Dobór próby – streszczenie

opracował Tomasz Żółtak wg wykładów Grzegorza Lissowskiego (i nie tylko)

Oznaczenia:

$E(X)-$ *średnia populacyjna X ;* $\quad \sigma^2-$ *wariancja populacyjna X ;*

$$\sigma^{*2}=\sigma^2\frac{N}{N-1}$$

$N-$ *liczba jednostek w populacji ;* $\quad n-$ *liczba jednostek w próbie ;*

$\bar{X}-$ *średnia X z próby (zm. losowa !)*

S^2- *wariancja X z próby (zm. losowa !)*

$$S^{*2}=S^2\frac{n}{n-1} \qquad\qquad D^{*2}(\bar{X})=D^2(X)\frac{N}{N-1}$$

Dobór prosty niezależny:

$\bar{X}$ jest nieobciążonym estymatorem średniej;

S^{*2} jest nieobciążonym estymatorem wariancji;

$$D^2(\bar{X})=\frac{D^2(X)}{n}$$

$$\text{oszac. } D^2(\bar{X})=\frac{S^{*2}}{n}$$

Dobór prosty zależny:

$\bar{X}$ jest nieobciążonym estymatorem średniej;

S^{*2} jest nieobciążonym estymatorem wariancji;

$$D^2(\bar{X})=\frac{D^2(X)}{n}\frac{N-n}{N-1}=\frac{D^{*2}(X)}{n}\frac{N-n}{N}$$

$$\text{oszac. } D^2(\bar{X})=\frac{S^{*2}}{n}\frac{N-n}{N}$$

$\frac{N-n}{N}$ - tzw. *poprawka dla skończonej populacji*;

Dobór warstwowy:

N_h- *liczba jednostek h−tej warstwy w populacji ;*

n_h- *liczba jedostek h−tej warstwy w próbie ;* $\qquad W_h=\frac{N_h}{N}$ *;*

nieobciążonym estymatorem średniej, o ile $\bar{X}_h$ są nieobciążonymi estymatorami średniej X w warstwach, jest:

$$\bar{X}_w=\sum W_h\bar{X}_h$$

ale $\bar{X}_w=\bar{X}$ tylko dla próby automatycznie wyważonej (prawdopodobieństwo proporcjonalne do liczebności warstwy);

$$D^2(\bar{X}_w)=\sum W_h^2\, D^2(\bar{X}_h)$$

$$\text{oszac. } D^2(\bar{X}_w)=\sum W_h^2\, \text{oszac. } D^2(\bar{X}_h)$$

alokacja proporcjonalna: $\frac{n_h}{n}=\frac{N_h}{N}$ a więc: $n_h=n\frac{N_h}{N}$;

alokacja optymalna: $\frac{n_h}{n}=\frac{N_h D^*(X_h)}{\sum N_h D^*(X_h)}$ a więc: $n_h=n\frac{N_h D^*(X_h)}{\sum N_h D^*(X_h)}$;

Dobór zespołowy:

zespoły o równej liczebności – losowanie proste zależne k zespołów:

$\bar{X}=\frac{1}{m}\sum_{1}^{k} E(X_k)$ jest nieobciążonym estymatorem średniej;

$$D^2(\bar{X})=\frac{M-m}{M}\frac{1}{m(m-1)}\sum_{r=1}^{M}\left[E(X_r)-\bar{X}\right]^2$$

$$\text{oszac. } D^2(\bar{X})=\frac{M-m}{M}\frac{1}{m(m-1)}\sum_{k=1}^{m}\left[E(X_k)-\bar{X}\right]^2=\frac{M-m}{M}\frac{WMZ}{m}$$

gdzie: $WMZ-\text{odciążona wariancja międzyzesposłowa}$

Warto zauważyć, że odpowiada to po prostu wyliczeniu średnich w ramach zespołów i przeprowadzeniu dalszych obliczeń tak, jakby to zespoły były jednostkami obserwacji.

