(Прост) случајан узорак

Из популације обима N бира се (прост) случајан узорак обима n. Нека је Y обележје од интереса, а y_k вредност обележја Y ентитета означеног са k. Када је приступ приликом закључивања заснован на методу одабира узорка сматра се да су $y_1, y_2, \ldots, y_N$ фиксиране, али непознате вредности.
Обележје Y је сл. величина $Y : \Omega \to \mathbb{R}$ са дискретном расподелом вероватноћа датом законом расподеле:

$$Y : \begin{pmatrix} \zeta_1 & \zeta_2 & \cdots & \zeta_m \\ \frac{N_1}{N} & \frac{N_2}{N} & \cdots & \frac{N_m}{N} \end{pmatrix}, \tag{1}$$

где су $\zeta_1, \zeta_2, \ldots, \zeta_m$ различите вредности из скупа $\{y_1, y_2, \ldots, y_N\}$, а N_k је број ентитета чија је вредност посматраног обележја једнака ζ_k, $k = \overline{1,m}$ $(m \leq N)$. Популацијска средња вредност, односно популацијска дисперзија, заправо су EY, односно DY.

Претпоставимо да се одабир ентитета у узорак врши сукцесивно – кроз n корака, са једнаком вероватноћом извлачења било које расположиве јединице у тренутном кораку, и то без враћања одабране јединице, односно са враћањем. Нека је са Y_k означена сл. величина која представља вредност обележја Y ентитета одабраног у k-том кораку. Без обзира на то да ли је (прост) случајан узорак без понављања или са понављањем, сл. величине $Y_1, Y_2, \ldots, Y_n$ имају исту расподелу вероватноћа као обележје Y (дата у (1)). Код узорка са понављањем ове сл. величине су и независне (свако извлачење је независно од осталих и увек се врши из комплетне популације), док су код узорка без понављања ове сл. величине зависне.

Надаље говоримо о (неуређеном) случајном узорку $\mathcal{S}$ без понављања обима n.

Лема 1. Узорачка средња вредност је непристрасна оцена непознате популацијске средње вредности. Дисперзија ове оцене једнака је:

$$\frac{\sigma^2}{n}\left(1 - \frac{n}{N}\right),$$

где је σ^2 непозната популацијска дисперзија.

Доказ.

Узорачка средина је сл. величина дата са:

$$\widehat{m}_Y := \frac{1}{n}\sum_{k \in \mathcal{S}} y_k,$$

а за реализован узорак s она узима конкретну реалну вредност:

$$\widehat{m}_Y(s) := \frac{1}{n}\sum_{k \in s} y_k.$$

Њено математичко очекивање рачуна се на основу њене узорачке расподеле, а суштински као тежинска средина њених реализација по свим могућим узорцима s обима n, са тежинама једнаким вероватноћама одабира конкретних узорака:

$$E\widehat{m}_Y = \sum_s \widehat{m}_Y(s)p(s) = \frac{1}{n}\sum_{k=1}^{N} y_k \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{1}{N}\sum_{k=1}^{N} y_k = m_Y.$$

Аналогно би се рачунала дисперзија оцене $\widehat{m}_Y$.

Алтернативни начин подразумевао би примену тзв. рандомизације, што је врло корисна идеја код сложенијих планова узорковања. Раније су уведени индикатори укључења I_k, дефинисани са:

$$I_k = \begin{cases} 1, & \text{ако је јединица означена са } k \text{ одабрана у узорак} \\ 0, & \text{иначе} \end{cases}.$$

То су сл. величине са законом расподеле:

$$I_k : \begin{pmatrix} 0 & 1 \\ 1 - \pi_k & \pi_k \end{pmatrix},$$

где је π_k одговарајућа вероватноћа укључења првог реда.

I_k је статистика, а π_k није ништа друго до:

$$\pi_k = \sum_{s \ni k} p(s)$$

где се сумирање врши по свим могућим узорцима s обима n који садрже јединицу означену са k.

Тада се узорачка средина може записати на следећи начин:

$$\widehat{m}_Y := \frac{1}{n} \sum_{k \in \mathcal{S}} y_k = \frac{1}{n} \sum_{k=1}^{N} I_k y_k \tag{2}$$

Приметити да су, при наведеним претпоставкама, у изразу (2) само I_k-ови сл. величине.

Већ је показано да је:

$$\pi_k = \frac{n}{N}, \quad \text{за свако } k = \overline{1, N},$$

па су основне нумеричке карактеристике сл. величине I_k дате формулама:

$$EI_k = \frac{n}{N}, \quad DI_k = \frac{n}{N}\left(1 - \frac{n}{N}\right).$$

Према томе је:

$$E\widehat{m}_Y = E\left(\frac{1}{n}\sum_{k=1}^{N} I_k y_k\right) = \frac{1}{n}\sum_{k=1}^{N} y_k E(I_k) = \frac{1}{n}\sum_{k=1}^{N} y_k \cdot \frac{n}{N} = \frac{1}{N}\sum_{k=1}^{N} y_k = m_Y, \tag{3}$$

чиме је показана непристрасност ове тачкасте оцене.

