

Стратификован случајан узорак

Мора се приметити да адекватна примена стратификације готово увек резултира мањим дисперзијама оцена непознатих популацијских параметара, у односу на оне које се добијају за упоредиви прост случајан узорак без понављања. У наставку се врши поређење прецизности простог случајног узорка без понављања и стратификованог случајног узорка.

Стандардан алгебарски идентитет који се користи за анализу дисперзије (**AN**alysis **Of** **VA**riance) стратификоване популације изведен је у наставку:

$$\begin{aligned}
 (N-1)\sigma_Y^2 &= \sum_{h=1}^L \sum_{k=1}^{N_h} (y_{hk} - m_Y)^2 = \sum_{h=1}^L \sum_{k=1}^{N_h} ((y_{hk} - m_h) - (m_Y - m_h))^2 \\
 &= \sum_{h=1}^L \sum_{k=1}^{N_h} (y_{hk} - m_h)^2 - 2 \sum_{h=1}^L (m_Y - m_h) \sum_{k=1}^{N_h} (y_{hk} - m_h) + \sum_{h=1}^L N_h (m_Y - m_h)^2 \\
 &= \underbrace{\sum_{h=1}^L (N_h - 1)\sigma_h^2}_{SSW} + \underbrace{\sum_{h=1}^L N_h (m_Y - m_h)^2}_{SSB}.
 \end{aligned} \tag{1}$$

Дакле, збир квадрата одступања вредности обележја Y за сваку јединицу у популацији од популацијске средње вредности m_Y може се разложити на две компоненте: SSW – збир квадрата одступања унутар стратума и SSB – збир квадрата одступања између стратума.

Поређење ће бити извршено за оцену популацијске средине, па се уводе следеће ознаке за дисперзије оцена:

$$D_{SRSWOR} := \frac{\sigma_Y^2}{n} \left(1 - \frac{n}{N}\right) \quad (\text{прост сл. узорак})$$

$$D_{STR} := \frac{1}{N^2} \sum_{h=1}^L \frac{N_h^2}{n_h} \sigma_h^2 \left(1 - \frac{n_h}{N_h}\right) \quad (\text{стратификован сл. узорак})$$

У зависности од технике распоређивања обима стратификованог сл. узорка појављују се и специјалне ознаке: $D_{PROP} := D_{STR}$ – код пропорционалног распореда, $D_{OPT} := D_{STR}$ – код Неуман-овог оптималног распореда.

Код пропорционалног распореда је $n_h = n \cdot \frac{N_h}{N}$, па је:

$$D_{PROP} = \frac{1}{N^2} \sum_{h=1}^L \frac{N}{n} \cdot N_h \sigma_h^2 \left(1 - \frac{n}{N}\right) = \frac{1}{nN} \left(1 - \frac{n}{N}\right) \sum_{h=1}^L N_h \sigma_h^2, \tag{2}$$

па је, на основу (1):

$$\begin{aligned}
 D_{SRSWOR} &= \frac{1}{N-1} \left(\sum_{h=1}^L (N_h - 1)\sigma_h^2 + \sum_{h=1}^L N_h (m_Y - m_h)^2 \right) \cdot \frac{1}{n} \left(1 - \frac{n}{N}\right) \\
 &= \frac{1}{N-1} \left(\frac{N-1}{N} \sum_{h=1}^L N_h \sigma_h^2 + \sum_{h=1}^L \left(\frac{N_h}{N} - 1\right) \sigma_h^2 + \sum_{h=1}^L N_h (m_Y - m_h)^2 \right) \cdot \frac{1}{n} \left(1 - \frac{n}{N}\right) \\
 &= \frac{1}{N-1} \left(\frac{N-1}{N} \cdot \frac{nN}{\left(1 - \frac{n}{N}\right)} \cdot D_{PROP} + \sum_{h=1}^L \left(\frac{N_h}{N} - 1\right) \sigma_h^2 + \sum_{h=1}^L N_h (m_Y - m_h)^2 \right) \cdot \frac{1}{n} \left(1 - \frac{n}{N}\right) \\
 &= D_{PROP} + \frac{1}{N-1} \cdot \frac{1}{n} \left(1 - \frac{n}{N}\right) \left(SSB - \sum_{h=1}^L \sigma_h^2 \left(1 - \frac{N_h}{N}\right) \right)
 \end{aligned} \tag{3}$$

