

STUDIA METODOLOGICZNE

Tomasz JÓZEFOWSKI, Marcin SZYMKOWIAK

Estymatory kalibracyjne w badaniach statystycznych

W badaniach prowadzonych przez urzędy statystyczne wielu krajów jednym z głównych źródeł błędów nielosowych są braki odpowiedzi. W przypadku gdy występują one na dużą skalę, mogą wpłynąć negatywnie na jakość estymacji. Dlatego szczególną rolę przypisuje się różnego rodzaju technikom, które mają niwelować ujemny wpływ nieuzyskania odpowiedzi na wyniki badań. Jedną

z takich metod jest kalibracja, czyli procedura estymacji parametrów w populacji generalnej. Polega ona na skorygowaniu wag wynikających ze schematu losowania próby z wykorzystaniem zmiennych pomocniczych tak, aby spełnione były odpowiednie równania kalibracyjne. W tradycyjnym podejściu, na etapie poszukiwania wag kalibracyjnych zakłada się, że powinny być one bliskie — w sensie przyjętej funkcji odległości — wyjściowym wagom wynikającym ze schematu losowania próby (Särndal, Lundström, 2005).

Kalibracja jako metoda szacowania parametrów w populacji generalnej jest techniką statystyczną wykorzystywaną w badaniach reprezentacyjnych. W badaniach pełnych występuje również problem braków danych, naturalne zatem wydaje się odpowiednie przeważenie danych. Okazuje się, że kalibracja może być metodą wykorzystywaną w badaniach pełnych (Wallgren, Wallgren, 2007). Punkt wyjścia do jej zastosowania stanowi odpowiedni sposób ustalania wag.

ESTYMATOR KALIBRACYJNY WARTOŚCI GLOBALNEJ

Założmy, że celem badania jest oszacowanie wartości globalnej zmiennej y , określonej wzorem:

$$Y = \sum_{i=1}^N y_i \quad (1)$$

gdzie:

y_i — wartość zmiennej y dla i -tej jednostki badania, $i = 1, \dots, N$,

N — liczebność populacji.

Zakładając, że z N -elementowej populacji $U = \{1, \dots, N\}$ losujemy zgodnie z określonym schematem losowania n -elementową próbę $s \subseteq U$, klasycznym estymatorem wartości globalnej (1) jest znany z metody reprezentacyjnej estymator Horvitz-Thompsona:

$$\hat{Y}_{HT} = \sum_{i \in s} d_i y_i = \sum_{i=1}^n d_i y_i \quad (2)$$

gdzie $d_i = \frac{1}{\pi_i}$ jest wagą odpowiadającą jednostce i , a π_i oznacza prawdopodobieństwo inkluzji i -tej jednostki do próby, tzn. $\pi_i = P(i \in s)$ dla $i = 1, \dots, N$.

Zazwyczaj w sytuacji, gdy nie są znane wszystkie wartości zmiennej y dla jednostek wylosowanych do próby (np. na skutek braków odpowiedzi), estymator Horvitz-Thompsona charakteryzuje się znacznym obciążeniem i wariancją. Wynika to na ogół z faktu, że braki odpowiedzi nie mają charakteru czysto lo-

sowego, a powstałe błędy wynikają z różnic pomiędzy respondentami i niere-spondentami. W konsekwencji ważona suma (2) jest zazwyczaj niedoszacowana w stosunku do wartości globalnej (1).

Zgodnie z ideą kalibracji opisaną przez Särndala i Lundströma (2005), wagi wynikające ze schematu losowania jednostek do próby powinny być skorygowane, aby zrekompensować utratę informacji związaną z brakami danych. Poszukuje się zatem nowych wag w_i dla wszystkich jednostek w próbie, dla których mamy informację o zmiennej y . Punkt wyjścia stanowi odpowiednie skonstruowanie wektora zmiennych pomocniczych, który stanowić będzie podstawę wyznaczania wag kalibracyjnych.

