Barbara Kowalczyk
Instytut Ekonometrii SGH w Warszawie

Estymacja ilorazu dwóch średnich w kolejnym okresie badania

Wprowadzenie

Wiele spośród ważnych badań metodą reprezentacyjną, wykonywanych zarówno przez instytucje rządowe, jak i firmy komercyjne przeprowadzające analizę rynku, powtarzanych jest w regularnych odstępach czasu. Do badań takich wykonywanych przez GUS należą m.in.: „Badania aktywności ekonomicznej ludności", „Badania budżetów gospodarstw domowych" oraz „Spisy rolne". Częstym przedmiotem zainteresowania w tego typu badaniach jest estymacja ilorazu wartości globalnych (średnich) dwóch cech. Szacuje się m.in.: stosunek liczby pracujących mężczyzn do liczby pracujących kobiet, przeciętne wydatki i spożycie w gospodarstwie domowym na osobę, przeciętny plon pszenicy z ha, stosunek powierzchni przeznaczonej pod uprawę pszenicy do powierzchni pod uprawę żyta itd.

Przy powtarzaniu badania reprezentacyjnego powstaje pytanie: w jaki sposób wykorzystać informacje z okresu poprzedniego, aby zwiększyć efektywność oceny dla okresu badanego? Istnieje szeroko rozwinięta teoria dotycząca badań powtarzalnych, ale związana raczej z szacunkiem takich parametrów, jak średnia cechy (wartość globalna) na okres bieżący, różnica średnich w dwóch kolejnych okresach czy też średnia z kilku okresów badania. Problem estymacji ilorazu wartości globalnych (średnich) dwóch cech w przypadku badań powtarzalnych jest bardziej skomplikowany i rzadziej poruszany w literaturze, choć z praktycznego punktu widzenia na pewno nie mniej ważny.

Schemat losowania dwustopniowego

Zakładamy, że N-elementowa populacja została podzielona na M rozłącznych, niepustych podzbiorów zwanych jednostkami losowania pierwszego stopnia (lps) (w praktyce jednostkami lps są np. rejony statystyczne czy terenowe punkty badań). Jednostka losowania pierwszego stopnia o numerze h

zawiera N_h jednostek badania, nazywanych jednostkami losowania drugiego stopnia (lds) (dla przykładu w „Badaniach budżetów gospodarstw domowych" jednostkami lds są mieszkania). Losowanie dwustopniowe polega na tym, że najpierw losujemy spośród M jednostek lps m jednostek według wybranego schematu. Na drugim stopniu losowanie przeprowadza się tylko w wylosowanych wcześniej jednostkach lps, przy czym zakłada się, że liczby jednostek lds n_h losowanych z jednostek lps ustalane są przed losowaniem.

Wartość cechy Y u j-tej jednostki lds należącej do h-tej jednostki lps oznaczamy przez Y_{hj} (h = 1, 2, ..., M, j = 1, 2, ..., N_h). Przyjmujemy poza tym następujące oznaczenia:

- wartość globalna cechy Y w h-tej jednostce lps: $Y_h = \sum_{j=1}^{N_h} Y_{hj}$,
- wariancja cechy Y dla h-tej jednostki lps:

$$S_{2h}^2 = \frac{1}{N_h - 1} \sum_{j=1}^{N_h} (Y_{hj} - \frac{1}{N_h} Y_h)^2,$$

- wartość globalna cechy Y dla całej populacji: $Y = \sum_{h=1}^{M} Y_h$,
- wartość cechy Y u jednostki lds wylosowanej w i-tym ciągnieniu z jednostki lps wylosowanej w g-tym ciągnieniu: y_{gi}.

W artykule rozpatrujemy schemat losowania dwustopniowego: 1-lppxzz, 2-lpbz. Skrót 1-lppxzz oznacza, że na pierwszym stopniu stosowany jest schemat: losowanie z prawdopodobieństwami proporcjonalnymi do wartości cechy X ze zwracaniem (lppxzz), tzn. prawdopodobieństwo wyciągnięcia h-tej jednostki w pojedynczym ciągnieniu wynosi $p_h = \frac{X_h}{\sum_{h=1}^{M} X_h}$, natomiast na drugim stopniu stosowany jest schemat: losowanie proste bez zwracania (lpbz). Cechą dodatkową w praktyce jest często wielkość jednostki lps mierzona liczbą jednostek lds w niej zawartych.