Efektywność doboru zespołowego (w ogólności, nie tylko przy prostym zależnym losowaniu zespołów o równej liczebności) zależy silnie od zróżnicowania wewnątrz zespołów – im większe, tym dobór bardziej efektywny. Problem w tym, że zwykle (przynajmniej w socjologii) jednostki wewnątrz zespołów są do siebie raczej zbliżone i w efekcie dobór zespołowy jest mniej efektywny niż dobór prosty. Zależność (dla wszystkich przypadków) przybliża wzór:

$$\frac{D^2(\bar{X}_Z)}{D^2(\bar{X})}\approx 1+(\frac{N}{M}-1)\varrho_{wz}$$

gdzie ϱ_{WZ} oznacza współczynnik korelacji wewnątrzzespołowej:

$$\varrho_{WZ}=\frac{2}{\sum_{r=1}^{M}N_r(N_r-1)}\sum_{r=1}^{M}\sum_{i=1}^{N_r-1}\sum_{g=i+1}^{N_r}\frac{\left[X_{ri}-E(X)\right]\left[X_{rg}-E(X)\right]}{D^2(X)}$$

dodatnie wartości wskazują na nikłe zróżnicowanie wewnątrzgrupowe;

zespoły o różnej liczebności losowane bezzwrotnie z równym prawdopodobieństwem:

$\bar{X}_Z=\frac{1}{\frac{m}{M}N}\sum_{k=1}^{m}\sum_{j=1}^{N_k}X_{kj}$ jest nieobciążonym estymatorem średniej;

gdzie: $M-\text{liczba zespołów w populacji};$ $m-\text{liczba zespołów w próbie};$
$N-\text{liczebność populacji};$

Warto zwrócić uwagę, że $\frac{m}{M}N$ to wartość oczekiwana liczby obserwacji w próbie.

$$D^2(\bar{X}_Z)=\frac{M-m}{M}\frac{M^2}{N^2}\frac{1}{m(M-1)}\sum_{r=1}^{M}\left[N_rE(X_r)-\frac{N}{M}E(X)\right]^2$$

gdzie: $\frac{1}{M-1}\sum_{r=1}^{M}\left[N_rE(X_r)-\frac{N}{M}E(X)\right]^2$

to odciążona wariancja międzyzespołowa sum wartości w ramach zespołu.

$$\text{oszac. } D^2(\bar{X}_Z)=\frac{M-m}{M}\frac{M^2}{N^2}\frac{1}{m(m-1)}\sum_{k=1}^{m}\left[N_kE(X_k)-\frac{N}{M}\bar{X}_Z\right]^2$$

zespoły o różnej liczebności losowane zwrotnie z prawdopodobieństwem proporcjonalnym do liczebności:

$\bar{X}_{ZP}=\frac{1}{m}\sum_{k=1}^{m} E(X_k)$ jest nieobciążonym estymatorem średniej;

$$D^2(\bar{X}_{ZP})=\frac{1}{mN}\sum_{r=1}^{M} N_r\left[E(X_r)-E(X)\right]^2$$

$$\text{oszac. } D^2(\bar{X}_{ZP})=\frac{1}{m}\frac{\sum_{k=1}^{m}\left[E(X_k)-\bar{X}_{ZP}\right]^2}{m-1}=\frac{WMZ}{m}$$

gdzie: $WMZ-$ *odciążona wariancja międzyzesposłowa*

Również w tym przypadku odpowiada to wyliczeniu średnich w ramach zespołów i przeprowadzeniu dalszych obliczeń tak, jakby to zespoły były jednostkami obserwacji.

Dobór systematyczny:

Można traktować jako specyficzny rodzaj doboru warstwowego – warstwę stanowi każdych k kolejnych jednostek obserwacji (dla $nk=N$).
Można go też traktować jako dobór zespołowy, w którym wszystkie co k-te elementy stanowią jeden zespół.

