Teoretyczne i praktyczne aspekty kalibracji w badaniach statystycznych

Marcin Szymkowiak

Katedra Statystyki
Uniwersytet Ekonomiczny w Poznaniu

07.01.2015

Plan prezentacji

1. Teoretyczne podstawy kalibracji
2. Kalibracja w badaniach pełnych
3. Kalibracja w NSP 2011

Kalibracja w ujęciu historycznym

- Kalibracja, w swych różnych formach, stała się na przestrzeni ostatnich lat ważną metodą wykorzystywaną w estymacji różnych parametrów w badaniach statystycznych z brakami odpowiedzi.
- Kalibracja — jako nowy termin w metodzie reprezentacyjnej pojawił się w literaturze około 20 lat temu.
- Podstawy teoretyczne kalibracji zostały sformułowane w pionierskiej pracy Särndala i Deville'a (1992) z początku lat 90-tych XX wieku.

Carl-Erik Särndal

- Szwedzki statystyk – wybitny znawca metody reprezentacyjnej
- Jeden z twórców podejścia kalibracyjnego
- Mistrz świata weteranów (75 lat i więcej) w skoku wzwyż i aktualny rekordzista świata w tej kategorii wiekowej (153 cm)

Jean Claude Deville

- Francuski statystyk – wybitny znawca metody reprezentacyjnej
- Jeden z twórców podejścia kalibracyjnego
- Pracownik francuskiego urzędu statystycznego INSEE

Formalne ujęcie kalibracji

- Zakładamy, że populacja $U = \{1, 2, \ldots, N\}$ składa się z N elementów.
- Z populacji tej losujemy zgodnie z określonym schematem losowania próbę $s \subseteq U$, składającą się z n elementów.
- Niech π_i oznacza prawdopodobieństwo inkluzji pierwszego rzędu tj. $\pi_i = P(i \in s)$ a $d_i = 1/\pi_i$ wagę przypisaną i–tej jednostce w procesie losowania.
- Zakładamy, że głównym celem jest oszacowanie wartości globalnej zmiennej y:

$$Y = \sum_{i=1}^{N} y_i, \tag{1}$$

gdzie y_i oznacza wartość zmiennej y dla i–tej jednostki, $i = 1, \ldots, N$.

Formalne ujęcie kalibracji

- Niech ponadto $x_1, \ldots, x_k$ oznaczają zmienne pomocnicze, a $\mathbf{X}_j$ oznacza wartość globalną zmiennej x_j, $j = 1, \ldots, k$, tj.

$$\mathbf{X}_j = \sum_{i=1}^{N} x_{ij}, \tag{2}$$

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

- Do oszacowania wartości globalnej zmiennej y wykorzystujemy estymator Horvitza-Thompsona:

$$\hat{Y}_{HT} = \sum_{i=1}^{n} d_i y_i. \tag{3}$$

- W praktyce bardzo często zdarza się, że:

$$\sum_{s} d_i x_{ij} \neq \mathbf{X}_j \tag{4}$$

co oznacza, że pewna korekta wag (kalibracja) jest pożądana.

Formalne ujęcie kalibracji

- Niech $\mathbf{d} = (d_1, \ldots, d_n)^T$ będzie wektorem wag wynikających ze schematu losowania próby, a $\mathbf{w} = (w_1, \ldots, w_n)^T$ poszukiwanym wektorem wag kalibracyjnych, gdzie n oznacza liczebność próby.
- Niech G będzie dowolną funkcją spełniającą następujące warunki:
  - $G(\cdot)$ jest dwukrotnie różniczkowalna,
  - $G(\cdot) \geq 0$,
  - $G(1) = 0$,
  - $G'(1) = 0$,
  - $G''(1) = 1$.
- Nowo wyznaczone wagi powinny nieznacznie się różnić od wag d_i oraz powinny spełniać warunek:

$$\sum_s w_i x_{ij} = \mathbf{X}_j . \tag{5}$$

Problem poszukiwania wag kalibracyjnych

(W1) Minimalizacja funkcji odległości:

$$D(\mathbf{w}, \mathbf{d}) = \sum_{i=1}^{n} d_i G\left(\frac{w_i}{d_i}\right) \longrightarrow \min, \tag{6}$$

(W2) Równania kalibracyjne:

$$\sum_{i=1}^{n} w_i x_{ij} = \mathbf{X}_j, \quad j = 1, \ldots, k, \tag{7}$$

(W3) Warunki ograniczające:

$$L \leq \frac{w_i}{d_i} \leq U, \quad \text{gdzie: } L < 1 \text{ i } U > 1, \ i = 1, \ldots, n. \tag{8}$$

Postać funkcji G

- Istnieje pewna dowolność przy wyborze funkcji $G(\cdot)$.
- Najczęściej rozważa się w literaturze następujące jej postacie:

$$G_1(x) = \frac{1}{2}(x-1)^2, \tag{9}$$

$$G_2(x) = \frac{(x-1)^2}{x}, \tag{10}$$

$$G_3(x) = x(\log x - 1) + 1, \tag{11}$$

$$G_4(x) = 2x - 4\sqrt{x} + 2, \tag{12}$$

$$G_5(x) = \frac{1}{2\alpha}\int_1^x \sinh\left[\alpha\left(t - \frac{1}{t}\right)\right] dt, \tag{13}$$

gdzie α jest dodatnim parametrem, pozwalającym sterować stopniem rozrzutu wag kalibracyjnych w stosunku do wag wynikających ze schematu losowania próby (domyślnie parametr przyjmuje wartość 1), a sinh jest funkcją sinusa hiperbolicznego zdefiniowanego jako $\sinh(x) = \frac{e^x - e^{-x}}{2}$.