Schemat losowania dwustopniowego należy do najczęściej stosowanych schematów w praktyce badań reprezentacyjnych. Poza tym, losowanie jednostopniowe jest szczególnym przypadkiem wyżej wymienionego, tak więc przedstawiona w artykule teoria przenosi się również na ten schemat losowania.

Jeżeli próba została wylosowana zgodnie ze schematem losowania dwustopniowego: 1-lppxzz, 2-lpbz, to estymatorem nieobciążonym wartości globalnej Y jest statystyka

$$y=\frac{1}{m}\sum_{g=1}^{m}\frac{N_g}{p_g n_g}\sum_{i=1}^{n_g}y_{gi}=\frac{1}{m}\sum_{g=1}^{m}\frac{N_g}{p_g}\bar{y}_g,$$

a jej wariancja dana jest wzorem

$$D^2(y)=\frac{1}{m}\left[\sum_{h=1}^{M}p_h\left(\frac{Y_h}{p_h}-Y\right)^2+\sum_{h=1}^{M}N_h(N_h-n_h)\frac{S_{2h}^2}{p_h n_h}\right].$$

Natomiast kowariancja

$$Cov(y,z)=\frac{1}{m}\left[\sum_{h=1}^{M}p_h\left(\frac{Y_h}{p_h}-Y\right)\left(\frac{Z_h}{p_h}-Z\right)+\sum_{h=1}^{M}N_h(N_h-n_h)\frac{S_{2hyz}}{p_h n_h}\right],$$

gdzie

$$S_{2hyz}=\frac{1}{N_h-1}\sum_{j=1}^{N_h}(Y_{hj}-\frac{1}{N_h}Y_h)(Z_{hj}-\frac{1}{N_h}Z_h).$$

W dalszym ciągu pracy stosujemy oznaczenia:

$$S^2(Y)=\left[\sum_{h=1}^{M}p_h\left(\frac{Y_h}{p_h}-Y\right)^2+\sum_{h=1}^{M}N_h(N_h-n_h)\frac{S_{2h}^2}{p_h n_h}\right],$$

$$C(Y,Z)=\left[\sum_{h=1}^{M}p_h\left(\frac{Y_h}{p_h}-Y\right)\left(\frac{Z_h}{p_h}-Z\right)+\sum_{h=1}^{M}N_h(N_h-n_h)\frac{S_{2hyz}}{p_h n_h}\right].$$

$$\rho(Y,Z)=\frac{C(Y,Z)}{S(Y)S(Z)}=\frac{Cov(y,z)}{D(y)D(z)}.$$

Rozważamy obecnie problem, w którym badanie reprezentacyjne dotyczące tej samej populacji przeprowadzane jest dwukrotnie. W pierwszym okresie losujemy m jednostek lps (następnie z każdej jednostki odpowiednią liczbę jednostek lds). W drugim okresie pewną część próby pozostawiamy do badania, a dokładniej pozostawiamy pm ($0 \leq p \leq 1$, pm $\in$ N) jednostek lps (wraz z wylosowanymi jednostkami lds), natomiast qm = (1 – p)m jednostek lps dolosowujemy z wyjściowej populacji, tj. spośród M jednostek lps i dalej z nich losujemy jednostki lds.

Interesuje nas odpowiedź na pytanie, jaką część próby (tj. frakcję jednostek lps) należy pozostawić do badania w drugim okresie oraz jak wykorzystać informacje z okresu poprzedniego w celu zwiększenia efektywności estymacji na okres bieżący.

