Arkadiusz KOZŁOWSKI

Losowanie zrównoważone i kalibracja

Streszczenie. *Losowanie zrównoważone polega na takim doborze próby, aby szacunki sum zmiennych pomocniczych estymatorem Horvitza-Thompsona równały się rzeczywistym sumom tych zmiennych. Kalibracja natomiast polega na modyfikacji wyjściowych wag wynikających z planu losowania w taki sposób, aby zmodyfikowane wagi odtwarzały znane sumy zmiennych pomocniczych. Ideą obu metod jest odwzorowanie wartości globalnych zmiennych dodatkowych. Celem artykułu jest przedstawienie i porównanie obu technik traktowanych jako alternatywa do osiągnięcia tego samego celu. Więcej uwagi poświęcono losowaniu zrównoważonemu. Algorytm doboru próby zilustrowano za pomocą dwóch prostych przykładów. Porównanie losowania zrównoważonego z kalibracją wypada korzystniej dla tej drugiej metody, jednak najlepszym rozwiązaniem jest zastosowanie obu metod jednocześnie.*

Słowa kluczowe: losowanie zrównoważone, kalibracja.

Odwzorowanie wartości globalnych zmiennych dodatkowych, będące istotą losowania zrównoważonego i kalibracji, ma na celu zwiększenie dokładności oraz zapewnienie spójności wyników badania próbkowego[1]. Podstawowa różnica między tymi technikami wynika z faktu, że pierwsza stosowana jest na etapie doboru próby do badania częściowego, natomiast druga na etapie estymacji parametrów populacji. W artykule więcej uwagi poświęcono losowaniu zrównoważonemu, ponieważ w polskiej literaturze jest to koncepcja rzadziej podejmowana niż kalibracja.

LOSOWANIE ZRÓWNOWAŻONE

Idea próby zrównoważonej (*balanced sample*) nie jest nowa, gdyż pojawiła się równocześnie z narodzinami metody reprezentacyjnej i związana jest z samym pojęciem reprezentatywności. Pierwsze zastosowanie tej koncepcji w praktyce dotyczy słynnego wyboru próby okręgów przez C. Giniego i L. Galvaniego w 1929 r. we Włoszech. Wybrano wówczas 29 okręgów w taki sposób, aby średnie z próby dla kilku zmiennych kontrolnych zgadzały się ze średnimi

[1] Cele te (dokładność i spójność) są wymieniane jako jedne z komponentów jakości badań statystycznych. Pisali na ten temat m.in. Platek, Särndal (2001).

w populacji, znanymi ze spisu powszechnego. Słynni statystycy — J. Neyman oraz F. Yates — stanowczo potępili takie postępowanie ze względu na to, że próba dobierana była w sposób celowy[2]. Jak jednak zauważono później, próba zrównoważona może być równie dobrze pobrana w sposób probabilistyczny. Istotnie, losowanie warstwowe jest doborem probabilistycznym i zarazem szczególnym przypadkiem losowania zrównoważonego, w którym cechami dodatkowymi są zmienne zero-jedynkowe definiujące warstwy.

W celu formalnego ujęcia losowania zrównoważonego rozważmy populację U, której elementy są identyfikowane przez etykiety $k \in \{1, ..., N\}$. Zakłada się, że wektor $\boldsymbol{x}_k = (x_{k1}, ..., x_{kj}, ..., x_{kp})^T$, z wartościami dla p zmiennych dodatkowych (równoważących), znany jest dla wszystkich jednostek w populacji. Zakłada się również, że zmienne dodatkowe są liniowo niezależne. Próba określona jest przez wektor $\boldsymbol{s} = (s_1, ..., s_k, ..., s_N)^T$, gdzie s_k przyjmuje wartość 1, jeżeli k-ty element jest w próbie oraz 0 w przeciwnym przypadku. Rozważane są próby dobierane w sposób losowy. O próbie $\boldsymbol{s}$ można powiedzieć, że jest zrównoważona, jeżeli zachodzi równość:

$$\sum_U \frac{s_k \boldsymbol{x}_k}{\pi_k} = \sum_U \boldsymbol{x}_k \tag{1}$$

gdzie π_k — prawdopodobieństwo inkluzji pierwszego rzędu.

Oznacza to, że estymator HT bezbłędnie szacuje sumy cech dodatkowych. Zrównoważony plan próbkowania (*balanced sampling design*) zapewnia spełnienie warunku (1) dla każdej możliwej próby[3].

Warto zwrócić uwagę na trzy szczególne wybory zmiennych równoważących:

1) wektor prawdopodobieństw inkluzji $\boldsymbol{\pi} = (\pi_1, ..., \pi_N)^T$, tj. $x_k = \pi_k$ dla $k \in U$, zapewnia stałą liczebność próby n, gdyż (1) sprowadza się do: $\sum_U \frac{s_k \boldsymbol{x}_k}{\pi_k} = \sum_U s_k = n$;
2) zmienna w formie jedynek, tj. $x_k = 1$ dla $k \in U$, zapewnia stałe oszacowanie liczebności populacji $\left(\hat{N} = \sum_U \frac{s_k}{\pi_k} \right)$ na poziomie znanej, rzeczywistej liczebności populacji, gdyż (1) sprowadza się do $\sum_U \frac{s_k \boldsymbol{x}_k}{\pi_k} = \sum_U \frac{s_k}{\pi_k} = N$ (ma to znaczenie w przypadku losowania z różnym prawdopodobieństwem wyboru jednostek);

[2] Langel, Tillé (2011), s. 51.
[3] Deville, Tillé (2004), s. 895.

3) zmienne zero-jedynkowe $\boldsymbol{\delta}_h$ definiujące warstwy, tj. $\delta_{kh}=1$ dla $k\in U_h$, $\delta_{kh}=0$ dla $k\notin U_h$, gdzie $U=\bigcup_{h=1}^{H}U_h$, oraz gdy $\pi_k=\frac{n_h}{N_h}$, prowadzą do klasycznego losowania warstwowego, w którym z każdej warstwy U_h o liczebności N_h losowana jest próba prosta bez zwracania o liczebności n_h, gdyż (1) sprowadza się do $\sum_U \frac{s_k\delta_{kh}}{\pi_k}=\sum_U s_k\delta_{kh}\frac{N_h}{n_h}=N_h$ dla $h=1,...,H$.

Jeżeli chodzi o praktyczną implementację zrównoważonego planu próbkowania, to istnieją dwie zasadnicze grupy schematów go realizujących. Schematy z pierwszej grupy określić można jako odrzucające (*rejective*), gdyż opierają się na losowaniu wielu prób według określonego schematu i odrzucaniu tych, które nie spełniają kryteriów zrównoważenia. W tym postępowaniu wyróżnić można dwa warianty — spośród kolekcji prób wybiera się do badania tę, która jest najlepiej zrównoważona (zgodnie z wybranym kryterium)[4] (Kozłowski, 2012) lub z góry zakłada się pewien poziom dobroci zrównoważenia i odrzuca się próby, które go nie osiągają, a finalna próba losowana jest spośród tych, które pozostały. W literaturze spotkać można rozwiązania, w których schemat losowania kolejnych prób jest prosty, systematyczny (wymagający posortowania populacji ze względu na zmienną x), warstwowy (z równolicznymi warstwami utworzonymi na podstawie zmiennej x) lub z różnym prawdopodobieństwem wyboru jednostek (również zależnym od zmiennej x, wobec której próba ma być zrównoważona)[5]. Określone warianty losowania odrzucającego są zwykle ograniczone do pewnego typu schematów, nie zawsze dają możliwość różnicowania prawdopodobieństwa inkluzji pierwszego rzędu oraz mają ograniczenia, co do liczby i typu zmiennych równoważących.

Druga grupa schematów polega na wyborze próby w jednym podejściu poprzez iteracyjną modyfikację wektora prawdopodobieństwa inkluzji pierwszego rzędu do momentu, w którym będzie on jednoznacznie definiował próbę. Metodą najbardziej generalną spośród nich jest zaproponowana przez Deville i Tillé (2004) metoda kostki (*cube method*), gdyż pozwala wylosować próbę co najmniej w przybliżeniu zrównoważoną dla dowolnego wektora prawdopodobieństwa inkluzji pierwszego rzędu i dowolnej liczby oraz typu zmiennych dodatkowych. Ze względu na swoją ogólność, w dalszej części artykułu metoda kostki traktowana będzie jako domyślny sposób uzyskania próby zrównoważonej. Przejdźmy do wyjaśnienia algorytmu doboru próby tą metodą oraz zaprezentowania dwóch przykładów liczbowych.