Wybór funkcji G

- W praktycznych zastosowaniach najczęściej wykorzystuje się funkcję G w postaci $G_1(x) = \frac{1}{2}(x-1)^2$. w tym przypadku mamy bowiem:

$$D(\mathbf{w}, \mathbf{d}) = \sum_{i=1}^{n} d_i G\left(\frac{w_i}{d_i}\right) = \sum_{i=1}^{n} d_i \frac{1}{2}\left(\frac{w_i}{d_i} - 1\right)^2 = \frac{1}{2}\sum_{i=1}^{n} \frac{(w_i - d_i)^2}{d_i}. \quad (14)$$

Estymator kalibracyjny wartości globalnej

Estymatorem kalibracyjnym wartości globalnej zmiennej Y jest:

$$\hat{Y}_{cal} = \sum_{i=1}^{n} w_i y_i, \tag{15}$$

gdzie wektor wag kalibracyjnych $\mathbf{w} = (w_1, w_2, \ldots, w_n)^T$ jest rozwiązaniem zadania minimalizacji:

$$\mathbf{w} = \operatorname{argmin}_v D(\mathbf{v}, \mathbf{d}), \tag{16}$$

$$\mathbf{X} = \tilde{\mathbf{X}}, \tag{17}$$

przy czym

$$D(\mathbf{v}, \mathbf{d}) = \frac{1}{2} \sum_{i=1}^{n} \frac{(v_i - d_i)^2}{d_i}, \tag{18}$$

$$\tilde{\mathbf{X}} = \left(\sum_{i=1}^{n} w_i x_{i1}, \sum_{i=1}^{n} w_i x_{i2}, \ldots, \sum_{i=1}^{n} w_i x_{ik} \right)^T, \quad \mathbf{X} = \left(\sum_{i=1}^{N} x_{i1}, \sum_{i=1}^{N} x_{i2}, \ldots, \sum_{i=1}^{N} x_{ik} \right)^T. \tag{19}$$

Twierdzenie o wagach kalibracyjnych

Rozwiązaniem zadania minimalizacji jest wektor wag kalibracyjnych $\mathbf{w} = (w_1, w_2, \ldots, w_n)^T$, którego składowe spełniają równanie

$$w_i = d_i + d_i \left(\mathbf{X} - \hat{\mathbf{X}}\right)^T \left(\sum_{i=1}^{n} d_i \underline{\mathbf{x}}_i \underline{\mathbf{x}}_i^T\right)^{-1} \underline{\mathbf{x}}_i \tag{20}$$

przy czym:

$$\hat{\mathbf{X}} = \left(\sum_{i=1}^{n} d_i x_{i1}, \sum_{i=1}^{n} d_i x_{i2}, \ldots, \sum_{i=1}^{n} d_i x_{ik}\right)^T, \tag{21}$$

$$\underline{\mathbf{x}}_i = (x_{i1}, x_{i2}, \ldots, x_{ik})^T. \tag{22}$$

Kalibracja w programach statystycznych

Wraz z rozwojem metodologii pojawiło się odpowiednie oprogramowanie, które wykorzystywane jest w praktycznych zastosowaniach.

- **G-Calib** - wykorzystywany w Belgii.
- **Calmar** - wykorzystywany we Francji.
- **Bascula** - wykorzystywany w Holandii.
- **GES** - wykorzystywany w Kanadzie.
- **CLAN 97** - wykorzystywane w Szwecji.
- Programy te w większości napisane zostały w języku 4GL w systemie SAS.
- Wyjątek stanowi Bascula oprogramowana w Delphi oraz G-Calib, którego kod został zaimplementowany w pakiecie SPSS.
- **Program R** - pakiet **survey** i zaimplementowana w nim funkcja **calibrate()**.

Przykład

- Zakładamy, że ze sztucznej populacji złożonej z $N = 1000$ przedsiębiorstw losujemy zgodnie ze schematem losowania prostego próbę o liczebności $n = 20$. Wagi wynikające ze schematu losowania próby wynoszą więc $d_i = N/n = 1000/20 = 50$.
- Weźmy pod uwagę zmienną ciągłą x_1 (na przykład miesięczny przychód przedsiębiorstwa) oraz jedną zmienną jakościową x_2 (na przykład wielkość przedsiębiorstwa duże - L i średnie - M).
- W przykładzie tym pokazane będzie jedynie jak wyznaczać wagi kalibracyjne. Nie są brane pod uwagę wartości zmiennej y, które nie są potrzebne do wyznaczania wag kalibracyjnych.

