Metoda reprezentacyjna

Stanisław Jaworski

Katedra Ekonometrii i Statystyki
Zakład Statystyki

Przedmiotem rozważań metody reprezentacyjnej są metody wyboru prób z populacji skończonych oraz metody szacowania nieznanych carakterystyk populacji.

Definicja (populacja)

Populacją generalną będziemy nazywać zbiór wszystkich jednostek badania. Populacja generalna będzie zapisywana w postaci:

$$\mathcal{U} = \{u_1, u_2, \dots, u_N\} \textit{ lub } \mathcal{U} = \{1, 2, \dots, N\}.$$

Liczbę $N \in \mathbb{N}$ nazywamy liczebnością populacji.

Przykłady

- Zbiór studentów, którz zdali egamin
- Zbiór gospodarstw domowych w Polsce w dniu 1 stycznia bieżącego roku
- Zbiór wyborców

Definicja (cecha statystyczna)

Cechą statystyczną nazywamy funkcję $\mathcal{Y}$*:*

$$\mathcal{Y} : \mathcal{U} \to \mathbb{R}$$

Wartość cechy dla j–tej jednostki badania, tzn. $\mathcal{Y}(u_j)$*, oznaczamy przez* Y_j

Przykłady

- ocena z egzaminu: $\mathcal{Y}(\text{student}) = \text{ocena z egzaminu}$
- dochody gospodarstwa domowego: $\mathcal{Y}(\text{gospodarstwo domowe}) = \text{dochód}$

Definicja (parametr populacji)

Parametrem populacji $\mathcal{U}$ *nazywamy wektor* $\mathbf{Y} = [Y_1, Y_2, \dots, Y_N]$

Definicja (przestrzeń parametrów)

Zbiór możliwych parametrów populacji $\mathcal{U}$ *nazywamy przestrzenią parametrów i oznaczamy przez* Ω

Definicja (Funkcja parametryczna)

Funkcją parametryczną nazywamy funkcję $T : \Omega \to \mathbb{R}$

Przykład

Populacja={student1, student2, student3}
Cecha=ocena z egzaminu
Parametr=$[2, 4, 5]$
Przestrzeń parametrów

$$\Omega = \{1, 2, 3, 4, 5\} \times \{1, 2, 3, 4, 5\} \times \{1, 2, 3, 4, 5\}$$

Zauważmy, że

$$[2, 4, 5] \in \Omega$$

Przykład funkcji parametrycznej $T : \Omega \to \mathbb{R}$:

$$T(Y_1, Y_2, Y_3) = \frac{1}{3} \sum_{i=1}^{3} Y_i$$

Przykłady

- Wartość globalna cechy $\mathcal{Y}$: $Y = \sum_{i=1}^{N} Y_i$
- Średnia cechy $\mathcal{Y}$: $\bar{Y} = \frac{1}{N} \sum_{i=1}^{N} Y_i$
- Wariancja cechy $\mathcal{Y}$: $S^2 = \frac{1}{N-1} \sum_{i=1}^{N} (Y_i - \bar{Y})^2$
- Minimalna wartość cechy $\mathcal{Y}$: $Y_{min} = \min\{Y_1, \ldots, Y_N\}$
- Maksymalna wartość cechy $\mathcal{Y}$: $Y_{max} = \max\{Y_1, \ldots, Y_N\}$
- Iloraz wartości globalnych cech $\mathcal{X}, \mathcal{Y}$: $R = \dfrac{Y}{X}$
- Kowariancja cech $\mathcal{X}, \mathcal{Y}$: $S_{xy} = \frac{1}{N-1} \sum_{j=1}^{N} (X_j - \bar{X})(Y_j - \bar{Y})$

Definicja (próba uporządkowana)

Próbą uporządkowaną s o liczebności n z populacji $\mathcal{U}$ nazywamy wektor

$$s = [j_1, j_2, \ldots, j_n] \in \mathcal{U}^n$$

Dla podkreślenia, że liczebność n dotyczy próby s będziemy ją oznaczać przez $n(s)$. Zbiór wszystkich uporządkowanych prób będziemy oznaczać przez $\mathscr{S}$

Definicja (próba nieuporządkowana)

Próbą nieuporządkowaną s' o liczebności ν z populacji $\mathcal{U}$ nazywamy zbiór

$$s' \subset \mathcal{U}$$

Dla podkreślenia, że liczebność ν dotyczy próby s będziemy ją oznaczać przez $\nu(s)$. Zbiór wszystkich nieuporządkowanych prób będziemy oznaczać przez $\mathscr{S}'$

Definicja (plan losowania)

Planem losowania nazywamy miarę prawdopodobieństwa p określoną na zbiorze $\mathscr{S}$:

$$p(s) \geq 0 \quad (\forall s \in \mathscr{S})$$

$$\sum_{s \in \mathscr{S}} p(s) = 1$$

Uwaga: Zapis $\sum_{s \ni j,k}$ będzie oznaczać, że sumowanie odbywa się potakich $s \in \mathscr{S}$, które zawierają jednostki j, k. Na przykład $s = (1, 2, 5), s = (1, 2)$ oraz $s = (6, 3, 1, 2)$ spełniają zapis $s \ni 1, 2$, a $s = (1, 3)$ już nie.