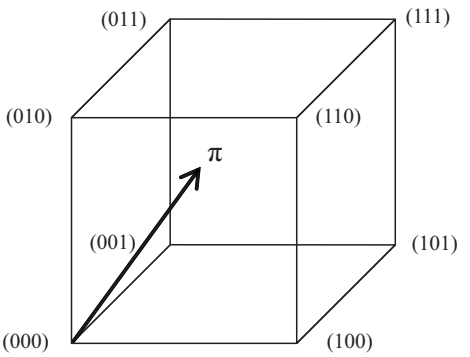
Punktem wyjścia w metodzie kostki jest geometryczne ujęcie wszystkich możliwych prób (w losowaniu bez zwracania) z populacji N-elementowej jako wierzchołków N-wymiarowej kostki jednostkowej $C=[0,1]^N$. Liczba wszystkich

[4] Wariantem tego postępowania jest procedura próbkowania związanego (*tied sampling procedure*).

[5] Valliant i in. (2000), s. 65—77.

możliwych prób (o dowolnej wielkości) jest równa liczbie wierzchołków kostki C, tj. 2^N. Dla prostego przypadku populacji 3-elementowej (N=3) przestrzeń prób przedstawić można jako wierzchołki sześcianu (wykres). Zaczynając od punktu określonego przez wektor prawdopodobieństwa inkluzji pierwszego rzędu $\boldsymbol{\pi}$, wybór próby może być zilustrowany jako losowe dotarcie do jednego z wierzchołków kostki[6].

GEOMETRYCZNA REPREZENTACJA PRZESTRZENI PRÓB DLA POPULACJI 3-ELEMENTOWEJ



Ź r ó d ł o: Deville, Tillé (2004), s. 896.

Warunek zrównoważenia (1) można zapisać równoważnie jako:

$$\sum_U \alpha_k s_k = \sum_U \alpha_k \pi_k \tag{2}$$

lub macierzowo:

$$\mathbf{A}_{\boldsymbol{S}} = \boldsymbol{X} \tag{3}$$

gdzie:

$\alpha_k = \dfrac{x_k}{\pi_k}$,

$\mathbf{A} = [\alpha_{jk}]$, gdzie $\alpha_{jk} = \dfrac{x_{jk}}{\pi_k}$,

$\boldsymbol{X} = \sum_U \boldsymbol{x}_k$ — wartości globalne zmiennych równoważących.

[6] Deville, Tillé (2004), s. 896.

Układ równań (3) ma nieskończenie wiele rozwiązań (naturalnie jeżeli $N>p$). Jednym z rozwiązań jest wektor $\boldsymbol{\pi}$, gdyż $\boldsymbol{X} = \mathbf{A}\boldsymbol{\pi}$, a zatem $\boldsymbol{s} = \boldsymbol{\pi}$ jest rozwiązaniem. Zbiór wszystkich rozwiązań każdego nieoznaczonego układu równań liniowych można zapisać jako wybrane rozwiązanie szczególne plus dowolnie wybrany element jądra macierzy współczynników. A więc zbiór wszystkich rozwiązań układu (3) zapiszemy jako $Q = \boldsymbol{\pi} + \ker\mathbf{A}$, gdzie ker**A** to jądro macierzy **A.** Ideą losowania zrównoważonego jest znalezienie szczególnego rozwiązania układu (3), które będzie składało się tylko z wartości 0 i 1, czyli będzie definiowało próbę. W metodzie kostki punktem wyjścia jest wektor $\boldsymbol{\pi}$, którego elementy są modyfikowane do momentu, w którym każdy z nich jest równy 0 lub 1. Proces ten podzielony jest na fazy lotu i lądowania (Deville, Tillé, 2004).

Faza lotu (*flight phase*) polega na losowej zamianie możliwie największej liczby elementów wektora $\boldsymbol{\pi}$ w 0 lub 1, zgodnie z zadanym prawdopodobieństwem inkluzji pierwszego rzędu, tak aby zmodyfikowany wektor $\boldsymbol{\pi}\ (\boldsymbol{\pi}^*)$ pozostawał rozwiązaniem układu (3). Po zainicjowaniu $\boldsymbol{\pi}\ (0) = \boldsymbol{\pi}$ powtarzane są trzy kroki[7], w czasie $t = 1, 2, \ldots, T$:

1) wygenerowanie dowolnego wektora $\boldsymbol{u}(t) = [u_k(t)] \neq 0$, będącego w jądrze macierzy **A**, takiego że $u_k(t) = 0$, jeżeli $\pi_k(t–1)$ jest liczbą całkowitą, tj. 0 lub 1;
2) obliczenie $\lambda_1^*(t)$ i $\lambda_2^*(t)$, które są największymi spośród dodatnich wartości $\lambda_1(t)$ i $\lambda_2(t)$ spełniających nierówności:

$$0 \le \pi_k(t-1) + \lambda_1(t)u_k(t) \le 1, \qquad 0 \le \pi_k(t-1) - \lambda_2(t)u_k(t) \le 1$$

dla $k = 1, \ldots, N$;

3) wybranie:

$$\boldsymbol{\pi}(t) = \begin{cases} \boldsymbol{\pi}(\text{t}–1) + \lambda_1^*(t)\boldsymbol{u}(t), \text{ z prawdopodobieństwem } q(t) \\ \boldsymbol{\pi}(\text{t}–1) - \lambda_2^*(t)\boldsymbol{u}(t), \text{ z prawdopodobieństwem } 1-q(t) \end{cases}$$

gdzie $q(t) = \dfrac{\lambda_2^*(t)}{\lambda_1^*(t) + \lambda_2^*(t)}$.

W każdym czasie t co najmniej jeden element $\boldsymbol{\pi}$ jest zamieniany na 0 lub 1 i w kolejnych iteracjach pozostaje na ustalonym poziomie. Procedura powtarzana jest do momentu, w którym nie można dłużej wykonywać kroku pierwszego. Jeżeli wektor wyjściowy $\boldsymbol{\pi}^* = \boldsymbol{\pi}(T)$ zawiera wyłącznie zera i jedynki, to losowanie jest zakończone. Próba $\boldsymbol{s} = \boldsymbol{\pi}^*$ jest idealnie zrównoważona. W przeciwnym przypadku

[7] Jest to podstawowy algorytm fazy lotu, jednak przy zapisie algorytmu do programu komputerowego lepiej jest skorzystać z efektywniejszego rozwiązania Tillé (2006), s. 162.

próba będzie tylko w przybliżeniu zrównoważona, a do jej ustalenia potrzebna będzie faza lądowania (*landing phase*). Polega ona na losowej zamianie niecałkowitych elementów $\boldsymbol{\pi}^*$ w 0 lub 1, tak aby prawdopodobieństwo inkluzji pierwszego rzędu było zachowane oraz wariancje estymatorów sum zmiennych dodatkowych były minimalizowane. Po fazie lotu pozostaje $q \leq p$ elementów niecałkowitych w wektorze $\boldsymbol{\pi}^*$. Możliwych jest zatem 2^q różnych prób (kolekcję tych prób oznaczmy przez $C(\boldsymbol{\pi}^*)$). Dobór próby w tej fazie jest enumeratywny, tzn. że rozważane są wszystkie możliwe próby i każdej przypisywane jest prawdopodobieństwo jej realizacji, a ostateczna próba wybierana jest w dowolnym eksperymencie losowym respektującym to prawdopodobieństwo.

Z każdą próbą związany jest koszt wynikający z niepełnego zrównoważenia próby. Ogólna postać funkcji kosztu jest następująca[8]:

$$C(\boldsymbol{s}) = (\boldsymbol{s} - \boldsymbol{\pi}^*)^T \mathbf{A}^T \mathbf{M} \mathbf{A} (\boldsymbol{s} - \boldsymbol{\pi}^*) \tag{4}$$

gdzie **M** — macierz $p \times p$, nieujemnie określona, która określa szczególną postać funkcji kosztu[9].

Ustalenie planu losowania $p(\boldsymbol{s}|\boldsymbol{\pi}^*)$ następuje w wyniku rozwiązania zadania programowania liniowego:

$$\min_{p(.|\boldsymbol{\pi}^*)} \sum_{s \in c(\boldsymbol{\pi}^*)} C(\boldsymbol{s}) p(\boldsymbol{s}|\boldsymbol{\pi}^*) \tag{5}$$

przy warunkach:

$$\begin{gathered} \sum_{s \in c(\boldsymbol{\pi}^*)} p(\boldsymbol{s}|\boldsymbol{\pi}^*) = 1 \\ \sum_{s \in c(\boldsymbol{\pi}^*) | s \ni k} p(\boldsymbol{s}|\boldsymbol{\pi}^*) = \pi_k^* \\ 0 \leq p(\boldsymbol{s}|\boldsymbol{\pi}^*) \leq 1 \end{gathered} \tag{6}$$

Takie zadanie można rozwiązać algorytmem simpleks. Jeżeli liczba elementów do ustalenia jest zbyt duża (q>20), to wyjściem jest wyrzucenie najmniej istotnej zmiennej równoważącej i powrót do fazy lotu. Czynność tę powtarza się z kolejnymi zmiennymi, aż faza lądowania będzie możliwa do przeprowadzenia[10].