Sztuczny zbiór danych

Numer przedsiębiorstwa	Miesięczny przychód x_1	Wielkość przedsiębiorstwa x_2	d_i
1	18	M	50
2	14	M	50
3	16	M	50
4	35	L	50
5	30	L	50
6	10	L	50
7	15	M	50
8	23	M	50
9	23	L	50
10	12	M	50
11	18	M	50
12	16	M	50
13	22	L	50
14	15	M	50
15	15	M	50
16	10	M	50
17	18	M	50
18	18	M	50
19	35	L	50
20	16	M	50

Przykład

- Ważona suma zmiennej x_1 jest równa 18950.
- Liczba średnich i dużych przedsiębiorstw wyznaczonych w oparciu o dane z próby jest równa 700 (14 średnich przedsiębiorstw x 50) i 300 (6 dużych przedsiębiorstw x 50) odpowiednio.
- **Założenie:** Zakładamy, że znany jest łączny przychód wszystkich przedsiębiorstw w populacji i wynosi on 19000. Zakładamy ponadto, że rzeczywista lista przedsiebiorstw średnich i dużych jest znana i wynosi odpowiednio 720 i 280.
- **Problem:** Należy zmienić wagi wynikające ze schematu losowania próby w taki sposób, aby odtworzone zostały znane wartości globalne zmiennych pomocniczych w populacji. Inaczej mówiąc, należy nieznacznie skalibrować wagi d_i, tak aby ważona suma zmiennej x_1 była równa 19000 a ważone sumy odpowiadające liczbie średnich i dużych przedsiębiorstw wynosiły odpowiednio 720 i 280.
- **Rozwiązanie:** Skorzystaj z kalibracji
- Kod programu SAS, który rozwiązuje powyższy problem tworząc niezbędne zbiory danych i wywołując makro CALMAR2 podany jest na następnych slajdach.

Rozwiązanie – kod makra CALMAR2

```
/**********Utworzenie sztucznego zbioru danych************/
data proba;
input przedsiebiorstwo $ wielkosc $ przychod waga;
cards;
p01     M           18   50
p02     M           14   50
p03     M           16   50
p04     L           35   50
p05     L           30   50
p06     L           10   50
p07     M           15   50
p08     M           23   50
p09     L           23   50
p10     M           12   50
p11     M           18   50
p12     M           16   50
p13     L           22   50
p14     M           15   50
p15     M           15   50
p16     M           10   50
p17     M           18   50
p18     M           18   50
p19     L           35   50
p20     M           16   50
;
run;
```

Rozwiązanie – kod makra CALMAR2

```
libname calm 'D:\Lamborghini\Calibration';
options mstored sasmstore=calm;
/**********Utworzenie zbioru z wartościami globalnymi***********/
data globalne;
input var $ n mar1 mar2;
cards;
wielkosc 2 280 720
przychod 0 19000 .
;
run;

/********************Bibliotek zawierająca makro CALMAR************************/
libname calm 'D:\Sopot';
options mstored sasmstore=calm;

/**********************Call to CALMAR*******************************/
%CALMAR2(DATAMEN=Proba, POIDS=waga, IDENT=przedsiebiorstwo,
MARMEN=Globalne, M=1,DATAPOI=wcal, POIDSFIN=cal_weights)
```

Przykład – wagi kalibracyjne

Numer przedsiębiorstwa	Miesięczny przychód x_1	Wielkość przedsiębiorstwa x_2	d_i	w_i
1	18	M	50	52,275
2	14	M	50	50,5821
3	16	M	50	51,4286
4	35	L	50	50,5462
5	30	L	50	48,4301
6	10	L	50	39,9657
7	15	M	50	51,0054
8	23	M	50	54,3911
9	23	L	50	45,4675
10	12	M	50	49,7357
11	18	M	50	52,275
12	16	M	50	51,4286
13	22	L	50	45,0443
14	15	M	50	51,0054
15	15	M	50	51,0054
16	10	M	50	48,8893
17	18	M	50	52,275
18	18	M	50	52,275
19	35	L	50	50,5462
20	16	M	50	51,4286

Braki odpowiedzi jako główne źródło błędów nielosowych

- Występują zarówno w badaniach pełnych jak i częściowych.
- Braki odpowiedzi są źródłem wielu „zaburzeń". Jest tak dlatego, że osoby, które odmawiają wzięcia udziału w badaniu bądź nie udzielają na niektóre pytania odpowiedzi, na ogół różnią się od tych, co biorą w nim udział i dostarczają niezbędnych danych.
- Zmniejsza się efektywny rozmiar badanej próby bądź populacji, co ma niekorzystny wpływ na wariancję estymatorów powodując ich zwiększenie.
- Uzyskane wyniki obciążone są zbyt dużymi błędami. Wyznaczone oceny parametrów znacznie odbiegają od ich „prawdziwych" wartości, a skonstruowane na podstawie próby przedziały ufności różnych parametrów, koncentrują się wokół „złych" wartości.
- Rozkłady wielu cech są zniekształcone i niemożliwe będzie zastosowanie wielu klasycznych metod statystycznych.
- Zbyt niski wskaźnik udzielonych odpowiedzi nie wpływa korzystnie na pozytywne postrzeganie badania przez jego użytkowników i w skrajnych przypadkach może się ono okazać dla nich całkowicie bezużyteczne.