Przyjmując, że $x_1, \dots, x_k$ oznaczają zmienne pomocnicze, które wykorzystane zostaną przy wyznaczaniu wag kalibracyjnych, a X_j oznacza wartość globalną zmiennej x_j , $j=1, \dots, k$, otrzymujemy:

$$X_j = \sum_{i=1}^N x_{ij} \quad (3)$$

gdzie x_{ij} oznacza wartość j -tej zmiennej pomocniczej dla i -tej jednostki badania.

Poszukiwanie wag kalibracyjnych w ujęciu matematycznym można przedstawić w następujący sposób:

- minimalizacja funkcji odległości:

$$D(w, d) = \sum_{i=1}^m d_i G\left(\frac{w_i}{d_i}\right) \rightarrow \min \quad (4)$$

- równania kalibracyjne:

$$\sum_{i=1}^m w_i x_{ij} = X_j \quad j = 1, \dots, k \quad (5)$$

- warunki ograniczające:

$$L \leq \frac{w_i}{d_i} \leq U \quad (6)$$

gdzie:

$L < 1$ i $U > 1$, $i=1, \dots, m$,

m — wielkość zbioru respondentów, czyli liczba jednostek, dla których znana jest wartość zmiennej y w próbie.

Zakłada się przy tym, że G jest funkcją ściśle wypukłą i dwukrotnie różniczkowalną, spełniającą warunki:

- $G(\cdot) \geq 0$,
- $G(1)=0$,
- $G'(1)=0$,
- $G''(1)=1$.

W odniesieniu do warunku ograniczającego najczęściej przyjmuje się, że $L=0$, a U jest dowolną liczbą większą od 1 (np. $U=1,5$).

Przedstawiona we wzorze (4) funkcja $D(\cdot)$ mierzy odległość między wagami wynikającymi ze schematu losowania próby a wagami kalibracyjnymi. Najczęściej na potrzeby badań, celem wyznaczenia postaci analitycznej funkcji $D(\cdot)$, przyjmuje się, że $G(\cdot)$ jest funkcją kwadratową postaci:

$$G(x) = \frac{1}{2}(x-1)^2 \quad (7)$$

W konsekwencji funkcja odległości $D(\cdot)$ może być przedstawiona następująco:

$$D(w, d) = \sum_{i=1}^m d_i G\left(\frac{w_i}{d_i}\right) = \sum_{i=1}^m d_i \frac{1}{2} \left(\frac{w_i}{d_i} - 1\right)^2 = \frac{1}{2} \sum_{i=1}^m \frac{(w_i - d_i)^2}{d_i}. \quad (8)$$

Przyjmując tę funkcję odległości do wyznaczania wag w_i można zdefiniować estymator kalibracyjny wartości globalnej (1).

Definicja. Estymatorem kalibracyjnym wartości globalnej (1) przy znanym wektorze wartości globalnych $X = \left(\sum_{i=1}^N x_{i1}, \sum_{i=1}^N x_{i2}, \dots, \sum_{i=1}^N x_{ik} \right)^T$ zmiennych pomocniczych jest:

$$\hat{Y}_X = \sum_{i=1}^m w_i y_i \quad (9)$$

gdzie wektor wag $w = (w_1, \dots, w_m)^T$ jest rozwiązaniem zadania optymalizacyjnego

$$w = \operatorname{argmin}_v D(v, d) \quad (10)$$

przy warunku

$$X = \tilde{X} \quad (11)$$

gdzie

$$\tilde{X} = \left(\sum_{i=1}^m w_i x_{i1}, \sum_{i=1}^m w_i x_{i2}, \dots, \sum_{i=1}^m w_i x_{ik} \right)^T \quad (12)$$

Korzystając z metody czynników nieoznaczonych Lagrange'a można udowodnić¹ twierdzenie, które przy funkcji odległości (8) podaje postać wag kalibracyjnych w_i .