[8] Deville, Tillé (2004), s. 900.
[9] Wybór różnych postaci macierzy **M** rozważają Deville, Tillé (2004), s. 911.
[10] Tillé (2011), s. 220.

W zależności od wyjściowego prawdopodobieństwa inkluzji pierwszego rzędu oraz zestawu zmiennych dodatkowych plan próbkowania może być[11]:

- dokładnie zrównoważony — każda możliwa próba jest idealnie zrównoważona;
- w przybliżeniu zrównoważony — wszystkie możliwe próby są tylko w przybliżeniu zrównoważone;
- czasami zrównoważony — istnieją próby, które są dokładnie zrównoważone, ale w celu zachowania prawdopodobieństwa inkluzji pierwszego rzędu będą pojawiać się też próby tylko w przybliżeniu zrównoważone.

Wybór próby zrównoważonej za pomocą metody kostki można zilustrować dwoma przykładami liczbowymi.

Przykład 1

Niech populacja U składać się będzie z N=10 osób i dla każdej osoby znany będzie przychód roczny (w tys. zł). Zamierza się pobrać próbę o liczebności n=4, zrównoważoną względem zmiennej przychód ($x1$, $x2$), z jednakowym prawdopodobieństwem wyboru jednostek. Dla zapewnienia stałego oszacowania N i stałej liczebności próby wprowadza się dodatkową zmienną równoważącą w postaci jedynek dla każdej jednostki. Dane wyjściowe są zatem następujące:

$$\boldsymbol{x1} = [1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1]$$

$$\boldsymbol{x2} = [66\ 60\ 79\ 70\ 53\ 69\ 40\ 59\ 59\ 63]$$

$$\boldsymbol{\pi} = [0{,}4\ 0{,}4\ 0{,}4\ 0{,}4\ 0{,}4\ 0{,}4\ 0{,}4\ 0{,}4\ 0{,}4\ 0{,}4]^T$$

$$\mathbf{A} = \begin{bmatrix} 2{,}5 & 2{,}5 & 2{,}5 & 2{,}5 & 2{,}5 & 2{,}5 & 2{,}5 & 2{,}5 & 2{,}5 & 2{,}5 \\ 165{,}0 & 150{,}0 & 197{,}5 & 175{,}0 & 132{,}5 & 172{,}5 & 100{,}0 & 147{,}5 & 147{,}5 & 157{,}5 \end{bmatrix}$$

$$\boldsymbol{X} = \begin{bmatrix} 10 \\ 618 \end{bmatrix}$$

Na początku przyjęto $\boldsymbol{\pi}(0)=\boldsymbol{\pi}$, a następnie w kolejnych iteracjach t wykonywane były trzy kroki zgodnie z algorytmem fazy lotu. W tabl. 1 podano wektory $\boldsymbol{u}(t)$[12] oraz $\boldsymbol{\pi}(t)$ dla kolejnych iteracji, a także wartości $\lambda_1^*(t)$ i $\lambda_2^*(t)$, z zaznaczeniem pogrubioną czcionką, która z tych wartości została losowo wybrana do wyliczenia $\boldsymbol{\pi}(t)$.

[11] Deville, Tillé (2004), s. 897.

[12] Wektor $\boldsymbol{u}(t)$ generowano z wykorzystaniem formuły $\boldsymbol{u}(t)=(\mathbf{I}-\mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^+\mathbf{A})\boldsymbol{v}$, gdzie: $\mathbf{I}$ to macierz jednostkowa, $\boldsymbol{v}$ to wektor wartości losowych, generowanych niezależnie z rozkładu $N(10, 1)$, a $\mathbf{D}^+$ oznacza macierz pseudoodwrotną Moore-Penrose do macierzy $\mathbf{D}$. Przed każdym zastosowaniem pomniejszano macierz $\mathbf{A}$ o kolumny, dla których $\pi_k(t-1)=1$ lub $\pi_k(t-1)=0$, a następnie przed krokiem 3 uzupełniano odpowiednie elementy $\boldsymbol{u}(t)$ zerami.

TABL. 1. PRZEBIEG FAZY LOTU DO PRZYKŁADU 1

$\boldsymbol{u}(t)$ i $\boldsymbol{\pi}(t)$	k										$\lambda_1^*(t)$ i $\lambda_2^*(t)$
	1	2	3	4	5	6	7	8	9	10	
$\boldsymbol{u}(1)$	–0,6251	–0,8677	1,1073	–1,3873	–1,1251	1,0220	1,0756	0,5569	–0,2126	0,4561	$\lambda_1^*(1)=0{,}2883$
$\boldsymbol{\pi}(1)$	0,6258	0,7135	0,0000	0,9012	0,8065	0,0308	0,0114	0,1988	0,4768	0,2352	$\boldsymbol{\lambda_2^*(1)=0{,}3613}$
$\boldsymbol{u}(2)$	–1,0268	1,1484	0,0000	–1,3045	1,3913	0,9624	–1,0105	–0,1373	–0,0009	–0,0219	$\boldsymbol{\lambda_1^*(2)=0{,}0113}$
$\boldsymbol{\pi}(2)$	0,6142	0,7264	0,0000	0,8864	0,8222	0,0417	0,0000	0,1973	0,4768	0,2350	$\lambda_2^*(2)=0{,}0320$
$\boldsymbol{u}(3)$	0,3602	0,4394	0,0000	0,4448	0,7292	0,0102	0,0000	–0,1117	–0,9770	–0,8951	$\lambda_1^*(3)=0{,}2438$
$\boldsymbol{\pi}(3)$	0,4213	0,4912	0,0000	0,6482	0,4317	0,0362	0,0000	0,2571	1,0000	0,7143	$\boldsymbol{\lambda_2^*(3)=0{,}5355}$
$\boldsymbol{u}(4)$	–2,0905	–0,1051	0,0000	0,2791	0,3053	1,1875	0,0000	0,0174	0,0000	0,4063	$\lambda_1^*(4)=0{,}2015$
$\boldsymbol{\pi}(4)$	0,4851	0,4944	0,0000	0,6397	0,4224	0,0000	0,0000	0,2565	1,0000	0,7019	$\boldsymbol{\lambda_2^*(4)=0{,}0305}$
$\boldsymbol{u}(5)$	0,0892	–0,5493	0,0000	0,2121	0,7194	0,0000	0,0000	–0,9485	0,0000	0,4771	$\boldsymbol{\lambda_1^*(5)=0{,}2705}$
$\boldsymbol{\pi}(5)$	0,5092	0,3458	0,0000	0,6971	0,6170	0,0000	0,0000	0,0000	1,0000	0,8310	$\lambda_2^*(5)=0{,}5871$
$\boldsymbol{u}(6)$	–0,1760	0,2942	0,0000	0,1766	–0,0175	0,0000	0,0000	0,0000	0,0000	–0,2773	$\boldsymbol{\lambda_1^*(6)=1{,}7157}$
$\boldsymbol{\pi}(6)$	0,2072	0,8505	0,0000	1,0000	0,5870	0,0000	0,0000	0,0000	1,0000	0,3552	$\lambda_2^*(6)=0{,}6096$
$\boldsymbol{u}(7)$	0,6855	–0,0889	0,0000	0,0000	0,2323	0,0000	0,0000	0,0000	0,0000	–0,8290	$\lambda_1^*(7)=0{,}4285$
$\boldsymbol{\pi}(7)$	0,0000	0,8774	0,0000	1,0000	0,5168	0,0000	0,0000	0,0000	1,0000	0,6058	$\boldsymbol{\lambda_2^*(7)=0{,}3023}$
$\boldsymbol{u}(8)$	0,0000	0,3232	0,0000	0,0000	–0,0970	0,0000	0,0000	0,0000	0,0000	–0,2262	$\boldsymbol{\lambda_1^*(8)=0{,}3793}$
$\boldsymbol{\pi}(8)$	0,0000	1,0000	0,0000	1,0000	0,4800	0,0000	0,0000	0,0000	1,0000	0,5200	$\lambda_2^*(8)=1{,}7423$

Ź r ó d ł o: opracowanie własne.