Metody niwelujące wpływ braków odpowiedzi

- W praktyce badań statystycznych stosuje się różnego rodzaju metody, których celem jest zwiększenie frakcji udzielonych odpowiedzi.
- Mają one zarówno zastosowanie na etapie zbierania danych (na przykład powtórne badanie jednostek, od których nie uzyskano danych, zastępowanie jednostek nie podejmujących badania innymi, stosowanie różnych bodźców — na przykład finansowych) oraz na etapie ich opracowywania (na przykład imputacja, kalibracja).
- Generalnie metody te można podzielić na trzy zasadnicze grupy: prewencyjne, redukujące frakcję braków odpowiedzi oraz korygujące.
- Granica pomiędzy poszczególnymi technikami w ramach wyróżnionych grup nie zawsze jest ostra, przy czym można jednak przyjąć w ogólności, że podejście prewencyjne ma miejsce na etapie planowania badania przed zebraniem danych, redukcja braków odpowiedzi odbywa się na etapie ich zbierania, a korygowanie odbywa się w procesie estymacji, kiedy zebrano już niezbędne informacje.

Metody prewencyjne

- Wywodzą się z nauk o zachowaniu się jednostek (psychologii, socjologii) — co jest naturalną konsekwencją faktu, że proces zbierania danych wymaga kontaktu z respondentem.
- Niezbędna jest tutaj znajomość technik mających przełamać sceptycyzm i niechęć respondenta do udzielania informacji oraz promujących pozytywne nastawienie do całego badania.
- Dużą rolę odgrywają w ramach tej grupy metod, czynniki motywacyjne mające przekonać jednostkę do wzięcia udziału w badaniu.
- Metody prewencyjne obejmują również zagadnienie konstrukcji kwestionariusza ankietowego, odpowiednie przeszkolenie ankietera, sposób zbierania danych oraz właściwe przygotowanie operatu losowania.

Metody redukujące frakcję braków odpowiedzi

- Obejmują m.in. wysyłanie monitów z prośbą o wzięcie udziału w badaniu, ponowny kontakt telefoniczny, stosowanie bodźców finansowych, zastępowanie jednostek, które nie wyrażają chęci wzięcia udziału w badaniu innymi itd.
- W przypadku stosowania zastępowania, zwykle jednostki zastępcze wybiera się z próby rezerwowej kierując się zasadą, aby miały one podobne cechy podstawowe jak jednostki nie podejmujące badań.
- Nie jest to jednak regułą, gdyż w badaniu budżetów gospodarstw domowych, jednostki zastępcze losuje się, a więc mogą się one diametralnie różnić od jednostek wylosowanych pierwotnie do próby.
- Podobnie jak metody prewencyjne, w znacznej mierze wywodzą się one z nauk o zachowaniu się jednostek.

Metody statystyczne

- Jest to grupa metod obejmująca różnego rodzaju techniki estymacji i metody ważenia danych, których celem jest zniwelowanie obciążenia będącego konsekwencją wystąpienia w badaniu braków odpowiedzi.
- Z racji tego, że w każdym — nawet najlepiej zaplanowanym badaniu — występują braki danych, metody statystyczne, rozwijane w ramach tej grupy, odgrywają coraz większą rolę.
- W światowej literaturze przedmiotu „nonresponse" przeszedł znamienną ewolucję – od podejścia polegającego na ograniczeniu się w procesie estymacji tylko do zbioru tych jednostek, dla których znane są wartości analizowanych cech, poprzez wysiłki uzyskania odpowiedzi „za wszelką cenę", aż do estymacji wykorzystującej alternatywne źródła informacji pośredniej.
- W literaturze przedmiotu przedstawia się dwie podstawowe metody stosowane w przypadku wystąpienia braków odpowiedzi w badaniach statystycznych: imputację i metody korygujące wyniki badań.

Imputacja

- Imputacja to metoda szacowania brakujących lub eliminowania niepoprawnych danych.
- Zastosowanie imputacji prowadzi do przypisania każdej jednostce w miejsce brakujących lub nieważnych danych jakiejś wartości.
- Brakujące dane uzupełniane są ich „substytutami" i są one z samej definicji „wartościami sztucznymi".
- W badaniach statystycznych, metody imputacji rozwinęły się na potrzeby spisów przeprowadzanych w różnych państwach na przełomie lat 50–tych XX wieku.
- Rozwój metod imputacji możliwy był dzięki postępowi, jaki miał miejsce w informatyce w latach 60–tych XX wieku.
- Intensywny rozwój teorii w zakresie imputacji miał miejsce w latach 80–tych XX wieku. Istotny wkład w tym zakresie miała pionierska praca Rubina (1976) oraz Little'a i Rubina (1987), w których przedstawiono po raz pierwszy w kompleksowy sposób metody imputacji, które następnie stosowano z powodzeniem w wielu badaniach statystycznych.

Założenia

Aby imputacja odegrała swoją rolę w badaniu muszą być spełnione trzy ważne założenia:

- imputacja nie powinna prowadzić do obciążeń bądź zmian rozkładów cech w zbiorze danych oraz do wzrostu wariancji stosowanych estymatorów,
- proces imputacji w większym stopniu powinien być uzależniony od danych pochodzących z próby aniżeli odwoływać się do założeń, co do natury brakujących danych,
- oszacowania ważnych statystyk z próby nie powinny „zbyt mocno" opierać się na imputowanych danych.

