ je roven $\hat{\delta} = s/\bar{x}$.

Míry tvaru: momentové charakteristiky tvaru poskytují informace o tvaru rozdělení. Užívá se *šikmost (asymetrie)* g_1 čili třetí normovaný centrální moment a *špičatost (exces)* g_2 čili čtvrtý normovaný centrální moment. Pro normální rozdělení platí hodnoty $g_1 = 0$ a $g_2 = 3$, pro rovnoměrné $g_1 = 0$ a $g_2 = 1.8$, pro Laplaceovo $g_1 = 0$ a $g_2 = 6$ a pro exponenciální $g_1 = 2$ a $g_2 = 9$.

Momentový odhad *šikmosti* g_1 je vyjádřen vztahem

$$\hat{g}_1 = \frac{\sqrt{n} \sum_{i=1}^{n} (x_i - \bar{x})^3}{\left[\sum_{i=1}^{n} (x_i - \bar{x})^2\right]^{3/2}} .$$

Střední hodnota pro výběry normálního rozdělení je $E(\hat{g}_1) = 0$. Pro asymptotický rozptyl tohoto odhadu platí $D(\hat{g}_1) = (n - 2)/[(n + 1)(n + 3)]$.

Momentový odhad *špičatosti* g_2 je vyjádřen vztahem

$$\hat{g}_2 = \frac{n \sum_{i=1}^{n} (x_i - \bar{x})^4}{\left[\sum_{i=1}^{n} (x_i - \bar{x})^2\right]^2} .$$

Střední hodnota tohoto odhadu pro výběry z normálního rozdělení je

$$E(\hat{g}_2) = 3 - 6/(n + 1).$$

Pro asymptotický rozptyl tohoto odhadu platí vztah

$$D(\hat{g}_2) \approx 24\, n\, (n - 2)\, (n - 3)/[(n + 1)^2\, (n + 3)\, (n + 5)].$$

Při stanovení libovolného bodového odhadu parametru je třeba určit vždy i jeho rozptyl. K docílení stejné "přesnosti" odhadů, vyjádřené jeho rozptylem, je třeba při užití méně efektivního odhadu provést větší počet měření n. Například u dat pocházejících z normálního rozdělení se musí při použití mediánu provést 1.6krát více měření než při použití aritmetického průměru, aby se docílilo stejné přesnosti odhadu. Naopak u dat pocházejících z Laplaceova rozdělení se k odhadu parametru polohy pomocí aritmetického průměru $\bar{x}$ musí použít dvojnásobný počet měření než u mediánu, aby bylo docíleno stejné přesnosti odhadu.

B. Kvantilové a robustní míry polohy a rozptýlení

Kvantilové a robustní charakteristiky jsou méně citlivé na vybočující hodnoty než momentové. Patří sem:

Modus, $\hat{x}_M$, který je definován jako lokální maximum na hustotě pravděpodobnosti. V praxi se vyskytují většinou rozdělení unimodální, jejichž hustota pravděpodobnosti má pouze jedno maximum. Módus je vždy robustní, není citlivý na vybočující měření.

Kvantily (kvartily, decily, percentily). *Výběrový α-kvantil* je hodnota, která rozděluje výběr prvků na dvě části, jedna obsahuje α% prvků, které jsou menší (nebo stejné) než tento kvantil, druhá část (1-α)% prvků, které jsou větší (nebo stejné) než kvantil. V případě kvartilů jde o kvantily, které dělí uspořádané prvky ve výběru na čtyři části, přičemž každá část obsahuje 25% prvků. Kvartily jsou celkem tři: *dolní kvartil* $\tilde{x}_{0.25}$ odděluje čtvrtinu nejmenších prvků. Prostřední kvartil se jmenuje *medián* $\tilde{x}_{0.5}$ a rozděluje výběr prvků na dvě stejné části, z nichž každá obsahuje 50% prvků. Jsou-li prvky výběru setříděny podle velikosti vzestupně $x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}$ (pořádkové statistiky), je medián pro n liché roven $\tilde{x}_{0.5} = x_{(k)}$, kde $k = (n + 1)/2$ a pro n sudé roven $\tilde{x}_{0.5} = [x_{(k)} + x_{(k+1)}]/2$ kde $k = n/2$. Medián patří mezi kvantilové odhady, které jsou robustní, tj. necitlivé na vybočující hodnoty. Medián je maximálně věrohodným odhadem polohy u Laplaceova (oboustranného exponenciálního) rozdělení a má pro toto rozdělení minimální rozptyl $D_L(\tilde{x}_{0.5}) = \sigma^2/2\, n$. Pro normální rozdělení však již medián nemá nejmenší rozptyl. Konečně třetím kvartilem je *horní kvartil* $\tilde{x}_{0.75}$, který odděluje 75% menších prvků od zbývajících 25% největších.

