Odhad nutného rozsahu výběru

B02907 Informační a komunikační technologie

Lubomír Štěpánek,
Ústav biofyziky a informatiky
1. LF UK

Upozornění!

- dole v poznámkách jsou u většiny snímků rozšiřující a vysvětlující komentáře
- u některých statistických metod budete odkazováni na statistické tabulky, které jsou volně přístupné online na adrese http://new.euromise.org/czech/tajne/ucebnice/html/html/node15.html
- (obvykle bude ještě na příslušném snímku odkaz zopakován; autor vynaložil značné úsilí, aby se symbolika v prezentacích shodovala se symbolikou v tabulkách, proto by neměla být orientace v tabulkách problémem)
- z předložených prezentací se můžete učit, můžete je kopírovat či jinak měnit, ale bez dovolení autora/autorů je nesmíte použít do svých publikací ☺
- předložené prezentace nejsou bezchybnou statistickou kuchařkou, proto ne zcela doporučuji se na ně ve svých pracích odkazovat, nebo je dokonce citovat ☺
- pokud se budu sám odkazovat na vhodnou literaturu, myslím tím nejspíše následující dvě knihy:
  - Zvára: Biostatistika. Karolinum, Praha 1988
  - Zvárová et al.: Biomedicínská statistika I. Základy statistiky pro biomedicínské obory
- dotazy a konzultace možné a vlastně i doporučeny

(Lubomír Štěpánek, stepanek.lub@seznam.cz)