Definicja

- *Prawdopodobieństwo pierwszego rzędu:* $\pi_j = \sum_{s \ni j} p(s)$
- *Prawdopodobieństwo drugiego rzędu:* $\pi_{j,k} = \sum_{s \ni j,k} p(s)$
- *Prawdopodobieństwo k–tego rzędu:* $\pi_{j_1,\ldots,j_k} = \sum_{s \ni j_1,\ldots,j_k} p(s)$

Definicja

- *Oczekiwana efektywna liczebność próby:* $\boldsymbol{\nu} = E[\nu(S)] = \sum_{s \in \mathscr{S}} \nu(s)p(s)$
- *Wariancja efektywnej liczebności próby:* $D^2[\nu(S)] = \sum_{s \in \mathscr{S}} (\nu(s) - \boldsymbol{\nu})^2 p(s)$

Uwaga: $\nu(S)$*–zmienna losowa, która przyjmuje wartość* $\nu(s)$ *z prawdopodobieństwem* $p(s)$

Definicja (schemat losowania)

Schematem losowania próby nazywamy proces wyboru (jedna po drugiej) jednostek z populacji $\mathcal{U}$ *ze zgóry ustalonym prawdopodobieństwem wyboru dla poszczególnych jednostek w każdym ciągnieniu.*

Losowanie proste ze zwracaniem (lpzz)

Niech $u, u_1, u_2, \ldots, u_{i-1} \in \mathcal{U}$ oraz $\mathcal{U}$ ma rozmiar N

A_u–zdarzenie oznaczające, że w $i-$tym ciągnięciu wylosowaliśmy element u

$A_{u_1,\ldots,u_{i-1}}$–zdarzenie oznaczające, że w poprzednich ciągnięciach (od 1 do $i-1$) wyciągnęliśmy elementy $u_1, \ldots, u_{i-1}$

W losowaniu ze zwracaniem zachodzi $P(A_u|A_{u_1,\ldots,u_{i-1}}) = \frac{1}{N}$ dla dowolnych u, oraz $u_1, \ldots, u_{i-1}$.

Losowanie proste bez zwracania (lpbz)

W losowaniu bez zwracania zachodzi

$$P(A_u|A_{u_1,\ldots,u_{i-1}}) = \frac{1}{N-(i-1)}$$

dla $u \notin \{u_1, \ldots, u_{i-1}\}$ oraz

$$P(A_u|A_{u_1,\ldots,u_{i-1}}) = 0$$

w przeciwnym przypadku.

Prawdopodobieństwa pierwszego rzędu w losowaniach prostych

π_j – prawdopodobieństwo wylosowania j–tego elementu w próbie
$1 - \pi_j$ – prawdopodobieństwo niewylosowania j–tego elementu w próbie

Pobieramy próbę n – elementową. Losujemy ze zwracaniem.

$$1 - \pi_j = (1 - \frac{1}{N})^n$$

Pobieramy próbę n – elementową. Losujemy bez zwracania.

$$\pi_j = \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{n}{N}$$

Losowanie z prawdopodobieństwami proporcjonalnymi do wartości cechu $\mathcal{X}$ i ze zwracaniem (lppxzz)

Niech składowe parametru $[X_1, \ldots, X_N]$ będą większe od zera. Określamy prawdopodobieństwo $p_j,\ j = 1, \ldots, N$, wylosowania j–tego elementu populacji następująco:

$$p_j = \frac{X_j}{X}$$

Zgodnie z zadanym rozkładem prawdopdobieństwa losujemy próbę $n-$elementową.
A_i-zdarzenie polegające na tym, że $i-$ty elment nie pojawi się w próbie.
Stąd

$$\begin{aligned}\pi_{ij} = 1 - P(A_i \cup A_j) &= 1 - P(A_i) - P(A_j) + P(A_i \cap A_j) = \\ &= 1 - (1 - p_i)^n - (1 - p_j)^n + (1 - p_i - p_j)^n\end{aligned}$$

Zauważmy: $A_i \cap A_j-$ zdarzenie polegające na tym, że $i-$ty oraz $j-$ty elment nie pojawi się w próbie.

Definicja (operat losowania)

Operatem losowania nazywamy wykaz jednostek badania lub ich zespołów, zwanych jednostkami losowania. Każdej jednostce jest przyporządkowany identyfikator. Jeżeli losujemy jednostki badania, mówimy o losowaniu indywidualnym, jeżeli zespoły, o zespołowym.

Definicja (przestrzeń prób)

Niech $\mathcal{U} = \{1, 2, \ldots, N\}$. *Możemy wówczas przyjąć, że* $\Omega = \Omega_1 \times \ldots \Omega_N$. *Niech* $\Omega_s = \Omega_{j_1} \times \ldots \times \Omega_{j_n}$ *dla* $s = (j_1, \ldots, j_n) \in \mathscr{S}$. *Zbiór*

$$\mathcal{P} = \bigcup_{s \in \mathscr{S}} \Omega_s$$

nazywamy przestrzenią prób.

Definicja (statystyka, estymator)

Funkcję $t : \mathcal{P} \to \mathbb{R}$ *nazywamy statystyką.*

Definicja (estymator)

Niech $T : \Omega \to \mathbb{R}$ *oznacza funkcję parametryczną oraz* $\Theta = T(\Omega)$. *Statystykę* $t : \mathcal{P} \to \Theta$ *nazywamy estymatorem funkcji parametrycznej* T. *Jeśli* $Y \in \mathcal{P}$, *to* $t(Y)$ *nazywamy oceną funkcji parametrycznej* T.