Klasyfikacja metod imputacji

Imputowane wartości można zaklasyfikować do jednej z trzech głównych kategorii:

- wartości imputowane z wykorzystaniem statystycznych reguł predykcyjnych,
- wartości imputowane uzyskiwane od jednostek badania mających podobne cechy,
- wartości imputowane w oparciu o opinię ekspertów.

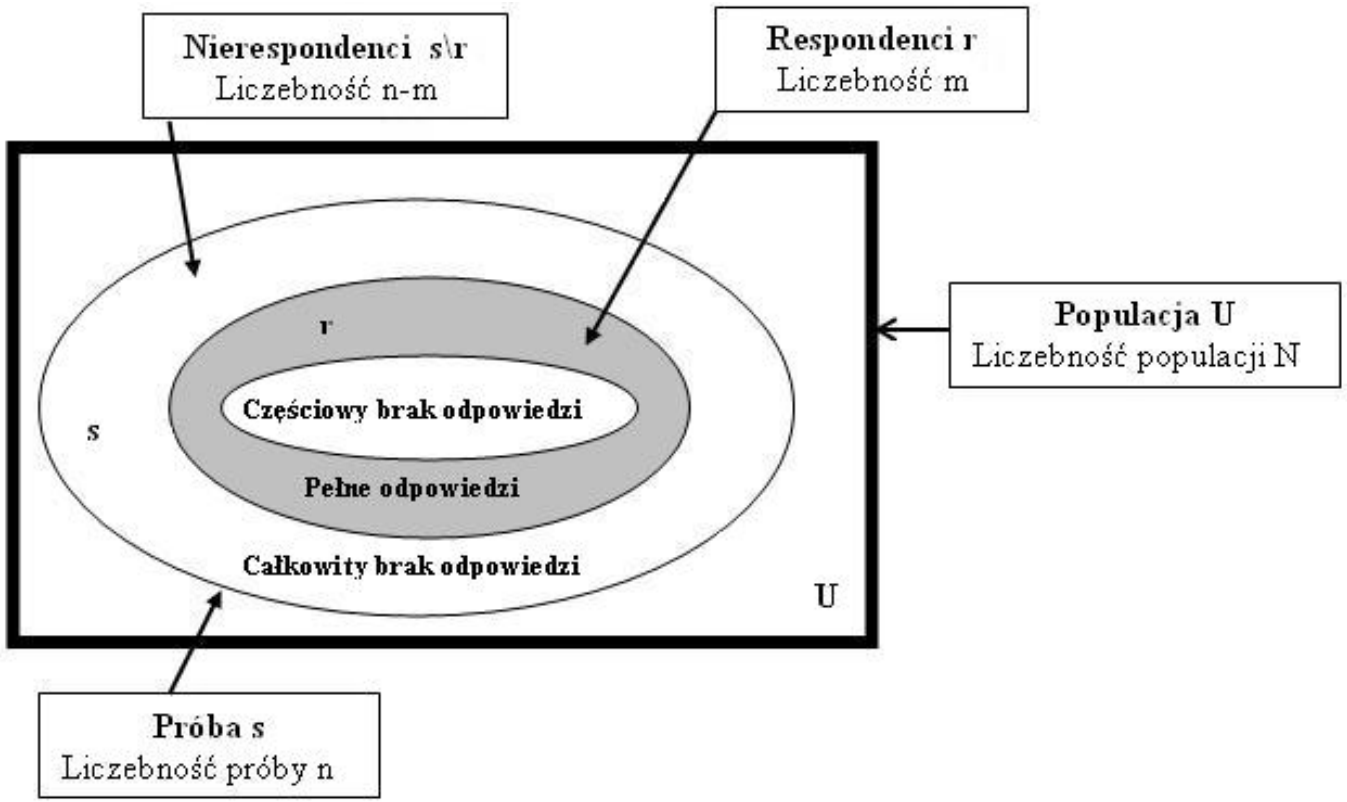
Kalibracja [C-E.Särndal, S. Lundström, 2005]

- Kalibracja – metoda polegająca na skorygowaniu wyjściowych wag celem redukcji obciążenia wynikającego z istnienia braków odpowiedzi.
- Wagi te obliczane są w oparciu o wykorzystanie informacji dodatkowych.
- W rezultacie uzyskuje się równowagę rozumianą w ten sposób, że po zastosowaniu kalibracji próba jest „wyglądem" zbliżona do całej populacji.

Podejście kalibracyjne [C-E.Särndal, 2007]

Podejście kalibracyjne w estymacji parametrów w odniesieniu do skończonych populacji składa się z :

- obliczenia wag z uwzględnieniem informacji dodatkowych, tak aby spełnione było odpowiednie równanie — tzw. równanie kalibracyjne,
- wykorzystania tych wag do estymacji wartości globalnej bądź innych parametrów, przy czym wartość zmiennej mnożona jest przez wagę, a sumowanie odbywa się po zbiorze wszystkich respondentów,
- uzyskania w ten sposób nieobciążonych oszacowań parametrów, w przypadku gdyby w badaniu nie wystąpiły braki odpowiedzi oraz inne błędy nielosowe.


