Cena zł 25000,—

Indeks 381306
PL ISSN 0043-518X

# WIADOMOŚCI STATYSTYCZNE

GŁÓWNY URZĄD STATYSTYCZNY

POLSKIE TOWARZYSTWO STATYSTYCZNE

MIESIĘCZNIK
ROK XXXIX
WARSZAWA
WRZESIEŃ 1994

9(400)

w numerze m.in.:

*Bronisław LEDNICKI, Jacek WESOŁOWSKI*

# Lokalizacja próby pomiędzy subpopulacje

Problemy lokalizacji próby stanowią ważny aspekt teoretyczny i praktyczny metod reprezentacyjnych. Podstawowe wyniki w tej dziedzinie weszły do kanonu literatury, zob. np. Cochran (1977), Zasępa (1972). Niniejsza praca dotyczy lokalizacji stałoprecyzyjnej.

W praktyce badań reprezentacyjnych zdarza się często, że ich celem jest oszacowanie parametrów nie tylko dla całej populacji, ale również dla pewnej liczby subpopulacji, z których jest ona złożona. Np. w badaniach prowadzonych przez GUS na ogół nie ograniczamy się do uzyskania informacji dla Polski ogółem, ale postulujemy jednocześnie uzyskanie informacji dla województw czy działów gospodarki względnie innych subpopulacji. Nie określa się przy tym subpopulacji szczególnie preferowanych, przeciwnie zakłada się, że wyniki badania powinny mieć, w przybliżeniu, jednakową precyzję dla wszystkich subpopulacji.

Przyjęcie powyższego założenia powoduje, że lokalizacja próby pomiędzy subpopulacje nie może być ani proporcjonalna, ani też optymalna według metod właściwych dla losowania warstwowego. Metody te stawiają sobie za cel minimalizację błędów losowych dla całej populacji, zaś precyzja dla poszczególnych subpopulacji (warstw) jest problemem wtórnym i z reguły nie jest stała.

Opisana niżej metoda lokalizacji próby umożliwia uzyskanie wyników o jednakowej precyzji dla wszystkich subpopulacji przy założonej liczebności próby dla całej populacji.

### *Algorytm lokalizacji próby*

Populacja o liczebności $N$ składa się z $L$ subpopulacji o licznościach $N_j$, $j=1, ..., L$. Losujemy bez zwracania frakcję $f=\frac{n}{N}$. Rozważamy problem lokalizacji próby zapewniającej jednakową precyzję estymatorów średniej i wartości globalnej w każdej subpopulacji.

Niech $n_j$ oznacza liczbę elementów losowanych z $j$-tej subpopulacji, $j=1, ..., L$. Oczywiście

$$\sum_{j=1}^{L} n_j = fN = A$$

Ponadto warunek stałości precyzji estymatora $t_j$ w subpopulacji implikuje

$$\frac{N_j - n_j}{n_j N_j}\gamma_j^2 = c = const, \quad j=1, ..., L$$

gdzie $\gamma_j$ oznacza współczynnik zmienności w $j$-tej subpopulacji, a $3\sqrt{c}$ jest maksymalnym błędem względnym estymatora $t_j$ niezależnym od $j$ (przy założeniu reguły trzysigmowej), $j=1, ..., L$. Wobec tego

$$n_j = \frac{\gamma_j^2 N_j}{cN_j + \gamma_j^2}, \quad j=1, ..., L$$

W konsekwencji dla stałoprecyzyjnej lokalizacji próby wystarczy znaleźć stałą $c$. Z ostatniego wzoru wynika, że jest ona rozwiązaniem równania

$$g(x) = A \tag{1}$$

gdzie funkcja $g$ określona jest wzorem

$$g(x) = \sum_{j=1}^{L} \frac{\gamma_j^2 N_j}{xN_j + \gamma_j^2}, \quad x \in [0, \infty)$$

Równanie powyższe rozwiązujemy numerycznie stosując następujący algorytm:

1° Wybieramy dowolną liczbę dodatnią $c_0$ — punkt startu, dostatecznie małą liczbę $\varepsilon$ — dokładność oraz $k=0$.

2° Obliczamy $g(c_k)$.

