Poznámky k předmětu Aplikovaná statistika, 8. téma

8. Odhady parametrů rozdělení pravděpodobnosti

Zaměříme se na odhad střední hodnoty a rozptylu a to dvěma způsoby. Předpokládejme, že máme náhodný výběr $X_1, \ldots, X_n$ z rozdělení náhodné veličiny X a chceme odhadnout parametr θ tohoto rozdělení.

(a) Bodový odhad parametru θ je funkce náhodného výběru

$$\hat{\theta} \equiv \hat{\theta}(X_1, \ldots, X_n).$$

Hodnotu $\hat{\theta}$ musí být možné smysluplně interpretovat jako odhad skutečné hodnoty parametru θ, např. je-li $\theta = \mu$, pak $\hat{\mu} = \bar{X} = \hat{\theta}(X_1, \ldots, X_n)$.

U bodového odhadu obvykle požadujeme, aby byl „nestranný“. V této třídě se pak snažíme vybrat ten „nejlepší“.

Definice 1. Bodový odhad $\hat{\theta}_N(X_1, \ldots, X_n)$ je **nestranný**, jestliže

$$\mathbb{E}\,\hat{\theta}_N = \theta.$$

Nestranný bodový odhad $\hat{\theta}_{NN}(X_1, \ldots, X_n)$ je **nejlepší nestranný odhad**, jestliže

$$\operatorname{var}(\hat{\theta}_{NN}) = \min_{\hat{\theta}_N} \operatorname{var}(\hat{\theta}_N).$$

Označme $\mathbb{E}X = \mu$ a $\operatorname{var}X = \sigma^2$. Jelikož ze sedmého tématu víme, že $\mathbb{E}\bar{X} = \mu$ a $\mathbb{E}S^2 = \sigma^2$, jsou $\bar{X}$ a S^2 nestranné odhady parametrů μ a σ^2. Nadto, pokud $X \sim \mathcal{N}(\mu, \sigma^2)$, pak $\bar{X}$ a S^2 jsou nejlepší nestranné odhady parametrů μ a σ^2.

Poznámka 1. *Odhadů parametru je nekonečně mnoho, záleží tedy na vkusu a zkušenosti každého z nás, jaký odhad použijeme.*

Např. pro symetrická rozdělení [1] *je nestranným odhadem parametru* μ

$$\hat{\mu} = \frac{X_{\max} + X_{\min}}{2},$$

kde $X_{\max} = \max\{X_i, i = 1, \ldots, n\}$, $X_{\min} = \min\{X_i, i = 1, \ldots, n\}$.

Anebo odhadem rozptylu σ^2 *je také (vedle výběrového rozptylu* S^2*)*

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Tento odhad ovšem není nestranný ($\mathbb{E}\hat{\sigma}^2 = \frac{n-1}{n}\sigma^2$)*, zato má menší rozptyl než* S^2.

[1] rozdělení je symetrické, jestliže existuje $x_0 \in \mathbb{R}$ tak, že pro funkci g platí

$$g(x_0 - x) = g(x_0 + x),\ x \in \mathbb{R},$$

kde

pro spojité rozdělení je g jeho hustota,

pro diskrétní rozdělení je g jeho pravděpodobnostní funkce.

(b) Intervalový odhad parametru θ

Definice 2. (i) **Oboustranný $(1-\alpha)100\%$ interval spolehlivosti** pro parametr θ je interval

$$\left[\hat{\theta}_1(X_1,\dots,X_n),\hat{\theta}_2(X_1,\dots,X_n)\right]$$

takový, že pro náhodné veličiny $\hat{\theta}_1(X_1,\dots,X_n),\hat{\theta}_2(X_1,\dots,X_n)$ platí

$$\mathbb{P}\left(\hat{\theta}_1(X_1,\dots,X_n)\le\theta\le\hat{\theta}_2(X_1,\dots,X_n)\right)=1-\alpha.$$

(ii) **Jednostranný $(1-\alpha)100\%$ dolní interval spolehlivosti** pro parametr θ je (zdola omezený) interval

$$\left[{}_j\hat{\theta}_1(X_1,\dots,X_n),+\infty\right)$$

s vlastností

$$\mathbb{P}\left({}_j\hat{\theta}_1(X_1,\dots,X_n)\le\theta\right)=1-\alpha.$$

(iii) **Jednostranný $(1-\alpha)100\%$ horní interval spolehlivosti** pro parametr θ je (shora omezený) interval

$$\left(-\infty,{}_j\hat{\theta}_2(X_1,\dots,X_n)\right]$$

s vlastností

$$\mathbb{P}\left(\theta\le{}_j\hat{\theta}_2(X_1,\dots,X_n)\right)=1-\alpha,$$

kde ${}_j\hat{\theta}_1(X_1,\dots,X_n),{}_j\hat{\theta}_2(X_1,\dots,X_n)$ jsou funkce náhodného výběru.