Nierespondenci s\r
Liczebność n-m
Respondenci r
Liczebność m
r
Częściowy brak odpowiedzi
s
Pełne odpowiedzi
Całkowity brak odpowiedzi
U
Populacja U
Liczebność populacji N
Próba s
Liczebność próby n

Wpływ braków odpowiedzi i wybór funkcji G

- W przypadku gdy w badaniu, w odniesieniu do zmiennej y, wystąpią braki odpowiedzi estymator Horvitza-Thompsona przyjmuje postać:

$$\hat{Y}_{HT} = \sum_{r} d_i y_i = \sum_{i=1}^{m} d_i y_i . \tag{23}$$

- W literaturze na potrzeby wyznaczania wag kalibracyjnych najczęściej wykorzystuje się funkcję kwadratową $G_1(x) = \frac{1}{2}(x-1)^2$. W tym przypadku:

$$D(\mathbf{w}, \mathbf{d}) = \sum_{i=1}^{m} d_i G\left(\frac{w_i}{d_i}\right) = \sum_{i=1}^{m} d_i \frac{1}{2}\left(\frac{w_i}{d_i} - 1\right)^2 = \frac{1}{2}\sum_{i=1}^{m} \frac{(w_i - d_i)^2}{d_i} . \tag{24}$$

Problem poszukiwania wag kalibracyjnych

(W1) Minimalizacja funkcji odległości:

$$D(\mathbf{w}, \mathbf{d}) = \frac{1}{2} \sum_{i=1}^{m} \frac{(w_i - d_i)^2}{d_i} \longrightarrow \min, \tag{25}$$

(W2) Równania kalibracyjne:

$$\sum_{i=1}^{m} w_i x_{ij} = \mathbf{X}_j, \quad j = 1, \ldots, k, \tag{26}$$

(W3) Warunki ograniczające:

$$L \leq \frac{w_i}{d_i} \leq U, \quad \text{gdzie: } L < 1 \text{ i } U > 1, \ i = 1, \ldots, m. \tag{27}$$

Estymator kalibracyjny wartości globalnej

Estymatorem kalibracyjnym wartości globalnej zmiennej Y jest:

$$\hat{Y}_{cal} = \sum_{i=1}^{m} w_i y_i, \tag{28}$$

gdzie wektor wag kalibracyjnych $\mathbf{w} = (w_1, w_2, \ldots, w_m)^T$ jest rozwiązaniem zadania minimalizacji:

$$\mathbf{w} = \operatorname{argmin}_v D(\mathbf{v}, \mathbf{d}), \tag{29}$$

$$\mathbf{X} = \tilde{\mathbf{X}}, \tag{30}$$

przy czym

$$D(\mathbf{v}, \mathbf{d}) = \frac{1}{2} \sum_{i=1}^{m} \frac{(v_i - d_i)^2}{d_i}, \tag{31}$$

$$\tilde{\mathbf{X}} = \left(\sum_{i=1}^{m} w_i x_{i1}, \sum_{i=1}^{m} w_i x_{i2}, \ldots, \sum_{i=1}^{m} w_i x_{ik} \right)^T, \quad \mathbf{X} = \left(\sum_{i=1}^{N} x_{i1}, \sum_{i=1}^{N} x_{i2}, \ldots, \sum_{i=1}^{N} x_{ik} \right)^T. \tag{32}$$

Twierdzenie o wagach kalibracyjnych

Rozwiązaniem zadania minimalizacji jest wektor wag kalibracyjnych $\mathbf{w} = (w_1, w_2, \ldots, w_m)^T$, którego składowe spełniają równanie

$$w_i = d_i + d_i \left(\mathbf{X} - \hat{\mathbf{X}}\right)^T \left(\sum_{i=1}^{m} d_i \underline{\mathbf{x}}_i \underline{\mathbf{x}}_i^T\right)^{-1} \underline{\mathbf{x}}_i \tag{33}$$

przy czym:

$$\hat{\mathbf{X}} = \left(\sum_{i=1}^{m} d_i x_{i1}, \sum_{i=1}^{m} d_i x_{i2}, \ldots, \sum_{i=1}^{m} d_i x_{ik}\right)^T, \tag{34}$$

$$\underline{\mathbf{x}}_i = (x_{i1}, x_{i2}, \ldots, x_{ik})^T. \tag{35}$$

▸ Wróć

Pożądane własności estymatora kalibracyjnego

- małe obciążenie,
- mała wariancja,
- systemem wag, który „naśladuje” informacje zawarte w wektorze zmiennych dodatkowych,
- system wag, który jest użyteczny do szacowania parametrów dla różnych zmiennych w danym badaniu.

Co umożliwia kalibracja?

- poprawę efektywności,
- „równowagę" – rozumiana w ten sposób, że po zastosowaniu kalibracji próba wyglądem jest zbliżona do całej populacji,
- „bardziej logiczne szacunki" – po zastosowaniu kalibracji wyniki są bardziej logiczne.

