

(Једноетапни) групни узорак

Примена групног узорковања готово увек резултира мањом прецизношћу оцена непознатих популацијских параметара, у односу на оцене које се добијају на основу простог случајног узорка без понављања са истим бројем ентитета. У наставку се врши поређење прецизности простог случајног узорка без понављања и групног узорка.

Из популације подељене на N примарних јединица (група) бира се (прост) случајан узорак без понављања група, обима n . Нека је

$$\bar{M} := \frac{M}{N}$$

просечна величина (у смислу броја секундарних јединица (ентитета)) групе. Очекивани број ентитета у простом случајном узорку без понављања група обима n , према томе, једнак је: $n\bar{M}$.

Мера добитка / губитка прецизности, до кога долази услед коришћења сложенијих метода одабира узорка уместо простог случајног узорка без понављања, дата је тзв. **ефектом дизајна DEFF** ('design effect'), дефинисаним са:

$$\text{DEFF}(\text{метод, статистика}) := \frac{D(\text{оцена добијена на основу узорка одабраног применом метода})}{D(\text{оцена добијена на основу SRSWOR узорка јединица посматрања истог обима})}$$

Ради једноставности, у наставку претпоставимо да је свака од N група сачињена од подједнаког броја ентитета, тј. $M_l = \bar{M}$, $l = \bar{1}, \bar{N}$. Овде је, нпр. за непристрасну оцену на основу групног узорка, ефекат дизајна дат са:

$$\text{DEFF}(clu, \hat{\tau}_Y^u) = \frac{\sigma_\tau^2}{\bar{M}\sigma_Y^2},$$

па је, дакле, групни узорак обима $n\bar{M}$ ентитета ефикаснији од простог случајног узорка без понављања истог обима ако је $\text{DEFF}(clu, \hat{\tau}_Y^u) < 1$, тј. $\sigma_\tau^2 < \bar{M}\sigma_Y^2$.

Стандардан алгебарски идентитет који се користи за анализу дисперзије (ANOVA) популације подељене на N примарних јединица изведен је у наставку:

$$\begin{aligned} (N\bar{M} - 1)\sigma_Y^2 &= \sum_{l=1}^N \sum_{k=1}^{\bar{M}} (y_{lk} - m_Y)^2 = \sum_{l=1}^N \sum_{k=1}^{\bar{M}} ((y_{lk} - m_l) - (m_Y - m_l))^2 \\ &= \underbrace{\sum_{l=1}^N \sum_{k=1}^{\bar{M}} (y_{lk} - m_l)^2}_{\text{SSW}} + \underbrace{\bar{M} \sum_{l=1}^N (m_Y - m_l)^2}_{\text{SSB}}. \end{aligned} \quad (1)$$

Уводе се и додатне ознаке:

$$\begin{aligned} \text{MSW} &:= \frac{1}{N(\bar{M} - 1)} \sum_{l=1}^N \sum_{k=1}^{\bar{M}} (y_{lk} - m_l)^2 = \frac{\text{SSW}}{N(\bar{M} - 1)} \\ \text{MSB} &:= \frac{\bar{M}}{N - 1} \sum_{l=1}^N (m_Y - m_l)^2 = \frac{\text{SSB}}{N - 1} \end{aligned}$$

Дакле, потпуно супротно него код стратификованог случајног узорка код кога је дисперзија $D\hat{\tau}_Y^{str}$ оцене непознатог популацијског тотала зависила искључиво од варијабилности посматраног обележја Y **унутар** стратума, дисперзија $D\hat{\tau}_Y^u$ непристрасне оцене непознатог популацијског тотала на основу групног узорка зависи искључиво од дела варијабилности који се односи на варијабилност **између** примарних јединица, јер је:

$$\sigma_\tau^2 = \bar{M} \cdot \text{MSB},$$

тј.

$$D\hat{\tau}_Y^u = \frac{N^2}{n} \left(1 - \frac{n}{N}\right) \bar{M} \cdot \text{MSB}. \quad (2)$$