Obdobně jsou definovány *decily* $\tilde{x}_{10}, \tilde{x}_{20}, \ldots, \tilde{x}_{90}$, které dělí výběr na 10 stejně obsazených částí, ve kterých je stejná relativní četnost a *centily* $\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_{99}$, které dělí výběr na 100 stejně obsazených částí.

Z kvantilových odhadů rozptýlení se používá *interkvartilové rozpětí*

$$R = (\tilde{x}_{0.75} - \tilde{x}_{0.25}),$$

kde $\tilde{x}_{0.75}$ je odhad horního kvartilu a $\tilde{x}_{0.25}$ odhad dolního kvartilu. S využitím R lze odhadnout směrodatnou odchylku σ_R podle vztahu $s_R = 0.7413\ R$.

Jedním z nejefektivnějších robustních a přitom jednoduchých odhadů parametru polohy je *uřezaný průměr* $\bar{x}(\vartheta)$, využívající lineární kombinace pořádkových statistik

$$\bar{x}(\vartheta) = \frac{1}{n - 2M} \sum_{i=M+1}^{n-M} x_{(i)} ,$$

kde M = int ($\vartheta\ n$ / 100). Parametr ϑ určuje procento "uřezaných" pořádkových statistik na každém konci, nejvyšších a nejnižších. Optimální hodnota bývá 10%. Vzniká tak 10%ní uřezaný průměr $\bar{x}(10)$. V případě očekávaného většího počtu vybočujících měření se uřezává až na hodnotu ϑ = 25%. Uřezaný průměr se užívá s odhadem směrodatné odchylky, určené z winsorizovaného součtu čtverců odchylek

$$S_w(\vartheta) = \sum_{i=M+2}^{n-M-1} (x_{(i)} - \bar{x}_w(\vartheta))^2 + (M + 1)\ [(x_{(M+1)} - \bar{x}_w(\vartheta))^2 + (x_{(n-M)} - \bar{x}_w(\vartheta))^2]$$

kde $\bar{x}_w(\vartheta)$ je *winsorizovaný průměr,* pro který platí definiční vztah

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

Pro nesymetrická, značně sešikmená rozdělení je doporučen *nesymetrický uřezaný průměr* $\bar{x}\ (\vartheta_1 , \vartheta_2)$, pro který platí

$$\bar{x}(\vartheta_1, \vartheta_2) = \frac{\sum_{i=n_1}^{n_2} x_i}{n_2 - n_1 + 1}$$

kde n_1 = int($\vartheta_1\ n$ / 100) a $n_2 = n$ - int($\vartheta_2\ n$ / 100). Pokud jsou hodnoty ϑ_1 a ϑ_2 zvoleny tak, že rozdělení uřezaného výběru je již symetrické, lze určit *rozptyl nesymetricky uřezaného* průměru $\bar{x}(\vartheta_1, \vartheta_2)$ vztahem

$$s_n^2 = \frac{1}{h\ (h + 1)}\ [n_1\ (x_{(n_1)} - \bar{x}(\vartheta_1, \vartheta_2))^2 + \sum_{i=n_1+1}^{n_2-1} (x_{(i)} - \bar{x}(\vartheta_1, \vartheta_2))^2 +$$

$$+ (n - n_2 + 1)\ (x_{(n_2)} - \bar{x}(\vartheta_1, \vartheta_2))^2 - ((n_1 - 1)\ (x_{(n_1)} -$$

$$- \bar{x}(\vartheta_1, \vartheta_2))^2 + (n - n_2)\ (x_{(n_2)} - \bar{x}(\vartheta_1, \vartheta_2))^2]$$

kde $h = n_2$ - n_1 + 1.