Estymatory

Estymator liniowy jednorodny

$$t(Y_s) = \sum_{i=1}^{s} w_i Y_{ji}$$

dla $s = (j_1, j_2, \ldots, j_n)$, $Y_s = (Y_{j_1}, \ldots, Y_{jn})$ oraz ustalonych $w_1, w_2, \ldots, w_n \in \mathbb{R}$

Estymator Horvitza–Thompsona: $\bar{y}_{HT}$ (plan lppbz)

$$t(Y_{s'}) = \frac{1}{N} \sum_{i=1}^{n} \frac{Y_{j_i}}{\pi_{j_i}}$$

dla $s' = \{j_1, j_2, \ldots, j_n\}$, $Y_{s'} = (Y_{j_1}, \ldots, Y_{j_n})$.

Oznaczenie: $\pi_i = \sum_{s' \ni i} p(s')$

Estymator Hansena–Hurwitza: $\bar{y}_{HH}$ (plan lppxzz)

$$t(Y_s) = \frac{1}{N}\sum_{i=1}^{n}\frac{Y_{ji}}{np_{j_i}} = \frac{\bar{X}}{n}\sum_{i=1}^{n}\frac{Y_{j_i}}{X_{j_i}}$$

dla $s = (j_1, j_2, \ldots, j_n)$, $Y_s = (Y_{j_1}, \ldots, Y_{jn})$

Definicja (estymator nieobciążony)

Niech $\mathbf{Y}$ *oznacza parametr populacji oraz* $\mathbf{S}$ *zmienną losową, która przyjmuje wartość* $s \in \mathscr{S}$ *z prawdopodobieństwem* $p(s)$*, gdzie* p *jest planem losowania. Estymator* t *nazywamy nieobciążonym dla funkcji parametrycznej* T*, jeżeli*

$$E_p(t(Y_{\mathbf{S}})) = \sum_{s\in\mathscr{S}} t(Y_s)p(s) = T(\mathbf{Y})$$

Wartość oczekiwaną $E_p(t(Y_{\mathbf{S}}))$ *będziemy dla uproszczenia oznaczać przez* $E_p(t)$ *lub przez* $E(t)$*.*

Jeżeli estymator jest obciążony, to różnica $E_p(t(Y_{\mathbf{S}})) - T(\mathbf{Y})$ *nazywa się obciążeniem estymatora. Różnicę tę będziemy oznaczać przez* $B(t)$*.*

Definicja (wariancja estymatora)

Niech $\mathbf{Y}$ *oznacza parametr populacji oraz* $\mathbf{S}$ *zmienną losową, która przyjmuje wartość* $s \in \mathscr{S}$ *z prawdopodobieństwem* $p(s)$*, gdzie* p *jest planem losowania. Wariancją estymatora* t *nazywamy wyrażenie*

$$D_p^2(t(Y_{\mathbf{S}})) = \sum_{s\in\mathscr{S}} [t(Y_s) - E_p(t(Y_{\mathbf{S}}))]^2 p(s)$$
$$= \sum_{s\in\mathscr{S}} t(Y_s)^2 p(s) - [E_p(t(Y_{\mathbf{S}}))]^2$$

Wariancję $D_p^2(t(Y_{\mathbf{S}}))$ *będziemy oznaczać przez* $D_p^2(t)$ *lub* $D^2(t)$*.*

Definicja (błąd średniokwadratowy)

Przy oznaczeniach, jak w powyższych definicjach, średnim błędem średniokwadratowym estymatora t *jest wyrażenie*

$$MSE_p(t) = \sum_{s\in\mathscr{S}} (t(Y_s) - T(\mathbf{Y}))^2 p(s)$$
$$= D_p^2(t(Y_{\mathbf{S}})) + [B(t)]^2$$

Wartość oczekiwana estymatora Horvitza-Thompsona

$$
\begin{aligned}
E_p(\bar{y}_{HT}) &= \sum_{s' \in \mathscr{S}'} \left[\left(\frac{1}{N} \sum_{j \in s'} \frac{Y_j}{\pi_j} \right) p(s') \right] \\
&= \sum_{j=1}^{N} \sum_{s' \ni j} \frac{1}{N} \frac{Y_j}{\pi_j} p(s') \\
&= \sum_{j=1}^{N} \frac{1}{N} \frac{Y_j}{\pi_j} \sum_{s' \ni j} p(s') = \bar{Y}
\end{aligned}
$$

Wariancja estymatora Horvitza-Thompsona

$$D_p^2(\bar{y}_{HT}) = E_p[(y_{HT})^2] - [E_p(y_{HT})]^2 = E_p\left[\frac{1}{N}\sum_{j\in\mathbf{S}'}\frac{Y_j}{\pi_j}\right]^2 - [\bar{Y}]^2$$

$$= \frac{1}{N^2}E\left[\sum_{j\in\mathbf{S}'}\left(\frac{Y_j}{\pi_j}\right)^2 + \sum_{\substack{j,k\in\mathbf{S}'\\ j\neq k}}\frac{Y_jY_k}{\pi_i\pi_j}\right] - [\bar{Y}]^2$$

$$= \frac{1}{N^2}\left[\sum_{j=1}^{N}\frac{Y_j^2}{\pi_j^2}\sum_{s'\ni j}p(s') + \sum_{\substack{1\leq j,k\leq N\\ j\neq k}}\frac{Y_jY_k}{\pi_j\pi_k}\sum_{s'\ni j,k}p(s')\right] - [\bar{Y}]^2$$

$$= \frac{1}{N^2}\left[\sum_{j=1}^{N}\frac{Y_j^2}{\pi_j} + \sum_{\substack{1\leq j,k\leq N\\ j\neq k}}\frac{\pi_{jk}}{\pi_j\pi_k}Y_jY_k\right] - [\bar{Y}]^2$$

$$= \frac{1}{N^2}\left[\sum_{j=1}^{N}Y_j^2\left(\frac{1}{\pi_j} - 1\right) + \sum_{1\leq j,k\leq N;\ j\neq k}\left(\frac{\pi_{jk}}{\pi_j\pi_k} - 1\right)Y_jY_k\right]$$

