Bodové a intervalové odhady

Supplementum ke cvičení 4ST210 Statistika pro finance

Lubomír Štěpánek[1, 2]

[1]Oddělení biomedicínské statistiky
Ústav biofyziky a informatiky
1. lékařská fakulta
Univerzita Karlova, Praha

[2]Katedra biomedicínské informatiky
Fakulta biomedicínského inženýrství
České vysoké učení technické v Praze

Obsah

1. Opakování
2. Úvod do odhadování
3. Bodový odhad
4. Intervalový odhad
5. Literatura

Příklad

- Pravděpodobnost narození chlapce je $0{,}515$. Jaká je pravděpodobnost, že mezi $10\,000$ narozenými novorozenci bude
  - (i) více děvčat než chlapců?
  - (ii) relativní četnost chlapců bude v mezích od $0{,}515$ do $0{,}517$?

Příklad

- Pravděpodobnost narození chlapce je $0{,}515$. Jaká je pravděpodobnost, že mezi $10\,000$ narozenými novorozenci bude
  (i) více děvčat než chlapců?
  (ii) relativní četnost chlapců bude v mezích od $0{,}515$ do $0{,}517$?
- *Řešení.*
  (i) $0{,}00135.$
  (ii) $0{,}15500.$ □

Příklad

- Zaměstnanec jezdí do zaměstnání pravidelně tam i zpět metrem. Doba čekání na příjezd soupravy metra se pohybuje rovnoměrně mezi 0 až 3 minutami. Jaká je pravděpodobnost, že celková doba čekání zaměstnance během 23 pracovních dní[1] bude kratší než 80 minut?

[1] zaměstnanec jede vždy dvakrát denně – do zaměstnání a zpět

Příklad

- Zaměstnanec jezdí do zaměstnání pravidelně tam i zpět metrem. Doba čekání na příjezd soupravy metra se pohybuje rovnoměrně mezi 0 až 3 minutami. Jaká je pravděpodobnost, že celková doba čekání zaměstnance během 23 pracovních dní[1] bude kratší než 80 minut?
- *Řešení.* $0{,}969$. $\square$

[1] zaměstnanec jede vždy dvakrát denně – do zaměstnání a zpět

Motivace

- ve výběru hodnot umíme spočítat výběrové charakteristiky
- obvykle nás ale zajímají *populační* charakteristiky
- pomocí výběrových charakteristik můžeme na populační charakteristiky odhadovat
  - jak se odhady provádí?
  - jaké vlastnosti mají takové odhady, jak jsou „přesné“?

Pojem *populace*

- **populace := základní soubor**
- úplná množina (statistický soubor) všech prvků (statistických jednotek), které spojuje určitá vlastnost a o kterých se snažíme statisticky něco zjistit
- prvky dány výčtem (je-li rozsah populace konečný), nebo společnou vlastností všech prvků (je-li rozsah populace nekonečný i konečný)
- rozsah konečně velké populace obvykle značíme N (u nekonečně velké populace $N \to \infty$)
- např. {T. G. Masaryk, E. Beneš, . . . , V. Klaus, M. Zeman}, {všichni dosavadní prezidenti českého státu}, {všichni obyvatelé Evropy}, apod.

Pojem *výběr*

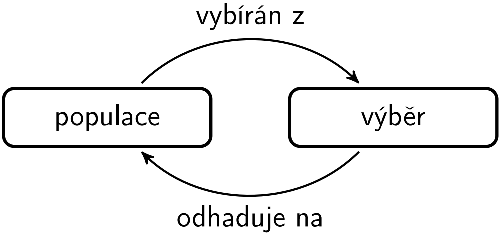
- vyšetřit celou populaci v praxi takřka nemožné
- nekonečně velké populace nelze celkově šetřit už z principu
- výběr := statistický soubor, obsahuje vybrané prvky z populace; je tedy podmnožinou populace
- výběr pořizujeme metodou náhodného, či záměrného výběru
- cílem získat reprezentativní výběr (vystihuje vlastnosti populace), nikoliv selektivní výběr