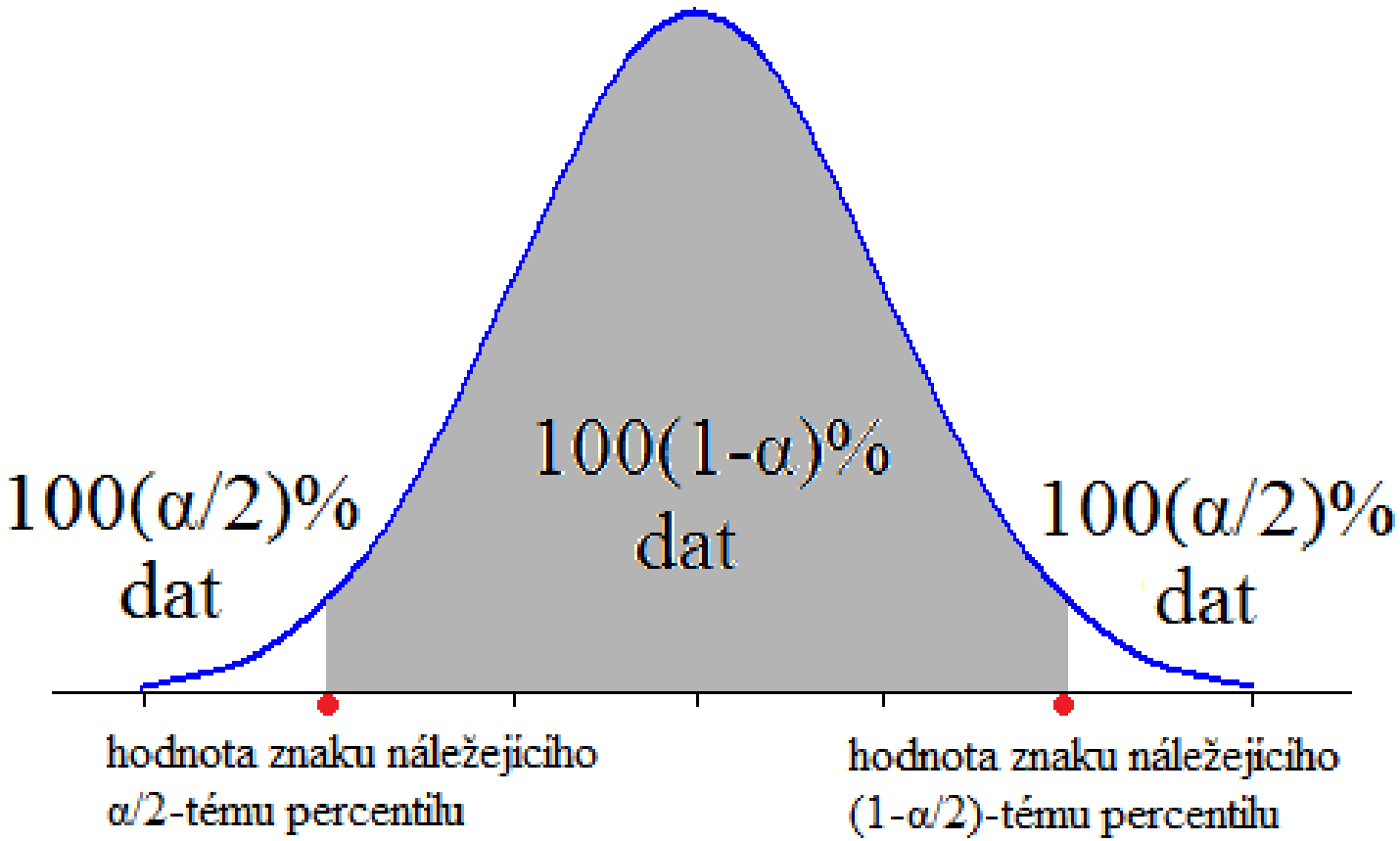
Intervaly spolehlivosti – opakování

Co známe:				Co chceme zjistit:
výběrová míra polohy	výběrová míra variability (spočítáme)	populační míra variability (známe)	rozsah souboru (známe)	interval, ve kterém leží populační míra polohy s pravděpodobností 1-α
průměr $\bar{x}$		odchylka σ	n	$\mu \in \left\langle \bar{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}; \bar{x} + z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \right\rangle$
průměr $\bar{x}$	odchylka s		$n > 31$	$\mu \in \left\langle \bar{x} - z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}; \bar{x} + z_{1-\alpha/2} \cdot \frac{s}{\sqrt{n}} \right\rangle$
průměr $\bar{x}$	odchylka s		$n \leq 30$	$\mu \in \left\langle \bar{x} - t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}; \bar{x} + t_{1-\alpha/2} \cdot \frac{s}{\sqrt{n}} \right\rangle$
četnost p		odchylka σ_p	n	$\pi \in \left\langle p - z_{\alpha/2} \cdot \sigma_p; p + z_{1-\alpha/2} \cdot \sigma_p \right\rangle$
četnost p	odchylka $\sqrt{p(1-p)/n}$		n	$\pi \in \left\langle p \pm z_{\alpha/2} \cdot \sqrt{\frac{p(1-p)}{n}} \right\rangle$

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

příslušné kvantily lze nalézt v tabulkách na:
http://new.euromise.org/czech/tajne/ucebnice/html/html/node15.html

Grafické znázornění



Odhady potřebného rozsahu výběru

- často s pomocí statistika je na začátku šetření určena *diference* (*d*)
- *diference* = maximální tolerovatelný rozdíl mezi skutečnou populační hodnotou (průměru, četnosti) a hodnotu odhadnutou z výběru na dané hladině spolehlivosti
- pomocí diference lze odhadnout minimální rozsah výběru
- někdy naopak chceme podle odhadnutého rozsahu výběrů určit, zda bude diference mezi nimi významná (tím se budeme zabývat až při testování hypotéz)

Minimální rozsah výběru pro odhad populačního průměru

- *„Kolik potřebujeme hodnot FVC plic astmatiků k odhadnutí průměrné hodnoty FVC všech pacientů s astma bronchiale, má-li se hodnota námi odhadnutého průměru FVC astmatiků od skutečné průměrné FVC lišit maximálně o 0,05 litru?“*

Minimální rozsah výběru pro odhad populačního průměru

- třeba určit diferenci d
- odhadnout populační směrodatnou odchylku (známe σ např. z jiné studie nebo odhadneme s z malého zkušebního výběru)
- zvolit hladinu spolehlivosti 1-α (obvykle $0{,}05$)

$$n_{\min} = \left\lceil \left(\frac{z_{1-\alpha/2} \cdot \sigma}{d} \right)^2 \right\rceil \approx \left\lceil \left(\frac{t_{1-\alpha/2} \cdot s}{d} \right)^2 \right\rceil$$