Faza lotu została zakończona po wykonaniu 8 iteracji. Nie jest możliwe wygenerowanie kolejnego wektora $\boldsymbol{u}(t)$, który pozostawałby w jądrze macierzy $\mathbf{A}$ i miał zera na odpowiednich miejscach. Wynikowy wektor $\boldsymbol{\pi}^*=\boldsymbol{\pi}(8)$ zawiera elementy niecałkowite, a zatem ostateczna próba nie jest jeszcze wybrana. Wiadomo, że do próby wejdą na pewno osoby o numerach 2, 4 i 9, zaś na pewno nie wejdą osoby o numerach 1, 3, 6, 7 i 8. Nieokreślony jest natomiast los osób o numerach 5 i 10. Wiadomo już, że próba nie będzie idealnie zrównoważona, a do jej ustalenia potrzebna będzie faza lądowania.

Po fazie lotu zostały q=2 elementy niecałkowite w wektorze $\boldsymbol{\pi}^*$, a zatem możliwe są 2^q=4 próby do wyboru w fazie lądowania. W tabl. 2 przedstawiono te próby wraz z kosztem wyboru każdej z nich oraz ostatecznym planem losowania minimalizującym łączny koszt wynikający z niedoskonałego zrównoważenia. Do policzenia kosztu $C(\boldsymbol{s})$ według formuły (4) wybrano $\mathbf{M}=(\mathbf{A}\mathbf{A}^{\mathrm{T}})^{-1}$, dzięki czemu koszt ten można interpretować jako kwadrat odległości w N-wymiarowej przestrzeni pomiędzy punktem $\boldsymbol{s}$ a jego rzutem na hiperpłaszczyznę Q.

TABL. 2. PLAN WYBORU PRÓBY W FAZIE LĄDOWANIA DO PRZYKŁADU 1

k	$\boldsymbol{\pi}^*$	$\boldsymbol{s}1$	$\boldsymbol{s}2$	$\boldsymbol{s}3$	$\boldsymbol{s}4$
1	0,00	0,00	0,00	0,00	0,00
2	1,00	1,00	1,00	1,00	1,00
3	0,00	0,00	0,00	0,00	0,00
4	1,00	1,00	1,00	1,00	1,00
5	0,48	1,00	1,00	0,00	0,00
6	0,00	0,00	0,00	0,00	0,00
7	0,00	0,00	0,00	0,00	0,00
8	0,00	0,00	0,00	0,00	0,00
9	1,00	1,00	1,00	1,00	1,00
10	0,52	1,00	0,00	1,00	0,00
$C(\boldsymbol{s})$		0,1159	0,0269	0,0229	0,1129
$p(\boldsymbol{s}\|\boldsymbol{\pi}^*)$		0,0000	0,4800	0,5200	0,0000

Ź r ó d ł o: jak przy tabl. 1.

Optymalny plan losowania daje zerową szansę na wylosowanie próby $\boldsymbol{s}\mathbf{1}$ i $\boldsymbol{s}\mathbf{4}$, co jest zgodne z oczekiwaniami, gdyż w przeciwnym razie liczebność próby byłaby inna niż zakładana, a zatem wybór ogranicza się do prób $\boldsymbol{s}\mathbf{2}$ i $\boldsymbol{s}\mathbf{3}$. Obie są mniej więcej tak samo prawdopodobne, a ostatecznego wyboru można dokonać za pomocą prostego eksperymentu losowego. Jakość zrównoważenia dla obu prób przedstawiono w tabl. 3.

TABL. 3. OCENA DOKŁADNOŚCI ZRÓWNOWAŻENIA DLA PRÓB Z PRZYKŁADU 1

Zmienne	X_j	$\boldsymbol{s}2$		$\boldsymbol{s}3$	
		$\hat{x}_{j,HT}$	błąd względny w %	$\hat{x}_{j,HT}$	błąd względny w %
$x1$	10	10	0,0	10	0,0
$x2$	618	605	–2,1	630	1,9

Ź r ó d ł o: jak przy tabl. 1.

Przykład 2

Niech populacja U składać się będzie z N=12 osób i znane będą płeć oraz miejsce zamieszkania (miasto/wieś) każdej osoby. Zamierza się pobrać próbę o liczebności n=4, z jednakowym prawdopodobieństwem wyboru jednostek, zrównoważoną względem obu zmiennych dodatkowych, ale nie koniecznie względem przekroju tych zmiennych. Przykład ten ilustruje przypadek wyboru próby warstwowej z nakładającymi się warstwami. W klasycznym losowaniu warstwowym coś takiego nie jest możliwe, gdyż warstwy muszą być rozłączne. Często jednak występuje wiele potencjalnych cech warstwujących i/lub wiele wariantów kilku cech i wylosowanie próby warstwowej na przekroju wszystkich cech jest praktycznie niemożliwe, gdyż powstaje bardzo dużo warstw (nierzadko więcej niż liczebność populacji), przy czym wiele z nich jest pustych lub bardzo mało licznych. Dzięki metodzie kostki można wylosować próbę, która będzie zapewniała proporcjonalną lokalizację próby względem rozkładu brzegowego każdej cechy warstwującej, bez konieczności tworzenia przekroju wszystkich zmiennych.

TABL. 4. DANE DOTYCZĄCE POPULACJI *U* DO PRZYKŁADU 2

x_j	k												Suma
	1	2	3	4	5	6	7	8	9	10	11	12	
Kobieta	1	1	1	1	1	0	0	0	0	0	1	0	6
Mężczyzna	0	0	0	0	0	1	1	1	1	1	0	1	6
Miasto	1	0	0	0	0	1	1	0	0	0	0	0	3
Wieś	0	1	1	1	1	0	0	1	1	1	1	1	9
π_k	1/3	1/3	1/3	1/3	1/3	1/3	1/3	1/3	1/3	1/3	1/3	1/3	4

Ź r ó d ł o: jak przy tabl. 1.

Dla zapewnienia niezależności zmiennych do losowania wykorzystano tylko trzy pierwsze zmienne dodatkowe. Macierz **A** wygląda zatem następująco:

$$\mathbf{A} = \begin{bmatrix} 3 & 3 & 3 & 3 & 3 & 0 & 0 & 0 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 3 & 3 & 3 & 3 & 3 & 0 & 3 \\ 3 & 0 & 0 & 0 & 0 & 3 & 3 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Analogicznie, jak w przykładzie 1, przeprowadzono fazę lotu, której przebieg przedstawiono w tabl. 5.

TABL. 5. PRZEBIEG FAZY LOTU DO PRZYKŁADU 2

$u(t)$ i $\pi(t)$	k												$\lambda_1^*(t)$ i $\lambda_2^*(t)$
	1	2	3	4	5	6	7	8	9	10	11	12	
$u(1)$	–0,487	–0,070	0,315	–0,598	0,096	0,534	–0,047	–0,951	0,159	0,243	0,745	0,062	$\lambda_1^*(1)=0,35$
$\pi(1)$	0,552	0,365	0,192	0,601	0,290	0,094	0,354	0,759	0,262	0,225	0,000	0,306	**$\lambda_2^*(1)=0,45$**
$u(2)$	–0,767	0,755	0,652	–0,478	–0,161	1,775	–1,008	–0,153	1,082	–0,361	0,000	–1,335	$\lambda_1^*(2)=0,23$
$\pi(2)$	0,592	0,325	0,158	0,626	0,299	0,000	0,408	0,767	0,205	0,244	0,000	0,377	**$\lambda_2^*(2)=0,05$**
$u(3)$	0,085	–0,864	0,663	1,301	–1,184	0,000	–0,085	–0,015	–0,430	0,709	0,000	–0,179	**$\lambda_1^*(3)=0,25$**
$\pi(3)$	0,614	0,107	0,325	0,955	0,000	0,000	0,386	0,764	0,096	0,423	0,000	0,331	$\lambda_2^*(3)=0,24$
$u(4)$	0,589	0,324	–0,910	–0,003	0,000	0,000	–0,589	–0,011	0,040	0,595	0,000	–0,035	**$\lambda_1^*(4)=0,36$**
$\pi(4)$	0,420	0,000	0,624	0,956	0,000	0,000	0,580	0,767	0,083	0,227	0,000	0,343	$\lambda_2^*(4)=0,33$
$u(5)$	0,508	0,000	0,802	–1,310	0,000	0,000	–0,508	–0,799	0,734	1,096	0,000	–0,523	**$\lambda_1^*(5)=0,47$**
$\pi(5)$	0,658	0,000	1,000	0,342	0,000	0,000	0,342	0,393	0,427	0,741	0,000	0,098	$\lambda_2^*(5)=0,03$
$u(6)$	0,103	0,000	0,000	–0,103	0,000	0,000	–0,103	0,005	–0,848	0,091	0,000	0,855	$\lambda_1^*(6)=0,50$
$\pi(6)$	0,646	0,000	1,000	0,354	0,000	0,000	0,354	0,392	0,524	0,730	0,000	0,000	**$\lambda_2^*(6)=0,11$**
$u(7)$	–0,974	0,000	0,000	0,974	0,000	0,000	0,974	–1,086	1,160	–1,049	0,000	0,000	**$\lambda_1^*(7)=0,36$**
$\pi(7)$	0,897	0,000	1,000	0,103	0,000	0,000	0,103	0,671	0,225	1,000	0,000	0,000	$\lambda_2^*(7)=0,26$
$u(8)$	–0,894	0,000	0,000	0,894	0,000	0,000	0,894	–0,270	–0,624	0,000	0,000	0,000	$\lambda_1^*(8)=0,36$
$\pi(8)$	1,000	0,000	1,000	0,000	0,000	0,000	0,000	0,703	0,298	1,000	0,000	0,000	**$\lambda_2^*(8)=0,12$**
$u(9)$	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,026	–0,026	0,000	0,000	0,000	$\lambda_1^*(9)=11,3$
$\pi(9)$	1,000	0,000	1,000	0,000	0,000	0,000	0,000	0,000	1,000	1,000	0,000	0,000	**$\lambda_2^*(9)=26,7$**