Reprezentativní výběr

- takový výběr, z kterého je induktivními metodami možné usuzovat na vlastnosti „mateřské“ populace
- pořizujeme *záměrným*, či *náhodným* výběrem
  - záměrný výběr – opírá se o expertízu, zatížen subjektivitou
  - náhodný výběr – náhodné, nezávislé vybírání prvků populace do výběru

Vztah populace a výběru

- z populace je vybírán výběr
- z charakteristik výběru jsou odhadovány charakteristiky populace (!)



Bodový odhad statistického znaku

- předpokládáme, že charakteristická hodnota výběru (průměr, četnost) odpovídá populační hodnotě
- populační hodnota se pokládá rovna dané charakteristické hodnotě výběru
- např. „je-li četnost hypertoniků mezi 20 náhodnými pacienty právě 7, je i četnost hypertoniků v populaci $\frac{7}{20} = 0{,}35 = 35$ %“
- s jakou „mírou jistoty“ jsme se „trefili“ do skutečné populační četnosti?
  - přirovnává se k lovu oštěpem

Intervalový odhad statistického znaku

- (interval spolehlivosti, konfidenční interval)
- interval, ve kterém leží charakteristická hodnota populace s určitou pravděpodobností (spolehlivostí)
- např. např. „je-li četnost hypertoniků mezi 20 náhodnými pacienty právě 7, pak průměrná populační četnost hypertoniků leží s pravděpodobností 95 % intervalu (15; 55) %“
- s jakou „mírou jistoty“ jsme se „trefili“ do skutečné populační četnosti?
  - přirovnává se k lovu sítí

Bodový odhad

- výběrová statistika T je odhadem charakteristiky Θ základního souboru (populace), tedy

$$T = \hat{\Theta}$$

- vlastnosti bodového odhadu
    - (*nezkreslenost*) systematicky nenadhodnocuje ani nepodhodnocuje populační charakteristiku, tedy při opakovaných výběrech

$$|\mathbb{E}(T) - \Theta| = 0$$

    - (*konzistentnost*) pro rostoucí velikost výběru n se odhad T blíží populačnímu protějšku, tedy pro libovolné $\varepsilon > 0$

$$\lim_{n\to\infty} P\left(|T - \Theta| < \varepsilon\right) = 1$$

    - (*vydatnost*) rozptyl odhadu T je co nejmenší, tedy

$$\lim_{n\to\infty} \mathrm{var}(T) = \lim_{n\to\infty} \mathbb{E}((T - \mathbb{E}(T))^2) = \lim_{n\to\infty} \mathbb{E}((T - \Theta)^2) = 0$$

Vybrané bodové odhady

- bodovým odhadem střední hodnoty je výběrový průměr výběru $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)^T$, tedy

$$\hat{\mu} = \bar{x} = \sum_{i=1}^{n} x_i$$

- bodovým odhadem pravděpodobnosti (našeho zájmu) je relativní četnost, tedy

$$\hat{p} = \frac{m}{n},$$

kde m je počet objektů našeho zájmu n je rozsah souboru

Intervalový odhad

- narozdíl od bodového lze určit míru jeho „spolehlivosti“
- $100(1-\alpha)$ % intervalový odhad charakteristiky Θ je

$$P(\Theta_{\text{dolní mez}} \leq \Theta \leq \Theta_{\text{horní mez}}) = 1 - \alpha,$$

  kde $\Theta_{\text{dolní mez}}$ a $\Theta_{\text{horní mez}}$ jsou náhodné veličiny a α se volí obvykle $\alpha \equiv 0{,}05$ nebo $\alpha \equiv 0{,}01$
- existují i jednostranné varianty