Twierdzenie. Jeżeli macierz $\sum_{i=1}^m d_i x_i \otimes x_i^T$ jest nieosobliwa, to rozwiązaniem zadania minimalizacji funkcji odległości (10), przy warunku (11), jest wektor wag kalibracyjnych postaci $w = (w_1, \dots, w_m)^T$, którego składowe spełniają równanie:

$$w_i = d_i + d_i \left(X - \hat{X} \right)^T \left(\sum_{i=1}^m d_i x_i \otimes x_i^T \right)^{-1} x_i \quad (13)$$

przy czym

$$\hat{X} = \left(\sum_{i=1}^m d_i x_{i1}, \sum_{i=1}^m d_i x_{i2}, \dots, \sum_{i=1}^m d_i x_{ik} \right)^T \quad (14)$$

a

$$x_i = (x_{i1}, \dots, x_{ik})^T \quad (15)$$

jest wektorem złożonym z wartości wszystkich k zmiennych pomocniczych dla i -tego respondenta, $i=1, \dots, m$, $\otimes$ oznacza symbol iloczynu Kroneckera macierzy.

ALGORYTM WYZNACZANIA WAG KALIBRACYJNYCH W BADANIACH PEŁNYCH

W badaniach pełnych istnieje również problem braków danych, więc zastosowanie właściwych metod ważenia jest szczególnie ważne z punktu widzenia końcowych wyników. Możliwość wykorzystania podejścia kalibracyjnego w badaniach pełnych, z racji tego że w ich przypadku nie dokonuje się losowania próby, wynika z odpowiedniego ustalenia wyjściowych wag. W dalszej czę-

¹ Dowody twierdzenia można znaleźć w pracy M. Szymkowiaka (2007).

ści artykułu przedstawimy algorytm wyznaczania wag kalibracyjnych oraz egemplifikację wykorzystującą przykładowy zbiór danych².

Założmy, że celem badania jest utworzenie dwuwymiarowej tabeli kontyngencji. Zawierać ma ona informacje o wartościach globalnych interesującej nas cechy Y w zależności od wariantów dwóch zmiennych X i Z (np. liczba osób w województwie o określonym stanie cywilnym w zależności od płci i grupy wiekowej, liczba bezrobotnych w zależności od wykształcenia i klasy miejscowości zamieszkania itd.). Zakładamy zatem, że wymiary konstruowanej tablicy wyznaczone są przez warianty zmiennych X i Z , przyjmujących odpowiednio k i n poziomów.

TABL. 1. DWUWYMIAROWA TABELA KONTYNGENCJI

X/Z	Z_1	Z_2	...	Z_n
X_1	$\sum_{i: X(i)=x_1, Z(i)=z_1} y_i$	$\sum_{i: X(i)=x_1, Z(i)=z_2} y_i$	...	$\sum_{i: X(i)=x_1, Z(i)=z_n} y_i$
X_2	$\sum_{i: X(i)=x_2, Z(i)=z_1} y_i$	$\sum_{i: X(i)=x_2, Z(i)=z_2} y_i$	...	$\sum_{i: X(i)=x_2, Z(i)=z_n} y_i$
$\vdots$	$\vdots$	$\vdots$	...	$\vdots$
X_k	$\sum_{i: X(i)=x_k, Z(i)=z_1} y_i$	$\sum_{i: X(i)=x_k, Z(i)=z_2} y_i$	...	$\sum_{i: X(i)=x_k, Z(i)=z_n} y_i$

Źródło: opracowanie własne.

Poszczególne części tabl. 1 zawierają wartości globalne zmiennej Y dla i -tego i j -tego wariantu zmiennych X i Z , czyli $\sum_{i: X(i)=x_k, Z(i)=z_n} y_i$, $i=1, 2, \dots, k$, $l=1, 2, \dots, n$.

W przypadku niektórych wariantów zmiennych X i Z nie będziemy jednak dysponowali informacjami o wartościach zmiennej Y dla wszystkich jednostek (braki danych). Wartość globalna będzie zazwyczaj niedoszacowana. Podejście opisane w artykule ma stanowić swego rodzaju remedium na istniejące w badaniach pełnych braki danych i oparte jest na metodologii kalibracji, której pionierami byli Deville i Särndal (1992).