$\bar{X}$ jest nieobciążonym estymatorem średniej jeśli $N=kn$;
W innym przypadku jest obciążony, ale niewiele (tym mniej, im mniejsze n/k);

$$D^2(\bar{X})=\frac{1}{k}\sum\left[\bar{X}-E(X)\right]^2$$

Problemem jest oszacowanie wariancji na podstawie pojedynczej próby.
Praktyczną sztuczką jest wylosowanie kilku mniejszych prób systematycznych zamiast jednej dużej i oszacowanie wariancji na tej podstawie:

Z populacji losujemy g podprób przy interwale $k=Ng/n$: z pierwszych k elementów populacji losujemy g elementów i na ich podstawie wybieramy podgrupy systematyczne;

$$\text{oszac. } D^2(\bar{X})=\left(1-\frac{g}{k}\right)\frac{1}{g(g-1)}\sum_{j=1}^{g}(\bar{X}_j-\bar{X})^2$$

Dobór dwustopniowy:

Dobór dwustopniowy można rozpatrywać jako połączenie doboru zespołowego z dodatkowym losowaniem w ramach wylosowanych zespołów (co doskonale widać we wzorach na wariancję estymatorów średniej). Klasycznie stosuje się dwa sposoby doboru próby automatycznie wyważonej: losowanie zespołów bezzwrotne z równym prawdopodobieństwem lub losowanie zwrotne z prawdopodobieństwem proporcjonalnym do wielkości zespołu. W użyciu bywa też dobór bezzwrotny z proporcjonalnym prawdopodobieństwem, ale nie będzie tu omawiany. Efektywność schematu zależy oczywiście bardzo silnie od wartości korelacji wewnątrzzespołowej (zwłaszcza gdy losuje się dużą liczbę obserwacji na dalszych stopniach losowania).
W przypadku doboru o większej liczbie stopni, procedurę przeprowadza się iteracyjnie, z dołu do góry (uzyskując kolejno oszacowania wariancji w ramach zespołów).

zespoły o różnej liczebności losowane bezzwrotnie z równym prawdopodobieństwem:

$$\bar{X}_{Db}=\frac{1}{\frac{m}{M}N}\sum_{k=1}^{m}N_k E(X_k)$$ jest nieobciążonym estymatorem średniej;

$$D^2(\bar{X}_{Db})=D^2(\bar{X}_Z)+\frac{M^2}{N^2}\frac{1}{mM}\sum_{r=1}^{M}\left(\frac{N_r-n_r}{N_r}\right)\frac{N_r^2\hat{D}^{*2}(X_r)}{n_r}$$

gdzie $D^2(\bar{X}_Z)$ jest wariancją doboru zespołowego odpowiadającemu pierwszemu poziomowi losowania (tj. losowanie bezzwrotne z równym prawdopodobieństwem);

$$oszac.\ D^2(\bar{X}_{Db})=oszac.\ D^2(\bar{X}_Z)+\frac{M^2}{N^2}\frac{1}{m(m-1)}\sum_{k=1}^{m}\left(\frac{N_k-n_k}{N_k}\right)\frac{N_k^2 S_k^{*2}}{n_k}$$

gdzie S_k^{*2} jest oszacowaniem $D^{*2}(X_k)$;

zespoły o różnej liczebności losowane zwrotnie z prawdopodobieństwem proporcjonalnym do liczebności:

$$\bar{X}_{Dp}=\frac{1}{m}\sum_{k=1}^{m}E(X_k)$$ jest nieobciążonym estymatorem średniej;

$$D^2(\bar{X}_{Dp})=D^2(\bar{X}_{ZP})+\frac{1}{m}\sum_{r=1}^{M}\left(\frac{N_r-n_r}{N_r}\right)\frac{N_r}{N}\frac{\hat{D}^{*2}(X_r)}{n_r}$$

gdzie $D^2(\bar{X}_{ZP})$ jest wariancją doboru zespołowego odpowiadającemu pierwszemu poziomowi losowania (tj. losowanie zwrotne z prawdopodobieństwem proporcjonalnym do liczebności);

$$oszac.\ D^2(\bar{X}_{Dp})=oszac.\ D^2(\bar{X}_{ZP})+\frac{1}{m}\sum_{k=1}^{m}\left(\frac{N_k-n_k}{N_k}\right)\frac{N_k}{N}\frac{S_k^{*2}}{n_k}$$

gdzie S_k^{*2} jest oszacowaniem $D^{*2}(X_k)$;

Szacowanie wielkości błędów standardowych dla współczynników regresji, średnich warunkowych, itp. przy złożonych schematach doboru próby[1)]:

Dla schematów zespołowego i wielopoziomowego szacowanie wielkości błędów standardowych tego typu współczynników jest możliwe tylko przy użyciu technik nieparametrycznych (jacknife, balanced repeated replication, bootstrapping), lub poprzez linearyzację Taylora (estymatory wariancji estymatorów uzyskane w wyniku zastosowania linearyzacji Taylora określa się też mianem *sandwich estimators*).