Узимајући у обзир све наведено, групни узорак обима $n\bar{M}$ ентитета ефикаснији је од простог случајног узорка без понављања истог обима ако је:

$$\text{MSB} < \sigma_Y^2.$$

Алтернативни начин за анализирање дисперзија оцена добијених на основу групног узорка могућ је записивањем формула за дисперзије преко коефицијента корелације између ентитета који сачињавају исту групу. **Унутаргрупни коефицијент корелације ICC** ('intraclass correlation coefficient') дефинисан је са

$$ICC := \frac{\sum_{l=1}^N \sum_{k=1}^{\overline{M}} \sum_{\substack{j=1 \\ j \neq k}}^{\overline{M}} (y_{lk} - m_Y)(y_{lj} - m_Y)}{(N\overline{M} - 1)(\overline{M} - 1)\sigma_Y^2} \quad (3)$$

и представља меру хомогености ентитета унутар исте групе указујући на то колико су ентитети из исте групе међусобно слични.

Може се искористити следеће:

$$\begin{aligned} (N-1)\sigma_\tau^2 &= \sum_{l=1}^N (\tau_l - \mu_Y)^2 = \sum_{l=1}^N (\tau_l - \overline{M} \cdot m_Y)^2 = \sum_{l=1}^N \left(\sum_{k=1}^{\overline{M}} (y_{lk} - m_Y) \right)^2 \\ &= \sum_{l=1}^N \sum_{k=1}^{\overline{M}} (y_{lk} - m_Y)^2 + \sum_{l=1}^N \sum_{k=1}^{\overline{M}} \sum_{\substack{j=1 \\ j \neq k}}^{\overline{M}} (y_{lk} - m_Y)(y_{lj} - m_Y) \\ &= (N\overline{M} - 1)\sigma_Y^2 + (N\overline{M} - 1)(\overline{M} - 1)\sigma_Y^2 \cdot ICC = (N\overline{M} - 1)\sigma_Y^2(1 + (\overline{M} - 1)ICC), \end{aligned}$$

па се добија:

$$ICC = 1 - \frac{\overline{M}}{\overline{M} - 1} \cdot \frac{SSW}{(N\overline{M} - 1)\sigma_Y^2},$$

одакле следе неједнакости:

$$-\frac{1}{\overline{M} - 1} \leq ICC \leq 1.$$

Такође, важи и једнакост:

$$MSB = \frac{M-1}{\overline{M}(N-1)}\sigma_Y^2(1 + (\overline{M} - 1)ICC)$$

на основу које се може анализирати $D\hat{\tau}_Y^u$ дата у (2).

Ако је број примарних јединица N довољно велики, тако да важи $N\overline{M} - 1 \approx \overline{M}(N - 1)$, приближна вредност ефекта дизајна је овде $1 + (\overline{M} - 1)ICC$, па ако је:

- $ICC = 0$ – дисперзије оцена на основу групног узорка приближно су једнаке дисперзијама оцена на основу простог случајног узорка без понављања истог обима ($n\overline{M}$)
- $ICC > 0$ – прост случајан узорак без понављања даје прецизније оцене
- $ICC < 0$ – групни узорак даје прецизније оцене

Унутаргрупни коефицијент корелације ICC дефинисан је само када су све групе сачињене од подједнаког броја ентитета. Аналогон ове мере хомогености ентитета унутар исте групе у општим популацијама је вредност ρ_a^2 дефинисана са:

$$\rho_a^2 = 1 - \frac{MSW}{\sigma_Y^2}.$$

Систематски узорак

Из популације обима $N = n \cdot K$, $n, K \in \mathbb{N}$, на случајан начин бира се K -периодичан узорак. Теоријска дисперзија оцене $\hat{m}_Y^{sys}$ непознате популацијске средње вредности m_Y посматраног обележја Y дата је формулом:

$$D\hat{m}_Y^{sys} = \frac{1}{n} \left(1 - \frac{1}{K} \right) \cdot MSB \approx \frac{1 + (n-1)ICC}{n} \cdot \sigma_Y^2.$$