Robustní odhady

Problémy analýzy:

- výběr nepochází z Gaussova (normál.) rozdělení
- v datech jsou vybočující hodnoty,

Robustní metody:

- umožňují pouze bodové odhady polohy,
- obvykle se neurčují odhady rozptylů,
- není známo jejich rozdělení,
- nelze konstruovat intervaly spolehlivosti,
- nelze provádět statistické testování.

Polosuma

Polosuma je efektivnější než $\bar{x}$ pro $g_2 > 2.2$.

- **Pro rozdělení s plochými vrcholy se doporučuje použití** kvartilové polosumy

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

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

Rozptyl odhadu polosumy

Rozptyl odhadu polosumy je pro normální rozdělení roven

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

Pro rovnoměrné rozdělení je $D(\hat{x}_P) = \frac{(6\sigma^2)}{[(N-1)(N-2)]}$

Arkussínové rozdělení je definováno v intervalu (-A, A) a pro jeho hustotu pravděpodobnosti platí $f(x) = \frac{1}{\pi\sqrt{A^2 - x^2}}$

Pro arkussínové rozdělení $D(\hat{x}_P) = \frac{(5\pi^4\sigma^2)}{N^4}$

Medián $\tilde{x}_{0.5}$

- patří k nejstarším robustním odhadům,
- má přesnou interpretaci jak pro symetrická, tak i nesymetrická rozdělení,
- jde vždy o 50%ní kvantil: polovina prvků leží pod a polovina nad hodnotou mediánu $\tilde{x}_{0.5}$,

- když rozdělení výběru není známé, jsou vybočující hodnoty a pak se vypočte neparametrický odhad směrodatné odchylky mediánu

$$s_M = \frac{x_{(n-k+1)} - x_{(k)}}{2\, u_{\alpha/2}},$$

kde $k = \frac{n+1}{2} - |u_{\alpha/2}| \sqrt{\frac{n}{4}}$, a pro $\alpha = 0.05$ je $|u_{0.025}| = 1.96$.

Uřezaný průměr $\bar{x}(\vartheta)$

Využívá lineární kombinace pořádkových statistik

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

kde M = int (ϑ n/100), a ϑ určuje procento "uřezaných" pořádkových statistik, a to nejvyšších a nejnižších.

Optimální hodnota bývá ϑ = 10%, a tak vznikne 10%ní uřezaný průměr $\bar{x}(10)$.

V případě většího počtu vybočujících měření se uřezává až na hodnotu ϑ = 25 %.

- **α-uřezaný průměr $\bar{x}(\alpha)$ je definován vztahem**

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

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

Winsorizovaný průměr $\bar{x}_w(\vartheta)$

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

Tukey-McLaughlinovo testační kritérium pro test správnosti

H_0: $\bar{x}(\vartheta) = \mu$ ***vs.*** H_A: $\bar{x}(\vartheta) \neq \mu$

$$t_R(\vartheta) = \frac{[\bar{x}(\vartheta) - \mu]\sqrt{h\,(h-1)}}{S_w(\vartheta)}$$, kde $h = n - 2\,M$.

Pro malé výběry od $n = 5$ se doporučuje $M = 1$, tj. $\vartheta = 20$

Testování: $t_R(\vartheta) < t_{1-\alpha/2}(h-1)$, pak H_0 je přijata.

Winsorizovaný rozptyl je

$$s_w^{\,2}(\vartheta) = \frac{S_w(\vartheta)}{(h-1)},$$

kde ***winsorizovaný součet čtverců odchylek*** se vypočte

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

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

Pro nesymetrická, silně sešikmená rozdělení užijeme

nesymetrický uřezaný průměr $\bar{x}(\vartheta_1, \vartheta_2)$

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

kde $n_1 = \mathrm{int}(\vartheta_1\, n/100)$ a $n_2 = n - \mathrm{int}(\vartheta_2\, n/100)$.

Rozptyl nesymetricky uřezaného průměru $\bar{x}(\vartheta_1, \vartheta_2)$ je

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

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

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

kde $h = n_2 - n_1 + 1$.

