Tomasz Bąk
Uniwersytet Ekonomiczny w Katowicach

OPTYMALIZACJA LICZBY WARSTW DLA ALOKACJI NEYMANA

Wprowadzenie

Losowanie warstwowe jest często wykorzystywaną w praktyce metodą doboru próby w przypadku estymacji wartości średniej pewnej cechy w populacji. Ten sposób selekcji umożliwia przejrzystą realizację badania w terenie. Dodatkowo pozwala na precyzyjnie ujęcie w próbie zróżnicowania charakterystyk badanej populacji. Ważne jest jednak dobre podzielenie populacji na warstwy – to, jak populacja zostanie podzielona na warstwy, bezpośrednio wpływa na wariancję estymatora wartości średniej pewnej cechy. Aby zminimalizować wariancję estymatora średniej, należy zmaksymalizować zróżnicowanie wartości średniej badanej zmiennej w warstwach (maksymalizacji wariancji międzygrupowej).

Do dobrego podziału populacji są potrzebne, co oczywiste, dane. Najlepiej wartości jakiejś zmiennej silnie skorelowanej ze zmienną objętą badaniem lub kilku takich zmiennych. Pozyskanie danych generuje jednak w sposób oczywisty koszty. W praktyce zazwyczaj ma się do czynienia z naturalnym podziałem populacji na warstwy lub podziałem wynikającym z dostępności operatu losowania (por. Wywiał, 1992). Wtedy wykorzystuje się gotowe warstwy. Oprócz takiego gotowego podziału populacji na warstwy badacz często ma dostęp do innych danych dotyczących populacji.

W przypadku optymalnej alokacji Neymana pożądane jest, aby występowało duże zróżnicowanie miar rozproszenia wewnątrzwarstwowego. Zagadnieniem warstwowania populacji na podstawie optymalnego podziału obszaru zmienności cechy zajmował się Dalenius (1957).

Opracowanie jest poświęcone optymalizacji liczb warstw w sytuacji, gdy badacz dysponuje gotowym podziałem na warstwy i planuje wykorzystanie optymalnej alokacji Neymana. Zakłada się, że badacz dysponuje również informacjami o liczności oraz wartości średniej i wariancji badanej cechy w każdej warstwie. Autor nie spotkał się z takim podejściem do optymalizacji warstw dla alokacji Neymana.

1. Warunek opłacalności łączenia warstw

Oznaczmy badaną cechę jako *X*. Dobór próby do estymacji tej cechy zostanie przeprowadzony losowaniem warstwowym. Wewnątrz każdej warstwy przeprowadzimy losowanie proste bezzwrotne. Liczność próby zostanie ustalona zgodnie z metodą optymalnej alokacji Neymana (1934), klasycznego podejścia w ustalaniu liczebności próby w losowaniu warstwowym.

Niech H oznacza liczbę warstw. Dalej niech μ_h oznacza średnią wartość cechy X w h-tej warstwie, a M_h ilość elementów w h-tej warstwie $h = 1, \dots, H$. Wtedy $\sum_{h=1}^{H} M_h \mu_h$ jest estymatorem globalnej wartości cechy *X*. Wariancja estymatora $\sum_{h=1}^{H} M_h \mu_h$ jest dana wzorem:

$$\sigma^2 = \sum\nolimits_{h=1}^{H} \frac{M_h^2}{m_h} \frac{M_h - m_h}{M_h - 1} \sigma_h^2, \tag{1}$$

gdzie σ_h^2 oznacza wariację cechy X w h-tej warstwie, $h = 1, \dots, H$.

Załóżmy, że z powodów niezależnych od badacza istnieje możliwość objęcia badaniem m_0 jednostek. Jest to wartość stała i niezmienna w badaniu. Neyman (1934) udowodnił, że wariancja (1) przyjmuje wartość minimalną, gdy liczności próby w warstwach są równe

$$m_h = m_0 \frac{M_h S_h}{\sum_{k=1}^{H} M_k S_k}, h = 1, \dots, H, \tag{2}$$

gdzie $S_h = \sqrt{\frac{M_h}{M_h - 1} \sigma_h^2}, h = 1, \dots, H$.