Máme-li realizaci náhodného výběru $x_1,\dots,x_n$, pak se často píše

$$\theta\in\left[\hat{\theta}_1(x_1,\dots,x_n),\hat{\theta}_2(x_1,\dots,x_n)\right].$$

Pokud bychom měli hodně realizací náhodného výběru, pak lze např. 95% interval spolehlivosti pro μ chápat tak, že 95 % intervalů sestavených z oněch realizací by obsahovalo μ.

Poznámka 2. *Uvědomme si, že intervaly spolehlivosti mají náhodné meze, např.*

$$\hat{\theta}_1(X_1,\dots,X_n)\equiv\hat{\theta}_1(X_1(\omega),\dots,X_n(\omega)),\ \omega\in\Omega.$$

Odvození oboustranného intervalu spolehlivosti pro μ,
je-li $X\sim\mathcal{N}(\mu,\sigma^2)$, kde μ neznáme a σ^2 známe.
Již víme, že pochází-li náhodný výběr z $\mathcal{N}(\mu,\sigma^2)$, má výběrový průměr $\bar{X}\sim\mathcal{N}(\mu,\frac{\sigma^2}{n})$, a tudíž normovaná náhodná veličina

$$\bar{X}_N=\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}\sim\mathcal{N}(0,1).$$

Ze sudosti hustoty normovaného normálního rozdělení plyne

$$1-\alpha=\mathbb{P}\left(|\bar{X}_N|\le u_{1-\frac{\alpha}{2}}\right).$$

Dále pak

$$\mathbb{P}\left(|\bar{X}_N|\le u_{1-\frac{\alpha}{2}}\right)=\mathbb{P}\left(-u_{1-\frac{\alpha}{2}}\le\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}\le u_{1-\frac{\alpha}{2}}\right)=\mathbb{P}\left(\bar{X}-\frac{\sigma}{\sqrt{n}}u_{1-\frac{\alpha}{2}}\le\mu\le\bar{X}+\frac{\sigma}{\sqrt{n}}u_{1-\frac{\alpha}{2}}\right)$$
$$=1-\alpha.$$

Definujeme-li

$$\hat{\theta}_1(X_1,\dots,X_n) = \bar{X} - \frac{\sigma}{\sqrt{n}}u_{1-\frac{\alpha}{2}},\ \hat{\theta}_2(X_1,\dots,X_n) = \bar{X} + \frac{\sigma}{\sqrt{n}}u_{1-\frac{\alpha}{2}},$$

našli jsme oboustranný $(1-\alpha)100\%$ interval spolehlivosti pro μ (při σ^2 známém)

$$\left[\bar{X} - \frac{\sigma}{\sqrt{n}}u_{1-\frac{\alpha}{2}}, \bar{X} + \frac{\sigma}{\sqrt{n}}u_{1-\frac{\alpha}{2}}\right].$$

Odvoďme si ještě např. **jednostranný dolní interval spolehlivosti pro $\boldsymbol{\mu}$** (známe-li σ^2):

$$1-\alpha = \mathbb{P}\left(\bar{X}_N \le u_{1-\alpha}\right) = \mathbb{P}\left(\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \le u_{1-\alpha}\right) = \mathbb{P}\left(\bar{X} - \frac{\sigma}{\sqrt{n}}u_{1-\alpha} \le \mu\right).$$

Označíme-li ${}_j\hat{\theta}_1(X_1,\dots,X_n) = \bar{X} - \frac{\sigma}{\sqrt{n}}u_{1-\alpha}$, je hledaný interval

$$\left[\bar{X} - \frac{\sigma}{\sqrt{n}}u_{1-\alpha}, +\infty\right).$$

Stejným postupem lze odvodit i jednostranný horní interval spolehlivosti pro μ (při σ^2 známém)

$$\left(-\infty, \bar{X} + \frac{\sigma}{\sqrt{n}}u_{1-\alpha}\right].$$

Příklad 1. *Náhodná veličina X popisuje výšku křováka. Předpokládáme, že $X \sim \mathcal{N}(\mu, 36)$, tj. $\sigma^2 = 36$ známe a chceme sestrojit 95% interval spolehlivosti pro μ, známe-li výšku pěti náhodně vybraných křováků*

x_i	170	160	165	155	160

Jelikož $\alpha = 0,05$, $n = 5$, $\bar{x} = 162$ a $u_{1-\frac{\alpha}{2}} = u_{0,975} = 1,960$, je hledaný interval

$$\left[162 - \frac{6}{\sqrt{5}}1,960; 162 + \frac{6}{\sqrt{5}}1,960\right] \doteq [155,09; 168,91].$$

Analogicky lze sestrojit 99% interval spolehlivosti pro μ ($u_{0,995} = 2,576$)

$$[154,55; 169,45].$$

Povšimněme si, že s rostoucí spolehlivostí (to je číslo $(1-\alpha)$ %) roste délka intervalu.