Pro malé výběry:

Marritzův-Jarretův odhad směrodatné odchylky mediánu

$$s_M^* = \sqrt{\sum_{i=1}^{n} w_i x_{(i)}^2 - \left(\sum_{i=1}^{n} w_i x_{(i)}\right)^2}$$

kde $w_i = \dfrac{J\left(\dfrac{i - 0.5}{n}\right)}{\sum_{j=1}^{n} J\left(\dfrac{j - 0.5}{n}\right)}$, $J(x) = \dfrac{n!}{(m!)^2} x^m (1 - x)^m$

a m = int((n - 1)/2) a int(x) je celočíselná část čísla x.

Interval spolehlivosti mediánu vychází z veličiny

$$T_M = \frac{\tilde{x}_{0.5} - \text{med}}{s_M} \approx \frac{\tilde{x}_{0.5} - \text{med}}{s_M^*}$$

která má Studentovo rozdělení s (n - 1) stupni volnosti.

Uřezaný výběr je již symetrický, a **testovací kritérium** má pro H_0: $\bar{x}(\vartheta_1, \vartheta_2) = \mu$ ***vs.*** H_A: $\bar{x}(\vartheta_1, \vartheta_2) \neq \mu$ tvar

$$t_a(\vartheta_1, \vartheta_2) = \frac{\bar{x}(\vartheta_1, \vartheta_2) - \mu}{s_n}$$

Testování: $t_a(\vartheta_1, \vartheta_2) < t_{1-\alpha/2}(n_2 - n_1 + 1 - 2)$, pak H_0 je přijata.

Velikost uřezání:

užívá se následujících selektorových kritérií,

a) pro ***relativní délky konců*** rozdělení kritérium Q_1

$$Q_1 = \frac{\bar{U}(0.05) - \bar{L}(0.05)}{\bar{U}(0.5) - \bar{L}(0.5)}$$

b) a pro ***relativní sešikmení*** se užívá kritérium Q_2

$$Q_2 = \frac{\bar{U}(0.05) - \bar{M}(0.5)}{\bar{M}(0.5) - \bar{L}(0.05)},$$

kde

$\bar{U}(ß)$ je průměr z nß největších pořádkových statistik,

$\bar{L}(ß)$ je průměr z nß nejmenších pořádkových statistik,

$\bar{M}(ß)$ je průměr z nß prostředních pořádkových statistik.

Pomocí Q_1 a Q_2 lze vybírat parametr ϑ pro $\bar{x}(\vartheta)$:

1. Je-li $Q_2 \approx 0$ a rozdělení výběru je symetrické, provádí se volba ϑ v závislosti na rozsahu výběru n podle tabulky

n	ϑ (%)	Q_1^*
<10	6.25	pro všechna Q_1^*
	12.5	$Q_1^* \leq 1.84$
10 - 20	25	$Q_1^* > 1.84$
	9.375	$Q_1^* < 1.81$
20 - 30	18.75	$1.81 \leq Q_1^* \leq 1.87$
	28.125	$Q_1^* > 1.87$

2. Pro rozdělení symetrické: volí se ϑ následovně

$\vartheta = 15\ \%$ pro $Q_1 < 2.9$,
$\vartheta = 25\ \%$ pro $2.9 \leq Q_1 < 3.5$,
$\vartheta = 35\ \%$ pro $Q_1 \geq 3.5$.

3. Pro Q_2 nenulové bude nesymetricky uřezaný průměr, a ϑ_1 a ϑ_2 se vybírají následovně:

pro $Q_2 \leq 1.4$, $Q_1 > 2.68$, $\vartheta_1 = 25\ \%$ a $\vartheta_2 = 25\ \%$,
pro $Q_2 > 1.4$, $Q_1 < 1.98$, $\vartheta_1 = 0\ \%$ a $\vartheta_2 = 50\ \%$,
pro $Q_2 > 1.4$, $Q_1 > 2.68$, $\vartheta_1 = 25\ \%$ a $\vartheta_2 = 25\ \%$.

Kombinovaný odhad centra

- **Pro symetrická rozdělení s možnými vybočujícími hodnotami je doporučeno volit jako odhad středu symetrie (centrální hodnoty) medián**

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

kde med{.} označuje medián z prvků v závorce.

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

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

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

Příklad 3.11 *Robustní analýza výběrů z pěti rozdělení*

Proveďte robustní analýzu výběrů z pěti rozdělení: normálního, rovnoměrného, exponenciálního, Laplaceova a lognormálního s využitím robustních měr.