W punkcie wyjścia do rozważań dysponujemy zatem podziałem badanego obszaru na H warstw oraz pewnymi dodatkowymi informacjami na temat każdej warstwy. Informacje te zostaną wykorzystane do oceny wariancji badanej cechy w warstwie powstałej z połączenia 2 warstw wyjściowych. Rozważmy zatem połączenie *i*-tej oraz *j*-tej warstwy w jedną. Dla przejrzystości dalszych zapisów wprowadźmy oznaczenie zbioru indeksów warstw, które nie ulegają połączeniu: $A = \{1, \dots, H\} \backslash \{i, j\}$. Wariację (nowego) estymatora $\sum_{h \in A \cup \{ij\}} M_h \mu_h$, gdzie ij jest indeksem nowo powstałej warstwy, można wtedy przedstawić w postaci:

$$\sigma^2 = \sum\nolimits_{h \in A} \frac{M_h^2}{m'_h} \frac{M_h - m'_h}{M_h - 1} \sigma_h^2 + \frac{\left(M_i + M_j\right)^2}{m'_{ij}} \frac{M_i + M_j - m'_{ij}}{M_i + M_j - 1} \sigma'^2_{ij}, \tag{3}$$

gdzie $m'_h, h \in A \cup \{ij\}$ są nowymi wielkościami prób w warstwach, a σ'^2_{ij} określa wariancję badanej cechy w nowo utworzonej warstwie. Wariancję σ'^2_{ij} można rozłożyć na sumę dwóch czynników: wariancji międzygrupowej oraz warian-

cji międzygrupowej (Fisher, 1925). Stąd wariancja nowo powstałej warstwy w największej ogólności przyjmuje postać:

$$\sigma'^2_{ij} = \frac{1}{M_i + M_i}\left[M_i\sigma_i^2 + M_j\sigma_j^2 + M_i\left(\bar{x}_i - \bar{x}_{ij}\right)^2 + M_j\left(\bar{x}_j - \bar{x}_{ij}\right)^2\right], \tag{4}$$

gdzie $\bar{x}_i$, $\bar{x}_j, \bar{x}_{ij}$ oznaczają odpowiednio średnią wartość cechy X w *i*-tej wyjściowej warstwie, *j*-tej wyjściowej warstwie oraz w nowo powstałej warstwie.

Zgodnie ze wzorem (2) liczność próby w nowo utworzonej warstwie definiuje zależność:

$$m'_{ij} = m_0 \frac{\left(M_i + M_j\right)\sqrt{\frac{M_i + M_j}{M_i + M_j - 1}\sigma'^2_{ij}}}{\sum_{k\in A} M_k\sqrt{\frac{M_h}{M_h - 1}\sigma_h^2} + \left(M_i + M_j\right)\sqrt{\frac{M_i + M_j}{M_i + M_j - 1}\sigma'^2_{ij}}}, \tag{5}$$

$h = 1, \ldots, H$, którą to prostymi środkami można przekształcić do postaci:

$$\sigma'^2_{ij} = \frac{m'^2_{ij}\left(\sum_{k\in A} M_k\sqrt{\frac{M_h}{M_h - 1}\sigma_h^2}\right)^2\left(M_i + M_j - 1\right)}{\left(m_0 - m'_{ij}\right)^2\left(M_i + M_j\right)^3}. \tag{6}$$

Połączenie *i*-tej i *j*-tej wyjściowej warstwy przypuszczalnie spowoduje zmianę liczności warstw, które nie zostały połączone. Wprowadźmy α jako niewiadomą spełniającą zależność $m'_{ij} = m_i + m_j + \alpha$. Innymi słowy, α określa różnicę pomiędzy wielkością próby w nowo powstałej warstwie a sumą wielkości prób w *i*-tej i j-tej wyjściowej warstwie. Warunek określający wartość α można zapisać inaczej jako $\sum_{h\in A} m_h = \sum_{h\in A} m'_h + \alpha$, a stąd otrzymujemy zależność:

$$m'_h = m_h\left(1 - \frac{\alpha}{\sum_{h\in A} m_h}\right), h \in A. \tag{7}$$

