

(Прост) случајан узорак без понављања – оцењивање потпопулацијске средње вредности

Нека је N обим посматране популације, а N_l број јединица те популације које припадају њеној l -тој потпопулацији. Нека је Y обележје од интереса, а y_{lk} вредност обележја Y ентитета означеног са k који потиче из l -те потпопулације. Бира се случајан узорак без понављања обима n из комплетне популације. Тачно n_l ентитета (од одабраних n) потиче из l -те потпопулације и они чине подузорак $\mathcal{S}_l$. Приметити да је n_l сл. величина са мат. очекивањем једаким $\frac{nN_l}{N}$.

Лема 1. Подузорачка средња вредност је непристрасна оцена непознате потпопулацијске средње вредности. Дисперзија ове оценое једнака је:

$$\sigma_l^2 \left(E \left(\frac{1}{n_l} \right) - \frac{1}{N_l} \right),$$

где је σ_l^2 непозната потпопулацијска дисперзија.

Доказ.

Подузорачка средина је случајна величина дата са:

$$\hat{m}_l := \frac{1}{n_l} \sum_{k \in \mathcal{S}_l} y_{lk}.$$

Непристрасност ове статистике, која се предлаже као тачкаста оцена за m_l , може се показати увођењем услова који се односи на сл. величину n_l . Наиме, ако је дата реализација ν (фиксирана вредност из скупа $\{0, 1, \dots, n\}$) сл. величине n_l , онда, на основу метода одабира узорка, свака могућа комбинација дужине ν ентитета из потпопулације обима N_l има једнаку вероватноћу да буде одабрана као подузорак. Дакле, при услову $n_l = \nu$, подузорачка средња вредност понаша се као узорачка средина случајног узорка без понављања обима ν из популације обима N_l , па је:

$$E(\hat{m}_l | n_l) = m_l \quad \Rightarrow \quad E\hat{m}_l = E(E(\hat{m}_l | n_l)) = m_l.$$

За одређивање дисперзије ове тачкасте оценое користи се:

$$D(\hat{m}_l | n_l) = \frac{\sigma_l^2}{n_l} \left(1 - \frac{n_l}{N_l} \right) \quad \Rightarrow \quad D\hat{m}_l = E(D(\hat{m}_l | n_l)) + D(E(\hat{m}_l | n_l)) = \sigma_l^2 \left(E \left(\frac{1}{n_l} \right) - \frac{1}{N_l} \right) + 0.$$

Коришћена је формула: $DZ = E(D(Z|W)) + D(E(Z|W))$, где су Z и W сл. величине дефинисане на истом простору вероватноћа.

Потпопулацијска дисперзија σ_l^2 дата је формулом:

$$\sigma_l^2 = \frac{1}{N_l - 1} \sum_{k=1}^{N_l} (y_{lk} - m_l)^2.$$

■

Подузорачка дисперзија је:

$$\bar{S}_l^2 = \frac{1}{n_l - 1} \sum_{k=1}^{n_l} (y_{lk} - \hat{m}_l)^2,$$

па, ако је дата реализација сл. величине n_l , онда је

$$\frac{\bar{S}_l^2}{n_l} \left(1 - \frac{n_l}{N_l} \right)$$

непристрасна оцена условне дисперзије оценое $\hat{m}_l$.

Дисперзија оценое $\hat{m}_l$ била би мања уколико би обим подузорка (из l -те потпулације) био фиксиран и пропорционалан обиму l -те потпопулације, тј. једнак $\frac{nN_l}{N}$, јер је:

$$E \left(\frac{1}{n_l} \right) > \frac{N}{nN_l}.$$

Коришћена је Јенсенова неједнакост: ако је $h : \mathbb{R} \rightarrow \mathbb{R}$ конвексна Борелова ф-ја и Z сл. величина са коначним мат. очекивањем. Тада важи неједнакост:

$$h(EZ) \leq E(h(Z)).$$

Узорак са неједнаким вероватноћама

Мотивација за Hansen-Hurwitz-ову оцену

Од раније је познато да када се узорковање врши сукцесивно са враћањем (тј. увек из комплетне популације), одабири јединица су међусобно независни. Посматра се једно (произвољно) извлачење у коме се јединица означена са k бира у узорак са вероватноћом ψ_k , $k = \overline{1, N}$. Резултат тог извлачења сматраћемо случајним узорком $\mathcal{S}$ обима 1. За реализован узорак $s = \{k\}$ са $T(s)$ означимо реализацију тачкасте оцене T популацијског тотала, дату са:

$$T(s) = \frac{y_k}{\psi_k}.$$

Тада важи:

$$ET = \sum_s T(s)p(s) = \sum_{k=1}^N \frac{y_k}{\psi_k} \cdot \psi_k = \tau_Y$$

и

$$DT = \sum_s (T(s) - \tau_Y)^2 p(s) = \sum_{k=1}^N \left(\frac{y_k}{\psi_k} - \tau_Y \right)^2 \cdot \psi_k.$$