Ź r ó d ł o: jak przy tabl. 1.

Tym razem już w fazie lotu udało się wybrać ostateczną próbę, gdyż w ostatniej iteracji wektor $\boldsymbol{\pi}^{*}=\boldsymbol{\pi}(9)$ zawiera wyłącznie zera i jedynki. Do próby wchodzą jednostki o numerach 1, 3, 9 i 10. Próba jest zatem idealnie zrównoważona, co potwierdzają wyliczenia zamieszczone w tabl. 6. Zrównoważenie względem zmiennej „miasto" automatycznie zapewnia zrównoważenie względem niewykorzystywanej zmiennej „wieś". Łatwo sprawdzić, że oszacowania liczebności z przekroju płci i miejsca zamieszkania nie pokrywają się z faktycznymi liczebnościami, gdyż przekrój nie był celem zrównoważenia i ewentualna zgodność mogłaby wynikać jedynie z przypadku.

TABL. 6. OCENA DOKŁADNOŚCI ZRÓWNOWAŻENIA DLA PRÓBY Z PRZYKŁADU 2

Zmienne	X_j	$\hat{x}_{j,HT}$	Błąd względny w %
Kobieta	6	6	0
Mężczyzna	6	6	0
Miasto	3	3	0

Ź r ó d ł o: jak przy tabl. 1.

PODEJŚCIE KALIBRACYJNE

Podobnie jak w przypadku losowania zrównoważonego, szczególne metody kalibracji znane są i praktykowane od dawna, a stosunkowo niedawno jedynie ta koncepcja doczekała się ogólnego sformułowania i nadania jej nazwy „podejście kalibracyjne". Za swoistą cezurę uważa się artykuł Deville'a, Särndala, (1992).

Standardowe zastosowanie podejścia kalibracyjnego dotyczy szacowania wartości globalnej zmiennej $y\left(Y=\sum_{U} y_k\right)$. Estymator kalibracyjny dla tego parametru zadany jest wzorem[13]:

$$\hat{y}_{cal}=\sum_{s} w_i y_i \tag{7}$$

gdzie w_i — waga kalibracyjna dla i-tej jednostki w próbie.

Konstrukcja estymatora kalibracyjnego jest analogiczna do estymatora HT $\left(\hat{y}_{HT}=\sum_{s} d_i y_i\right)$. W przeciwieństwie jednak do wag wynikających z planu losowania $\left(d_i=1/\pi_i\right)$,wagi kalibracyjne w_i nie są znane *a priori*, lecz zależą od wylosowanej próby. Ustalenie wag kalibracyjnych, które są kluczowym elemen-

[13] Särndal, Lundström (2005), s. 57.

tem całego podejścia następuje w taki sposób, aby spełnione było tzw. równanie kalibracyjne (*calibration equation*)[14]:

$$\sum_s w_i \boldsymbol{x}_i = \boldsymbol{x} \tag{8}$$

Równanie to oznacza, że wagi kalibracyjne to takie, które zastosowane do zmiennych pomocniczych odtwarzają (tj. szacują bez błędu) znane wartości globalne tych zmiennych. Dodatkowo wagi wyznaczane są w taki sposób, aby różnica pomiędzy wektorem ostatecznych wag kalibracyjnych $[w_i]_{i\in s}$ a wektorem wag z planu losowania $[d_i]_{i\in s}$ była minimalna. Motywacją takiego postępowania jest zmniejszenie obciążenia estymatora[15].

Możliwe jest wyznaczenie wielu różnych zestawów wag, które przy danych zmiennych pomocniczych będą spełniały równanie kalibracyjne. Uzyskane wagi zależą od metody wyznaczania oraz parametrów wybranej metody. W literaturze dominują dwa podejścia. Pierwsze polega na przyjęciu pewnej funkcji odległości między w_i a d_i, a następnie wyznaczeniu minimum tej funkcji ze względu na w_i, przy warunku (8). Najczęściej spotykaną postacią funkcji odległości jest[16]:

$$G(w_i, d_i) = \frac{(w_i - d_i)^2}{2d_i} \tag{9}$$

Suma odległości (9) po elementach w próbie, tj. $\sum_s \frac{(w_i - d_i)^2}{2d_i}$, osiąga minimum przy ograniczeniach wynikających z równania kalibracyjnego dla wag[17]:

$$w_i = d_i + d_i \left(\boldsymbol{X} - \sum_s d_i \boldsymbol{x}_i \right)^T \left(\sum_s d_i \boldsymbol{x}_i \boldsymbol{x}_i^T \right)^{-1} \boldsymbol{x}_i \tag{10}$$

Drugim podejściem jest metoda instrumentu. Punktem wyjścia jest zapisanie wagi kalibracyjnej jako iloczynu wagi wyjściowej ($d_{\alpha i}$) oraz czynnika korygującego (υ_i), tj. $w_i = d_{\alpha i}\upsilon_i$. Wartości υ_i powinny odzwierciedlać informację dodatkową dla i-tej jednostki. Najczęściej przyjmuje się $\upsilon_i = 1 + \boldsymbol{\lambda}^T \boldsymbol{z}_i$, gdzie $\boldsymbol{\lambda}$ jest wektorem

[14] Ibid., s. 58.

[15] Estymator HT z wagami d_i jest nieobciążony, a obciążenie estymatora kalibracyjnego jest dodatnią funkcją różnicy $w_i - d_i$ — Särndal (2007), s. 105.

[16] Inne funkcje podają Deville, Särndal (1992), s. 378. Autorzy konkludują, że wybór funkcji odległości nie ma istotnego znaczenia dla oszacowań oraz wariancji estymatora w przypadku co najmniej średnio licznych prób.

[17] Deville, Särndal (1992), s. 377. Szczegółowe wyprowadzenie podaje Szymkowiak (2009), s. 101—103.

wartości zapewniającym spełnienie równania kalibracyjnego, a $\boldsymbol{z}_i$ jest tzw. wektorem—instrumentem, który jest pewną funkcją na wartościach $\boldsymbol{x}_i$. Standardowym wyborem jest $d_{\alpha i}=d_i$ oraz $\boldsymbol{z}_i=\boldsymbol{x}_i$, co prowadzi do wag identycznych z wzoru (10)[18]. W zależności zatem od wyboru funkcji odległości $G(w_i, d_i)$ bądź postaci czynnika korygującego υ_i oraz instrumentu $\boldsymbol{z}_i$ można uzyskać inne postaci wag kalibracyjnych dopasowane do konkretnego badania (przede wszystkim do charakteru informacji dodatkowej).

Kalibracja może skutkować tym, że niektóre wagi będą bardzo duże, a także, że niektóre będą ujemne[19]. Obie sytuacje są postrzegane jako niepożądane[20]. W celu ich uniknięcia proponowanych jest kilka rozwiązań. Jednym z nich jest odpowiednia konstrukcja funkcji odległości, która będzie zawierała dolną i górną wartość ograniczającą[21]. Innym jest dołączenie do równania kalibracyjnego kolejnego warunku utrzymującego wagi w ustalonych granicach ($L_i \le w_i \le U_i$) i zastosowanie programowania matematycznego w celu minimalizacji (9)[22].

Pożądaną własnością wag jest to, że sumują się one do liczebności populacji. W wielu badaniach dokładna liczebność populacji nie jest znana. Jeżeli jednak jest ona znana lub istnieją co do niej wiarygodne szacunki, to aby uzyskać sumowalność do N, wystarczy do zbioru zmiennych dodatkowych dołączyć zmienną z wartościami 1 dla każdej jednostki.