Pewne przekształcenia

Dla zmiennej losowej **S** o wartościach z $\mathscr{S}$ lub $\mathscr{S}'$ i rozkładzie p zachodzi:

$$\nu(s) = \sum_{j=1}^{N} \mathcal{I}_s(j)$$

$$E_p\nu(\mathbf{S}) = \sum_{j=1}^{N} E_p(\mathcal{I}_{\mathbf{S}}(j)) = \sum_{j=1}^{N} \sum_{s \ni j} p(s) = \sum_{j=1}^{N} \pi_j$$

$$\sum_{\substack{1 \le j,k \le N \\ j \ne k}} \pi_{ij} = \sum_{\substack{1 \le j,k \le N \\ j \ne k}} E_p\left(\mathcal{I}_{\mathbf{S}}(j)\mathcal{I}_{\mathbf{S}}(k)\right) = E_p\left(\sum_{\substack{1 \le j,k \le N \\ j \ne k}} \mathcal{I}_{\mathbf{S}}(j)\mathcal{I}_{\mathbf{S}}(k)\right) =$$

$$= E_p\left(\left(\sum_{j=1}^{N} \mathcal{I}_{\mathbf{S}}(j)\right)^2 - \sum_{j=1}^{N} \mathcal{I}_{\mathbf{S}}^2(j)\right) = E_p\left(\nu^2(\mathbf{S}) - \sum_{j=1}^{N} \mathcal{I}_{\mathbf{S}}(j)\right) =$$

$$= E_p\left(\nu^2(\mathbf{S}) - \nu(\mathbf{S})\right) = E_p(\nu^2(\mathbf{S})) - E_p(\nu(\mathbf{S})) =$$

$$= D_p^2(\nu(\mathbf{S})) + [E_p(\nu(\mathbf{S}))]^2 - E_p(\nu(\mathbf{S}))$$

Uwaga: Wartość $E_p(\nu(\mathbf{S}))$ nazywamy oczekiwanym efektywnym rozmiarem próby. Oznaczmy $\nu = E_p(\nu(\mathbf{S}))$. Dla $\nu(S) \equiv \nu$ mamy zatem tożsamości:

$$\nu = \sum_{j=1}^{N} \pi_j, \qquad \sum_{\substack{1 \le j,k \le N \\ j \neq k}} \pi_{ij} = \nu^2 - \nu$$

dodatkowo

$$\sum_{\substack{j=1 \\ j \neq k}}^{N} \pi_{jk} = \sum_{\substack{j=1 \\ j \neq k}}^{N} E_p(\mathcal{I}_{\mathbf{S}}(j)\mathcal{I}_{\mathbf{S}}(k)) = E_p(\mathcal{I}_{\mathbf{S}}(k) \sum_{\substack{j=1 \\ j \neq k}}^{N} \mathcal{I}_{\mathbf{S}}(j)) =$$

$$= E_p(\mathcal{I}_{\mathbf{S}}(k)[\nu(S) - \mathcal{I}_{\mathbf{S}}(k)]) = (\nu - 1)\pi_k$$

$$\frac{1}{N^2}\sum_{\substack{1\le j,k\le N\\ j\neq k}}(\pi_i\pi_j-\pi_{ij})\left(\frac{Y_j}{\pi_j}-\frac{Y_k}{\pi_k}\right)^2=$$

$$=\frac{1}{N^2}\sum_{\substack{1\le j,k\le N\\ j\neq k}}\left(\pi_k\frac{Y_j^2}{\pi_j}+\pi_j\frac{Y_k^2}{\pi_k}-\pi_{jk}\frac{Y_j^2}{\pi_j^2}-\pi_{jk}\frac{Y_k^2}{\pi_k^2}-2Y_jY_k+2\pi_{ij}\frac{Y_jY_k}{\pi_j\pi_k}\right)$$

$$=\frac{2}{N^2}\sum_{\substack{1\le j,k\le N\\ j\neq k}}\left(\pi_k\frac{Y_j^2}{\pi_j}-\pi_{jk}\frac{Y_k^2}{\pi_k^2}-Y_jY_k+\pi_{ij}\frac{Y_jY_k}{\pi_j\pi_k}\right)=^{\star}$$