Řešení:

Rozdělení	$\tilde{x}_{0.5}$	R_F^2	s_M	s_M^*	L_1	L_2
normální, N(0; 1)	0.031	1.334	0.032	0.423	-0.328	0.391
rovnoměrné, R(0.5; 0.083)	0.506	0.242	0.006	0.279	0.349	0.663
exponenciální, E(1; 1)	0.705	0.923	0.024	0.393	0.395	1.016
Laplaceovo, L(0; 2)	0.011	1.964	0.032	0.423	-0.348	0.371
lognormální, LN(2.71; 47.21)	1.293	4.109	0.19	0.66	0.416	2.169

a) pro norm. a rovnom. rozdělení poskytují mediánové míry nevýznamně horší výsledky;
b) int. spol. jsou pro exp. rozdělení zhruba stejné, zatímco medián je vůči aritmetickému průměru příliš robustní;
c) mediánové míry jsou lepší pro Laplaceovo rozdělení;
d) pro lognorm. rozd. dochází u mediánových mír ke znač-nému zkreslení a 95%ní interval spolehlivosti neobsahuje skutečnou střední hodnotu.

Závěr: Mediánové a ostatní robustní charakteristiky nejsou vhodné pro sešikmené výběry, protože "uřezávají" správné hodnoty a provádějí tak symetrizaci výběru.

Příklad 3.12 *Uřezaný průměr a winsorizovaný rozptyl výběrů z pěti rozličných rozdělení*

Proveďte robustní analýzu výběrů pěti teoretických rozdělení: normálního, rovnoměrného, exponenciálního, Laplaceova a lognormálního, a to s využitím uřezaného průměru a winsorizovaného rozptylu.

Řešení:

Robustní analýza výběrů z pěti teoretických rozdělení ($\vartheta = 5$)

Rozdělení souboru	$\bar{x}(\vartheta)$	$s_w^2(\vartheta)$	L_1	L_2
normální N(0; 1)	-0.046	0.938	-0.337	0.244
rovnoměrné R(0.5; 0.083)	0.486	0.088	0.397	0.575
exponenciální E(1; 1)	0.836	0.587	0.607	1.066
Laplaceovo L(0; 2)	-0.088	1.409	-0.444	0.268
lognormální LN(2.71; 47.21)	2.551	46.38	0.507	4.595

Robustní analýza výběrů pěti teoretických rozdělení ($\vartheta = 10$)

Rozdělení souboru	$\bar{x}(\vartheta)$	$s_w^2(\vartheta)$	L_1	L_2
normální N(0; 1)	-0.0528	0.787	-0.340	0.234
rovnoměrné R(0.5; 0.083)	0.489	0.092	0.391	0.587
exponenciální E(1; 1)	0.805	0.512	0.573	1.036
Laplaceovo L(0; 2)	-0.049	1.119	-0.392	0.293
lognormální LN(2.71; 47.21)	1.796	7.302	0.921	2.670

Je patrné, že

a) pro symetrická rozdělení N, R a L vede uřezání k malému zkrácení intervalů spolehlivosti,

b) pro nesymetrická rozdělení E a LN se uřezání pro ϑ = 10% projevuje negativně.
   U lognormálního rozdělení již 95%ní interval spolehlivosti neobsahuje skutečnou hodnotu.

Selektorové statistiky pro výběry pěti teoretických rozdělení

Rozdělení souboru	Q_1	Q_1^*	Q_2
normální N(0; 1)	2.856	1.795	0.919
rovnoměrné R(0.5; 0.083)	1.922	1.571	0.054
exponenciální E(1; 1)	3.576	1.784	6.175
Laplaceovo L(0; 2)	3.66	1.885	1.358
lognormální LN(2.71; 47.21)	5.02	2.064	25.182

Malé uřezání ϑ = 5% vede vždy k úzkým int. spol., které obsahují skutečnou střední hodnotu ("pravdu").

Selektorové statistiky Q_1 a Q_2 dobře indikují rozdíly ve špičatosti a šikmosti jednotlivých výběrů.

Závěr: Uřezané průměry ve spojení se selektorovými statistikami umožňují určení robustních odhadů také pro nesymetrická rozdělení.

Robustní M-odhady

Jsou maximálně věrohodné odhady parametrů speciálních rozdělení.

Maximalizace věrohodnostní funkce podle μ_M 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.

Mezi M-odhady patří i medián a průměr.

Derivací vztahu vyjde **robustní M-odhad polohy**

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

kde $w_i = W\left(\frac{x_i - \mu_M}{\sigma}\right)$ a $W(u) = \frac{d\rho(u)}{du}$

a funkce W(u) musí být ohraničená.

