KGG/STG Statistika pro geografy

5. Odhady parametrů základního souboru

Mgr. David Fiedor
16. března 2015

Osnova

1. Úvod, pojmy
   - Vztahy mezi výběrovým a základním souborem
2. Bodový odhad
   - Bodový odhad parametrů základního souboru
   - Vlastnosti odhadů
   - Výběrové rozdělení
3. Intervalový odhad
   - Přesnost a spolehlivost odhadu
   - Interval spolehlivosti
   - Odhad parametru μ
   - Odhad parametru σ^2
   - Určení rozsahu n náhodného výběru

Opakování základních pojmů

- jak lze odhadnout z výběrového souboru charakteristiky souboru základního
- základní soubor
- výběrový soubor

Náhodný výběr

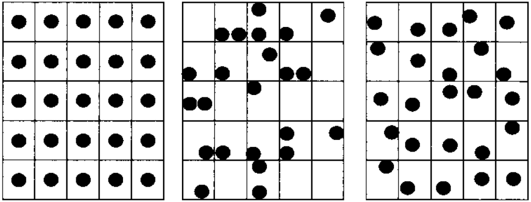
Definice

Náhodným výběrem rozumíme výběr, ve kterém má každý prvek základního souboru stejnou šanci být vybrán.

Prostorová specifikace:

- stratifikovaný - z každého okresu jeden prvek
- systematický - seřadit a každá pátá obec
- prostý náhodný

Grafické znázornění náhodných výběrů



Obrázek: Systematický, náhodný a stratifikovaný náhodný výběr

Statistiky = funkce odvozené z náhodného výběru

- výběrový aritmetický průměr

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

- výběrová směrodatná odchylka

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

Základní x výběrový soubor

Tabulka: Označení charakteristik základního a výběrového souboru

	Základní soubor	Výběrový soubor
rozsah	N	n
i-tý prvek	a_i	x_i
střední hodnota	μ	$\bar{x}$
směrodatná odchylka	σ	s
rozptyl	σ^2	s^2

Bodový a intervalový odhad

- ***bodový odhad*** - realizací tohoto odhadu získáme pro hledanou charakteristiku základního souboru konkrétní číslo
- ***intervalový odhad*** - realizací tohoto odhadu získáme pro hledanou charakteristiku základního souboru interval hodnot, který s předem danou pravděpodobností bude pokrývat hledanou charakteristiku

Bodový odhad

- slouží k odhadu parametrů základního souboru, ze kterého daný náhodný výběr pochází
- realizací tohoto odhadu získáme pro konkrétní charakteristiku určitou hodnotu
- pro různé náhodné výběry ze stejného základního souboru se bodový odhad téže charakteristiky zpravidla liší

Bodový odhad střední hodnoty

Definice

Bodovým odhadem střední hodnoty $\hat{\mu}$ základního souboru rozumíme aritmetický průměr daného náhodného výběru, tj.

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

- bodový odhad střední hodnoty základního souboru ztotožňujeme z výběrovým průměrem

Bodový odhad směrodatné odchylky

Definice

Bodový odhad směrodatné odchylky $\hat{\sigma}$ základního souboru určujeme pomocí kvadrátů odchylek jednotlivých prvků od výběrového průměru, tj.

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

(v případě, že rozsah výběru $n \geq 2$).

Bodový odhad směrodatné odchylky

- bodový odhad směrodatné odchylky základního souboru ztotožňujeme z výběrovou směrodatnou odchylkou
- pozor: v popisné statistice jsme počítali hodnotu směrodatné odchylky jinak (odlišíme čárkou):

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

Stupně volnosti

- udávají, jaký je potřeba minimální počet nezávislých výběrů k tomu, abychom dostali výslednou statistiku
- kvůli položení rovnosti mezi bodový odhad a výběrovou charakteristiku „ztrácíme“ jeden nezávislý výběr - proto $n-1$
- například pokud průměr vypočtený ze tří měření je 5 a první dvě měření budou 4 a 5, poslední měření již není nezávislé, neboť se musí rovnat 6

Vlastnosti odhadů

- vždy platí $\hat{\sigma} > \acute{s}$

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

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

- neboli

$$\hat{\sigma} = \acute{s} \cdot \sqrt{\frac{n}{n-1}}$$

Vlastnosti odhadů

1. *konzistentnost odhadu* – označuje trend zmenšování se rozdílu mezi skutečnou a odhadnutou hodnotou charakteristiky s rostoucím rozsahem výběru; tj. s rostoucím rozsahem výběru se realizace odhadu dané charakteristiky přibližuje hodnotě dané charakteristiky základního souboru (s pravděpodobností 1)
2. *nestrannost odhadu* – při opakovaných výběrech kolísá odhad kolem teoretické hodnoty symetricky na obě strany

Vlastnosti odhadů

3. *vydatnost odhadu* – rozptyl odhadů při opakovaných výběrech je malý
4. *rezistence odhadu* – odlehlé hodnoty vzniklé chybou měření, či špatným zápisem, nemají velký vliv na hodnotu odhadu

Výběrové rozdělení

- pravděpodobnostní rozdělení hodnot, které statistika nabývá ve všech možných výběrech z daného základního souboru
- následující statistiky jsou definovány jako vážené - z důvodu různých rozsahů výběrů

Statistiky výběrového rozdělení

1. *Vážený průměr výběrových průměrů*

$$\begin{aligned}\mu_{\bar{x}} &= (n_1 \cdot \bar{x}_1 + n_2 \cdot \bar{x}_2 + \cdots + n_{r-1} \cdot \bar{x}_{r-1} + n_r \cdot \bar{x}_r)/r \\ &= \frac{1}{n} \textstyle\sum_{i=1}^{r} n_i \cdot \bar{x}_i\end{aligned}$$

2. *Vážený průměr výběrových rozptylů*

$$\begin{aligned}s_*^2 &= \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \cdots + (n_r - 1)s_r^2}{n_1 + n_2 + \cdots + n_r - r} \\ &= \frac{\sum_{i=1}^{r}(n_i - 1)s_i^2}{n - r},\end{aligned}$$

kde $n = n_1 + \cdots + n_r$ a r je počet výběrů.