Dla schematów warstwowych, gdy znane są wagi dla poszczególnych jednostek obserwacji, sprawę da się rozwiązać prościej, bo przez wykorzystanie regresji ważonej (chociaż ma to inne niemiłe konsekwencje, jak np. brak rozsądnej interpretacji R^2 i w ogóle sum kwadratów, a więc i niemożliwość posługiwania się np. testem F). Najogólniej rzecz biorąc, w takim przypadku, poza użyciem ważonych średnich i wariancji we wszystkich miejscach gdzie występuje, zastępuje się liczbę obserwacji *n* sumą wag *W*.

[1)] Za: L. Kish „Inference from Complex Samples". *Journal of the Royal Statistical Society Series B (Methodological)* 36 (1974), s. 1-37.

Testy χ^2 dla złożonych schematów doboru próby[2)]:

Test zgodności (H_0: rozkładu X w próbie z rozkładem teoretycznym) dokonuje się w oparciu o zmodyfikowaną statystykę testową:

$$\chi^2_* = \frac{\chi^2}{\hat{\lambda}}$$

$$\hat{\lambda} = \frac{n}{k-1} \sum_{i=1}^{k} \frac{\hat{D}^2(p_i)}{p_{i0}}$$

gdzie:

n – liczba obserwacji, k – liczba komórek rozkładu;

p_{i0} – częstość występowania i-tej wartości w rozkładzie teoretycznym;

$\hat{D}^2(p_i)$ - estymator wariancji dla częstości występowania i-tej wartości w rozkładzie empirycznym;

przykładowo, dla doboru prostego $\hat{D}^2(p_i) = \frac{p_{i0}(1-p_{i0})}{n}$;

Test jednorodności (H_0: r niezależnie dobranych, ale niekoniecznie w ten sam sposób, rozkładów empirycznych $X_1, X_2, \ldots X_r$ zostało wylosowanych z tej samej populacji) przeprowadza się w oparciu o statystykę testową:

$$\chi^2_* = \frac{1}{\hat{\lambda}} \sum_{i=1}^{r} n_i \sum_{j=1}^{k} \frac{(p_{ij} - p_j)^2}{p_j}$$

$$\hat{\lambda} = \frac{1}{(r-1)(k-1)} \sum_{i=1}^{r} n_i \left(1 - \frac{n_i}{n}\right) \sum_{j=1}^{k} \frac{\hat{D}^2(p_{ij})}{p_j}$$

gdzie:

r – liczba porównywanych rozkładów empirycznych;

n_i – liczba obserwacji w i-tym rozkładzie empirycznym;

p_{ij} – częstość występowania j-tej wartości w i-tym rozkładzie empirycznym;

p_j – częstość występowania j-tej wartości w hipotetycznym rozkładzie wspólnym (należy sobie założyć jakiś);

Test niezależności (H_0: zmienne X i Y są niezależne stochastycznie) przeprowadza się w oparciu o zmodyfikowaną statystykę testową:

$$\chi^2_* = \frac{\chi^2}{\hat{\lambda}_0} \qquad \text{lub} \qquad \chi^2_{**} = \frac{\chi^2}{min(\hat{\lambda}_X, \hat{\lambda}_Y)}$$

dla wartości poprawek równych:

$$\hat{\lambda}_0 = \frac{n}{rs} \sum_{i=1}^{r} \sum_{i=1}^{s} \frac{\hat{D}^2(p_{ij})^2}{p_{ij}} \qquad \hat{\lambda}_X = \frac{n}{r-1} \sum_{i=1}^{r} \frac{\hat{D}^2(p_{i.})^2}{p_{i.}} \qquad \hat{\lambda}_Y = \frac{n}{s-1} \sum_{j=1}^{s} \frac{\hat{D}^2(p_{.j})^2}{p_{.j}}$$

gdzie:

r – liczba komórek rozkładu X; s - liczba komórek rozkładu Y;

p_{ij} – hipotetyczna częstość występowania obserwacji o wartościach X=i i Y=j;

$p_{i.}$ - szacowane częstości X; $p_{.j}$ - szacowane częstości Y;

[2)] Za: Cz. Domański, K. Pruska *Nieklasyczne metody statystyczne*. Polskie Wydawnictwa Ekonomiczne, Warszawa 2000.