Výběrové rozdělení

- předpokládejme, že výběr $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)^T$ je konkrétní realizací vektoru náhodných veličin $(X_1, X_2, \ldots, X_n)^T$ a že $\forall i \in \{1, 2, \ldots, n\}$ je $X_i \sim \mathcal{N}(\mu, \sigma^2)$
- pak pro odhad střední hodnoty výběrového průměru platí

$$\mathbb{E}(\bar{X}) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i) = \frac{1}{n}\cdot n\mu = \mu$$

- a pro odhad rozptylu výběrového průměru platí

$$\mathrm{var}(\bar{X}) = \mathrm{var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{var}(X_i) = \frac{1}{n^2}\cdot n\sigma^2 = \frac{\sigma^2}{n}$$

- zřejmě tedy variabilita výběrového průměru klesá s rostoucím rozsahem výběru n

Výběrové rozdělení

- pro výběrový průměr tedy platí

$$\mathbb{E}(\bar{X}) = \mu$$

$$\mathrm{var}(\bar{X}) = \frac{\sigma^2}{n}$$

- ať je $U \equiv \frac{\bar{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}}$, pak pro u_p-tý kvantil normálního rozdělení je

$$P\left(\frac{\bar{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}} \leq u_p\right) = p$$

- je dále i

$$P\left(\frac{\bar{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}} \leq u_{1-p}\right) = 1 - p$$

Interval spolehlivosti pro střední hodnotu

- dále je i

$$P\left(u_{p/2} \leq \frac{\bar{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}} \leq u_{1-p/2}\right) = 1-(p/2+p/2) = 1-p$$

- a nakonec

$$P\left(-\bar{X}+u_{p/2}\cdot\sqrt{\frac{\sigma^2}{n}} \leq -\mu \leq -\bar{X}+u_{1-p/2}\cdot\sqrt{\frac{\sigma^2}{n}}\right) = 1-p$$

$$P\left(\bar{X}-u_{p/2}\cdot\sqrt{\frac{\sigma^2}{n}} \geq \mu \geq \bar{X}-u_{1-p/2}\cdot\sqrt{\frac{\sigma^2}{n}}\right) = 1-p$$

$$P\left(\bar{X}-u_{1-p/2}\cdot\sqrt{\frac{\sigma^2}{n}} \leq \mu \leq \bar{X}-u_{p/2}\cdot\sqrt{\frac{\sigma^2}{n}}\right) = 1-p$$

Interval spolehlivosti pro střední hodnotu

- pro odhad populační střední hodnoty tak dostáváme finální tvar intervalu spolehlivosti

$$P\left(\bar{X} - u_{1-p/2} \cdot \sqrt{\frac{\sigma^2}{n}} \leq \mu \leq \bar{X} + u_{1-p/2} \cdot \sqrt{\frac{\sigma^2}{n}}\right) = 1 - p$$

- lze odvodit i jednostranné varianty
  - levostranný interval spolehlivosti

$$P\left(\bar{X} - u_{1-p} \cdot \sqrt{\frac{\sigma^2}{n}} \leq \mu\right) = 1 - p$$

  - pravostranný interval spolehlivosti

$$P\left(\mu \leq \bar{X} + u_{1-p} \cdot \sqrt{\frac{\sigma^2}{n}}\right) = 1 - p$$

Interval spolehlivosti pro střední hodnotu

- pokud bychom tedy z rozdělení se střední hodnotou μ a rozptylem σ^2 pořídili mnohokrát opakovaně výběr tak, že v prvním výběru bychom spočítali jeho průměr $\bar{x}$, leží populační střední hodnota μ v intervalu

$$\left\langle \bar{x} - u_{1-\alpha/2} \cdot \sqrt{\frac{\sigma^2}{n}};\ \bar{x} + u_{1-\alpha/2} \cdot \sqrt{\frac{\sigma^2}{n}} \right\rangle$$

  s pravděpodobností $1 - \alpha$

Interval spolehlivosti pro střední hodnotu

- kvantil standardního normálního rozdělení se někdy nahrazuje kvantilem Studentova t-rozdělení
  - je-li rozsah souboru $n < 30$ nebo
  - není-li známý populační rozptyl σ^2
- pak je

$$P\left(\bar{X} - t_{1-p/2}(n-1) \cdot \sqrt{\frac{s_x^2}{n}} \leq \mu \leq \bar{X} + t_{1-p/2}(n-1) \cdot \sqrt{\frac{s_x^2}{n}}\right) = 1 - p$$