Przyjmujemy następujące oznaczenia:

– estymator wartości globalnej Y_t w okresie t (t = 1, 2) obliczony na podstawie wspólnej dla obu okresów części próby:

$$y_{tp} = \frac{1}{mp} \sum_{g=1}^{mp} \frac{N_g}{p_g} \bar{y}_{tg},$$

– estymator wartości globalnej Y_t w okresie t (t = 1, 2) obliczony na podstawie części próby badanej tylko w jednym okresie:

$$y_{tq} = \frac{1}{mq} \sum_{g=mp+1}^{m} \frac{N_g}{p_g} \bar{y}_{tg}.$$

Proponowane estymatory

Klasyczny estymator $R_2 = \frac{Y_2}{X_2} = \frac{\bar{Y}_2}{\bar{X}_2}$ ilorazu wartości globalnych (średnich) dwóch cech w drugim okresie badania:

$$t = \frac{y_2}{x_2}.$$

Estymator ten nie wykorzystuje informacji z okresu poprzedniego. Przybliżony błąd średniokwadratowy tego estymatora wynosi:

$$MSE(t) \approx \frac{R_2^2}{m}\left[\frac{S^2(Y_2)}{Y_2^2}+\frac{S^2(X_2)}{X_2^2}-2\frac{C(Y_2,X_2)}{Y_2X_2}\right]=\frac{R_2^2G_2}{m},$$

gdzie

$$G_i = V^2(Y_i)+V^2(X_i)-2V(Y_i)V(X_i)\rho(Y_i,X_i),$$

$$V(Y_i)=\frac{S(Y_i)}{Y_i}, \quad V(X_i)=\frac{S(X_i)}{X_i}, \quad R_i=\frac{Y_i}{X_i}, \qquad i = 1, 2.$$

Estymator R_2 zaproponowany przez Tripathiego i Sinhę [2]:

$$t_{TS} = Qr'_{2p} + (1-Q)r_{2q},$$

gdzie

$$r_{2q}=\frac{y_{2q}}{x_{2q}}, \qquad r'_{2p}=\frac{y_{2p}+b(y_1-y_{1p})}{x_{2p}+b^*(x_1-x_{1p})},$$

$$b=\frac{C(Y_2,Y_1)}{S^2(Y_1)}, \qquad b^*=\frac{C(X_2,X_1)}{S^2(X_1)}.$$

Q jest tak dobrane, żeby minimalizowało błąd średniokwadratowy MSE(t_{TS}), tzn.

$$Q=\frac{MSE(r_{2q})}{MSE(r'_{2p})+MSE(r_{2q})}.$$

Przybliżony błąd średniokwadratowy tego estymatora wynosi:

$$MSE(t_{TS}) \approx \frac{R_2^2G_2}{m}\frac{G_2-qF}{G_2-q^2F}=\frac{R_2^2G_2}{m}\frac{1-q\frac{F}{G_2}}{1-q^2\frac{F}{G_2}},$$

gdzie

$$F = V^2(Y_2)\rho^2(Y_2,Y_1)+V^2(X_2)\rho^2(X_2,X_1)-2V(Y_2)V(X_2)\rho(X_2,X_1)\rho(Y_2,X_1)$$
$$-2V(Y_2)V(X_2)\rho(Y_2,Y_1)\rho(X_2,Y_1)+2V(Y_2)V(X_2)\rho(Y_2,Y_1)\rho(X_2,X_1)\rho(Y_1,X_1)$$

Optymalna wartość q (minimalizująca MSE(t_{TS})) przedstawia się następująco:

$$q_{opt} = 1, \quad gdy \quad F < 0,$$

$$q_{opt} = \frac{1}{1+\sqrt{1-\frac{F}{G_2}}}, \quad gdy \quad F \geq 0.$$

Esrtymator R_2 zaproponowany przez Okafora i Arnaba [1]:

$$t_{OA} = \frac{y_2'}{x_2'} = \frac{Q_y y_{2p}' + (1-Q_y) y_{2q}}{Q_x x_{2p}' + (1-Q_x) x_{2q}},$$

gdzie

$$y_{2p}' = y_{2p} + b_y(y_1 - y_{1p}), \quad x_{2p}' = x_{2p} + b_x(x_1 - x_{1p}).$$

Autorzy zaproponowali przyjąć takie wartości współczynników Q_y oraz b_y, które minimalizują $D^2(y_2)$ i analogicznie takie Q_x oraz b_x, które minimalizują $D^2(x_2)$, tzn.