Estymatory – ogólnie

Estymatory różnych parametrów populacji *tradycyjnie* wyprowadza się od estymatorów sumy wartości zmiennej w populacji. Od tych estymatorów można łatwo przejść do estymatorów innych parametrów, w szczególności średniej, ale także wariancji (kowariancji), kwantyli, itd. Tak więc, jeśli chodzi o estymację punktową, sprawa jest względnie prosta i sprowadza się do obliczenia odpowiedniego estymatora(ów) sumy i przekształcania ich aż do uzyskania pożądanych efektów. Co więcej, w praktyce sprowadza się to po prostu do policzenia (odpowiednio) ważonego parametru z próby!
Znacznie trudniejsze jest wyliczanie błędów standardowych. Dla parametrów będących liniowymi funkcjami estymatorów sumy (np. średnia) sprawa jest prosta – można je wyliczyć, stosując to samo przekształcenie liniowe. Dla innych parametrów jak np. wariancja (kowariancja), czy współczynniki regresji trzeba stosować złożone metody: albo linearyzację Taylora, albo metody symulacyjne (BRR, jacknife, bootstrap). Zresztą było już o tym wcześniej.
Tu uwaga, w literaturze anglojęzycznej wyrażenie „estimation of variance" właściwie bez wyjątków odnosi się do estymacji wariancji estymatorów, a nie do estymacji wariancji jako parametru populacji!

Dwa podstawowe typy estymatorów sum w populacji to **estymator Hansena-Hurvitza** i **estymator Horvitza-Thompsona**. Pierwszy ma zastosowanie tylko do schematów zwrotnych, drugi do wszystkich bez wyjątku (tyle że nie zawsze łatwo go policzyć).

Oznaczmy:

- n *liczba elementów w próbie*
- π_i *prawd. znalezienia się i-tego el. populacji w próbie*
- π_{ij} *prawd. jednoczesnego znalezienia się w próbie i-tego i j-tego el. populacji*
- p_i *prawd. wybrania i-tego el. populacji w pojedynczym losowaniu zwrotnym*

Estymatory punktowe sumy w populacji Hansena-Hurvitza i Horvitza-Thompsona to:

$$\hat{T}_{HH}=\frac{1}{n}\sum_{i=1}^{n}\frac{x_i}{p_i} \qquad \sum_{i=1}^{n}\frac{x_i}{\pi_i}=\hat{T}_{HT}$$

i są one estymatorami nieobciążonymi.
Warto przy tym zauważyć, że w przypadku doboru zwrotnego oba estymatory są tożsame, bo w takiej sytuacji: $\pi_i = n\,p_i$

Wariancja estymatora Hansena-Hurvitza:

$$D^2(\hat{T}_{HH})=\frac{1}{n}\sum_{i=1}^{N}p_i\left(\frac{x_i}{p_i}-X\right)^2$$

gdzie:

$$X=\sum_{i}^{N}x_i$$

a jej nieobciążony estymator:

$$\hat{D}^2(\hat{T}_{HH})=\frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{x_i}{p_i}-\frac{1}{n}\sum_{j=1}^{n}\frac{x_j}{p_j}\right)^2$$