Idea wykorzystania estymatorów kalibracyjnych w badaniach pełnych polega na odpowiednim skonstruowaniu wag, które podlegać będą kalibracji. W badaniach częściowych wagi podlegające temu procesowi odpowiadają wagom wynikającym ze schematu losowania próby d_i . W badaniach pełnych, które obejmują swym zasięgiem całą populację, wagi takie należy utworzyć w sposób sztuczny, a następnie należy je wykalibrować zgodnie z wcześniej opisaną metodologią. Algorytm konstruowania tablicy kontyngencji, który uwzględnić będzie występowanie braków danych w odniesieniu do interesującej badacza cechy Y , można opisać w trzech zasadniczych punktach:

1. Ustalenie sztucznych — wyjściowych wag d_i , które podlegać będą procesowi kalibracji.

² Algorytm ten został opisany również w publikacji Wallgrenów (2007).

- Wybór zmiennych pomocniczych $X_1, X_2, \dots, X_k$, które wykorzystane zostaną do konstrukcji wektora X złożonego z ich wartości globalnych.
- Ustanowienie nowych wag kalibracyjnych i zbudowanie odpowiedniej tablicy kontyngencji.

W kroku pierwszym opisanego algorytmu wagi d_i ustala się w sposób sztuczny, zgodnie z następującą zasadą:

$$d_i = \begin{cases} 1 & \text{dla } i \in r \\ 0 & \text{dla } i \notin r \end{cases} \quad (16)$$

gdzie r oznacza zbiór respondentów, a więc zbiór jednostek, dla których znana jest wartość zmiennej Y .

W kroku drugim dokonuje się wyboru zmiennych pomocniczych. Zalecane jest, aby były to zmienne, o których mamy pełną informację dla wszystkich jednostek z badania. Wykorzystuje się tutaj najczęściej różnego rodzaju zmienne demograficzne (płeć, wiek, klasa miejscowości zamieszkania itd.). Następnie, wykorzystując zestaw zmiennych pomocniczych, buduje się wektor X złożony z ich wartości globalnych. Wektor ten będzie zawierał, w zależności od rodzaju zmiennych, m.in. informacje o liczbie mężczyzn i kobiet, osób w określonej grupie wiekowej itd. Wagi kalibracyjne wyznaczane będą zatem w taki sposób, aby ich sumy odpowiednich jednostek odpowiadały wartościom globalnym zmiennych pomocniczych. Przykładowo, obliczając sumę wag kalibracyjnych dla wszystkich mężczyzn z badania uzyskamy wartość, która odpowiadać będzie łącznej liczbie mężczyzn. Tak określone wagi kalibracyjne odtwarzać będą wartości globalne zmiennych pomocniczych. Z dużym prawdopodobieństwem można zatem przyjąć, że po ich zastosowaniu do zmiennej Y uzyskamy oszacowaną wartość globalną bliższą prawdziwej wartości. W ostatnim, trzecim kroku, wykorzystując twierdzenie (13), wyznaczamy wagi kalibracyjne, a następnie konstruujemy odpowiednią tablicę kontyngencji, która zawierać będzie odpowiednio przeważone dane.

PRZYKŁAD EMPIRYCZNY

Założmy, że celem badania jest stworzenie tablicy kontyngencji zawierającej informacje o liczbie osób w zależności od ich stanu cywilnego i grupy wieku. Na potrzeby przykładu przyjęto, że dysponujemy informacjami z hipotetycznego rejestru, który zawiera m.in. zmienne: płeć (k — kobieta, m — mężczyzna), klasa (klasa miejscowości zamieszkania: m — miasto, w — wieś), wiek (wyróżniono dwie grupy wiekowe: 15—34 lata i 35 lat i więcej) oraz stan cywilny (k — kawalerowie/panny, z — żonaci/zamężne, w — wdowcy/wdowy, r — rozwiedzeni/rozwidzione).

Ze względu na braki danych występujące w rejestrze, zarówno w odniesieniu do stanu cywilnego jak i wieku, uzyskana tablica kontyngencji nie będzie właściwa.