Za W(u) je doporučovaná bikvadratická funkce

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

kde konstanta 4.69 zajišťuje, že pro normálně rozdělená data bude asymptotická efektivnost odhadu μ_M rovna 0.95.

M-odhad směrodatné odchylky se vypočte

$$s_M = \sqrt{\frac{\sum_{i=1}^{n} V_i (x_i - \hat{\mu}_M)^2}{\sum_{i=1}^{n} V_i}}$$

kde

$$V_i = W\left(\sqrt{\Delta\left(\frac{x_i - \hat{\mu}_M}{s_M}\right)}\right)$$

a $\Delta(u)$ je odchylková funkce, pro kterou platí

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

Postup určení odhadů $\hat{\mu}$ a s_M:

1. Za počáteční odhady $\hat{\mu}_M^{(0)}$ a $s_M^{(0)}$ se berou $\tilde{x}_{0.5}$ a $\hat{\sigma}_R = 0.75(\tilde{x}_{0.75} - \tilde{x}_{0.25})$.

2. Pro tyto počáteční odhady se vypočtou váhy w_i a V_i a zpřesněné odhady $\hat{\mu}_M^{(1)}$ a $s_M^{(1)}$.

3. V druhé iteraci se z prvních odhadů určí nové váhy w_i a V_i a vyčíslí nové odhady $\hat{\mu}_M^{(2)}$ a $s_M^{(2)}$.

4. Iterační proces pokračuje, dokud se odhady ve dvou následujících iteracích významně liší.

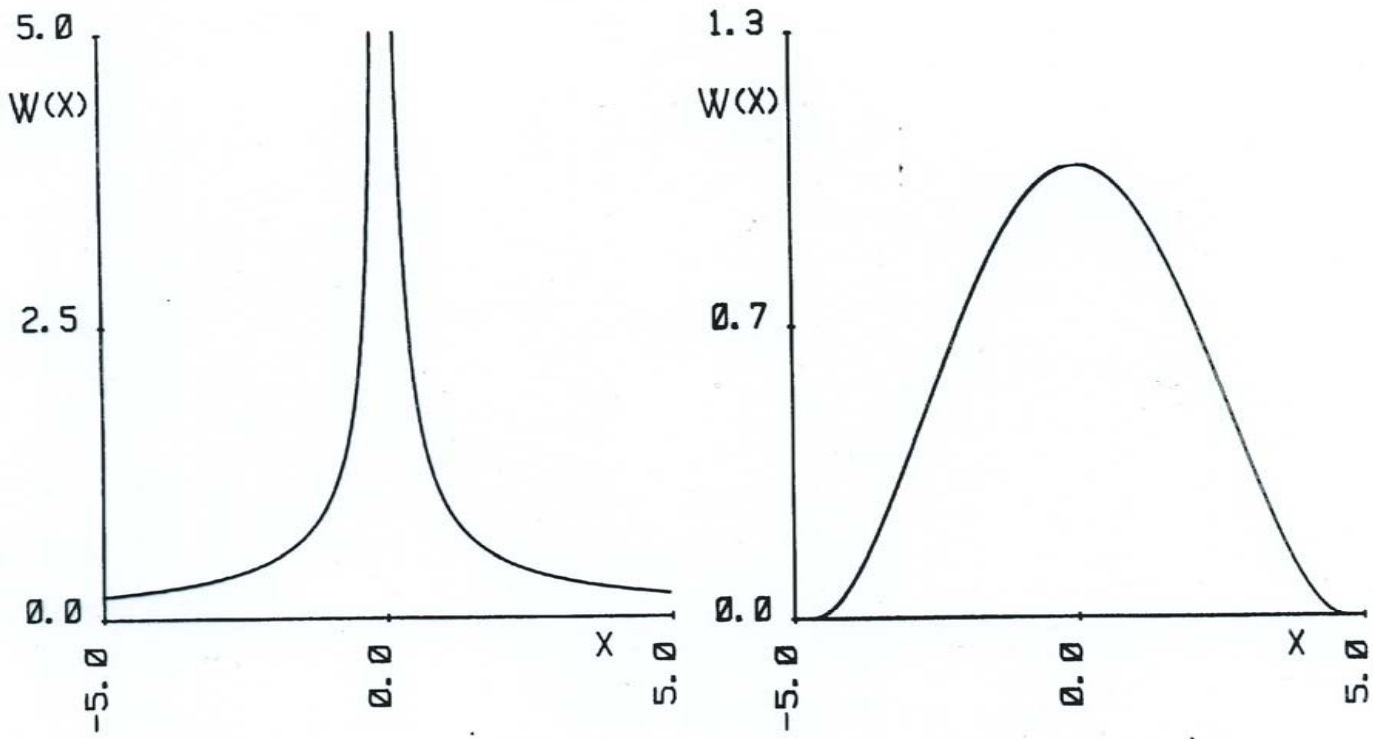
Robustní M-odhad $\hat{\mu}_M$ má **rozptyl** daný vztahem