За одређивање дисперзије ове тачкасте оцене такође се користе својства сл. величина $I_1, I_2, \ldots, I_N$. У наставку требало би приметити да је, за $k \neq l$:

$$E(I_k I_l) = \frac{n(n-1)}{N(N-1)}$$

што је, заправо, вероватноћа укључења другог реда π_{kl}, па су ове сл. величине међусобно зависне о чему јасно говори и коваријација:

$$\operatorname{cov}(I_k, I_l) = E(I_k I_l) - EI_k \cdot EI_l = -\frac{1}{N-1} \cdot \frac{n}{N}\left(1 - \frac{n}{N}\right).$$

Према томе је:

$$\begin{aligned}
D\widehat{m}_Y &= \frac{1}{n^2} D\left(\sum_{k=1}^{N} I_k y_k\right) = \frac{1}{n^2}\left(\sum_{k=1}^{N} y_k^2 D(I_k) + \sum_{k=1}^{N}\sum_{\substack{l=1 \\ l \neq k}}^{N} y_k y_l \cdot \operatorname{cov}(I_k, I_l)\right) \\
&= \frac{1}{n^2} \cdot \frac{n}{N}\left(1 - \frac{n}{N}\right)\left(\sum_{k=1}^{N} y_k^2 - \frac{1}{N-1}\sum_{k=1}^{N}\sum_{\substack{l=1 \\ l \neq k}}^{N} y_k y_l\right) \\
&= \frac{1}{n} \cdot \frac{1}{N(N-1)}\left(1 - \frac{n}{N}\right)\left((N-1)\sum_{k=1}^{N} y_k^2 - \sum_{k=1}^{N}\sum_{\substack{l=1 \\ l \neq k}}^{N} y_k y_l\right) \\
&= \frac{1}{n} \cdot \frac{1}{N(N-1)}\left(1 - \frac{n}{N}\right)\left((N-1)\sum_{k=1}^{N} y_k^2 - \left(\left(\sum_{k=1}^{N} y_k\right)^2 - \sum_{k=1}^{N} y_k^2\right)\right) \\
&= \frac{1}{n} \cdot \frac{1}{N(N-1)}\left(1 - \frac{n}{N}\right)\left(N\sum_{k=1}^{N} y_k^2 - \left(\sum_{k=1}^{N} y_k\right)^2\right)
\end{aligned}$$

$$= \frac{1}{n}\left(1-\frac{n}{N}\right)\cdot\frac{1}{(N-1)}\left(\sum_{k=1}^{N}y_k^2-\frac{1}{N}\left(\sum_{k=1}^{N}y_k\right)^2\right)$$
$$= \frac{1}{n}\left(1-\frac{n}{N}\right)\cdot\underbrace{\frac{1}{(N-1)}\left(\sum_{k=1}^{N}y_k^2-Nm_Y^2\right)}_{\stackrel{\text{def}}{=}\ \sigma^2} \tag{4}$$

Коришћена је формула: $D\left(\sum_{k=1}^{N}Z_k\right)=\sum_{k=1}^{N}DZ_k+\sum_{k=1}^{N}\sum_{\substack{l=1\\l\neq k}}^{N}\operatorname{cov}(Z_k,Z_l)$, где су $Z_1,Z_2,\ldots,Z_N$ сл. величине дефинисане на истом простору вероватноћа. ■

Лема 2. Узорачка дисперзија је непристрасна оцена непознате популацијске дисперзије.

Доказ.

Идеја је слична као у доказу Леме 1. Може се нпр. прво одредити следеће мат. очекивање:

$$\begin{aligned}
E\left(\sum_{k\in\mathcal{S}}(y_k-\widehat{m}_Y)^2\right) &= E\left(\sum_{k\in\mathcal{S}}((y_k-m_Y)-(\widehat{m}_Y-m_Y))^2\right)\\
&= E\left(\sum_{k\in\mathcal{S}}(y_k-m_Y)^2-n(\widehat{m}_Y-m_Y)^2\right)\\
&= E\left(\sum_{k=1}^{N}I_k(y_k-m_Y)^2\right)-nD\widehat{m}_Y\\
&= \frac{n}{N}\sum_{k=1}^{N}(y_k-m_Y)^2-\left(1-\frac{n}{N}\right)\sigma^2\\
&= \frac{n(N-1)}{N}\sigma^2-\left(1-\frac{n}{N}\right)\sigma^2=(n-1)\sigma^2.
\end{aligned} \tag{5}$$