3° Jeżeli $|g(c_k) - A| < \varepsilon$, to przyjmujemy $c = c_k$ i kończymy procedurę. W przeciwnym przypadku zwiększamy $k$ o 1 przyjmując

$$c_k = c_{k-1}\frac{g(c_{k-1})}{A} \tag{2}$$

i wracamy do punktu 2°.

Zbieżność algorytmu zapewnia następujące

**Twierdzenie 1.** *Dane są liczby dodatnie* $A, B_1, ..., B_L, D_1, ..., D_4, c_0$, *przy czym*

$$A < \sum_{j=1}^{L} B_j^{-1}$$

*Wtedy istnieje dokładnie jedna liczba dodatnia* $c$ *taka, że*

$$A = \sum_{j=1}^{L} \frac{D_j}{c + D_j B_i}$$

*Liczba ta jest granicą ciągu* $(c_k)$:

$$c_k = c_{k-1}\frac{A_{k-1}}{A}, \quad k=1, 2, ...$$

*gdzie*

$$A_k = \sum_{j=1}^{L} \frac{D_j}{c_k + D_j B_j}, \quad k=0, 1, ...$$

*oraz*

$$\lim_{k\to\infty} A_k = A$$

**Dowód.** Rozważmy funkcję $g$ określoną wzorem:

$$g(x) = \sum_{j=1}^{L} \frac{D_j}{x + D_j B_j}, \quad x \in [0, \infty)$$

Jest to funkcja malejąca, przy czym

$$g(0) > A$$

oraz

$$\lim_{x\to\infty} g(x) = 0$$

Z ciągłości $g$ wynika więc, iż istnieje dokładnie jeden punkt $c$, dla którego

$$g(c)=A$$

W zależności od wyboru punktu $c_0$ wyróżnimy trzy przypadki: (a) $c_0<c$, (b) $c_0>c$ oraz (c) $c_0=c$.

(a) Stosując indukcję wykażemy następujące ciągi nierówności dla dowolnego $k=0, 1, ...$

$$c_0<c_1<...<c_k<c \tag{3}$$

oraz

$$A_0>A_1>...>A_k>A \tag{4}$$

Dla $k=0$ z założenia $c_0<c$. Ponieważ $g$ jest malejąca, więc $A_0=g(c_0)>g(c)=A$. Załóżmy teraz, że (3) i (4) są spełnione dla $k=m-1$, gdzie $m$ jest dowolną liczbą naturalną. Wykażemy, że nierówności te zachodzą również dla $k=m$.

Z definicji $c_m$ i (4) mamy

$$c_m=c_{m-1}\frac{A_{m-1}}{A}>c_{m-1}$$

ponieważ $A_{m-1}>A$. Z drugiej strony zauważmy, że $A_j=1, ..., L$

$$\frac{c_{m-1}D_j}{c_{m-1}+D_jB_j}<\frac{cD_j}{c+D_jB_j}$$

ponieważ $c_{m-1}<c$. Wobec tego

$$c_{m-1}A_{m-1}<cA$$

a zatem $c_m<c$. Kończy to dowód (3).

Nierówności (4) są teraz prostą konsekwencją monotoniczności funkcji $g$. Ponieważ $c>c_m>c_{m-1}$, więc

$$g(c)<g(c_m)<g(c_{m-1})$$

czyli

$$A<A_m<A_{m-1}$$

Z twierdzenia o zbieżności ciągów monotonicznych ograniczonych wynika istnienie liczb $\hat{c}$ i $\hat{A}$ takich, że

$$\lim_{k\to\infty} c_k=\hat{c}$$
$$\lim_{k\to\infty} A_k=\hat{A}$$

oraz $g(\hat{c})=\hat{A}$. Przechodząc we wzorze rekurencyjnym dla $c_k$ obustronnie z $k$ do $\infty$ otrzymujemy

$$\hat{c}=\hat{c}\frac{\hat{A}}{A}$$

W konsekwencji $\hat{A}=A$ i różnowartościowość funkcji $g$ implikuje $\hat{c}=c$.

(b) Analogicznie jak w przypadku (a) można wykazać, że dla dowolnego $k=0, 1, ...$

$$c_0>c_1>...>c_k>c$$

oraz

$$A_0<A_1<...<A_k<A$$

(c) Przypadek trywialny.