$$b_z = \frac{C(Z_2, Z_1)}{S^2(Z_1)},$$

$$Q_z = \frac{D^2(z_{2q})}{D^2(z_{2q}) + D^2(z_{2p}')}, \qquad z = y,\ x.$$

Autorzy założyli, że

(i) $$\rho(y_1, y_2) = \rho(x_1, x_2) = \rho_0$$

i przy tym założeniu przyjęli $q = \frac{1}{1+\sqrt{1-\rho_0^2}}$, które minimalizuje jednocześnie $D^2(y_2')$ oraz $D^2(x_2')$.

Przyjmując dalsze założenia

(ii) $\rho(y_i, x_j) = \rho, \quad i, j = 1, 2,$

(iii) $V(Y_2) = V(X_2) = V_2$

autorzy przedstawili wzór na przybliżony błąd średniokwadratowy swojego estymatora.

$$MSE_{opt}(t_{OA}) \approx \frac{R_2^2 V_2^{\,2}}{m} \frac{1}{\sqrt{1-\rho_0^2}} \left\{ (1+\sqrt{1-\rho_0^2})(1-\rho) + \rho_0(\rho-\rho_0) \right\}.$$

Zaproponowana nowa postać estymatora R_2:

$$t_A = Q r_{2p}^{"} + (1-Q) r_{2q},$$

gdzie

$$r_{2q} = \frac{y_{2q}}{x_{2q}}, \qquad r_{2p}^{"} = r_{2p} + \tilde{b}(r_1 - r_{1p}) = \frac{y_{2p}}{x_{2p}} + \tilde{b}\left(\frac{y_1}{x_1} - \frac{y_{1p}}{x_{1p}}\right),$$

$$\tilde{b} = \frac{E(r_2 - R_2)(r_1 - R_1)}{E(r_1 - R_1)^2}.$$

Wykonując odpowiednie obliczenia, otrzymujemy:

$$\tilde{b} = \frac{I_Y}{I_X} \frac{H}{G_1},$$

gdzie

$$H = V(Y_1)V(Y_2)\rho(Y_1, Y_2) + V(X_1)V(X_2)\rho(X_1, X_2) + \\ -V(Y_1)V(X_2)\rho(Y_1, X_2) - V(Y_2)V(X_1)\rho(Y_2, X_1),$$

$$I_Y = \frac{Y_2}{Y_1}, \quad I_X = \frac{X_2}{X_1}.$$

Q wybieramy tak, aby minimalizowało $MSE(t_A)$, tzn.

$$Q = \frac{MSE(r_{2q})}{MSE(r_{2p}^{"}) + MSE(r_{2q})}.$$

Błąd średniokwadratowy zaproponowanego przeze mnie estymatora wynosi:

$$MSE(t_A) = \frac{R_2^2 G_2}{m} \frac{1 - q\frac{H^2}{G_1 G_2}}{1 - q^2 \frac{H^2}{G_1 G_2}}.$$

$MSE(t_A)$ osiąga najmniejszą wartość dla

$$q_{opt} = \frac{1}{1 + \sqrt{1 - \frac{H^2}{G_1 G_2}}}.$$

Zatem optymalna frakcja jednostek lps, które powinny być pozostawione do badania w okresie następnym, wynosi: $p_{opt} = 1 - q_{opt}$.

Podstawiając q_{opt} do wzoru na błąd średniokwadratowy otrzymujemy:

$$MSE_{opt}(t_A) \approx \frac{R_2^2 G_2}{m} \frac{1}{2}\left(1 + \sqrt{1 - \frac{H^2}{G_1 G_2}}\right).$$

Estymator zaproponowany przeze mnie jest trochę bardziej skomplikowany w konstrukcji niż estymatory proponowane wcześniej (trudno jest oszacować $\tilde{b}$). Ma on jednak tę zaletę, iż wszystkie współczynniki liczone były tak, aby minimalizowały błąd średniokwadratowy całego estymatora (w przypadku estymatora Tripathiego-Sinha współczynniki b, b^* nie spełniały tego kryterium).