PRZYKŁADOWY REJESTR OSÓB

Lp.	Płeć	Klasa	Wiek	Stan cywilny	d_i	x_1	x_2	x_3	w_i
1	m	m	.	k	0	0	1	1	0
2	k	w	15—34	z	1	1	0	0	1,1738
3	k	w	35 i więcej	.	0	1	0	0	0
4	m	w	35 i więcej	w	1	0	1	0	1,5364
5	m	m	15—34	k	1	0	1	1	1,4844
6	m	m	.	r	0	0	1	1	0
7	k	m	35 i więcej	r	1	1	0	1	1,1218
8	k	w	35 i więcej	w	1	1	0	0	1,1738
9	k	m	15—34	z	1	1	0	1	1,1218
10	m	m	15—34	z	1	0	1	1	1,4844
11	k	w	15—34	z	1	1	0	0	1,1738
12	m	w	35 i więcej	r	1	0	1	0	1,5364
13	k	m	15—34	k	1	1	0	1	1,1218
14	m	w	15—34	.	0	0	1	0	0
15	m	m	15—34	z	1	0	1	1	1,4844
16	k	w	35 i więcej	z	1	1	0	0	1,1738
17	k	m	15—34	k	1	1	0	1	1,1218
18	m	w	35 i więcej	k	1	0	1	0	1,5364
19	m	m	35 i więcej	.	0	0	1	1	0
20	k	m	35 i więcej	z	1	1	0	1	1,1218
21	m	m	15—34	z	1	0	1	1	1,4844
22	k	w	15—34	r	1	1	0	0	1,1738
23	m	m	15—34	r	1	0	1	1	1,4844
24	k	w	15—34	r	1	1	0	0	1,1738
25	m	m	.	z	0	0	1	1	0
26	k	w	.	z	0	1	0	0	0
27	m	m	35 i więcej	z	1	0	1	1	1,4844
28	k	w	35 i więcej	k	1	1	0	0	1,1738
29	m	m	35 i więcej	k	1	0	1	1	1,4844
30	k	w	15—34	w	1	1	0	0	1,1738

Źródło: jak przy tabl. 1.

Łączna liczba osób w tabl. 2 nie sumuje się do liczby wszystkich osób znajdujących się w rejestrze, wynoszącej 30. Zachodzi zatem potrzeba odpowiedniego przeważenia zawartych w tablicy danych tak, aby wartości w poszczególnych komórkach sumowały się do liczby osób w całym rejestrze.

TABL. 2. LUDNOŚĆ WEDŁUG STANU CYWILNEGO I WIEKU

Stan cywilny	Ogółem	15—34 lata	35 lat i więcej
Ogółem	23	13	10
Kawalerowie/panny	6	3	3
Rozwiedzeni/rozwidzione	5	3	2
Wdowcy/wdowy	3	1	2
Żonaci/zamężne	9	6	3

Źródło: jak przy tabl. 1.

Punktem wyjścia do stworzenia odpowiednich wag jest, zgodnie z opisanym algorytmem, utworzenie sztucznych wag wyjściowych, które podlegać będą

kalibracji. Tworzona tablica będzie mieć dwa wymiary, odpowiadające stanowi cywilnemu oraz grupie wieku. Wagi należy więc ustanowić w następujący sposób:

$$d_i = \begin{cases} 1 & \text{dla } i \in r \\ 0 & \text{dla } i \notin r \end{cases}$$

gdzie r oznacza zbiór osób, dla których znany jest stan cywilny oraz grupa wieku.