Ponieważ

$$\sum_{\substack{1\le j,k\le N\\ j\neq k}} \pi_k \frac{Y_j^2}{\pi_j} = \sum_{j=1}^{N} \frac{Y_j^2}{\pi_j} \sum_{\substack{k=1\\ k\neq j}}^{N} \pi_k = \sum_{j=1}^{N} \frac{Y_j^2}{\pi_j}(\nu - \pi_j) =$$

$$= \nu \sum_{j=1}^{N} \frac{Y_j^2}{\pi_j} - \sum_{j=1}^{N} Y_j^2$$

$$\sum_{\substack{1\le j,k\le N\\ j\neq k}} \pi_{jk} \frac{Y_j^2}{\pi_j^2} = \sum_{j=1}^{N} \frac{Y_j^2}{\pi_j^2} \sum_{\substack{k=1\\ k\neq j}}^{N} \pi_{jk} = \sum_{j=1}^{N} \frac{Y_j^2}{\pi_j^2}\pi_j(\nu - 1) =$$

$$= (\nu - 1) \sum_{j=1}^{N} \frac{Y_j^2}{\pi_j}$$

$$\sum_{\substack{1\le j,k\le N\\ j\neq k}} Y_j Y_k = Y^2 - \sum_{j=1}^{N} Y_j^2$$

mamy

$$=^{\star} \frac{2}{N^2}\left(\nu\sum_{j=1}^{N}\frac{Y_j^2}{\pi_j}-\sum_{j=1}^{N}Y_j^2-(\nu-1)\sum_{j=1}^{N}\frac{Y_j^2}{\pi_j}-Y^2+\sum_{j=1}^{N}Y_j^2+\sum_{\substack{1\le j,k\le N\\ j\ne k}}\pi_{ij}\frac{Y_jY_k}{\pi_j\pi_k}\right)$$

$$=\frac{2}{N^2}\left[\sum_{j=1}^{N}\frac{Y_j^2}{\pi_j}+\sum_{\substack{1\le j,k\le N\\ j\ne k}}\frac{\pi_{jk}}{\pi_j\pi_k}Y_jY_k-Y^2\right]$$

$$=\frac{2}{N^2}\left[\sum_{j=1}^{N}\frac{Y_j^2}{\pi_j}+\sum_{\substack{1\le j,k\le N\\ j\ne k}}\frac{\pi_{jk}}{\pi_j\pi_k}Y_jY_k\right]-2[\bar{Y}]^2=2D^2(y_{HT})$$

Z powyższych rachunków wynika, że dla $\nu(\mathbf{S}) \equiv \nu$ wariancja estymatora Hurwitza–Thompsona wynosi

$$D^2(y_{HT}) = \frac{1}{2N^2} \sum_{\substack{1 \le j,k \le N \\ j \neq k}} (\pi_i \pi_j - \pi_{ij}) \left(\frac{Y_j}{\pi_j} - \frac{Y_k}{\pi_k} \right)^2$$

Twierdzenie

Jeżeli $\pi_j > 0$ dla $j = 1, \ldots, N$ oraz $_{ij} > 0$ dla $i, j = 1, \ldots, N, j \neq k$, to statystyka

$$\hat{D}^2(\bar{y}_{HT}) = \frac{1}{N^2} \left[\sum_{j \in \mathbf{S}'} \frac{Y_j^2}{\pi_j} \left(\frac{1}{\pi_j} - 1 \right) + \sum_{\substack{1 \le j,k \in \mathbf{S}' \\ j \neq k}} \left(\frac{\pi_{jk}}{\pi_j \pi_k} - 1 \right) \frac{Y_j Y_k}{\pi_{j,k}} \right]$$

jest nieobciążonym estymatorem wariancji $D^2(\bar{y}_{HT})$. Dla $\nu(\mathbf{S}') \equiv \nu$ ma on postać

$$\hat{D}^2(\bar{y}_{HT}) = \frac{1}{2N^2} \sum_{\substack{j,k \in \mathbf{S}' \\ j \neq k}} \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{jk}} \left(\frac{Y_j}{\pi_j} - \frac{Y_k}{\pi_k} \right)^2$$

Wniosek

Dla n−elementowej próby wylosowanej według planu lpbz (losowanie proste bez zwracania)

$$\hat{D}^2(\bar{y}_{HT}) = \left(1 - \frac{n}{N}\right)\frac{s^2}{n},$$

gdzie

$$s^2 = \frac{1}{n-1}\sum_{i\in\mathbf{S}'}(Y_i - \bar{Y}_{\mathbf{S}'})^2, \quad \bar{Y}_{\mathbf{S}'} = \frac{1}{n}\sum_{i\in\mathbf{S}'} Y_i$$

Uwaga. Jeżeli zmienna $\mathbf{S}'$ zrealizuje się jako $s' = \{j_1, \ldots, j_n\}$, to dla uproszczenia będziemy oznaczać $(Y_{j_1}, \ldots, Y_{j_n})$ przez $(y_1, \ldots, y_n)$. Wtedy możemy zapisać

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Wartość oczekiwana estymatora Hansena–Hurwitza (plan Ippxzz, $n(s) \equiv n$)

Ze względu na sposób losowania estymator ten można zapisać w postaci

$$\bar{y}_{HH} = \frac{1}{N} \sum_{i=1}^{n} \frac{Y_{J_i}}{n p_{J_i}}$$

gdzie $J_1, J_2, \ldots, J_n$ są niezależnymi zmiennymi losowymi o tym samym rozkładzie

$$p : \begin{array}{c|c|c|c} 1 & 2 & \cdots & N \\ \hline p_1 & p_2 & \cdots & p_N \end{array}$$