Minimální rozsah výběru pro odhad populačního průměru

- *„Kolik potřebujeme hodnot FVC plic astmatiků k odhadnutí průměrné hodnoty FVC všech pacientů s astma bronchiale, má-li se hodnota námi odhadnutého průměru FVC astmatiků od skutečné průměrné FVC lišit maximálně o 0,05 litru?“*
- $\sigma = 0{,}20\ \mathrm{l}$

$$n_{\min} = \left\lceil \left(\frac{z_{1-\alpha/2} \cdot \sigma}{d} \right)^2 \right\rceil$$

Minimální rozsah výběru pro odhad populačního průměru

- *„Kolik potřebujeme hodnot FVC plic astmatiků k odhadnutí průměrné hodnoty FVC všech pacientů s astma bronchiale, má-li se hodnota námi odhadnutého průměru FVC astmatiků od skutečné průměrné FVC lišit maximálně o 0,05 litru?“*
- $d = 0{,}05\ \mathrm{l};\ \sigma = 0{,}20\ \mathrm{l};\ \alpha = 0{,}05$

$$n_{\min} = \left\lceil \left(\frac{z_{1-\alpha/2} \cdot \sigma}{d} \right)^2 \right\rceil = \left\lceil \left(\frac{1{,}96 \cdot 0{,}20}{0{,}05} \right)^2 \right\rceil = \lceil 61{,}47 \rceil = 62$$

Minimální rozsah výběru pro odhad populační četnosti

- *„Kolik potřebujeme náhodně vybraných pacientů s obstrukčními plicními chorobami, chceme-li odhadnout četnost astmatiků mezi nimi tak, aby se náš odhad od skutečné hodnoty relativní populační četnosti astmatiků lišil maximálně o 3 %?“*

Minimální rozsah výběru pro odhad populační četnosti

- třeba určit diferenci d (rozdíl je maximálně d % z π)
- odhadnout populační relativní četnost π (např. z jiné studie)
- zvolit hladinu spolehlivosti α

$$n_{\min} = \left\lceil \left(\frac{z_{1-\alpha/2}}{d/100} \right)^2 \frac{\pi(1-\pi)}{\pi^2} \right\rceil$$

- pro hodnoty populační četnosti kolem $0{,}5$ je obecně potřeba nejpočetnější výběr (největší n)

Minimální rozsah výběru pro odhad populační četnosti

- *„Kolik potřebujeme náhodně vybraných pacientů s obstrukčními plicními chorobami, chceme-li odhadnout četnost astmatiků mezi nimi tak, aby se náš odhad od skutečné hodnoty relativní populační četnosti astmatiků lišil maximálně o 3 %?“*
- $\pi = 42\% = 0{,}42$

$$n_{\min} = \left\lceil \left(\frac{z_{1-\alpha/2}}{d/100} \right)^2 \frac{\pi(1-\pi)}{\pi^2} \right\rceil$$

Minimální rozsah výběru pro odhad populační četnosti

- *„Kolik potřebujeme náhodně vybraných pacientů s obstrukčními plicními chorobami, chceme-li odhadnout četnost astmatiků mezi nimi tak, aby se náš odhad od skutečné hodnoty relativní populační četnosti astmatiků lišil maximálně o 3 %?“*
- $d = 3;\ \pi = 42\% = 0{,}42;\ \alpha = 0{,}05$

$$n_{\min} = \left\lceil \left(\frac{1{,}96}{3/100} \right)^2 \frac{0{,}42(1-0{,}42)}{0{,}42^2} \right\rceil = \lceil 5894{,}5 \rceil = 5895$$

Velký soubor a jeho rizika

- velký soubor nezajistí nutně správně signifikantní výsledek!
- obecně v testech hypotéz pro velká *n* jsou „volnější“ kritické hodnoty → roste tedy chyba I. typu → lze přijmout i hypotézu, která ve skutečnosti neplatí

Malý soubor a jeho rizika

- omezené možnosti stratifikace, randomizace, výběry metod, kontrol signifikance
- pro malá *n* jsou kritické hodnoty testů hypotéz „přísné“→ roste možnost chyby II. typu → lze zavrhnout hypotézu, která ve skutečnosti platí
- malá přesnost odhadů

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