Када се врши n међусобно независних одабира (у сваком се јединица означена са k бира са вероватноћом ψ_k , $k = \overline{1, N}$) Hansen-Hurwitz-ова оцена непознатог популацијског тотала формира се као аритметичка средина n независних копија $T_1, T_2, \dots, T_n$ сл. величине T , тј:

$$\hat{\tau}_\psi := \frac{1}{n} \sum_{l=1}^n T_l = \frac{1}{n} \sum_{k \in \mathcal{R}} \frac{y_k}{\psi_k},$$

где је $\mathcal{R}$ случајан узорак, који се посматра као неуређен скуп ознака јединица одабраних у узорак, кардиналности n , са укљученим свим евентуалним понављањима ознака.

Користећи чињеницу да су сл. величине $T_1, T_2, \dots, T_n$ независне и једнако расподеле, лако се добијају својства тачкасте оцене $\hat{\tau}_\psi$:

$$\begin{aligned} E\hat{\tau}_\psi &= \frac{1}{n} \cdot nET = \tau_Y && \text{(непристрасност)} \\ D\hat{\tau}_\psi &= \frac{1}{n^2} \cdot nDT = \frac{1}{n} \sum_{k=1}^N \psi_k \left(\frac{y_k}{\psi_k} - \tau_Y \right)^2. \end{aligned} \quad (1)$$

Лема 2. Непристрасна оцена дисперзије $D\hat{\tau}_\psi$ из (1) дата је формулом:

$$\frac{1}{n(n-1)} \sum_{k \in \mathcal{R}} \left(\frac{y_k}{\psi_k} - \hat{\tau}_\psi \right)^2. \quad (2)$$

Доказ.

Уведу се сл. величине Q_k , $k = \overline{1, N}$, где Q_k представља број појављивања јединице означене са k у узорку (оне су, код узорковања са понављањем, присутне као аналогони индикатора укључења код узорковања без понављања). Оцена $\hat{\tau}_\psi$ може се записати и на следећи начин:

$$\hat{\tau}_\psi = \frac{1}{n} \sum_{k=1}^N Q_k \frac{y_k}{\psi_k}. \quad (3)$$

Приметити да важи: $\sum_{k=1}^N Q_k = n$ и $EQ_k = n\psi_k$, $k = \overline{1, N}$.

Рачуна се мат. очекивање предложене оцене из (2):

$$E \left(\frac{1}{n(n-1)} \sum_{k \in \mathcal{R}} \left(\frac{y_k}{\psi_k} - \hat{\tau}_\psi \right)^2 \right) = E \left(\frac{1}{n(n-1)} \sum_{k=1}^N Q_k \left(\frac{y_k}{\psi_k} - \hat{\tau}_\psi \right)^2 \right)$$

$$\begin{aligned}
&= \frac{1}{n(n-1)} \sum_{k=1}^N E \left(Q_k \left(\frac{y_k}{\psi_k} - \hat{\tau}_\psi \right)^2 \right) \\
&= \frac{1}{n(n-1)} \sum_{k=1}^N E \left(Q_k \left(\frac{y_k}{\psi_k} - \tau_Y - (\hat{\tau}_\psi - \tau_Y) \right)^2 \right) \\
&= \frac{1}{n(n-1)} \sum_{k=1}^N E \left(Q_k \left(\left(\frac{y_k}{\psi_k} - \tau_Y \right)^2 - 2 \cdot \left(\frac{y_k}{\psi_k} - \tau_Y \right) (\hat{\tau}_\psi - \tau_Y) + (\hat{\tau}_\psi - \tau_Y)^2 \right) \right) \\
&= \frac{1}{n(n-1)} \left\{ \sum_{k=1}^N \left(\frac{y_k}{\psi_k} - \tau_Y \right)^2 E Q_k - 2E \left((\hat{\tau}_\psi - \tau_Y) \sum_{k=1}^N Q_k \left(\frac{y_k}{\psi_k} - \tau_Y \right) \right) + nE(\hat{\tau}_\psi - \tau_Y)^2 \right\} \\
&= \frac{1}{n(n-1)} \left(n \cdot \sum_{k=1}^N \psi_k \left(\frac{y_k}{\psi_k} - \tau_Y \right)^2 - 2nE(\hat{\tau}_\psi - \tau_Y)^2 + nE(\hat{\tau}_\psi - \tau_Y)^2 \right) \\
&= \frac{1}{n-1} \left(\sum_{k=1}^N \psi_k \left(\frac{y_k}{\psi_k} - \tau_Y \right)^2 - E(\hat{\tau}_\psi - \tau_Y)^2 \right) \\
&= \frac{1}{n-1} (nD\hat{\tau}_\psi - D\hat{\tau}_\psi)
\end{aligned}$$

чиме је показана непристраност ове тачкасте оценое. ■

Мотивација за Horvitz-Thompson-ову оценоу

Бира се случајан узорак $\mathcal{S}$ без понављања обима n са неједнаким вероватноћама укључења π_k , $k = \overline{1, N}$. Претпоставља се да су познате и строго позитивне вероватноће укључења првог и другог реда.