Wtedy $E_p(\bar{y}_{HH}) = \frac{1}{Nn} \sum_{i=1}^{n} E_p\left(\frac{Y_{J_i}}{p_{J_i}}\right) = \frac{1}{Nn} \sum_{i=1}^{n} \sum_{j=1}^{N} \frac{Y_j}{p_j} p_j = \bar{Y}$

Wariancja estymatora Hansena–Hurwitza (plan lppxzz, $n(s) \equiv n$)

$$D^2(\bar{y}_{HH}) = E_p(y_{HH})^2 - (E_p(y_{HH}))^2 = E_p\left(\sum_{i=1}^{n} \frac{Y_{J_i}}{p_{J_i}}\right)^2 - (\bar{Y})^2 =$$

$$= \frac{1}{(Nn)^2} E_p\left(\sum_{i=1}^{n} \frac{Y_{J_i}^2}{p_{J_i}^2} + \sum_{\substack{1\le j,k\le n \\ j\neq k}} \frac{Y_{J_i} Y_{J_k}}{p_{J_i} p_{J_k}}\right) - (\bar{y})^2 =$$

$$= \frac{1}{(Nn)^2}\left(\sum_{i=1}^{n}\sum_{k=1}^{N} p_k \frac{Y_k^2}{p_k^2} + \sum_{\substack{1\le j,k\le n \\ j\neq k}} E\left(\frac{Y_{J_i}}{p_{J_i}}\right) E\left(\frac{Y_{J_k}}{p_{J_k}}\right)\right) - (\bar{y})^2 =$$

$$= \frac{1}{(Nn)^2}\left(n\sum_{k=1}^{N} \frac{Y_k^2}{p_k} + n(n-1)N^2(\bar{Y})^2\right) - (\bar{Y})^2 =$$

$$= \frac{1}{n}\sum_{k=1}^{N} p_k \left(\frac{Y_k}{Np_k} - \bar{Y}\right)^2$$

- Zauważmy, że dla $p_k = \frac{Y_k}{Y} = \frac{Y_k}{\sum_k Y_k}$ zachodzi $D^2(\bar{y}_{HH}) = 0$.
- Jeżeli dla $k = 1, 2, \ldots, N$ mamy $p_k = 1/N$, to

$$\bar{y}_{HH} = \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

oraz

$$D^2(\bar{y}_{HH}) = D^2(\bar{y}) = \frac{1}{n} \cdot \frac{1}{N} \sum_{k=1}^{N} (Y_k - \bar{Y})^2$$

Uogólniony estymator różnicy dla średniej

$$\bar{y}_{GD} = \frac{1}{N} \sum_{i \in \mathbf{S}'} \frac{Y_i - E_i}{\pi_i} + \bar{E}$$

gdzie $E_1, \ldots, E_N$ są dowolnymi stałymi. W szczególności dla $E_i = cX_i,\ i = 1, \ldots, N$, gdzie X_i są wartościami cechy dodatkowej $\mathcal{X}$ oraz c jest stałą, estymator ten przyjmuje postać:

$$\bar{y}_{GD} = \frac{1}{N} \sum_{i \in \mathbf{S}'} \frac{Y_i}{p_i} + c\bar{X} - \frac{1}{N} \sum_{i \in \mathbf{S}'} \frac{X_i}{\pi_i} = \bar{y}_{HT} + c(\bar{X} - \bar{x}_{HT})$$

Zatem $E(\bar{y}_{GD}) = \bar{Y}$ oraz

$$D^2(\bar{y}_{GD}) = D^2(\bar{y}_{HT}) + c^2 D^2(\bar{x}_{HT}) - 2cCov(\bar{x}_{HT}, \bar{y}_{HT}),$$

która jest minimalizowana dla $c = \frac{Cov(\bar{x}_{HT}, \bar{y}_{HT})}{D^2(\bar{x}_{HT})}$.
Zatem najmniejsza wariancja wynosi:

$$D(\bar{y}_{GD}) = D^2(\bar{y}_{HT})[1 - \varrho^2(\bar{x}_{HT}, \bar{y}_{HT})]$$

Estymator Khamisa $\bar{y}_K$ (losowanie zgodnie ze schematem lpzz)

Niech $s' = r(s)$, gdzie r jest funkcją redukcyjną

$$\bar{y}_K = \frac{1}{\nu(\mathbf{S}')} \sum_{j \in \mathbf{S}'} Y_j$$

Estymator Khamisa jest nieobciążony: $E(\bar{y}_K) = E(E(\bar{y}_K | \nu(\mathbf{S}'))) = E(\bar{Y}) = \bar{Y}$

Wariancja

$$D^2(\bar{y}_K) = D^2\left(\frac{1}{\nu(\mathbf{S}')}\sum_{j\in\mathbf{S}'} Y_j\right) =$$

$$= D^2\left(E\left[\frac{1}{\nu(\mathbf{S}')}\sum_{j\in\mathbf{S}'} Y_j \middle| \nu(\mathbf{S}')\right]\right) + E\left(D^2\left[\frac{1}{\nu(\mathbf{S}')}\sum_{j\in\mathbf{S}'} Y_j \middle| \nu(\mathbf{S}')\right]\right) =$$

$$= D^2(\bar{Y}) + E\left[\left(\frac{1}{\nu(\mathbf{S}')} - \frac{1}{N}\right)\frac{N\sigma^2}{N-1}\right]$$

gdzie $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \bar{Y})^2$

$$D^2(\bar{y}_K) = \left(E\left[\frac{1}{\nu(\mathbf{S}')}\right] - \frac{1}{N}\right)\frac{N\sigma^2}{N-1} > \left(\frac{1}{n} - \frac{1}{N}\right)S^2 = D^2(\bar{y}_{HT})$$

gdzie $D^2(\bar{y}_{HT})$ wyznaczone w przypadku: lpbz, $\nu \equiv n$

W przypadku lpzz i rozmiaru próby n mamy $D^2(\bar{y}_{HH}) = \sigma^2/n$
Zauważmy:

$$\left(E\left[\frac{1}{\nu(\mathbf{S}')}\right] - \frac{1}{N}\right)\frac{N}{N-1} < \frac{1}{n} \iff \frac{NE\left[\frac{n}{\nu(\mathbf{S}')}\right] - n}{N-1} < 1$$