Jeśli chodzi o wariancję estymatora Horvitza-Thompsona[3)], to przyjmując:

$$R_i = \begin{cases} 1 & \text{gdy i-ty el. populacji znalazł się w próbie} \\ 0 & \text{gdy i-ty el. populacji nie znalazł się w próbie} \end{cases}$$

i wiedząc, że:

$$E(R_i)=\pi_i \qquad D^2(R_i)=\pi_i(1-\pi_i) \qquad Cov(R_i, R_j)=\pi_{ij}-\pi_i\pi_j$$

możemy rozpisać ją jako:

$$D^2(\hat{T}_{HT})=D^2\left(\sum_{i=1}^{N} R_i\frac{x_i}{\pi_i}\right)$$

$$D^2(\hat{T}_{HT})=\sum_{i=1}^{N} D^2\left(R_i\frac{x_i}{\pi_i}\right)+\sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N} Cov\left(R_i\frac{x_i}{\pi_i}, R_j\frac{x_j}{\pi_j}\right)$$

$$D^2(\hat{T}_{HT})=\sum_{i=1}^{N}\frac{x_i^2}{\pi_i^2}\pi_i(1-\pi_i)+\sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N}\frac{x_i x_j}{\pi_i\pi_j}(\pi_{ij}-\pi_i\pi_j)$$

Ponieważ dla $i=j$:

$$\sum_{i=1}^{N}\sum_{j=i}^{N}\frac{x_i x_j}{\pi_i\pi_j}(\pi_{ij}-\pi_i\pi_j)=\sum_{i=1}^{N}\frac{x_i^2}{\pi_i^2}(\pi_i-\pi_i^2)=\sum_{i=1}^{N}\frac{x_i^2}{\pi_i^2}\pi_i(1-\pi_i^2)$$

($\pi_{ij}=\pi_i$ bo losowanie jest bezzwrotne)

możemy zapisać:

$$D^2(\hat{T}_{HT})=\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{x_i x_j}{\pi_i\pi_j}(\pi_{ij}-\pi_i\pi_j)$$

zaś estymatorem wariancji estymatora jest:

$$\hat{D}^2(\hat{T}_{HT})=\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{x_i x_j}{\pi_i\pi_j}(\pi_{ij}-\pi_i\pi_j)\frac{1}{\pi_{ij}}$$

Jest on nieobciążony jeżeli $\forall_{i,j\in\{1,\ldots,N\}}: \pi_{ij}>0$.

Alternatywny (równoważny) wzór na wariancję estymatora Horvitza-Thompsona zaproponowali Yates i Grundy oraz Sen:

$$D^2(\hat{T}_{HT})=\sum_{i=1}^{N}\sum_{j>i}^{N}(\pi_i\pi_j-\pi_{ij})\left(\frac{x_i}{\pi_i}-\frac{x_j}{\pi_j}\right)^2$$

Ze wzoru tego wynika również alternatywny estymator wariancji estymatora:

$$\hat{D}^2_{YGS}(\hat{T}_{HT})=\sum_{i=1}^{n}\sum_{j>i}^{n}\frac{\pi_i\pi_j-\pi_{ij}}{\pi_{ij}}\left(\frac{x_i}{\pi_i}-\frac{x_j}{\pi_j}\right)^2$$

Estymator ten ma duże znaczenie, gdyż posiada lepsze własności (raczej nie zdarza mu się przyjmować wartości ujemnych) niż ten zaproponowany przez Horvitza i Thompsona. Jest nieobciążony dla zadanej wielkości próby (tzn. wtedy, gdy zastosowany schemat losowania zawsze zwraca próbę o tej samej, założonej liczebności).

Oczywistą uciążliwością estymatora Horvitza-Thompsona jest konieczność uwzględnienia przy obliczaniu wariancji estymatorów prawdopodobieństw jednoczesnego znalezienia się w próbie π_{ij}. W praktyce z reguły stosuje się różnego rodzaju aproksymacje (obliczenie takich jednoczesnych prawdopodobieństw może być koszmarnie trudne, a przy dużej liczebności próby staje się ich na dodatek potwornie dużo). Dodatkowo dla dużych prób bardzo obciążająca może okazać się już sama konieczność przejścia w sumowaniu przez wszystkie nieuporządkowane pary jednostek obserwacji.

[3)] Za: T. J. Rao „Five Decades of the Horvitz·Thompson Estimator and Furthermore...". *Journal of the Indian Society of Agricultural Statistics 58(2)* (2004), s. 177-189.