Приметити да важи: $\sum_{k=1}^N \pi_k = n$, као и $\sum_{\substack{l=1 \\ l \neq k}}^N \pi_{kl} = (n-1)\pi_k$.

Коришћено је:

$$\begin{aligned}
\sum_{k=1}^N I_k = n &\Rightarrow n = \sum_{k=1}^N EI_k = \sum_{k=1}^N \pi_k \\
\sum_{\substack{l=1 \\ l \neq k}}^N \pi_{kl} &= \sum_{\substack{l=1 \\ l \neq k}}^N E(I_k I_l) = E(I_k(n - I_k)) = nEI_k - E(I_k^2) = n\pi_k - \pi_k = (n-1)\pi_k,
\end{aligned}$$

где су I_k индикатори укључења.

Управо с обзиром на чињеницу да је збир вероватноћа укључења првог реда једнак обиму узорка n , количник $\frac{\pi_k}{n}$ може се посматрати као "просечна" вероватноћа да ће јединица означена са k бити одабрана у узорак у неком од корака. За разлику од узорковања са понављањем, где је оцена тотала била формирана као аритметичка средина вредности $\frac{y_k}{\psi_k}$ за јединице које су се нашле у узорку, овде (код узорковања без понављања) се y_k не скалира дељењем са ψ_k него дељењем са просечном вероватноћом избора јединице у једном кораку, тј. са $\frac{\pi_k}{n}$ (што је интуитивно јасно с обзиром да се вероватноће одабира мењају од извлачења до извлачења). Тако се добија Horvitz-Thompson-ова оцена непознатог популацијског тотала:

$$\hat{\tau}_\pi := \sum_{k \in \mathcal{S}} \frac{y_k}{\pi_k} = \sum_{k=1}^N I_k \frac{y_k}{\pi_k}.$$

Својства тачкасте оценое $\hat{\tau}_\pi$:

$$\begin{aligned}
E\hat{\tau}_\pi &= \sum_{k=1}^N \frac{y_k}{\pi_k} EI_k = \sum_{k=1}^N \frac{y_k}{\pi_k} \cdot \pi_k = \tau_Y && \text{(непристрасност)} \\
D\hat{\tau}_\pi &= D \left(\sum_{k=1}^N I_k \frac{y_k}{\pi_k} \right) = \sum_{k=1}^N \left(\frac{y_k}{\pi_k} \right)^2 D(I_k) + \sum_{k=1}^N \sum_{\substack{l=1 \\ l \neq k}}^N \frac{y_k y_l}{\pi_k \pi_l} \cdot \text{cov}(I_k, I_l) \\
&= \sum_{k=1}^N \left(\frac{y_k}{\pi_k} \right)^2 \cdot \pi_k (1 - \pi_k) + \sum_{k=1}^N \sum_{\substack{l=1 \\ l \neq k}}^N \frac{y_k y_l}{\pi_k \pi_l} \cdot (\pi_{kl} - \pi_k \pi_l)
\end{aligned}$$

$$= \sum_{k=1}^N \left(\frac{1 - \pi_k}{\pi_k} \right) y_k^2 + \sum_{k=1}^N \sum_{\substack{l=1 \\ l \neq k}}^N \frac{\pi_{kl} - \pi_k \pi_l}{\pi_k \pi_l} y_k y_l \quad (4)$$

$$= \frac{1}{2} \sum_{k=1}^N \sum_{\substack{l=1 \\ l \neq k}}^N (\pi_k \pi_l - \pi_{kl}) \left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l} \right)^2. \quad (5)$$

Полазећи од израза за $D\hat{\tau}_\pi$ из (4) предлаже се њена тачкаста оцена:

$$\sum_{k \in S} \left(\frac{1 - \pi_k}{\pi_k^2} \right) y_k^2 + \sum_{k \in S} \sum_{\substack{l \in S \\ l \neq k}} \frac{\pi_{kl} - \pi_k \pi_l}{\pi_k \pi_l \pi_{kl}} y_k y_l,$$

а ако се крене од израза из (5) добија се следећа тачкаста оцена дисперзије:

$$\frac{1}{2} \sum_{k \in S} \sum_{\substack{l \in S \\ l \neq k}} \frac{\pi_k \pi_l - \pi_{kl}}{\pi_{kl}} \left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l} \right)^2.$$

Обе оцене су непристрасне.