Porównanie estymatorów oraz końcowe wnioski

W tej części opracowania założymy spełnienie warunków (i) – (iii). Dla różnych wartości ρ, ρ_0 porównamy efektywność estymatora zaproponowanego przeze mnie z efektywnością pozostałych estymatorów wykorzystujących informacje z okresu poprzedniego.

$MSE_{opt}(t_A)/MSE_{opt}(t_{TS})$:

ρ_0 \ ρ	0,1	0,3	0,5	0,7	0,9
0,1	1,000	0,979	0,800	--	--
0,3	0,998	1,000	0,958	--	--
0,5	0,999	0,988	1,000	0,873	--
0,7	0,9996	0,995	0,976	1,000	--
0,9	0,9999	0,999	0,995	0,980	1,000

$MSE_{opt}(t_A)/MSE_{opt}(t_{OA})$:

ρ_0 \ ρ	0,1	0,3	0,5	0,7	0,9
0,1	0,997	0,963	0,767	--	--
0,3	0,998	0,976	0,882	--	--
0,5	0,999	0,984	0,928	0,687	--
0,7	0,999	0,989	0,954	0,833	--
0,9	0,9995	0,994	0,974	0,910	0,607

Jak widać, estymator skonstruowany jako estymator liniowy od r_{2p}, r_{2q}, r_1, r_{1p} okazał się nieco lepszy od estymatorów proponowanych wcześniej (-- oznacza, że dana kombinacja ρ, ρ_0 jest niedopuszczalna). Zobaczymy teraz, jaki jest zysk na efektywności przy zastosowaniu estymatora t_A w porównaniu z zastosowaniem estymatora klasycznego, który nie wykorzystuje informacji z okresu poprzedniego.

$100\% [MSE(t) - MSE_{opt}(t_A)]/MSE_{opt}(t_A)$:

ρ_0 \ ρ	0,1	0,3	0,5	0,7	0,9
0,1	0,0	2,1	**25,0**	--	--
0,3	1,3	0,0	4,4	--	--
0,5	5,5	2,1	0,0	**14,6**	--
0,7	**14,6**	9,9	4,4	0,0	--
0,9	**37,2**	**32,0**	**25,0**	**14,6**	0,0

W zaznaczonych rubrykach (gdy odpowiednie współczynniki korelacji, jak można było się spodziewać, są wysokie) zysk na efektywności jest znaczny i przekracza 10%.

Na koniec powiemy jeszcze kilka słów o estymatorze zaproponowanym przez Okafora i Arnaba. Jego zaletą jest prosta konstrukcja. Jak widzieliśmy, okazał się on jednak mniej efektywny od estymatora zaproponowanego przeze mnie. W niektórych przypadkach okazuje się on mniej efektywny nawet od estymatora klasycznego t, który żadnych informacji z okresu poprzedniego nie wykorzystuje.

$MSE_{opt}(t_{OA})/MSE(t)$:

ρ_0 \ ρ	0,1	0,3	0,5	0,7	0,9
0,1	**1,003**	**1,017**	**1,043**	--	--
0,3	0,989	**1,024**	**1,087**	--	--
0,5	0,949	0,995	**1,077**	**1,270**	--
0,7	0,873	0,920	**1,004**	**1,200**	--
0,9	0,729	0,762	0,821	0,959	**1,647**

Literatura

1. OKAFOR F.C., ARNAB R., 1987: Some strategies of two-stage sampling for estimating population ratios over two occasions. *Austral. J. Statist.* 29, 128–142.
2. TRIPATHI T.P., SINHA S.K.P., 1976: Estimation of ratio on successive occasions. Paper presented at the „Symposium on Recent Developement in Survey Methodology" held at Indian Statistical Institute.

Estimation of population ratios on the most recent occasion

Abstract

The paper concerns survey sampling on repeated occasions. We discuss the estimation of the ratio of two population totals (means) on the second occasion using the two stage sampling. We present three estimators (including one new), based on informations from previous and recent occasions, and we compare the estimators with the ordinary ratio estimator.