- neznámý populační rozptyl σ^2 tak nahrazujeme výběrovým rozptylem $s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$, neboť $\hat{\sigma}^2 = s_x^2$

Interval spolehlivosti pro střední hodnotu

- pokud bychom tedy z rozdělení se střední hodnotou μ a rozptylem σ^2 pořídili mnohokrát opakovaně výběr tak, že v prvním výběru bychom spočítali jeho průměr $\bar{x}$ a vyběrový rozptyl s_x^2, leží populační střední hodnota μ v intervalu

$$\left\langle \bar{x} - t_{1-\alpha/2}(n-1) \cdot \sqrt{\frac{s_x^2}{n}};\ \bar{x} + t_{1-\alpha/2}(n-1) \cdot \sqrt{\frac{s_x^2}{n}} \right\rangle$$

  s pravděpodobností $1 - \alpha$

Příklad

- V zásilce 100 kusů součástek byl vždy změřen dlouhý rozměr součástky. Průměrná hodnota dlouhého rozměru součástky byla 156 cm, směrodatná odchylka dlouhého rozměru součástky byla 2 cm. Určeme 95 % interval spolehlivosti pro průměr dlouhého rozměru součástky.

Odhad minimálního nutného rozsahu výběru

- označme symbolem Δ výraz $u_{1-\alpha/2} \cdot \sqrt{\frac{\sigma^2}{n}}$, tedy $\Delta \equiv u_{1-\alpha/2} \cdot \sqrt{\frac{\sigma^2}{n}}$
- snadno nahlédneme, že délka intervalu spolehlivosti je 2Δ
- požadujeme-li, aby byla poloviční délka intervalu spolehlivosti (též zvaná *přípustná chyba odhadu*) maximálně Δ, musí platit

$$u_{1-\alpha/2} \cdot \sqrt{\frac{\sigma^2}{n}} \leq \Delta \qquad \text{anebo} \qquad t_{1-\alpha/2}(n-1) \cdot \sqrt{\frac{\sigma^2}{n}} \leq \Delta,$$

odkud odvodíme, že pak je minimální nutný rozsah výběru k zajištění takové délky alespoň

$$n \geq \left\lceil \frac{u^2_{1-\alpha/2} \cdot \sigma^2}{\Delta^2} \right\rceil \qquad \text{anebo} \qquad n \geq \left\lceil \frac{t^2_{1-\alpha/2}(n-1) \cdot \sigma^2}{\Delta^2} \right\rceil,$$

kde symbol $\lceil x \rceil$ vždy značí horní celou část čísla $x \in \mathbb{R}$, tedy nejnižší celé číslo alespoň rovné x

Interval spolehlivosti pro relativní četnost

- předpokládejme, že výběr $\boldsymbol{x} = (x_1, x_2, \dots, x_n)^T$ je konkrétní realizací vektoru náhodných veličin $(X_1, X_2, \dots, X_n)^T$ a že $\forall i \in \{1, 2, \dots, n\}$ je $X_i \sim \mathrm{alt}(\pi)$
- pak pro odhad střední hodnoty výběrového průměru, který značme p, platí

$$p \equiv \mathbb{E}(\bar{X}) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i) = \frac{1}{n}\cdot n\pi = \pi$$

- a pro odhad rozptylu výběrového průměru platí

$$\mathrm{var}(\bar{X}) = \mathrm{var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{var}(X_i) = \frac{1}{n^2}\cdot n\cdot\pi(1-\pi) = \frac{\pi(1-\pi)}{n}$$