W dalszym etapie wyznaczania wag należy wybrać odpowiednie zmienne pomocnicze. Jest to szczególnie ważny krok w podejściu kalibracyjnym, gdyż właściwy dobór zmiennych wspomagających w istotny sposób rzutuje na jakość wyników końcowych. W praktycznych zastosowaniach wykorzystuje się na tym etapie zmienne, dla których znane są wartości na poziomie całego rejestru. Są to zazwyczaj cechy demograficzne odnoszące się do analizowanych osób. W rozważanym przykładzie w charakterze zmiennych pomocniczych zdecydowano się wybrać płeć oraz klasę miejscowości zamieszkania. Ostatecznie zmienne pomocnicze określono w następujący sposób:

$$x_{i1} = \begin{cases} 1 & \text{jeżeli } i\text{-ta osoba jest kobietą} \\ 0 & \text{jeżeli } i\text{-ta osoba jest mężczyzną} \end{cases}$$

$$x_{i2} = \begin{cases} 1 & \text{jeżeli } i\text{-ta osoba jest mężczyzną} \\ 0 & \text{jeżeli } i\text{-ta osoba jest kobietą} \end{cases}$$

$$x_{i3} = \begin{cases} 1 & \text{jeżeli } i\text{-ta osoba mieszka w mieście} \\ 0 & \text{jeżeli } i\text{-ta osoba mieszka na wsi} \end{cases}$$

Utworzone w ten sposób trzy zmienne pomocnicze zagwarantują sumowalność odpowiednich wag kalibracyjnych w_i do faktycznej liczby kobiet i mężczyzn w rejestrze, a także do liczby osób zamieszkujących miasto. Z formalnego punktu widzenia do opisu płci wystarczyłaby jedna dychotomiczna zmienna pomocnicza, która przyjmowałaby wartość 1 w sytuacji, gdy badana osoba jest kobietą oraz wartość 0, gdy rozważana osoba jest mężczyzną. Takie podejście stosowane jest np. przy budowie modeli regresji.

W podejściu kalibracyjnym przyjęcie tylko jednej (np. pierwszej) zmiennej dychotomicznej zagwarantowałoby jedynie sumowalność wag kalibracyjnych do liczby kobiet w rejestrze. Wagi kalibracyjne dla mężczyzn, w przypadku utworzenia tylko jednej zmiennej dychotomicznej, nie musiałyby się jednak sumować do łącznej liczby mężczyzn w rejestrze. Byłoby to konsekwencją braku odpowiedniego równania kalibracyjnego, odpowiedzialnego za sumowalność wag dla mężczyzn do łącznej liczby mężczyzn w rejestrze. Utworzenie dwóch zmien-

nych dychotomicznych w odniesieniu do płci zagwarantuje sumowalność wag dla kobiet i mężczyzn odpowiednio do łącznej rzeczywistej liczby kobiet i mężczyzn. Wymusi to dodatkowo sumowalność wag kalibracyjnych do łącznej liczby wszystkich osób w rejestrze.

Gdy wszystkie wagi sumować się będą do łącznej liczby osób w rejestrze, w odniesieniu do klasy miejscowości zamieszkania konieczne należy utworzyć tylko jedną zmienną dychotomiczną. W naszym przypadku wagi kalibracyjne dla mieszkańców miast będą się sumowały do łącznej liczby osób z rejestru zamieszkujących miasta. Sumowalność wag dla wszystkich osób w rejestrze i wag dla osób zamieszkujących miasta do łącznej liczby osób z miast wymusi sumowalność wag dla mieszkańców wsi do łącznej liczby osób pochodzących ze wsi. Konieczność utworzenia tylko jednej zmiennej dychotomicznej w odniesieniu do zmiennej klasa miejscowości zamieszkania wynika również z postaci wag kalibracyjnych, danych wzorem (13). Utworzenie dwóch zmiennych dychotomicznych w odniesieniu do zmiennej klasa miejscowości zamieszkania doprowadziłoby do współliniowości zmiennych pomocniczych, co w konsekwencji spowodowałoby, że macierz $\sum_{i=1}^m d_i x_i \otimes x_i^T$ byłaby osobliwa.

We wzorze (13) sumowanie odbywa się po zbiorze respondentów. Formalnie można byłoby rozszerzyć sumowanie na zbiór wszystkich jednostek z rejestru (N), gdyż ostatecznie do wyrażenia $\sum_{i=1}^N d_i x_i \otimes x_i^T$ wejdą i tak te elementy, dla których $d_i=1$.