Zatem

$$D^2(\bar{y}_{HH}) > D^2(\bar{y}_K) > D^2(\bar{y}_{HT})$$

Definicja (strategia losowania)

Strategią losowania nazywamy parę (p, t), gdzie p jest planem losowania, natomiast t jest estymatorem funkcji parametrycznej T. Strategię losowania będziemy oznaczać przez $H(p, t)$. Jeśli estymator t jest nieobciążony, powiemy o strategii, że jest nieobciążona.

Definicja (porównanie strategii)

Powiemy, że startegia $H_1(p_1, t_1)$ jest co najmniej tak dobra jak strategia $H_2(p_2, t_2)$, jeżeli

$$MSE_{p_1}(t_1) \leq MSE_{p_2}(t_2)$$

dla wszystkich $\mathbf{Y} \in \Omega$. Jeżeli dodatkowo

$$MSE_{p_1}(t_1) < MSE_{p_2}(t_2)$$

dla pewnego $\mathbf{Y} \in \Omega$, to mówimy, że strategia $H_1(p_1, t_1)$ jest lpesza od strategii $H_2(p_2, t_2)$

Przypomnienie: $\mathbf{MSE}_p(t) = \sum_{s \in \mathscr{S}} (t(Y_s) - T(\mathbf{Y}))^2 p(s)$

Przykład

Rozmiar próby: n

$$MSE_{lpzz}(\bar{y}_{HH}) > MSE_{lpzz}(\bar{y}_{K}) > MSE_{lpbz}(\bar{y}_{HT})$$

Problem

Jak ustalić liczebność próby n, aby błąd szacunku nie przekroczył zadanej wielkości d ze z góry ustalonym prawdopodobieństwem $1-\alpha$?

$$P(|t_n - T| < d) = 1 - \alpha$$

Definicja (minimalna liczebność próby)

Wielkość d nazywana jest maksymalnym dopuszczalnym błędem szacunku, natomiast $\delta = d/T$ maksymalnym dopuszaczalnym względnym błędem szacunku

Przykład

W przypadku szacowania $\bar{Y}$ na podstawie próby wylosowanej według schematu lpbz mamy

$$P(|\bar{y} - \bar{Y}| < d) = P(\bar{y} - d < \bar{Y} < \bar{y} + d) = 1 - \alpha$$

$$P(\bar{y} - u_{1-\alpha/2} D(\bar{y}) < \bar{Y} < \bar{y} + u_{1-\alpha/2} D(\bar{y})) \approx 1 - \alpha$$

gdzie $u_{1-\alpha/2}$ jest kwantylem rozkładu normalnego rzędu $(1-\alpha/2)$ oraz

$$D^2(\bar{y}) = \left(\frac{1}{n} - \frac{1}{N}\right) S^2$$

Zatem

$$d = u_{1-\alpha/2} D(\bar{y})$$

$$\delta = u_{1-\alpha/2} \frac{D(\bar{y})}{\bar{Y}}$$

oraz minimalna liczebność próby $n = [n^\star] + 1$

$$n^\star = \frac{N u_{1-\alpha/2}^2 S^2}{N d^2 + u_{1-\alpha/2}^2 S^2} = \frac{N u_{1-\alpha/2}^2 V^2}{N \delta^2 + u_{1-\alpha/2}^2 V^2}$$

gdzie $V = S/\bar{Y}$– współczynnik zmienności
Przy wyznaczaniu minimalnej próby parametry, których nie znamy zastępujemy ich oszacowaniami

Definicja (estymator produktowy (iloczynowy))

Statystykę

$$\bar{y}_p = \frac{\bar{x}\bar{y}}{\bar{X}}, \quad \bar{X} > 0$$

nazywamy estymatorem produktowym (iloczynowym) średniej $\bar{Y}$

Twierdzenie

Jeżeli n–elementowa próba wylosowana została według schematu lpbz z N–elementowej populacji, to

$$E(\bar{y}_p) = \bar{Y} + (1 - n/N)\frac{S_{xy}^2}{n\bar{X}} + O(n^{-2})$$

$$MSE(\bar{y}_p) = \left(1 - \frac{n}{N}\right)\frac{S_y^2 + 2RS_{xy} + R^2S_x^2}{n} + O(n^{-2})$$

gdzie

$$S_{xy} = \frac{1}{N-1}\sum_{j=1}^{N}(X_j - \bar{X})(Y_j - \bar{Y})$$

$$R = \frac{\bar{Y}}{\bar{X}}$$

Definicja (estymator ilorazowy)

Statystykę

$$\bar{y}_q = \frac{\bar{y}}{\bar{x}}\bar{X} = r\bar{X}$$

gdzie

$$r = \frac{\bar{y}}{\bar{x}}$$

nazywamy estymatorem ilorazowym średniej $\bar{Y}$.