Warto jeszcze zwrócić uwagę na podobieństwo podejścia kalibracyjnego i podejścia związanego z uogólnionym estymatorem regresyjnym (*generalized regression estimator*, GREG). Generalnie są to dwa różne podejścia do wykorzystania zmiennych dodatkowych na etapie estymacji statystycznej, jednak w niektórych przypadkach dają ten sam rezultat. W szczególności liniowy GREG, tj. estymator postaci:

$$\hat{y}_{reg} = \hat{y}_{HT} + (\boldsymbol{X}^{\boldsymbol{T}} - \hat{\boldsymbol{x}}_{HT}^{T})\hat{\boldsymbol{B}} \tag{11}$$

gdzie $\hat{\boldsymbol{B}} = \left(\sum_s d_i q_i \boldsymbol{x}_i \boldsymbol{x}_i^T \right)^{-1} \left(\sum_s d_i q_i \boldsymbol{x}_i y_i \right)$ może być przedstawiony w formie ważonej sumy wartości z próby[23]:

$$\hat{y}_{reg} = \sum_s d_i g_i y_i \tag{12}$$

[18] Särndal, Lundström (2005), s. 58 i 62.

[19] Gdyby przyjąć, że waga określa liczbę jednostek w populacji, które są reprezentowane przez daną obserwację w próbie, jedynie wagi większe lub równe jedności byłyby do zaakceptowania.

[20] Deville i in. (1993), s. 1019.

[21] Deville, Särndal (1992), s. 378.

[22] Särndal (2007), s. 108.

[23] Ibid., s. 103.

gdzie:

$$g_i = 1 + q_i \boldsymbol{\lambda}^T \boldsymbol{x}_i,$$

$$\boldsymbol{\lambda}^T = \left(\boldsymbol{X} - \sum_s d_i \boldsymbol{x}_i \right)^T \left(\sum_s d_i q_i \boldsymbol{x}_i \boldsymbol{x}_i^T \right)^{-1},$$

q_i — czynnik skalujący, zwykle $q_i = 1$ dla $i \in s$.

Łatwo sprawdzić, że w tym przypadku wagi $d_i g_i$ z (12) są identyczne z wagami w_i z (10), a zatem liniowy GREG jest tym samym, co estymator kalibracyjny z funkcją odległości (9). Możliwość przedstawienia estymatora w formie ważonej sumy wartości z próby jest ważną własnością, szczególnie z praktycznego punktu widzenia. W rzeczywistym badaniu szacuje się zwykle wiele parametrów, więc budowanie estymatora osobno dla każdego parametru wymaga dużego nakładu pracy. Stosowanie estymatora wagowego znacznie upraszcza cały proces, gdyż wagi w przedstawionej formie nie zależą od wartości cechy y, a zatem mogą być wykorzystane wielokrotnie do szacowania innych parametrów. Poza tym system wag zapewnia addytywność szacunków, co oznacza, że przy szacowaniu parametrów dla podpopulacji, sumy ocen są równe ocenie dla całej populacji[24].

Ciekawy przykład liczbowy dotyczący kalibracji podają Józefowski i Szymkowiak (2012).

LOSOWANIE ZRÓWNOWAŻONE I KALIBRACJA

Zarówno losowanie zrównoważone, jak i kalibracja są metodami ogólnymi w swej dziedzinie, tzn. większość metod losowania może być postrzegana jako szczególne przypadki losowania zrównoważonego, a większość metod estymacji może być postrzegana jako szczególne przypadki podejścia kalibracyjnego[25]. Pomijając kilka specyficznych wyborów, np. zrównoważenie lub kalibracja względem zmiennej złożonej z samych jedynek, obie metody dążą do zachowania spójności z rzeczywistymi zmiennymi dodatkowymi znanymi z innych źródeł. W tej części artykułu dokonano porównania losowania zrównoważonego i kalibracji, skupiając się na kryteriach związanych z wykorzystaniem informacji dodatkowej. Zestawiono ze sobą dwa alternatywne badania — losowanie zrównoważone z estymatorem HT oraz losowanie proste z estymatorem kalibracyjnym. Oba sposoby dotyczą szacowania wartości globalnej.

[24] Bracha i in. (2004), s. 30.
[25] Tillé (2011), s. 222.

Skuteczność odwzorowania parametrów zmiennych dodatkowych

W podejściu kalibracyjnym zawsze da się wykalibrować wagi w taki sposób, aby idealnie odtwarzały sumy zmiennych dodatkowych. W losowaniu zrównoważonym przeważnie nie da się wylosować próby idealnie zrównoważonej, najczęściej będzie to próba zrównoważona jedynie w przybliżeniu. Wynika to z tzw. problemu zaokrąglenia (*rounding problem*) — wielkość próby musi być liczbą całkowitą, co znacznie ogranicza pole manewru. Dla większości zmiennych nie będzie numerycznie możliwe wylosowanie próby, z której szacunek sumy estymatorem HT byłby taki sam, jak rzeczywista suma. Ilustracją tego problemu niech będzie następujący przykład:

Populacja U liczy 10 jednostek. Znane są wartości cechy x, którymi są kolejne liczby naturalne od 1 do 10. Zamierza się wylosować próbę 3-elementową, z jednakowym prawdopodobieństwem wyboru jednostek (π_k=0,3 dla $k \in U$). Warunek zrównoważenia (1) sprowadza się do tego, że prosta średnia z próby musi równać się średniej z populacji, czyli 5,5. Niestety żadna kombinacja 3 elementów nie wygeneruje takiego szacunku, gdyż suma z próby musiałaby być równa 16,5, co jest niemożliwe przy wszystkich wartościach całkowitoliczbowych.

Problem zaokrągleń staje się mało znaczący wraz ze wzrostem liczebności próby. Dla planów próbkowania z ustaloną liczebnością próby i w których suma prawdopodobieństwa inkluzji pierwszego rzędu jest liczbą całkowitą można wykazać, że dla dowolnej zmiennej równoważącej zachodzi[26]:

$$\left|\hat{x}_{HT} - X\right| \leq (p-1) \cdot \max_{k \in U} \left| \frac{x_k}{\pi_k} - \frac{X}{n} \right| \tag{13}$$

gdzie $X = \sum_U x_k$.

Taka nierówność pokazuje górną granicę niedokładności zrównoważenia, czyli dla najgorszego możliwego przypadku. Należy jednak pamiętać, że faza lądowania, która jest wykonywana, jeżeli występuje problem zaokrągleń, ma na celu wybór próby najmniej odległej od idealnie zrównoważonej, więc zwykle różnica nie będzie sięgać tej górnej granicy.

Wpływ liczby zmiennych dodatkowych

Ogólnie rzecz biorąc, im większa liczba zmiennych pomocniczych, tym lepiej dla badania. Naturalnie zmienne te powinny być nieskorelowane między sobą, a skorelowane ze zmiennymi badanymi. W podejściu kalibracyjnym liczba

[26] Tillé (2006), s. 165.

zmiennych dodatkowych może być w zasadzie dowolna (o ile nie jest większa od liczebności próby), ponieważ zawsze da się dokładnie wykalibrować wagi względem każdej ze zmiennych. W losowaniu zrównoważonym, co wynika z nierówności (13), im więcej cech równoważących, tym jakość zrównoważenia dla poszczególnych zmiennych może być gorsza. Dodatkowo, jeżeli po fazie lotu zostanie zbyt dużo elementów niecałkowitych w wektorze $\boldsymbol{\pi}^*$, to część najmniej istotnych zmiennych równoważących musi zostać odrzucona.

Elastyczność w doborze zmiennych dodatkowych

Kalibracja wag dokonywana jest po wylosowaniu próby, a zatem nic nie stoi na przeszkodzie, aby zmieniać zestaw cech dodatkowych, a także dokonywać ich transformacji, osobno dla każdej cechy badanej[27]. W przypadku losowania zrównoważonego możliwy jest tylko jeden zestaw zmiennych pomocniczych, obowiązujący dla całego badania.

Wymogi co do znajomości zmiennych dodatkowych

W celu wykalibrowania wag wystarczająca jest znajomość wartości zmiennych dodatkowych dla jednostek wylosowanych do próby oraz wartości globalnych dla całej populacji. W losowaniu zrównoważonym wymogi są większe, gdyż niezbędna jest znajomość *a priori* wartości cech dodatkowych dla wszystkich jednostek w populacji. W przypadku badań statystyki publicznej oznacza to konieczność korzystania z rejestrów administracyjnych.