Код Неуман-овог оптималног распореда је $n_h = n \cdot \frac{N_h \sigma_h}{\sum_{l=1}^L N_l \sigma_l}$, па је:

$$D_{OPT} = \frac{1}{N^2} \sum_{h=1}^L \frac{N_h}{n} \sigma_h \sum_{l=1}^L N_l \sigma_l \left(1 - \frac{n \sigma_h}{\sum_{l=1}^L N_l \sigma_l}\right) = \frac{1}{nN^2} \left(\sum_{h=1}^L N_h \sigma_h \right)^2 - \frac{1}{N^2} \sum_{h=1}^L N_h \sigma_h^2 \tag{4}$$

На основу (2) је:

$$\begin{aligned}
D_{PROP} &= \frac{1}{nN} \left(1 - \frac{n}{N}\right) \sum_{h=1}^L N_h \sigma_h^2 = \frac{1}{nN} \sum_{h=1}^L N_h \sigma_h^2 - \frac{1}{N^2} \sum_{h=1}^L N_h \sigma_h^2 \\
&= \frac{1}{nN} \sum_{h=1}^L N_h \sigma_h^2 - \frac{1}{nN^2} \left(\sum_{h=1}^L N_h \sigma_h \right)^2 + D_{OPT} \\
&= D_{OPT} + \frac{1}{n} \sum_{h=1}^L \frac{N_h}{N} \left(\sigma_h - \sum_{l=1}^L \frac{N_l}{N} \sigma_l \right)^2
\end{aligned} \tag{5}$$

Користећи једнакости (3) и (5) коначно се добија:

$$D_{SRSWOR} = D_{OPT} + \frac{1}{nN} \sum_{h=1}^L N_h \sigma_h \left(\sigma_h - \sum_{l=1}^L \frac{N_l}{N} \sigma_l \right) + \frac{1}{N-1} \cdot \frac{1}{n} \left(1 - \frac{n}{N}\right) \left(\text{SSB} - \sum_{h=1}^L \sigma_h^2 \left(1 - \frac{N_h}{N}\right) \right) \tag{6}$$

Под претпоставком да је број јединица N_h у h -том стратуму велики за $h = \overline{1, L}$ (обим популације N је тада, такође, велики) а да је обим узорка n_h из h -тог стратума релативно мали, у смислу $\frac{n_h}{N_h} \approx 0$ (могу се занемарити фактори корекције због коначности стратума $1 - \frac{n_h}{N_h}$, као и фактор корекције због коначности популације $1 - \frac{n}{N}$), поређење дисперзија оцена резимирано је следећим низом неједнакости:

$$D_{OPT} \leq D_{PROP} \leq D_{SRSWOR},$$

што заиста оправдава примену стратификације.

Наиме, ту је:

$$\begin{aligned}
D_{SRSWOR} &= \frac{\sigma_Y^2}{n} \approx \frac{1}{n} \left(\sum_{h=1}^L \frac{N_h}{N} \sigma_h^2 + \sum_{h=1}^L \frac{N_h}{N} (m_h - m_Y)^2 \right) \\
&\approx D_{PROP} + \frac{1}{n} \sum_{h=1}^L \frac{N_h}{N} (m_h - m_Y)^2 \\
&= D_{OPT} + \frac{1}{n} \sum_{h=1}^L \frac{N_h}{N} \left(\sigma_h - \sum_{l=1}^L \frac{N_l}{N} \sigma_l \right)^2 + \frac{1}{n} \sum_{h=1}^L \frac{N_h}{N} (m_h - m_Y)^2
\end{aligned}$$

Последњи израз доводи до закључка да се, преласком са простог случајног узрока на стратификован случајан узорак са оптималним распоредом за познат и фиксиран обим целог узорка n , може значајно добити на прецизности оцена уколико се стратуми формирају тако да међусобна одступања средњих вредности, односно стандардних девијација (тј. дисперзија), по стратумима буду "велика".