Kalibracja w badaniach pełnych

- Kalibracja jako metoda szacowania parametrów w populacji generalnej polegająca na korygowaniu wag wynikających ze schematu losowania próby – z samej definicji – jest techniką statystyczną wykorzystywaną w badaniach reprezentacyjnych.
- Ponieważ w badaniach pełnych (spisy, różnego rodzaju rejestry administracyjne) występuje również problem braków odpowiedzi, naturalnym wydaje się odpowiednie przeważenie danych celem uzyskania zadowalających wyników.
- Podejście kalibracyjne znane z badań reprezentacyjnych może być wykorzystywane w badaniach pełnych.
- Punkt wyjścia do jej zastosowania stanowi sposób ustalania wag, według następującego algorytmu:
  - ustalenie sztucznych - wyjściowych wag,
  - wybór zmiennych wspomagających,
  - wyznaczenie wag kalibracyjnych.

Przykładowy rejestr

L.p.	Płeć	Zamieszkanie	Zatrudnienie	Wykształcenie
1	M	M	·	P
2	K	W	P	S
3	K	W	B	·
4	M	W	P	W
5	M	M	·	W
6	M	M	B	W
7	K	M	B	W
8	K	W	P	W
9	K	M	P	S
10	M	M	P	S
11	K	W	B	S
12	M	W	P	W
13	K	M	P	P
14	M	W	P	·
15	M	M	B	S
16	K	W	P	S
17	K	M	B	W
18	M	W	B	P
19	M	M	B	·
20	K	M	P	S

Tworzenie tablic – wpływ braków danych

- Załóżmy, że celem badania jest stworzenie tabeli ukazującej status zatrudnienia w zależności od wykształcenia osoby. Ze względu na występujące w rejestrze braki danych uzyskana tabela nie będzie odpowiednia.
- Opis zmiennych (Płeć: M-mężczyzna, K-kobieta; Zamieszkanie: M-miasto, W-wieś; Zatrudnienie: P-osoba pracująca, B-osoba bezrobotna; Wykształcenie: P-podstawowe, S-średnie, W-wyższe)

	Status zatrudnienia		
Wykształcenie	Pracująca	Bezrobotna	Razem
Wyższe	3	3	6
Średnie	5	2	7
Podstawowe	1	1	2
Razem	9	6	15

Algorytm wyznaczania wag kalibracyjnych – krok 1

- Wyjściowe wagi przypisujemy sztucznie w ten sposób, że dla jednostki dla której nie jest znana wartość co najmniej jednej z interesujących nas cech $d_i = 0$. Z kolei gdy znane są wartości wszystkich cech, w oparciu o które tworzona będzie tabela przyjmujemy $d_i = 1$.

L.p.	Płeć	Zamieszkanie	Zatrudnienie	Wykształcenie	d_i
1	M	M	·	P	0
2	K	W	P	S	1
3	K	W	B	·	0
4	M	W	P	W	1
5	M	M	·	W	0
6	M	M	B	W	1
7	K	M	B	W	1
8	K	W	P	W	1
9	K	M	P	S	1
10	M	M	P	S	1
...	...	...	...	...	...

Algorytm wyznaczania wag kalibracyjnych – krok 2

- W następnym kroku dokonujemy wyboru zmiennych, dla których znane są wartości dla wszystkich jednostek w rejestrze. Ponieważ informacja o płci osoby jak i jej miejscu zamieszkania znana jest dla wszystkich osób, zmienne te można wykorzystać celem ustalenia nowych wag w_i. Na potrzeby przykładu przyjęte zostały trzy zmienne pomocnicze: $\mathbf{x}_{i1}$, $\mathbf{x}_{i2}$, $\mathbf{x}_{i3}$.

$$\mathbf{x}_{i1} = \begin{cases} 1 & \text{jeżeli } i\text{-ta osoba jest kobietą,} \\ 0 & \text{jeżeli } i\text{-ta osoba jest mężczyzną,} \end{cases} \tag{36}$$

$$\mathbf{x}_{i2} = \begin{cases} 1 & \text{jeżeli } i\text{-ta osoba jest mężczyzną,} \\ 0 & \text{jeżeli } i\text{-ta osoba jest kobietą,} \end{cases} \tag{37}$$

$$\mathbf{x}_{i3} = \begin{cases} 1 & \text{jeżeli } i\text{-ta osoba mieszka w mieście,} \\ 0 & \text{jeżeli } i\text{-ta osoba mieszka na wsi.} \end{cases} \tag{38}$$

L.p.	Płeć	Zamieszkanie	Zatrudnienie	Wykształcenie	d_i	$\mathbf{x}_{i1}$	$\mathbf{x}_{i2}$	$\mathbf{x}_{i3}$
1	M	M	·	P	0	0	1	1
2	K	W	P	S	1	1	0	0
3	K	W	B	·	0	1	0	0
4	M	W	P	W	1	0	1	0
...	...	...	...	...	...	...	...	...

Algorytm wyznaczania wag kalibracyjnych – krok 3

- Tworzymy wektor złożony z wartości globalnych wszystkich zmiennych pomocniczych $\mathbf{X}$ oraz wektor oszacowanych wartości globalnych $\hat{\mathbf{X}}$.
- W naszym przykładzie mamy: $\mathbf{X} = (10, 10, 11)^T$, $\hat{\mathbf{X}} = (9, 6, 8)^T$.
- Następnie wyznaczamy wagi kalibracyjne w_i korzystając ze wzoru (33).