Wariancja estymatora sumy

Zarówno w kalibracji, jak i losowaniu zrównoważonym nie istnieją dokładne wzory na wariancję estymatora sumy. W obu przypadkach korzysta się z formuł przybliżonych i dla każdego z nich wariancje zwykle są niedoszacowane. Odnośnie kalibracji niedoszacowanie wynika z tego, że nie bierze się pod uwagę zmienności wag, a w losowaniu zrównoważonym z tego, że zakłada się idealne zrównoważenie. Wariancja estymatora kalibracyjnego jest w przybliżeniu taka sama, jak wariancja estymatora GREG i wyraża się wzorem[28]:

$$D^2(\hat{y}_{cal}) \approx \sum\sum_{U} (\pi_{kl} - \pi_k \pi_l) \left(\frac{y_k - y_k^*}{\pi_k} \right) \left(\frac{y_l - y_l^*}{\pi_l} \right) \tag{14}$$

[27] W praktyce raczej nie zmienia się zestawu zmiennych dodatkowych, szczególnie jeżeli chodzi o statystykę publiczną, gdyż pociągałoby to za sobą różne zestawy wag, a tym samym możliwość uzyskiwania różnych rozkładów brzegowych tej samej zmiennej, co jest sprzeczne z jednym z celów kalibracji, jakim jest uzyskanie spójności szacunków.

[28] Deville, Särndal (1992), s. 379.

gdzie:

π_{kl} — prawdopodobieństwo inkluzji drugiego rzędu,

$y_k^* = \boldsymbol{x}_k^T \boldsymbol{B}$ — wartość teoretyczna zmiennej *y* dla *k*-tej jednostki,

$$\boldsymbol{B} = \left(\sum_U q_k \boldsymbol{x}_k \boldsymbol{x}_k^T \right)^{-1} \left(\sum_U q_k \boldsymbol{x}_k y_k \right),$$

q_k — waga ustalana przez badacza (zwykle $q_k = 1$ dla $k \in U$).

Estymatorem tej wariancji jest wyrażenie[29]:

$$\hat{D}^2(\hat{y}_{cal}) \approx \sum\sum_s \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} w_i (y_i - y_i^*) w_j (y_j - y_j^*) \tag{15}$$

gdzie:

$$y_k^* = \boldsymbol{x}_k^T \hat{\boldsymbol{B}}_w,$$

$$\hat{\boldsymbol{B}}_{\boldsymbol{w}} = \left(\sum_s w_i q_i \boldsymbol{x}_i \boldsymbol{x}_i^T \right)^{-1} \left(\sum_s w_i q_i \boldsymbol{x}_i y_i \right)$$

W przypadku losowania zrównoważonego trudno jest określić prawdopodobieństwa inkluzji drugiego rzędu, szuka się więc rozwiązań, które będą pomijały ten problem. Jednym z nich jest technika resztowa (*residual technique*), opierająca się na resztach z regresji zmiennej badanej względem zmiennych równoważących. Punktem wyjścia jest tu przyjęcie, że losowanie zrównoważone może być postrzegane jako warunkowe próbkowanie Poissona[30]. Pozwala to na wyrażenie wariancji w następujący sposób[31]:

$$D_{bal}^2(\hat{y}_{HT}) = \sum_U b_k (\breve{y}_k - \breve{y}_k^*)^2 \tag{16}$$

gdzie:

$b_k = \tilde{\pi}_k (1 - \tilde{\pi}_k)$,

$\tilde{\pi}_k$ — prawdopodobieństwo inkluzji pierwszego rzędu dla próbkowania Poissona,

[29] Ibid, s. 380.

[30] Próbkowanie/losowanie Poissona (*Poisson sampling*) polega na dokonaniu *N* niezależnych eksperymentów losowych, w wyniku których każdej jednostce w populacji zostaje przypisana wartość 1 lub 0, oznaczająca odpowiednio wylosowanie lub niewylosowanie jednostki do próby, w taki sposób, aby prawdopodobieństwo wylosowania jednostki było równe π_k (Särndal i in., 1997, s. 85). Warunkowe próbkowanie Poissona pozwala wylosować próbę o założonej wielkości, np. poprzez powtarzanie próbkowania do uzyskania założonego *n* (Tillé, 2006, s. 79—98).

[31] Tillé (2006), s. 170.

$\breve{y}_k = \frac{y_k}{\pi_k}$,

$\breve{y}_k^* = \breve{\boldsymbol{x}}_K^T \left(\sum_U b_l \breve{\boldsymbol{x}}_l \breve{\boldsymbol{x}}_l^T \right)^{-1} \sum_U b_l \breve{\boldsymbol{x}}_l \breve{y}_l$ — wartość teoretyczna dla regresji $\breve{y}_k$ względem $\breve{\boldsymbol{x}}_k$,

$\breve{\boldsymbol{x}}_k = \frac{\boldsymbol{x}_k}{\pi_k}$.

Ze względu na to, że losowanie jest zrównoważone, nie są znane wartości $\widetilde{\pi}_k$ (nie są tożsame z π_k) i tym samym możliwe jest jedynie przybliżenie wariancji estymatora sumy. Deville i Tillé (2005) podają cztery warianty doboru wartości b_k, które w większości przypadków dają przybliżenie wariancji z błędem względnym nie większym niż 10%.

W rozważanym tutaj przypadku (losowanie proste z estymatorem kalibracyjnym *vs*. losowanie zrównoważone z estymatorem HT) wariancja estymatora sumy jest w przybliżeniu taka sama i wynosi:

$$D^2(\hat{y}_{cal}) \approx D^2_{bal}(\hat{y}_{HT}) \approx N^2 \left(1 - \frac{n}{N}\right) \frac{S_e^2}{n} \tag{17}$$

gdzie $S_e^2 = \frac{1}{N-p} \sum_U (y_k - y_k^*)^2$ — wariancja resztowa regresji y względem zmiennych x.

Bez względu na dokładną postać wzoru, wariancja estymatora sumy w obu podejściach zależy od stopnia wyjaśnienia zmiennej badanej przez zmienne dodatkowe. Im pełniejsze będzie to wyjaśnienie, tym zmienność estymatora będzie mniejsza.

Stosunek do podejścia modelowego

Zarówno kalibracja, jak i losowanie zrównoważone nie wymagają formułowania modelu, gdyż są to techniki wywodzące się z podejścia randomizacyjnego. Przyjmując jednak punkt widzenia wspomagany modelem (*model-assisted*) można wykazać, że dla prostego modelu liniowego postaci: $y_k = \boldsymbol{x}_k^T \boldsymbol{\beta} + \varepsilon_k$ optymalną strategią badania jest losowanie próby zrównoważonej (z prawdopodobieństwami inkluzji pierwszego rzędu proporcjonalnymi do odchylenia standardowego błędów modelu) i estymator HT. Strategia taka będzie optymalna również ze *stricte* modelowego punktu widzenia, jeżeli heteroskedastyczność modelu jest w pełni wyjaśniona[32]. Bez względu na podejście, zawsze rozsądne

[32] Nedyalkova, Tillé (2008), s. 533.

jest wylosowanie próby w sposób zrównoważony, gdyż praktycznie zawsze przyniesie ono wzrost efektywności pod względem randomizacyjnym, natomiast z punktu widzenia nadpopulacyjnego ochroni przed błędną specyfikacją modelu.

Występowanie braków odpowiedzi

Kalibracja jest bardzo dobrym narzędziem do redukcji obciążenia wynikającego z braków odpowiedzi (też z błędów pomiaru i pokrycia). Särndal i Lundström (2005) poświęcili znaczną część swojej książki o estymacji w badaniach z brakami odpowiedzi technikom kalibracji. Z kolei w przypadku losowania zrównoważonego braków odpowiedzi nie powinno być wcale. Stąd losowanie zrównoważone dobrze się sprawdza w sytuacjach, w których nie występuje zagrożenie brakami odpowiedzi, np.: w losowaniu jednostek na pierwszych stopniach w losowaniu wielostopniowym, w losowaniu zespołów jednostek, statystycznej kontroli jakości, badaniach obowiązkowych itp.

LOSOWANIE ZRÓWNOWAŻONE I KALIBRACJA W JEDNYM BADANIU

Nic nie stoi na przeszkodzie, aby stosować losowanie zrównoważone i kalibrację w jednym badaniu. Z takiego łącznego zastosowania wynikają dodatkowe korzyści. Jak już napisano, losowanie zrównoważone prawie nigdy nie zapewnia dokładnego odwzorowania parametrów zmiennych dodatkowych, dlatego celowe jest w takiej sytuacji dokonanie kalibracji wag, aby uzyskać dokładne odwzorowanie. Zastosowanie wyłącznie kalibracji może skutkować dużą zmiennością wag (nawet wagami ujemnymi), co jest problematyczne, szczególnie przy szacowaniu wariancji. Kalibracja zastosowana do próby w przybliżeniu zrównoważonej zmieni wagi jedynie w niewielkim stopniu, ponieważ korekta dotyczyć będzie jedynie problemu zaokrągleń. Dzięki takiemu rozwiązaniu zmienność ostatecznych wag będzie znacznie mniejsza niż w przypadku zastosowania wyłącznie kalibracji[33].