Wróćmy do wzorów (1) i (3) określających wariancję estymatora globalnego cechy X przed i po połączeniu warstw. Skupmy uwagę na warstwach, w których po połączeniu nie nastąpiła zmiana wariancji badanej cechy X. Zauważmy, że jest spełniona zależność w postaci:

$$\begin{aligned}\sum_{h\in A}\frac{M_h^2}{m_h}\frac{M_h - m_h}{M_h - 1}\sigma_h^2 - \frac{M_h^2}{m'_h}\frac{M_h - m'_h}{M_h - 1}\sigma_h^2 = \\ = \sum_{h\in A}\frac{-\alpha M_h^3}{(M_h - 1)m_h(\sum_{k\in A} m_k - \alpha)}\sigma_h^2.\end{aligned} \tag{8}$$

Aby połączenie *i*-tej i *j*-tej wyjściowej warstwy dało pożądany efekt (nie większą wariancję estymatora), musi zostać spełniona następująca nierówność:

$$\sum_{h\in A}\frac{M_h^2}{m'_h}\frac{M_h-m'_h}{M_h-1}\sigma_h^2+\frac{\left(M_i+M_j\right)^2}{m'_{ij}}\frac{M_i+M_j-m'_{ij}}{M_i+M_j-1}\sigma'^2_{ij}\leq \\ \leq\sum_{h\in A}\frac{M_h^2}{m_h}\frac{M_h-m_h}{M_h-1}\sigma_h^2+\sum_{h\in\{i,j\}}\frac{M_h^2}{m_h}\frac{M_h-m_h}{M_h-1}\sigma_h^2. \tag{9}$$

Nierówność (9), korzystając z zależności (8), można zapisać w postaci:

$$\frac{\left(M_i+M_j\right)^2}{m'_{ij}}\frac{M_i+M_j-m'_{ij}}{M_i+M_j-1}\sigma'^2_{ij}\leq \\ \leq\sum_{h\in A}\frac{\left(m_i+m_j-m'_{ij}\right)M_h^3}{M_h-1}\sigma_h^2+\sum_{h\in\{i,j\}}\frac{M_h^2}{m_h}\frac{M_h-m_h}{M_h-1}\sigma_h^2. \tag{10}$$

Korzystając z zależności (6) i dokonując odpowiednich przekształceń, nierówność (10) można zapisać w następującej postaci:

$$m'^2_{ij}\left[\frac{\left(\sum_{k\in A}M_k\sqrt{\frac{M_h}{M_h-1}\sigma_h^2}\right)^2}{M_i+M_j}+\sum_{k\in A}\frac{M_k^3}{(M_k-1)m_k}\sigma_h^2+\sum_{k\in\{i,j\}}\frac{M_k^2}{m_k}\frac{M_k-m_k}{M_k-1}\sigma_k^2\right]- \\ -m'_{ij}\left[\left(\sum_{k\in A}M_k\sqrt{\frac{M_h}{M_h-1}\sigma_h^2}\right)^2+\sum_{k\in A}\frac{M_k^3\left(m_0+m_i+m_j\right)}{(M_k-1)m_k}\sigma_k^2+\right. \\ \left.+2m_0\sum_{k\in\{i,j\}}\frac{M_k^2}{m_k}\frac{M_k-m_k}{M_k-1}\sigma_k^2\right]+ \\ +\sum_{k\in A}\frac{M_k^3\left(m_i+m_j\right)m_0}{(M_k-1)m_k}\sigma_k^2+m_0^2\sum_{k\in\{i,j\}}\frac{M_k^2}{m_k}\frac{M_k-m_k}{M_k-1}\sigma_k^2 \\ \geq 0. \tag{11}$$

Dla zmniejszenia ilości zmiennych w nierówności (11), ale równocześnie skomplikowania zapisu, można skorzystać z zależności (2)[1].

Przykład

Rozpatrzmy sytuację, w której badacz dysponuje podziałem badanej populacji na 3 warstwy. Charakterystykę tych warstw zawiera tabela 1.

[1] Opisana forma nierówności nie została przedstawiona explicite właśnie ze względu na skomplikowanie zapisu.