W kolejnym kroku tworzy się wektor wartości globalnych X , przechowujący, przy tak określonych zmiennych pomocniczych, informacje z rejestru o łącznej liczbie kobiet, mężczyzn i osób zamieszkujących miasta. W rozważanym przykładzie wektor ten będzie miał postać: $X=(15,15,16)^T$. Podobnie stworzymy wektor $\hat{X}$, który odpowiednio przechowuje informacje z rejestru o liczbie kobiet, mężczyzn oraz osób zamieszkujących miasta, w przypadku których znany jest zarówno stan cywilny, jak i grupa wieku. W rozpatrywanym przykładzie wektor ten będzie miał postać: $\hat{X} = (13,10,12)^T$.

W ostatnim kroku rozważanego algorytmu, przy użyciu wzoru (13), wyznaczone są wagi kalibracyjne dla każdej osoby z rejestru. Przykładowo, wagę kalibracyjną dla drugiej osoby policzono na podstawie formuły:

$$w_2 = 1 + 1 \cdot (2 \ 5 \ 4) \begin{pmatrix} 13 & 0 & 5 \\ 0 & 10 & 7 \\ 5 & 7 & 12 \end{pmatrix}^{-1} \cdot \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = 1,1738$$

W podobny sposób wyznaczamy pozostałe wagi kalibracyjne, otrzymując ostatecznie wektor wag kalibracyjnych w . Warto zwrócić przy tym uwagę, że

suma wszystkich wag kalibracyjnych wynosi 30. Z kolei sumy wag kalibracyjnych odnoszących się do kobiet, mężczyzn oraz osób mieszkających w mieście i na wsi wynoszą odpowiednio: 15, 15, 16 i 14. Spełnione są więc odpowiednie równania kalibracyjne, stanowiące fundament podejścia kalibracyjnego. Zgodnie z ideą kalibracji można z dużym prawdopodobieństwem przypuszczać, że skoro wyznaczone wagi kalibracyjne odtwarzają dokładnie wartości globalne zmiennych pomocniczych, to powinny również odtwarzać wartość globalną zmiennej Y . Ostatecznie tablica wynikowa (tabl. 3), w której uwzględniono wagi kalibracyjne w_i celem zredukowania obciążenia będącego konsekwencją braków danych, jest postaci³:

TABL. 3. LUDNOŚĆ WEDŁUG STANU CYWILNEGO I WIEKU (dane przeważone)

Stan cywilny	Ogółem	15—34 lata	35 lat i więcej
O g ó ł e m	30	16,656760	13,343239
Kawalerowie/panny	7,922734	3,728083	4,194651
Rozwiedzeni/rozwidzione	6,490342	3,832095	2,658247
Wdowcy/wdowy	3,884101	1,173848	2,710253
Żonaci/zamężne	11,702820	7,922734	3,780089

Ź r ó d ł o: jak przy tabl. 1.

Podsumowanie

W badaniach statystycznych prowadzonych przez GUS podejście kalibracyjne nie było dotąd wykorzystywane na szeroką skalę. W zasadzie jedynym badaniem, w którym wykorzystuje się kalibrację, celem redukcji obciążenia wynikającego z braków odpowiedzi, jest badanie EU-SILC. W badaniu tym wykorzystano tzw. metodę kalibracji zintegrowanej — w wersji sinusa hiperbolicznego, ze względu na wykazaną w praktyce własność uzyskania wag kalibracyjnych bardzo blisko skupionych wokół wag wyjściowych. Wagi początkowe — wynikające z przyjętego dwustopniowego schematu doboru jednostek do próby — korygowane były na podstawie informacji o liczbie gospodarstw domowych oraz o liczbie osób według płci i wieku. Informacje te na poziomie województw (NUTS2) z dodatkowym podziałem na obszary miejski i wiejski, pochodziły z Narodowego Spisu Powszechnego Ludności i Mieszkań 2002 oraz z bieżących szacunków demograficznych (*Dochody...*, 2008).