$$D(\hat{\mu}_M) = \frac{s_M^2}{\sum_{i=1}^{n} w_i}$$

Při konstrukci intervalu spolehlivosti a testování hypotéz se využívá náhodné veličiny

$$T_M = \frac{(\mu_M - \hat{\mu}_M)\sqrt{\sum_{i=1}^{n} w_i}}{s_M}$$

která má Studentovo rozdělení s (n - 1) stupni volnosti.



Porovnání váhových funkcí M-odhadů pro medián (vlevo) a bikvadratickou funkci

Příklad 3.13 *Robustní bikvadratické odhady pro výběry z pěti teoretických rozdělení*

Proveďte robustní analýzu výběrů pěti teoretických rozdělení: normálního, rovnoměrného, exponenciálního, Laplaceova a lognormálního, a to s využitím robustních bikvadratických odhadů.

Rozdělení souboru	$\hat{\mu}_M$	$D(\mu_M)$	L_1	L_2
normální N(0; 1)	-0.0458	1.039	-0.349	0.257
rovnoměrné R(0.5; 0.083)	0.488	0.089	0.399	0.577
exponenciální E(1; 1)	0.762	0.442	0.561	0.964
Laplaceovo L(0; 2)	-0.124	1.464	-0.490	0.242
lognormální LN(2.71; 47.21)	1.375	2.378	0.893	1.858

Řešení:

- V tabulce jsou uvedeny robustní odhady $\hat{\mu}_M$, rozptyly $D(\hat{\mu}_M)$ a meze 95%ního intervalu spolehlivosti střední hodnoty.
- Pro všechna tři symetrická rozdělení N, R a L poskytuje odhady blízké teoretickým a konfidenční interval je poměrně úzký.
- Horší jsou výsledky pro sešikmená rozdělení, kde pro exponenciální i lognormální rozdělení neobsahuje 95%ní konfidenční interval teoretickou hodnotu.

Závěr: robustní M-odhady se nehodí pro analýzu sešikmených rozdělení.

Neparametrické odhady rozptylů

Metoda Bootstrap vychází ze známého odhadu $\hat{\Theta} = g(x_1, ..., x_n)$, který je funkcí n nezávislých náhodných veličin $x_1, ..., x_n$ s neznámým rozdělením F_E, např. uřezaný průměr $\bar{\mathbf{x}}(\vartheta)$.

Postup metody:

1. Z původních dat $\{x_i\}$, i = 1, ..., n, se náhodným výběrem s vracením sestaví B výběrů $\mathbf{x}_j$ velikosti n. Některý prvek může být v Bootstrap výběru zastoupen i několikrát, zatímco jiný prvek vůbec ne. Vybírání je na základě náhodného indexu i

$$\mathbf{i} \quad = \quad \mathbf{int(rnd(0)\,n \;+\; 1)}$$

kde rnd(0) je pseudonáhodné číslo s rozdělením rovnoměrným na intervalu $\langle 0, 1\rangle$. Vyjde-li rnd(0) = 1, dosazuje se i = n.

2. Vypočítají se odhady parametru polohy $\hat{\Theta}_j = g(\mathbf{x}_j)$, j = 1, ..., B, kde $\mathbf{x}_j$ označuje j-tý Bootstrap výběr velikosti n.

3. **Odhad rozptylu** se vyčíslí dle $\hat{\sigma}_B^2 = \dfrac{\sum_{j=1}^{B} (\hat{\Theta}_j - \bar{\Theta}_B)^2}{B - 1}$,

kde **odhad střední hodnoty** je $\bar{\Theta}_B = \sum_{j=1}^{B} \dfrac{\hat{\Theta}_J}{B}$.