Vlastnosti parametrů

1. v případě náhodného výběru z normálního rozdělení s parametry μ a σ platí, že rozdělení výběrových průměrů je také normální s parametry

$$\mu_{\bar{x}} = \mu$$

$$\sigma_{\bar{x}} = \sigma / \sqrt{n}$$

2. směrodatná odchylka rozdělení výběrových průměrů $\sigma_{\bar{x}}$ (také používáme název *směrodatná chyba odhadu průměru*) je menší než směrodatná odchylka základního souboru a to tím menší, čím je větší rozsah výběru

Vlastnosti parametrů

3. pro rozsah výběru jdoucí do nekonečna se výběrové rozdělení průměrů dokonce blíží normálnímu rozdělení bez ohledu na tvar původního rozdělení

Intervalový odhad

- realizací tohoto odhadu získáme pro konkrétní charakteristiku interval hodnot, v němž se zadanou pravděpodobností leží odhadovaný neznámý parametr
- liší se podle rozsahu výběru
- záleží na tom, které parametry známe a které nikoliv

Kvalita odhadu

Přesnost odhadu je dána násobkem směrodatné chyby odhadu (tj. poloviční šířkou intervalu spolehlivosti).

Spolehlivost odhadu je určena pravděpodobností, se kterou je možné určitý odhad považovat za správný.

Interval spolehlivosti

Princip konstrukce pro střední hodnotu:

- středem intervalu spolehlivosti je bodový odhad dané charakteristiky (tj. výběrový průměr)
- na obě strany od výběrového průměru ve vzdálenosti určené násobkem směrodatné chyby odhadu - hranice intervalu spolehlivosti
- označení: nalevo dolní hranice (q_1), napravo horní hranice (q_2)
- rozdíl horní a dolní hranice určuje šířku intervalu spolehlivosti
- polovina šířky je mírou přesnosti odhadu

Základní pojmy

- míra spolehlivosti odhadu $(1 - \alpha)$ (pravděpodobnost, že se skutečný parametr základního souboru nachází uvnitř intervalu)
- riziko α - pravděpodobnost, že skutečný parametr není pokryt intervalem spolehlivosti

Tabulka: Nejčastěji používané intervaly spolehlivosti

Násobky *s*	Spolehlivost	Riziko
1,960	0,95	0,05
2,576	0,99	0,01
3,291	0,999	0,001

Charakteristiky

- výběry s velkým rozsahem jsou přesnější
- výběry s rozsahem nad 30 - jejich výběrové rozdělení lze považovat za normální
- navíc lze ve vzorci pro intervalový odhad dosadit místo σ^2 výběrový rozptyl s^2
- platí

$$P(q_1 \leq \mu \leq q_2) = 1 - \alpha$$

- rozdělení na případy v závislosti na tom, zda známe rozptyl σ^2, nebo nikoliv

Odhad parametru μ při známém rozptylu σ^2

- pro hranice intervalu spolehlivosti platí:

$$q_1 = \bar{x} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}},$$

$$q_2 = \bar{x} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}},$$

  přičemž $z_{1-\frac{\alpha}{2}}$ je příslušný kvantil normovaného normálního rozdělení
- intervalový odhad lze přepsat do tvaru:

$$\bar{x} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

Odhad parametru μ při neznámém rozptylu σ^2

- nutnost nahradit hodnotu jistého kvantilu normovaného normálního rozdělení kvantilem *Studentova t rozdělení* pro $\nu = n - 1$ stupňů volnosti

$$\bar{x} - t_{1-\frac{\alpha}{2}}(n-1)\frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{1-\frac{\alpha}{2}}(n-1)\frac{s}{\sqrt{n}}$$

Přípustná chyba Δ

$$\Delta = z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

S tímto novým označením tak můžeme přepsat obecný tvar intervalového odhadu do podoby:

$$\bar{x} - \Delta \leq \mu \leq \bar{x} + \Delta$$

Odhad parametru σ^2

- obecně bude mít intervalový odhad parametru σ^2 tvar:

$$P(q_1 \leq \sigma^2 \leq q_2) = 1 - \alpha$$

- omezení se na variantu, kdy neznáme parametr μ
- výsledný interval má tvar

$$\frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2}}(n-1)} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{\frac{1-\alpha}{2}}(n-1)}$$

Určení rozsahu *n* náhodného výběru

- situace: potřebujeme předem vědět, jak velký soubor vybrat (předem si zvolit přesnost intervalového odhadu)
- úpravou vzorce přípustné chyby dostáváme vztah

$$n \geq \left(\frac{\sigma \cdot z_{1-\frac{\alpha}{2}}}{\Delta} \right)^2,$$

kde za *n* zvolíme nejmenší přirozené číslo vyhovující této podmínce

Děkuji za pozornost...