W większości badań statystycznych występuje jednak problem braków odpowiedzi. Wydaje się więc zasadne zastosowanie podejścia kalibracyjnego celem redukcji obciążenia i poprawy efektywności — w sensie mniejszej wariancji stosowanych estymatorów. Z doświadczeń państw stosujących kalibrację w badaniach statystycznych wynika, że jest to metoda, która może stanowić reme-

³ Wartości w tabl. 3 nie są liczbami naturalnymi ze względu na małą wielkość przykładowego zbioru. Celem pokazania techniki przeważania danych z uwzględnieniem podejścia kalibracyjnego zestawiono je w oryginalnej postaci, nie dokonując żadnych zaokrągleń.

dium na ten rodzaj błędów nielosowych. Metoda ta jest bardzo intuicyjna w zastosowaniu i może być wykorzystana zarówno w badaniach częściowych prowadzonych metodą reprezentacyjną, jak i w badaniach pełnych. Może ona przy tym nabrać szczególnego znaczenia, zwłaszcza w tych badaniach, które opierać się będą na maksymalnym wykorzystaniu danych pochodzących z różnych źródeł, w tym z rejestrów administracyjnych. Jej wykorzystanie będzie również ważne z punktu widzenia badań pełnych, jakimi są spisy powszechne, w których także występuje problem braków danych. Opisana metoda kalibracji, polegająca na korygowaniu sztucznie utworzonych wag, może stanowić cenną technikę statystyczną, wykorzystywaną w tego typu badaniach na etapie tworzenia tablic wynikowych. W konsekwencji przyczyni się to do poprawy jakości publikowanych wyników, z punktu widzenia ich dokładności, a także zapewni ich większą akceptowalność przez odbiorców informacji statystycznych.

mgr Tomasz Józefowski — Ośrodek Statystyki Małych Obszarów, Urząd Statystyczny w Poznaniu,

dr Marcin Szymkowiak — Uniwersytet Ekonomiczny w Poznaniu, Ośrodek Statystyki Małych Obszarów, Urząd Statystyczny w Poznaniu

LITERATURA

Deville J.-C., Särndal C.-E. (1992), *Calibration estimators in survey sampling*, „Journal of the American Statistical Association”, vol. 87

Dochody i warunki życia ludności (raport z badania EU-SILC 2006 r.) (2008), GUS

Särndal C.-E., Lundström S. (2005), *Estimation in surveys with nonresponse*, John Wiley & Sons, Ltd

Szymkowiak M. (2007), *Przyczynek do kalibracji w badaniach statystycznych z brakami odpowiedzi*, [w:] *Kapitał ludzki i wiedza w gospodarce. Wyzwania XXI wieku*, Zeszyty Naukowe Wydziału Informatyki i Gospodarki Elektronicznej AE w Poznaniu

Wallgren A., Wallgren B. (2007), *Register-based statistics. Administrative data for statistical purposes*, Wiley Series in Survey Methodology, John Wiley & Sons Ltd

SUMMARY

The literature proposes two basic statistical techniques: imputation and weighting. They mitigate the negative impact of non-response test results. One of the weighing methods, used by some foreign statistical offices, is calibration, which consists of correcting the output weights resulting from the sampling scheme using a variety of auxiliary variables. Calibration can also be used in the complete surveys, and the starting point in its practical application is to establish an appropriate baseline — artificial weights. This article presents an example of how to use the calibration approach. With it one can create contingency tables based on the data from complete surveys (censuses, administrative records).

РЕЗЮМЕ

В литературе предлагается два основных статистических метода: импутацию и взвешивание. Они должны нивелировать отрицательное влияние отсутствия ответов на результаты обследований. Одним из методов взвешивания, используемым некоторыми зарубежными статистическими управлениями является калибровка, которая заключается в коррекции исходных весов, являющихся результатом схемы выборки с использованием разных вспомогательных переменных. Калибровка может также использоваться в полных обследованиях, а отправную точку в ее практическом использовании составляет соответствующее определение исходных — искусственных весов.

В статье представляется примерный способ использования калибрационного подхода. Благодаря этому, можно создавать таблицы ассоциации качественных признаков (контингенции) на основе данных из полных обследований (переписи, административные регистры).