Дељењем обе стране једнакости (5) са $(n-1)$ добија се:

$$E\left(\underbrace{\frac{1}{n-1}\sum_{k\in\mathcal{S}}(y_k-\widehat{m}_Y)^2}_{\stackrel{\text{def}}{=}\ \overline{S}^2}\right)=\sigma^2.$$

■

Када је приступ заснован на моделу, као што је већ речено, врло једноставан модел популације био би модел у коме су сл. величине $Y_1,Y_2,\ldots,Y_N$ независне и имају исту расподелу вероватноћа као посматрано обележје Y (овде је са Y_k означена вредност обележја Y ентитета означеног са k). Дакле, ове сл. величине су независне копије обележја Y са истим (али непознатим) мат. очекивањем $m_Y:=E_MY=EY$ и дисперзијом $\sigma_Y^2:=D_MY=DY$ (слово M у индексу указује на то да се нумеричке карактеристике обележја Y одређују на основу претпостављеног модела).

За конкретан узорак $\mathcal{S}$ без понављања обима n (без обзира на то да ли је он случајан или неслучајан узорак) узорачка средња вредност:

$$\overline{Y}=\frac{1}{n}\sum_{k\in\mathcal{S}}Y_k$$

је непристрасна оцена непознатог мат. очекивања EY, али и непристрасна оцена (боље речено: предиктор) сл. величине

$$\frac{1}{N}\sum_{k=1}^{N}Y_k=\frac{1}{N}\sum_{k\in\mathcal{S}}Y_k+\frac{1}{N}\sum_{k\notin\mathcal{S}}Y_k.$$

Идеја за овакав избор предиктора је да се издвоје сл. величине које су у вези са одабраним узорком $\mathcal{S}$ и на основу њих изврши "предвиђање" преосталих Y_k-ова, користећи чињеницу да је узорачка средња вредност најбољи линеарни непристрасни предиктор (при претпостављеном моделу) вредности обележја сваког ентитета који није одабран у узорак:

$$\frac{1}{N}\sum_{k\in\mathcal{S}} Y_k + \frac{1}{N}\sum_{k\notin\mathcal{S}} \widehat{Y}_k = \frac{1}{N}\sum_{k\in\mathcal{S}} Y_k + \frac{N-n}{N}\cdot\frac{1}{n}\sum_{k\in\mathcal{S}} Y_k = \frac{1}{n}\sum_{k\in\mathcal{S}} Y_k.$$

Непристрасност би овде требало схватити у смислу: "непристрасност у односу на модел", што једноставно значи да је за реализовани узорак s тачна једнакост:

$$E\left(\overline{Y}|\,\mathcal{S}=s\right) = E\left(\frac{1}{N}\sum_{k=1}^{N} Y_k|\,\mathcal{S}=s\right) \quad (= m_Y \quad \text{код овог метода узорковања}).$$

Лема 3. Средње квадратна грешка предвиђања (предикције) дата је са:

$$E\left(\overline{Y} - \frac{1}{N}\sum_{k=1}^{N} Y_k\right)^2 = \frac{\sigma_Y^2}{n}\left(1-\frac{n}{N}\right).$$

Доказ.

Због непристрасности предиктора, средње квадратна грешка своди се на:

$$D\left(\frac{1}{n}\sum_{k\in\mathcal{S}} Y_k - \frac{1}{N}\sum_{k=1}^{N} Y_k\right). \tag{6}$$

Прво се дисперзија из (6) запише као:

$$D\left(\left(\frac{1}{n}-\frac{1}{N}\right)\sum_{k\in\mathcal{S}} Y_k - \frac{1}{N}\sum_{k\notin\mathcal{S}} Y_k\right),$$

а затим је, на основу претпостављене независности и једнаке расподељености вредности обележја на узорку и ван њега, она даље једнака:

$$\left(\frac{1}{n}-\frac{1}{N}\right)^2\sum_{k\in\mathcal{S}} DY_k + \frac{1}{N^2}\sum_{k\notin\mathcal{S}} DY_k = \left(\frac{1}{n}-\frac{1}{N}\right)^2 nDY + \frac{1}{N^2}(N-n)DY$$
$$= \frac{DY}{n}\left(1-\frac{n}{N}\right).$$

■

Непристрасна оцена средње квадратне грешке предвиђања дата је са:

$$\frac{\overline{S}^2}{n}\left(1-\frac{n}{N}\right),$$

где је $\overline{S}^2$ узорачка дисперзија дата са:

$$\overline{S}^2 = \frac{1}{n-1}\sum_{k\in\mathcal{S}}(Y_k-\overline{Y})^2.$$