Obdobnými postupy lze odvodit i další intervaly spolehlivosti.

Přehled intervalů spolehlivosti

- pro μ, σ^2 známé

$$\left[\bar{X} - \frac{\sigma}{\sqrt{n}}u_{1-\frac{\alpha}{2}}, \bar{X} + \frac{\sigma}{\sqrt{n}}u_{1-\frac{\alpha}{2}}\right],\ \left[\bar{X} - \frac{\sigma}{\sqrt{n}}u_{1-\alpha}, +\infty\right),\ \left(-\infty, \bar{X} + \frac{\sigma}{\sqrt{n}}u_{1-\alpha}\right],$$

- pro μ, σ^2 neznámé

$$\left[\bar{X} - t_{1-\frac{\alpha}{2}}(n-1)\frac{S}{\sqrt{n}}, \bar{X} + t_{1-\frac{\alpha}{2}}(n-1)\frac{S}{\sqrt{n}}\right],$$
$$\left[\bar{X} - t_{1-\alpha}(n-1)\frac{S}{\sqrt{n}}, +\infty\right), \left(-\infty, \bar{X} + t_{1-\alpha}(n-1)\frac{S}{\sqrt{n}}\right],$$

- pro σ^2

$$\left[\frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)}, \frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2}}(n-1)}\right], \left[\frac{(n-1)S^2}{\chi^2_{1-\alpha}(n-1)}, +\infty\right), \left(0, \frac{(n-1)S^2}{\chi^2_{\alpha}(n-1)}\right].$$

Cvičení k tomuto tématu

(i) Ve výrobně rumu se náhodně odebralo 7 vzorků, ve kterých se změřil obsah alkoholu

x_i (v %)	42	44	43	39	40	40	39

Předpokládejme, že koncentrace alkoholu v rumu je náhodná veličina s normálním rozdělením $\mathcal{N}(\mu, \sigma^2)$.

(a) Na základě bodových odhadů μ, σ^2 odhadněte pravděpodobnost, že vzorek odebraný při zítřejší kontrole bude splňovat normu (která zní, že koncentrace je nejméně 40 %).

(b) Kontrola odebere 5 vzorků. Jaká je pravděpodobnost, že

(ba) budou všechny v normě?

(bb) budou aspoň 3 v normě?

(bc) budou všechny pod normou?

(bd) průměr obsahu alkoholu bude nejméně 40 %?

(ii) Během jedné minuty se změřilo množství oxidu siřičitého v 1 m^3 v pěti náhodně vybraných místech dané lokality

x_i (μg)	85	90	80	105	90

Předpokládá se, že množství oxidu siřičitého v 1 m^3 se řídí normálním rozdělením $\mathcal{N}(\mu, \sigma^2)$. Norma říká, že průměrné množství oxidu siřičitého v dané lokalitě nesmí překročit 100 μg na 1 m^3. Sestrojte

(a) 95% a 99% intervaly spolehlivosti (oboustranné, horní i dolní) pro μ, znáte-li $\sigma^2 = 87,5$;

(b) 95% a 90% intervaly spolehlivosti (oboustranné, horní i dolní) pro μ, je-li σ^2 neznámé;

(c) 95% intervaly spolehlivosti (oboustranné, horní i dolní) pro σ^2.

(iii) Uvažme úlohu z předchozího příkladu.

(a) 95% interval spolehlivosti pro μ při σ^2 známém je moc široký. Jaký by musel být rozsah náhodného výběru, aby šířka intervalu byla nejvýše 10 μg?

(b) Řešte úlohu (a) pro σ^2 neznámé.

(dcv) Vezměme si úlohu z Příkladu (i) a sestrojme 95% intervaly spolehlivosti (oboustranné i jednostranné) pro

(a) μ při $\sigma^2 = 4$ známém,

(b) μ při σ^2 neznámém,

(c) σ^2.

Může výrobce na základě vhodných intervalů spolehlivosti pro μ tvrdit, že splnil normu?