Twierdzenie

Jeżeli n–elementowa próba wylosowana została według schematu lpbz z N–elementowej populacji, to

$$E(\bar{y}_q) = \bar{Y} + (1 - n/N)\frac{RS_x^2 - S_{xy}}{n\bar{X}} + O(n^{-2})$$

$$MSE(\bar{y}_q) = \left(1 - \frac{n}{N}\right)\frac{S_y^2 - 2RS_{xy} + R^2S_x^2}{n} + O(n^{-2})$$

Definicja (estymator liniowy)

Statystykę

$$\bar{y}_{lr} = \bar{y} + b(\bar{X} - \bar{x})$$

gdzie

$$b = \frac{s_{xy}^2}{s_x^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

nazwywamy estymatorem liniowym regresyjnym

Twierdzenie

Jeżeli n–elementowa próba wylosowana została według schematu lpbz z N–elementowej populacji, to

$$E(\bar{y}_{lr}) = \bar{Y} + \left(1 - \frac{n}{N}\right) C + O(n^{-2})$$

gdzie

$$C = \frac{B\sum_{j=1}^{N}(Y_j - \bar{Y})^3 - \sum_{j=1}^{N}(X_j - \bar{X})^2(Y_j - \bar{Y})}{(n-1)(N-1)S_x^2}$$

$$B = \frac{\sum_{j=1}^{N}(X_j - \bar{X})(Y_j - \bar{Y})}{\sum_{j=1}^{N}(X_j - \bar{X})^2}$$

oraz

$$MSE(\bar{y}_{lr}) = \left(1 - \frac{n}{N}\right)\frac{S_y^2(1-\rho_{xy}^2)}{n} + O(n^{-2})$$

gdzie

$$\rho_{xy} = \frac{S_{xy}}{S_x S_y} = B\frac{S_x}{S_y}$$

Zagadnienie (estymacja wartości średniej w losowaniu warstwowym)

Badamy cechę mierzalną Y. Populacja generalna o liczności N podzielona jest na L warstw o licznościach N_h, $h = 1, 2, \ldots, L$, przy czym

$$\sum_{h=1}^{L} N_h = N.$$

Frakcja elementów w warstwie h wynosi $W_h = N_h/N$. Z każdej warstwy oddzielnie losujemy n_h elementów do próby. Dla próby n elementowej spełnione jest

$$n_h = \frac{N_h}{N} n = W_h n, \quad h = 1, 2, \ldots, L.$$

Z próby otrzymujemy wyniki

$$y_{ih},\ i = 1, 2, \ldots, n_h,\ h = 1, 2, \ldots, L.$$

Na podstawie próby szacujemy średnią wartość $\bar{Y}$ populacji w następujący sposób:

$$\bar{y}_w = \frac{1}{N}\sum_{h=1}^{L} N_h \bar{y}_h = \sum_{h=1}^{L} W_h \bar{y}_h,$$

gdzie

$$\bar{y}_h = \frac{1}{n_h}\sum_{i=1}^{n_h} y_{ih}$$

Wariancja tego estymatora wynosi:

$$D^2(\bar{y}_w) = \left(\frac{1}{n} - \frac{1}{N}\right)\sum_{h=1}^{L} W_h S_h^2,$$

gdzie S_h^2 jest wariancją w h–tej warstwie (jeżeli jej nie znamy, to z dużej próby można ją oszacować za pomocą s_h^2).

Jeżeli badana cecha w populacji ma rozkład zbliżony do normalnego lub gdy przy innym rozkładzie próba jest duża, to przybliżony przedział ufności dla średniej ma postać

$$\left(\bar{y}_w - u_{1-\alpha/2} D(\bar{y}_w), \bar{y}_w + u_{1-\alpha/2} D(\bar{y}_w)\right)$$

Minimalna liczebność próby potrzebna do oszacowania średniej z maksymalnym dopuszczalnym błędem szacunku d wynosi

$$n = \frac{\sum_{h=1}^{L} W_h S_h^2}{\frac{d^2}{u_{1-\alpha/2}^2} + \frac{1}{N} \sum_{h=1}^{L} W_h S_h^2}$$

Zagadnienie („optymalna" estymacja wartości średniej w losowaniu warstwowym)

Założenia są identyczne, jak w poprzednim zagadnieniu, z tą różnicą, że

- *liczba wylosowanych elementów z h-tej warstwy wynosi*

$$n_h = \frac{W_h S_h}{\sum_{h=1}^{L} W_h S_h} n, \quad h = 1, 2, \ldots, L$$

- *wariancja estymatora wynosi*

$$D^2(\bar{y}_w) = \frac{1}{n} \left(\sum_{h=1}^{L} W_h S_h \right)^2 - \frac{1}{N} \sum_{h=1}^{L} W_h S_h^2$$

- *minimalna liczebność próby wynosi*

$$n = \frac{(\sum_{h=1}^{L} W_h S_h)^2}{\frac{d^2}{u_{1-\alpha/2}^2} + \frac{1}{N}\sum_{h=1}^{L} W_h S_h^2}$$

Losowanie w zagadnieniu pierwszym nazywamy *proporcjonalnym* (do wielkości warstwy), a w zagadnieniu drugim *optymalnym* (zapewnia najmniejszą wariancję estymatora)