Głębsza analiza przedstawionego dowodu prowadzi do następujących wniosków:

1. Proponowany algorytm znajdowania pierwiastka równania (1), gdzie $g: \mathbf{R}_+\to\mathbf{R}_+$ jest funkcją malejącą i $A<g(0)$ jest zbieżny jeśli funkcja $h$ określona wzorem

$$h(x)=xg(x), \quad x\in\mathbf{R}_+$$

jest rosnąca.

2. Analogiczny algorytm można zastosować do rozwiązania równania (1), w przypadku gdy funkcja $g$ jest rosnąca, $A>g(0)$ oraz funkcja $h$ określona wzorem

$$h(x)=\frac{g(x)}{x}, \quad x\in(0,\infty)$$

jest malejąca. Jedyna zmiana polega na zastąpieniu rekurencji (2) wzorem

$$c_k=c_{k-1}\frac{A}{g(c_{k-1})}$$

Tak zmodyfikowany algorytm można zastosować na przykład do rozwiązania równania

$$\sum_{j=1}^{L}(x+c_j)^{bj}=A$$

gdzie $A$, $b_j$, $c_j$ są danymi liczbami dodatnimi, $b_j<1$, $j=1, ..., L$, przy czym $A>\sum_{j=1}^{L} c_j^{bj}$.

3. Nietrudno podać przykład funkcji malejącej $g$, dla której istnieje punkt początkowy $c_0$ nie dający zbieżności proponowanej procedury. Niech $c$ będzie szukanym pierwiastkiem równania (1). Jeżeli $c_0$ wybierzemy w taki sposób, że:

$$g(c_0)g\left(c_0\frac{g(c_0)}{A}\right)=A^2 \tag{5}$$

to $c_{2k}=c_0$ i $c_{2k+1}=c_0\frac{g(c_0)}{A}$, $k=0, 1, ...$ Jako przykład rozważmy funkcję $g$ określoną wzorem

$$g(x)=\begin{cases}\frac{-8x+74}{5} & x\in[0,8]\\ \frac{-8x+134}{35} & x\in[8,\frac{67}{4}]\\ 0 & x\geqslant\frac{67}{4}\end{cases}$$

Szukamy rozwiązania równania $g(x)=2$. Zastosowanie algorytmu z $c_0=3$ prowadzi do następującego ciągu $c_k$:

$$c_{2k}=3, \quad c_{2k+1}=15, \quad k=0, 1, ...$$

Oczywiście rozwiązaniem jest punkt $c=8$. Zauważmy, że w podanym przykładzie zachodzi równość (5).

***Przykład zastosowania algorytmu***

Omówiony algorytm został wykorzystany m.in. do wstępnej lokalizacji próby pierwszego stopnia w reprezentacyjnym spisie rolnym. Zgodnie z przyjętymi założeniami liczebność próby pierwszego stopnia wynosiła około 40% populacji rolniczych obwodów spisowych w skali całego kraju. Dokonując lokalizacji próby pomiędzy województwa, traktowane jako subpopulacje, przyjęto również założenie, że współczynniki zmienności są jednakowe we wszystkich województwach. Prowadzi to do uproszczonej wersji równania (1)

$$\sum_{j=1}^{49}\frac{N_j}{xN_j+1}=0{,}4N$$

gdzie:

$N=68253$ — liczba rolniczych obwodów spisowych w całej Polsce,

$N_j$ — liczba rolniczych obwodów spisowych w $j$-tym województwie, $j=1, ..., 49$.

Stosując opisany algorytm przyjęto punkt startu $c_0=\frac{49}{N}=$ $=0{,}0007179172$ oraz dokładność $\varepsilon=2$. Wyniki kolejnych iteracji zamieszczone są w tablicy 1. Należy podkreślić, że algorytm realizowano wielokrotnie dla różnych wartości $c_0$ i $\varepsilon$ każdorazowo otrzymując założoną dokładność po nie więcej niż dwudziestu iteracjach. Obserwacja ta prowadzi do praktycznego wniosku o szybkiej zbieżności proponowanego algorytmu. Odpowiedni wynik teoretyczny dotyczący szybkości zbieżności pozostaje problemem otwartym.