M-odhady jsou maximálně věrohodné odhady parametrů pro speciálně vybraná rozdělení. Maximalizace věrohodnostní funkce podle parametru μ_M zde vede k minimalizaci funkce $\sum_{i=1}^{n} \varrho \left\| \frac{x_i - \mu_M}{\sigma} \right\| \rightarrow \min$. Tvar funkce $\varrho(u)$ určuje vlastnost odhadu. Derivací tohoto vztahu a úpravou vyjde výraz pro *M-odhad střední hodnoty*

$$\hat{\mu}_M = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i},$$

kde *statistická váha* je $w_i = W \left\| \frac{x_i - \mu_M}{\sigma} \right\|$ a $W(u) = \frac{d\rho(u)}{du}$. Mezi *M*-odhady patří však i medián a aritmetický průměr. Pro *robustní M*-odhady platí, že *váhová funkce W(u)* musí být ohraničená. Protože statistická váha w_i je funkcí μ_M, musí se výpočet provést iterativně a počáteční hodnotou může být aritmetický průměr. Mezi doporučované váhové funkce *W(u)* patří bikvadratická funkce typu

$$W(u) = \begin{cases} (1 - \left(\frac{u}{4.69}\right)^2)^2 & \text{pro} \quad |u| < 4.69 \\ 0 & \text{pro} \quad |u| \geq 4.69 \end{cases}$$

kde konstanta 4.69 zajišťuje, že pro normálně rozdělená data bude asymptotická efektivnost odhadu μ_M rovna 0.95. Doporučuje se použít i robustní *M-odhad směrodatné odchylky* dle výrazu

$$s_M = \sqrt{(\sum_{i=1}^{n} V_i [x_i - \hat{\mu}_M]^2) \; / \; \sum_{i=1}^{n} V_i},$$

kde $V_i = W\left(\sqrt{\Delta\left(\frac{x_i - \hat{\mu}_M}{s_M}\right)}\right)$, a kde $\Delta(u)$ je *odchylková funkce*, pro kterou platí

$$\Delta(u) = \begin{cases} u^2 - \ln(u^2) - 1 & \text{pro} \quad u \neq 0 \\ \infty & \text{pro} \quad u = 0 \end{cases}$$

Protože robustní *M*-odhad $\hat{\mu}_M$ představuje vlastně vážený aritmetický průměr, je jeho rozptyl vyjádřen vztahem $D(\hat{\mu}_M) = s_M^2 \; / \sum_{i=1}^{n} w_i$.

Analýza jednorozměrného výběru *norm* (ADSTAT)

(1) Odhady klasických parametrů:	
Odhad aritmetického průměru $\bar{x}$	10.012
Odhad rozptylu s^2	5.2232E-03
Odhad směrodatné odchylky s	0.0723
Odhad šikmosti $\hat{g}_1$	-0.04
Odhad špičatosti $\hat{g}_2$	3.08
Dolní mez 95.0% intervalu spolehlivosti L_D	9.998
Horní mez 95.0% intervalu spolehlivosti L_H	10.026
(2) Odhady ostatních parametrů:	
Odhad modu $\hat{x}_M$	10.000
Odhad polosumy $\hat{x}_P$	10.011
(3) Robustní odhady parametrů:	
Medián $\tilde{x}_{0.5}$	10.011
Odhad směrodatné odchylky mediánu $s(\tilde{x}_{0.5})$	0.0780
Dolní mez 95.0% intervalu spolehlivosti L_D	9.991
Horní mez 95.0% intervalu spolehlivosti L_H	10.031
Odhad 10%ního uřezaného průměru $\bar{x}(10\%)$	10.013
Odhad směrodatné odchylky $s(10\%)$	0.0719
Odhad winsorizovaného průměru $\bar{x}_w(10\%)$	10.013
Odhad směrodatné odchylky $s_w(10\%)$	0.0648
Dolní mez 95.0% intervalu spolehlivosti L_D	9.998
Horní mez 95.0% intervalu spolehlivosti L_H	10.027
Odhad 40%ního uřezaného průměru $\bar{x}(40\%)$	10.011
Odhad směrodatné odchylky $s(40\%)$	0.0726
Odhad winsorizovaného průměru $\bar{x}_w(40\%)$	10.013
Odhad směrodatné odchylky $s_w(40\%)$	0.0316
Dolní mez 95.0% intervalu spolehlivosti L_D	9.996
Horní mez 95.0% intervalu spolehlivosti L_H	10.026
Odhad M-odhadu střední hodnoty $\hat{\mu}_M$	10.012
Odhad směrodatné odchylky s_M	0.0719
Dolní mez 95.0% intervalu spolehlivosti L_D	9.997
Horní mez 95.0% intervalu spolehlivosti L_H	10.027
(4) Hoggovy adaptivní odhady parametrů:	
Hoggův průměr $\hat{\mu}_M$	10.012
Odhad směrodatné odchylky s_M	0.0716
Dolní mez 95.0% intervalu spolehlivosti L_D	9.998
Horní mez 95.0% intervalu spolehlivosti L_H	10.027