Dodatkowo losowanie zrównoważone może się odbywać względem określonego zestawu zmiennych dodatkowych, natomiast kalibracja może być dokonana w stosunku do innego zestawu zmiennych dodatkowych. Zwykle po wylosowaniu próby istnieje możliwość kalibracji względem większej liczby zmiennych, gdyż wymagania co do zakresu znajomości zmiennych dodatkowych są mniejsze. W takiej sytuacji konieczne jest uwzględnienie w kalibracji również zmiennych wykorzystywanych do losowania zrównoważonego, ponieważ bez tego efekt zrównoważenia mógłby zostać utracony[34].

[33] Deville, Tillé (2004), s. 907.

[34] Wyjątkiem od tej zasady jest sytuacja, w której dane użyte do losowania zrównoważonego zdezaktualizowały się i na etapie estymacji dostępne są nowsze dane dotyczące tych samych zmiennych (Tillé, 2011, s. 223).

Poza zestawem zmiennych dodatkowych, inny może być również zakres populacji, do którego odnosi się dana technika. Czasami na etapie doboru próby celowo pomijane są niektóre jednostki populacji ze względów organizacyjnych (np. trudny dostęp, wysokie koszty) lub dane jednostkowe odnośnie zmiennych pomocniczych nie obejmują pewnego typu jednostek. W takich sytuacjach próba może być zrównoważona względem części populacji, natomiast aby można było wnioskować o całej populacji, kalibracja dokonywana jest w stosunku do znanych sum dla całej populacji. Przykładowo w badaniu *exit poll*[35] jako zmienne dodatkowe wykorzystać można wyniki przeszłych wyborów, znane dla każdego obwodu głosowania. Ze względu na to, że populacja obwodów zmienia się nieco pomiędzy wyborami, a także dlatego, że badanie nie jest przeprowadzane w niektórych obwodach (np. niepowszechnych), nie można wylosować próby zrównoważonej w stosunku do całej populacji. Ale próba może być zrównoważona w stosunku do znacznej części populacji, a następnie kalibracja wag w stosunku do tych samych zmiennych, ale odnoszących się do wszystkich jednostek, zapewni możliwość wnioskowania o całej populacji[36].

Podsumowanie

Losowanie zrównoważone oraz kalibracja wag umożliwiają osadzenie badania próbkowego w konkretnej rzeczywistości dzięki powiązaniu ze znanymi zmiennymi dodatkowymi. Dzięki temu wnioskowanie o populacji jest dokładniejsze, a szacunki cech pomocniczych są spójne z informacjami pochodzącymi z innych źródeł. Wpływa to na poprawę jakości statystyki z punktu widzenia użytkownika.

W artykule przedstawiono sposób doboru próby zrównoważonej za pomocą metody kostki. Algorytm postępowania zilustrowano za pomocą dwóch przykładów liczbowych. Zwrócono uwagę na to, że próba przeważnie nie będzie mogła być idealnie zrównoważona, ale zawsze możliwe jest zrównoważenie w przybliżeniu, co znacznie zmniejsza zmienność oszacowań sum cech dodatkowych.

Porównanie losowania zrównoważonego z kalibracją wypada korzystniej dla tej drugiej metody — głównie ze względu na zawsze dokładne odwzorowanie sum zmiennych dodatkowych bez względu na ich liczbę, mniejsze wymogi informacyjne i lepsze radzenie sobie z brakami odpowiedzi. Jednak kalibracja również ma swoje wady, dlatego postuluje się stosowanie obu metod w jednym badaniu, co połączy korzyści z nich wynikające, eliminując jednocześnie ich wady.

dr Arkadiusz Kozłowski — *Uniwersytet Gdański*

[35] Badanie przeprowadzane w dniu wyborów, w którym respondenci (wyborcy) opuszczający lokal wyborczy odpowiadają m.in. na kogo oddali swój głos.

[36] Szerzej na ten temat Kozłowski (2014).

LITERATURA

Bracha C., Jakubowski J., Szarkowski A. (2004), *Analiza porównawcza estymatorów regresyjnych w reprezentacyjnych badaniach statystycznych*, GUS ZBSE, Warszawa.

Deville J.-C., Särndal C.-E. (1992), *Calibration estimators in survey sampling*, „Journal of the American Statistical Association", Vol. 87, No. 418.

Deville J.-C., Särndal C.-E., Sautory O. (1993), *Generalized raking procedures in survey sampling*, „Journal of the American Statistical Association", Vol. 88, No. 423.

Deville J.-C., Tillé Y. (2004), *Efficient balanced sampling: The cube method*, „Biometrika", Vol. 91, No. 4.

Deville J.-C., Tillé Y. (2005), *Variance approximation under balanced sampling*, „Journal of Statistical Planning and Inference", Vol. 128, No. 2.

Józefowski T., Szymkowiak M. (2012), *Estymatory kalibracyjne w badaniach statystycznych*, „Wiadomości Statystyczne", nr 1.

Kozłowski A. (2012), *The usefulness of past data in sampling design for exit poll surveys*. „Studia Ekonomiczne", t. 120.

Kozłowski A. (2014), *The use of non-sample information in exit poll surveys in Poland*. „Statistics in Transition — new series", Vol. 15, No. 1.

Langel M., Tillé, Y. (2011), *Corrado Gini, a pioneer in balanced sampling and inequality theory*, „METRON — International Journal of Statistics", Vol. LXIX, No. 1.

Nedyalkova D., Tillé Y. (2008), *Optimal sampling and estimation strategies under the linear model*, „Biometrika", Vol. 95, No. 3.

Platek R., Särndal C.-E. (2001), *Czy statystyk może dostarczyć dane wysokiej jakości?* „Wiadomości Statystyczne", nr 4.

Särndal C.-E. (2007), *The calibration approach in survey theory and practice*, „Survey Methodology", Vol. 33, No. 2.

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

Särndal C.-E., Swensson B., Wretman J. H. (1997), *Model assisted survey sampling*, Springer, New York.

Szymkowiak M. (2009), *Imputacja i kalibracja — nowe możliwości estymacji w badaniach statystycznych z brakami odpowiedzi*, „Zeszyty Naukowe, Uniwersytet Ekonomiczny w Poznaniu", nr 116.

Tillé Y. (2006), *Sampling algorithms*, Springer, New York.

Tillé Y. (2011), *Ten years of balanced sampling with the cube method: An appraisal*, „Survey Methodology", Vol. 37, No. 2.

Valliant R., Dorfman A. H., Royall R. M. (2000), *Finite population sampling and inference: A prediction approach*, John Wiley&Sons, New York.

Summary. *A balanced sampling design is a design in which Horvitz-Thompson estimators of population totals for a set of auxiliary variables equal the known totals of these variables. On the other hand, calibration is a technique where the modification of design weights occurs in such a way that the new weights, when applied to auxiliary variables, reproduce, i.e. estimate without error, the known totals for these variables. The general idea behind balanced sampling and calibration is thus the same — both techniques tend to reproduce known totals of the auxiliary variables. The purpose of the paper is to describe*

and compare both techniques, considering them as alternatives in achieving the same goal. More attention was devoted to balanced sampling. The algorithm for selecting a sample was illustrated with two numerical examples. The comparison between balanced sampling and calibration, as alternatives, indicates calibration, but the best strategy is to use both methods simultaneously.
Keywords: balanced sampling, calibration.

Резюме. *Сбалансированная выборка заключается в такой выборке, чтобы оценки вспомогательных сумм величин оценкой Хорвица-Томсона были равны фактическим суммам этих величин. В то же время калибровка состит в модификации выходных весов являющихся результатом плана выборки так, чтобы модифицированные весы воссоздавали известные суммы вспомогательных величин. Идея обоих методов состоит в копировании значения глобальных дополнительных величин. Целью статьи является представление и сопоставление двух методов, которые считаются альтернативой для достижения той же цели. Больше внимание было уделено сбалансированной выборке. Алгоритм выборки представлен с помощью двух простых примеров. Сравнение сбалансированной выборки с калибровкой является выгодным для другого метода, но самым хорошим решением является одновременное использование обоих методов.*
Ключевые слова: сбалансированная выборка, калибровка.