Po zaokrągleniu obliczonych wartości $n_j$, $j=1, ..., 49$, do liczb całkowitych próba liczyła 27298 obwodów. W tablicy 2 podano ostateczne wyniki dla wszystkich województw: dla $j$-tego województwa $n_j$ — liczba losowanych obwodów, $N_j$ — całkowita liczba obwodów, $f_j=\frac{n_j}{N_j}$ — frakcja losowania, $c(j)=\frac{1-f_j}{n_j}$ — rzeczywista poprawka na bezzwrotność losowania, $j=1, ..., 49$. Wartości w ostatniej kolumnie różnią się nieznacznie od $c_{10}$ wyliczonego w tablicy 1 ze względu na zaokrąglenia liczby obwodów w poszczególnych województwach do liczb całkowitych podane w trzeciej kolumnie. Wielkości te dotyczą wstępnej lokalizacji próby. Schemat losowania próby pierwszego stopnia w spisie rolnym w 1994 r. był bardziej złożony i zakładał np. warstwowanie obwodów w poszczególnych województwach. Z tego względu ostateczne wartości $n_j$, $j=1, ..., 49$, różnią się nieco od podanych w tablicy 2.

Analogiczną metodę zastosować można przy takiej lokalizacji próby w subpopulacjach, która zapewnia stałą wariancję $c$ średniej:

$$\frac{N_j-n_j}{n_jN_j}S_j^2=c, \quad j=1, ..., L$$

lub wartości globalnej:

$$\frac{N_j\,(N_j-n_j)}{n_j}S_j^2=c, \quad j=1, ..., L$$

gdzie $S_j^2$ oznacza wariancję w $j$-tej subpopulacji, $j=1, ..., L$. Problemy te prowadzą do poszukiwania rozwiązania równania (1) z funkcją $g$ określoną wzorem

$$g(x)=\sum_{j=1}^{L}\frac{S_j^2N_j}{xN_j+S_j^2}$$

lub

$$g(x)=\sum_{j=1}^{L}\frac{N_j^2S_j^2}{x+N_jS_j^2}$$

odpowiednio. Zgodnie z przedstawionym twierdzeniem można do tego celu wykorzystać omawiany algorytm.

TABL. 1. PRZEBIEG ITERACJI DLA $c_0=0{,}0007179172$ i $\varepsilon=2$

| $k$ | $c_k$ | $f(c_k)$ |
|---|---|---|
| 0 | 7,179172E-04 | 32770,01 |
| 1 | 8,619215E-04 | 29788,67 |
| 2 | 9,406664E-04 | 28383,00 |
| 3 | 9,781621E-04 | 27760,49 |
| 4 | 9,948436E-04 | 27492,47 |
| 5 | 1,002041E-03 | 27378,45 |
| 6 | 1,005104E-03 | 27330,23 |
| 7 | 1,006402E-03 | 27309,86 |
| 8 | 1,006949E-03 | 27301,26 |
| 9 | 1,007180E-03 | 27297,64 |
| 10 | 1,007278E-03 | 27296,11 |