- zřejmě tedy variabilita výběrového průměru klesá s rostoucím rozsahem výběru n

Interval spolehlivosti pro relativní četnost

- pro výběrový průměr tedy platí

$$\mathbb{E}(\bar{X}) = p$$

$$\operatorname{var}(\bar{X}) = \frac{\pi(1-\pi)}{n}$$

- ať je $U \equiv \frac{\bar{X}-\pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} = \frac{p-\pi}{\sqrt{\frac{\pi(1-\pi)}{n}}}$, pak pro u_q-tý kvantil normálního rozdělení je

$$P\left(\frac{p-\pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \leq u_q\right) = 1-q$$

- je dále i

$$P\left(\frac{p-\pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \leq u_{1-q}\right) = 1-q$$

Interval spolehlivosti pro relativní četnost

- pro odhad populační relativní četnosti lze nakonec odvodit (jak bylo podrobně ukázáno pro střední hodnotu) s využitím faktu, že $p \equiv \mathbb{E}(\bar{X}) = \pi$, tedy $p = \hat{\pi}$, finální tvar intervalu spolehlivosti

$$P\left(p - u_{1-q/2} \cdot \sqrt{\frac{p(1-p)}{n}} \leq \pi \leq p + u_{1-q/2} \cdot \sqrt{\frac{p(1-p)}{n}}\right) = 1-q$$

- lze odvodit i jednostranné varianty
  - levostranný interval spolehlivosti

$$P\left(p - u_{1-q} \cdot \sqrt{\frac{p(1-p)}{n}} \leq \pi\right) = 1-q$$

  - pravostranný interval spolehlivosti

$$P\left(\pi \leq p + u_{1-q} \cdot \sqrt{\frac{p(1-p)}{n}}\right) = 1-q$$

Interval spolehlivosti pro relativní četnost

- pokud bychom tedy ze základního souboru s (populační) relativní četností π znaku našeho zájmu pořídili mnohokrát opakovaně výběr tak, že v prvním výběru bychom odhadli populační relativní četnost pomocí výběrové relativní četnosti p, leží populační relativní četnost π v intervalu

$$\left\langle p - u_{1-\alpha/2} \cdot \sqrt{\frac{p(1-p)}{n}};\ p + u_{1-\alpha/2} \cdot \sqrt{\frac{p(1-p)}{n}} \right\rangle$$

s pravděpodobností $1 - \alpha$

Odhad minimálního nutného rozsahu výběru

- označme symbolem Δ výraz $u_{1-\alpha/2} \cdot \sqrt{\frac{p(1-p)}{n}}$, tedy
  $\Delta \equiv u_{1-\alpha/2} \cdot \sqrt{\frac{p(1-p)}{n}}$
- požadujeme-li, aby byla poloviční délka intervalu spolehlivosti (též zvaná *přípustná chyba odhadu*) maximálně Δ, musí platit

$$u_{1-\alpha/2} \cdot \sqrt{\frac{p(1-p)}{n}} \leq \Delta,$$

odkud odvodíme, že pak je minimální nutný rozsah výběru k zajištění takové délky alespoň

$$n \geq \left\lceil \frac{u_{1-\alpha/2}^2 \cdot p(1-p)}{\Delta^2} \right\rceil,$$

kde symbol $\lceil x \rceil$ vždy značí horní celou část čísla $x \in \mathbb{R}$, tedy nejnižší celé číslo alespoň rovné x

Literatura

- Hindls, Richard, Stanislava Hronová, Jan Seger a Jakub Fischer. *Statistika pro ekonomy*. Praha: Professional Publishing, 2007. ISBN: 978-80-86946-43-6.
- Marek, Luboš. *Statistika v příkladech*. Praha: Professional Publishing, 2015. ISBN: 978-80-7431-153-6.

Děkuji za pozornost!

lubomir.stepanek@vse.cz
lubomir.stepanek@lf1.cuni.cz
lubomir.stepanek@fbmi.cvut.cz

https://github.com/LStepanek/4ST210_Statistika_pro_finance