Tabela 1

Przykładowe dane 1

Wyszczególnienie	Warstwa 1	Warstwa 2	Warstwa 3
Liczność warstwy	M	M	M
Średnia badanej cechy	μ	αμ	βμ
Wariancja badanej cechy	σ	σ	σ

Niech m_0 oznacza wielkość próby w badaniu. Wtedy, zgodnie z alokacją Neymana, wielkość próby w każdej z warstw będzie równa $\frac{m_0}{3}$.

Rozważmy połączenie warstwy 2 oraz warstwy 3. Dla omawianego przykładu nierówność (11) można uprościć do postaci:

$$m'^2_{23}\left(\frac{9M-1\frac{1}{2}m_0}{m_0}\right)-m'_{23}(18M-4m_0)+2m_0(4M-m_0)\geq 0. \qquad (12)$$

Dalej, przy założeniu, że $m_0=\frac{3}{5}M$, zbiór wartości m_{23}, dla których połączenie 2 i 3 warstwy będzie opłacalne, ma postać:

$$m_{23}\in\left[\frac{2}{3}m_0,\frac{34}{27}m_0\right]. \qquad (13)$$

Połączenie *2* i *3* warstwy spowoduje zmniejszenie wariancji estymatora globalnej wartości cechy *X,* gdy wartości średnie cechy *X* będą stosunkowo sobie bliskie. Przykład takiej sytuacji przedstawiono w tabeli 2.

Tabela 2

Przykładowe dane 2

Wyszczególnienie	Warstwa 1	Warstwa 2	Warstwa 3
Liczność warstwy	50	50	50
Średnia badanej cechy	1	2	2,5
Wariancja badanej cechy	1	1	1
Liczność próby	30		

2. Algorytm łączenia warstw

Podsumujmy zakres posiadanych informacji przed uruchomieniem algorytmu: dysponujemy podziałem badanej populacji na H warstw – podział wyjściowy; posiadamy dodatkowe informacje nt. każdej z warstw, które pozwalają na ocenę wariancji badanej cechy X w warstwie powstałej z połączenia dowolnych dwóch warstw wyjściowych. Przyjmiemy, że dla każdej z warstw jest znana liczba elementów $(\mathrm{M_h}, \mathrm{h}=1,\dots,\mathrm{H})$, średnia wartość zmiennej X $(\mu_\mathrm{h}, \mathrm{h}=1,\dots,\mathrm{H})$ oraz warian-

cja tej zmiennej $\left(\sigma_{\mathrm{h}}^2, \mathrm{h}=1,\ldots,\mathrm{H}\right)$. Celem łączenia warstw jest zmniejszenie wariancji estymatora $\sum_{\mathrm{h}=1}^{\mathrm{H}} \mathrm{M_h}\mu_{\mathrm{h}}$ danej wzorem (1). Dokonując odpowiednich przekształceń, otrzymujemy warunek (11), który określa opłacalność łączenia warstw.

Skonstruujmy odpowiedni algorytm, który pozwoli na określenie optymalnej liczby i charakterystyki warstw. Określmy zbiór:

$$D = \left\{\sigma'^2_{ij}\ i,j=1,\ldots H; i\neq j\right\}, \tag{14}$$

gdzie σ'^2_{ij} jest wariancją cechy X w nowo utworzonej warstwie, powstałej z połączenia wyjściowej *i*-tej oraz *j*-tej warstwy. Wariancja σ'^2_{ij} będzie wyznaczana na podstawie posiadanych danych, zgodnie ze wzorem (4).

Dysponując zbiorem wartości D, można określić liczności potencjalnie nowo utworzonych warstw $\left\{m'_{ij}, i,j=1,\ldots H; i\neq j\right\}$, a zatem zweryfikować nierówność (11). Prawdziwość (dla przyjętego m'_{ij}) nierówności (11) nie pozwala na wybór spośród H warstw pary, której połączenie najsilniej wpłynie na zmniejszenie wariancji estymatora globalnej wartości cechy X. W związku z tym proponuje się konstrukcję, opierając się na nierówności (11), funkcji celu pozwalającej na wybór pary warstw, która powinna zostać połączona:

$$f(i,j) = m'^2_{ij}\left[\frac{\left(\sum_{k\in A} M_k\sqrt{\frac{M_h}{M_h-1}\sigma_h^2}\right)^2}{M_i+M_j} + \sum\nolimits_{k\in A}\frac{M_k^3}{(M_k-1)m_k}\sigma_h^2 + \sum\nolimits_{k\in\{i,j\}}\frac{M_k^2}{m_k}\frac{M_k-m_k}{M_k-1}\sigma_k^2\right] -$$

$$-m'_{ij}\left[\left(\sum\nolimits_{k\in A} M_k\sqrt{\frac{M_h}{M_h-1}\sigma_h^2}\right)^2 + \sum\nolimits_{k\in A}\frac{M_k^3(m_0+m_i+m_j)}{(M_k-1)m_k}\sigma_k^2 + 2m_0\sum\nolimits_{k\in\{i,j\}}\frac{M_k^2}{m_k}\frac{M_k-m_k}{M_k-1}\sigma_k^2\right] +$$

$$+\sum\nolimits_{k\in A}\frac{M_k^3(m_i+m_j)m_0}{(M_k-1)m_k}\sigma_k^2 + m_0^2\sum\nolimits_{k\in\{i,j\}}\frac{M_k^2}{m_k}\frac{M_k-m_k}{M_k-1}\sigma_k^2. \tag{15}$$

Ze względu na typ funkcji celu (funkcja kwadratowa) oraz jej złożoność proponuje się następującą metodę postępowania:

1. Przekształćmy zbiór D z użyciem zależności (5) do zbioru:

$$M = \{m'_{ij}\ i, j = 1, \dots H; i \neq j\}. \tag{16}$$

2. W zbiorze M znajdźmy wartość minimalną $m'_{i_{min}j_{min}}$ oraz maksymalną $m'_{i_{max}j_{max}}$. Niech (i_0, j_0) oznaczają indeksy warstw, których połączenie jest najbardziej opłacalne dla wariancji estymatora $\sum_{h=1}^{H} M_h \mu_h$ ($(i_0, j_0) = (i_{min}, j_{min})$, gdy $f(i_{min}, j_{min}) > f(i_{max}, j_{max})$ lub $(i_0, j_0) = (i_{max}, j_{max})$, gdy $f(i_{min}, j_{min}) \leq f(i_{max}, j_{max})$).
3. Jeżeli $f(i_0, j_0) \geq 0$, to dokonujemy połączenia warstwy i_0-tej oraz j_0-tej i wykonujemy ponownie krok 1, jednak już dla *H-1* warstw. W pozostałych przypadkach algorytm kończy działanie.

Podsumowanie

Przestawione wyniki dają możliwość optymalizacji ilości i charakteru (zawartości) warstw, którymi często w praktyce badacz dysponuje jeszcze przed rozpoczęciem badania. Dzięki takiemu zabiegowi w efekcie badania uzyskuje się estymator globalny badanej cechy o mniejszej wariancji. Skonstruowany algorytm pozwala na automatyzację opisanych procedur.

Literatura

Dalenius T. (1957): *Sampling in Sweden. Contributions to Methods and Theories of Sample Survey Practice*. Almvist & Wiksells, Stockholm.

Fisher R. (1925): *Statisticial Methods for Research Workers*. Oliver and Boyd, Edynburg.

Neyman J. (1934): *On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection*. „Journal of the Royal Statistical Society", Vol. 97, No. 4, s. 558-625.

Wywiał J. (1992): *Statystyczna metoda reprezentacyjna w badaniach ekonomicznych*. Akademia Ekonomiczna im. Karola Adamieckiego, Katowice, s. 255-272.

OPTIMIZATION OF THE NUMBER OF THE STRATA FOR NEYMAN OPTIMAL ALLOCATION

Summary

In this paper the optimization of the number of the strata in a situation where the researcher has a previously known stratification of the population is presented. Usage of Neyman optimal allocation is assumed. It is also assumed that the researcher has the information about the number of elements, the mean value and the variance of the characteristic under study in each strata. Under these assumptions the condition of efficiency (in terms of reduction of the variance of the estimator) is defined. This condition is used for construction of the strata-merging algorithm.