TABL. 2. STAŁOPRECYZYJNA LOKALIZACJA PRÓBY SPISOWYCH OBWODÓW ROLNICZYCH W WOJEWÓDZTWACH

| $j$ | Województwo | $n_j$ | $N_j$ | $f_j$ | $c(j)$ |
|---|---|---|---|---|---|
| 1 | Warszawskie | 570 | 1337 | 0,426 | 0,001006 |
| 2 | Bialskopodlaskie | 504 | 1023 | 0,493 | 0,001007 |
| 3 | Białostockie | 641 | 1812 | 0,354 | 0,001008 |
| 4 | Bielskie | 651 | 1893 | 0,344 | 0,001008 |
| 5 | Bydgoskie | 613 | 1601 | 0,383 | 0,001007 |
| 6 | Chełmskie | 435 | 775 | 0,561 | 0,001009 |
| 7 | Ciechanowskie | 592 | 1469 | 0,403 | 0,001008 |
| 8 | Częstochowskie | 625 | 1686 | 0,371 | 0,001007 |
| 9 | Elbląskie | 428 | 752 | 0,569 | 0,001007 |
| 10 | Gdańskie | 537 | 1168 | 0,460 | 0,001006 |
| 11 | Gorzowskie | 454 | 835 | 0,544 | 0,001005 |
| 12 | Jeleniogórskie | 433 | 768 | 0,564 | 0,001007 |
| 13 | Kaliskie | 605 | 1548 | 0,391 | 0,001007 |
| 14 | Katowickie | 775 | 3541 | 0,219 | 0,001008 |
| 15 | Kieleckie | 741 | 2918 | 0,254 | 0,001007 |
| 16 | Konińskie | 562 | 1294 | 0,434 | 0,001007 |
| 17 | Koszalińskie | 441 | 795 | 0,555 | 0,001010 |
| 18 | Krakowskie | 593 | 1472 | 0,403 | 0,001007 |
| 19 | Krośnieńskie | 595 | 1482 | 0,401 | 0,001006 |
| 20 | Legnickie | 389 | 640 | 0,608 | 0,001008 |
| 21 | Leszczyńskie | 464 | 872 | 0,532 | 0,001008 |
| 22 | Lubelskie | 656 | 1930 | 0,340 | 0,001006 |
| 23 | Łomżyńskie | 535 | 1158 | 0,462 | 0,001006 |
| 24 | Łódzkie | 289 | 408 | 0,708 | 0,001009 |
| 25 | Nowosądeckie | 651 | 1891 | 0,344 | 0,001007 |
| 26 | Olsztyńskie | 557 | 1268 | 0,439 | 0,001007 |
| 27 | Opolskie | 650 | 1883 | 0,345 | 0,001007 |
| 28 | Ostrołęckie | 566 | 1318 | 0,429 | 0,001008 |
| 29 | Pilskie | 427 | 748 | 0,571 | 0,001005 |
| 30 | Piotrkowskie | 619 | 1641 | 0,377 | 0,001006 |
| 31 | Płockie | 553 | 1248 | 0,443 | 0,001007 |
| 32 | Poznańskie | 593 | 1474 | 0,402 | 0,001008 |
| 33 | Przemyskie | 514 | 1067 | 0,482 | 0,001008 |
| 34 | Radomskie | 665 | 2011 | 0,331 | 0,001006 |
| 35 | Rzeszowskie | 670 | 2059 | 0,325 | 0,001007 |
| 36 | Siedleckie | 702 | 2391 | 0,294 | 0,001006 |
| 37 | Sieradzkie | 546 | 1214 | 0,450 | 0,001008 |
| 38 | Skierniewickie | 541 | 1187 | 0,456 | 0,001006 |
| 39 | Słupskie | 357 | 558 | 0,640 | 0,001009 |
| 40 | Suwalskie | 562 | 1297 | 0,433 | 0,001008 |
| 41 | Szczecińskie | 461 | 862 | 0,535 | 0,001009 |
| 42 | Tarnobrzeskie | 626 | 1697 | 0,369 | 0,001008 |
| 43 | Tarnowskie | 658 | 1951 | 0,337 | 0,001007 |
| 44 | Toruńskie | 498 | 998 | 0,499 | 0,001006 |
| 45 | Wałbrzyskie | 502 | 1014 | 0,495 | 0,001006 |
| 46 | Włocławskie | 534 | 1154 | 0,463 | 0,001006 |
| 47 | Wrocławskie | 586 | 1431 | 0,410 | 0,001008 |
| 48 | Zamojskie | 621 | 1660 | 0,374 | 0,001008 |
| 49 | Zielonogórskie | 511 | 1054 | 0,485 | 0,001008 |

Należy zwrócić uwagę, że zagadnienia stałoprecyzyjnej lokalizacji próby w subpopulacjach mimo ich znaczenia dla praktyki badań reprezentacyjnych w zasadzie nie były do tej pory rozważane w literaturze. Jedynie w monografii Cochran (1977) (roz. 4.9) przedstawiony został podobny problem określenia całkowitej wielkości próby tak, aby estymator średniej w każdej z $k$ subpopulacji miał jednakową wariancję $V$. Przyjęte tam upraszczające założenia prowadzą do wniosku, że wielkość próby powinna być w przybliżeniu równa $kn$, gdzie $n$ oznacza wielkość próby zapewniającej wariancję średniej równą $V$ w całej populacji.

**mgr Bronisław Lednicki** — *ZBSE*, **Jacek Wesołowski** — *Politechnika Warszawska*

LITERATURA

[1] Cochran W. G. (1977): *Sampling Techniques* Wiley, New York
[2] Zasępa R. (1972): *Metoda reprezentacyjna*, PWE, Warszawa