Obyčejně se volí B = 200 až 1000 Bootstrap výběrů.

4. Ke konstrukci intervalu spolehlivosti a testování hypotézy o parametru polohy μ se užívá **náhodná veličina**

$$T_B = \frac{(\bar{\Theta}_B - \mu)\sqrt{B}}{\hat{\sigma}_B}$$

která má pro velká B přibližně standardizované normální rozdělení.

Metoda Jackknife využívá tzv. pseudohodnot y_i

$$y_i = n\hat{\Theta} - (n-1)\hat{\Theta}_{(i)}$$

kde $\hat{\Theta}_{(i)}$ jsou odhady ze všech prvků výběru kromě i-tého.

Střední hodnota Jackknife odhadu $\bar{\Theta}_J$ se vypočte dle

$$\bar{\Theta}_J = \frac{1}{n}\sum_{i=1}^{n} y_i = n\hat{\Theta} - \frac{n-1}{n}\sum_{i=1}^{n}\hat{\Theta}_{(i)}$$

Pro střední a větší výběry lze pseudohodnoty y_i považovat za přibližně normálně rozdělené.

Náhodná veličina Jackknife odhadu $T_J = \dfrac{\bar{\Theta}_J - \Theta}{\hat{\sigma}_J}$

má normované normální rozdělení.

Rozptyl Jackknife odhadu $\hat{\sigma}_J^2$ se počítá ze vztahu

$$\hat{\sigma}_J^2 = \frac{1}{n(n-1)} \sum_{i=1}^{n} (y_i - \bar{\Theta}_J)^2$$

Příklad 3.16 *Interval spolehlivosti metodou Bootstrap a Jackknife u Laplaceova rozdělení*

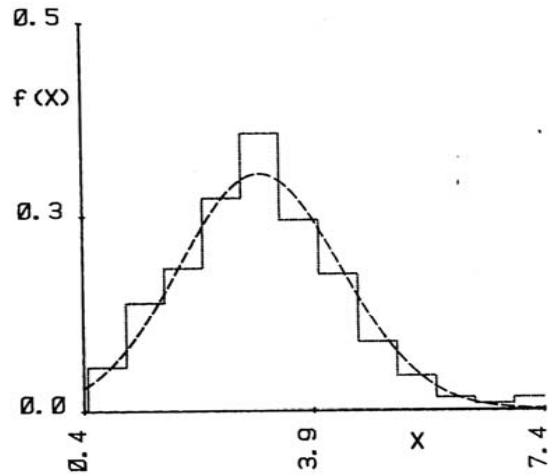
Vypočtěte 95%ní interval spolehlivosti výběrového rozptylu s^2 pro výběr rozsahu n = 30 z Laplaceova rozdělení metodou Bootstrap a Jackknife.

Data:

0.064	0.436	-1.072	-2.036	1.192	-3.162	-0.275	0.734	0.049	-0.569
1.144	-1.070	0.304	0.343	0.144	0.209	-1.269	4.452	5.477	-0.862
-0.026	0.210	0.695	-0.774	-3.723	-0.026	1.879	0.887	-0.333	1.038

Řešení:

1. ***Metoda Bootstrap:*** Pro B = 400 Bootstrap výběrů byla vyčíslena střední hodnota $\bar{s}_B^2 = 3.12$ a její rozptyl $\hat{\sigma}_B^2 = 1.53$. Meze 95%ního intervalu spolehlivosti rozptylu platí $L_{1,2} = \bar{s}_B^2 \pm 1.95\ \hat{\sigma}_B = 3.12 \pm 2.410$.



Histogram Bootstrap odhadů B = 400 rozptylu pro výběr n = 30 z Laplaceova rozdělení

2. ***Metoda Jackknife:*** střední hodnota $\bar{s}_J^2 = 3.2$ a její rozptyl $\hat{\sigma}_J^2 = 1.289$, a meze 95%ního int. spol. $L_{1,2} = \bar{s}_J^2 \pm 1.95\ \hat{\sigma}_J^2 = 3.2 \pm 2.510$.

Závěr: Intervaly spolehlivosti výběrového rozptylu metodou Bootstrap i metodou Jackknife jsou dosti blízké.

Příklad 3.15 *Analýza malých výběrů z pěti teoretických rozdělení*

Proveďte analýzu malých výběrů (n = 5) z pěti teoretických rozdělení: normálního, rovnoměrného, exponenciálního, Laplaceova a lognormál-ního, a to s využitím Hornova postupu.

Data:

N(0, 1)	-1.008	-0.500	0.749	1.723	0.076
R(0.5, 0.083)	0.531	0.677	0.171	0.065	0.848
E(1, 1)	0.757	1.129	0.188	0.067	1.885
L(0, 2)	0.064	0.436	-1.072	-2.036	1.192
LN(2.71, 47.21)	0.191	2.118	0.380	0.264	3.374

Řešení:

Tabulka 3.12 Hornův postup analýzy malých výběrů (n = 5) z pěti teoretických rozdělení

Rozdělení $X(\mu, \sigma^2)$	$\bar{x}$	s^2	L_1	L_2	P_L	R_L	L_1	L_2
normální N(0; 1)	0.208	1.146	-1.121	1.537	0.125	1.249	-0.084	0.333
rovnoměrné R(0.5; 0.083)	0.458	0.111	0.046	0.871	0.424	0.506	-1.332	2.180
exponenciální E(1; 1)	0.805	0.550	-0.115	1.726	0.658	0.941	-0.806	2.123
Laplaceovo L(0; 2)	-0.283	1.628	-1.867	1.300	-0.318	1.508	-0.760	0.124
lognormální LN(2.71; 47.21)	1.265	2.029	-0.503	3.034	1.190	1.855	-0.154	2.536

Závěr: Pivotové charakteristiky umožňují rychlé určení robustních odhadů polohy a rozptýlení parametrů u velmi malých výběrů.

Příklad 3.22 ***Kalibrace pipety z malého počtu naměřených dat***

Pipeta o objemu 25 ml byla kalibrována metodou vážení a bylo získáno 7 hodnot. Určete bodové a intervalové odhady skutečného objemu pipety.

Data: n = 7, přepočtený objem pipety v ml :

24.96439, 24.97758, 24.96809, 24.97409, 24.96880, 24.94759, 24.97119.

Řešení:

1. Bodové odhady parametrů z původních dat:

$\bar{x}$ = 24.967 ml, s = .00972 ml, δ = 0.039 %, $\hat{g}_1$ = -1.25, $\hat{g}_2$ = 3.64.

2. Ověření základních předpokladů:
   a) nezávislost prvků výběru,
   b) nelze zamítnout nulovou hypotézu o normalitě dat;
   c) mimo vnitřní hradby leží hodnota 24.94759 ml.

3. Interval spolehlivosti aritmetického průměru pro $\alpha = 0.05$ je

$$24.958 \leq \mu \leq 24.976$$

4. Robustní interval spolehlivosti z "biweight" odhadů je

$$24.960 \leq \mu \leq 24.977.$$

5. Výraznější rozdíly jsou však v odhadech rozptylu:

$$D(\hat{\mu}_M) = 8.55 \cdot 10^{-5} \text{ a } \sigma^2(x) = 9.44 \cdot 10^{-5}.$$

6. Hornův postup malému počtu dat:
   - hloubka pivotu $H_L = 2$,
   - dolní pivot je $x_{(2)} = 24.96439$ a horní pivot $x_{(6)} = 24.97409$,
   - pro $1 - \alpha/2 = 0.975$ je kvantil $t_L = 0.72$,
   - odhadem střední hodnoty je pivotová polosuma $P_L = 24.9692$ ml,
   - odhadem rozptýlení pivotové rozpětí $R_L = 0.0097$ ml, $\delta = 0.039$ %,
   - pro 95%ní interval spolehlivosti střední hodnoty je

$$24.962 \leq \mu \leq 24.976.$$

Závěr:

1. S ohledem na malý rozsah výběru je postačující Hornův postup, tj. $P_L = 24.9692$ ml a interval spolehlivosti i. s. $\mathbf{24.962 \leq \mu \leq 24.976.}$
2. Předběžná analýza indikuje jedno vybočující měření, a proto je vhodné použít robustních odhadů, tj. z "biweight" bude i. s. $\mathbf{24.960 \leq \mu \leq 24.977.}$

Příklad 3.23 ***Vyšetření obsahu Na^+, Ca^{2+}, Mg^{2+} a močoviny v krvi***

Ze stáda 180 dojnic byla 32 náhodně vybraným jedincům odebrána krev a vyšetřen obsah tří biogenních prvků a močoviny.

Stanovte, zda obsahy těchto čtyř látek leží v intervalu zdravotní normy:

$$134.5 \leq Na^+ \leq 150 \text{ mmol dm}^{-3},\ 2.24 \leq Ca^{2+} \leq 3.0 \text{ mmol dm}^{-3},$$

$$0.77 \leq Mg^{2+} \leq 1.07 \text{ mmol dm}^{-3},\ 2.5 \leq \text{močovina} \leq 5.1 \text{ mmol dm}^{-3}.$$

Zvolte hladinu významnosti $\alpha = 0.05$.

Data:

1. Výběr hodnot Na^+ [mmol dm^{-3}],
2. Výběr hodnot Ca^{2+} [mmol dm^{-3}],
3. Výběr hodnot Mg^{2+} [mmol dm^{-3}],
4. Výběr hodnot močoviny [mmol dm^{-3}].

Řešení: Vzhledem k rozsahu dat byly vyčísleny 95%ní intervaly spolehlivosti pro očekávané hodnoty aritmetického průměru $E(\bar{x})$, mediánu $E(\tilde{x}_{0.5})$ a 40%ního uřezaného průměru $E(\bar{x}(40))$:

1. ***Pro Na^+:*** $136.27 \le E(\bar{x}) \le 138.89$, $135.63 \le E(\tilde{x}_{0.5}) \le 139.77$ a $135.74 \le E(\bar{x}(40)) \le 139.71$.

Obsah Na^+ leží na hladině významnosti $\alpha = 0.05$ v rozmezí zadané normy.

2. ***Pro Ca^{2+}:*** $2.434 \le E(\bar{x}) \le 2.487$, $2.429 \le E(\tilde{x}_{0.5}) \le 2.491$ a $2.438 \le E(\bar{x}(40)) \le 2.486$.

Obsah Ca^{2+} leží na hladině významnosti $\alpha = 0.05$ v rozmezí zadané normy.

3. ***Pro Mg^{2+}:*** $0.743 \le E(\bar{x}) \le 0.810$, $0.741 \le E(\tilde{x}_{0.5}) \le 0.811$ a $0.746 \le E(\bar{x}(40)) \le 0.802$.

Obsah Mg^{2+} leží na hladině významnosti $\alpha = 0.05$ v rozmezí zadané normy.

4. ***Pro močovinu:*** $4.797 \le E(\bar{x}) \le 5.583$, $4.765 \le E(\tilde{x}_{0.5}) \le 5.635$ a $4.816 \le E(\bar{x}(40)) \le 5.568$.

Obsah močoviny leží na hladině významnosti $\alpha = 0.05$ v intervalu předepsaném normou.

Závěr: Robustní a silně uřezané odhady poskytují v tomto případě prakticky stejné výsledky jako klasická analýza.

Příklad 3.28 ***Horní hranice intervalu spolehlivosti průměrného obsahu fluoru v hnojivu***

Fluor představuje nežádoucí nečistotu ve fosfátových hnojivech. Určete horní hranici 95%ního intervalu spolehlivosti průměrného obsahu fluoru na základě výsledků analýzy 20 vybraných vzorků fosfátového hnojiva.

Data: n = 20, [obsah v %]:
0.16, 0.16, 0.15, 0.13, 0.18, 0.19, 0.13, 0.19, 0.18, 0.14, 0.29, 0.14, 0.12, 0.10, 0.16, 0.13, 0.16, 0.16, 0.13, 0.14.

Řešení:

Vzhledem k heterogenitě hnojiva lze očekávat výskyt vybočujících hodnot, a proto byly vyčísleny intervaly spolehlivosti nejen aritmetického průměru $E(\bar{x})$ ale také mediánu $E(\tilde{x}_{0.5})$ a robustního "biweight" odhadu $E(\hat{\mu}_M)$:

$$\mathbf{0.139 \leq E(\bar{x}) \leq 0.175},$$

$$\mathbf{0.134 \leq E(\tilde{x}_{0.5}) \leq 0.176},$$

$$\mathbf{0.139 \leq E(\hat{\mu}_M) \leq 0.163}.$$

Závěr:

Vzhledem k vybočující hodnotě (0.29) je interval spolehlivosti očekávané hodnoty aritmetického průměru příliš široký a sám aritmetický průměr je vychýlen k vyšším hodnotám.

Horní mez 95%ního intervalu spolehlivosti průměrného obsahu fluoru je proto stanovena na základě robustního odhadu jako 0.163 %.