▸ Twierdzenie o wagach kalibracyjnych

Wagi kalibracyjne w_i

L.p.	Płeć	Zamieszkanie	Zatrudnienie	Wykształcenie	d_i	$\mathbf{x}_{i1}$	$\mathbf{x}_{i2}$	$\mathbf{x}_{i3}$	w_i
1	M	M	·	P	0	0	1	1	0
2	K	W	P	S	1	1	0	0	1,0447761
3	K	W	B	·	0	1	0	0	0
4	M	W	P	W	1	0	1	0	1,6069652
5	M	M	·	W	0	0	1	1	0
6	M	M	B	W	1	0	1	1	1,7263682
7	K	M	B	W	1	1	0	1	1,1641791
8	K	W	P	W	1	1	0	0	1,0447761
9	K	M	P	S	1	1	0	1	1,1641791
10	M	M	P	S	1	0	1	1	1,7263682
11	K	W	B	S	1	1	0	0	1,0447761
12	M	W	P	W	1	0	1	0	1,6069652
13	K	M	P	P	1	1	0	1	1,1641791
14	M	W	P	·	0	0	1	0	0
15	M	M	B	S	1	0	1	1	1,7263682
16	K	W	P	S	1	1	0	0	1,0447761
17	K	M	B	W	1	1	0	1	1,1641791
18	M	W	B	P	1	0	1	0	1,6069652
19	M	M	B	·	0	0	1	1	0
20	K	M	P	S	1	1	0	1	1,1641791

Tworzenie tablic – bez uwzględniania wag kalibracyjnych

	Status zatrudnienia		
Wykształcenie	Pracująca	Bezrobotna	Razem
Wyższe	3	3	6
Średnie	5	2	7
Podstawowe	1	1	2
Razem	9	6	15

Tworzenie tablic – z uwzględnieniem wag kalibracyjnych

	Status zatrudnienia		
Wykształcenie	Pracująca	Bezrobotna	Razem
Wyższe	4,27	4,05	8,32
Średnie	6,14	2,77	8,91
Podstawowe	1,16	1,61	2,77
Razem	11,57	8,43	20

Założenia badania symulacyjnego

Celem oceny własności estymatorów kalibracyjnych w rejestrze PESEL badanie zrealizowano w następujących krokach:

- Utworzono bazę danych o wszystkich osobach powiatu chodzieskiego, które miały ukończone co najmniej 15 lat. Bazę tą w dalszej kolejności ograniczono do tych osób, dla których istniałą informacja o ich stanie cywilnym. Tak skonstruowana baza stanowiła punkt wyjścia do utworzenia tablicy o ludności w wieku 15 lat i więcej według stanu cywilnego, wieku i płci.
- W drugim etapie usuwano według różnych wariantów część informacji o stanie cywilnym osób i zastępowano ją brakiem danych. Następnie wyznaczano wagi kalibracyjne, w oparciu o które dokonywano utworzenia analogicznej tablicy o ludności w wieku 15 lat i więcej według stanu cywilnego, wieku i płci.
- W ostatnim etapie badań dokonywano analizy uzyskanych wyników. W tym celu porównywano na ile liczby osób w poszczególnych komórkach utworzonej tablicy z wykorzystaniem wag kalibracyjnych zbieżne były z prawdziwymi wartościami znanymi dla sztucznie utworzonej bazy.

Wybór zmiennych pomocniczych

- Jako zmienne pomocnicze w przeprowadzonym badaniu przyjęto płeć oraz wiek osoby.
- W odniesieniu do wieku utworzono jego 10 grup zgodnie z podziałem, który stosowany jest w roczniku demograficznym na potrzeby konstrukcji tablic o liczbie ludności w wieku 15 lat i więcej według stanu cywilnego, wieku i płci. Przyjęto zatem następujące grupy wiekowe: 15–19 lat, 20–24, 25–29, 30–34, 35–39, 40–44, 45–49, 50–54, 55-59, 60 lat i więcej.
- W odniesieniu do stanu cywilnego przyjęto podział zgodny z tym jaki występuje w PESEL-u tj., wyodrębniono 8 jego kategorii: kawaler, panna, żonaty, zamężna, rozwiedziony, rozwiedziona, wdowiec oraz wdowa.
- Braki danych w odniesieniu do stanu cywilnego tworzono na cztery różne sposoby. Trzy pierwsze podejścia polegały na tym, że losowano po 5%, 10% oraz 20% osób z wcześniej utworzonej bazy danych (obejmującej osoby z powiatu chodzieskiego w wieku 15 lat i więcej dla których stan cywilny był znany) i zmieniano faktyczny stan cywilny takiej osoby na brak danych. Czwarte podejście polegało na losowym zastąpieniu po 20% informacji o stanie cywilnym w dwóch grupach wiekowych 20–24 oraz 25–29 lat oraz 5% informacji we wszystkich pozostałych grupach wiekowych rozpatrywanych łącznie. Miało to na celu większe, aniżeli w trzech pierwszych podejściach, zniekształcenie rozkładu analizowanych cech i sprawdzenie jak w takim przypadku zachowują się estymatory kalibracyjne.

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

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

$$x_{i3} = \begin{cases} 1 & \text{jeżeli } i\text{-ta osoba ma wiek z przedziału 15–19 lat,} \\ 0 & \text{w przeciwnym wypadku,} \end{cases}$$

$$x_{i4} = \begin{cases} 1 & \text{jeżeli } i\text{-ta osoba ma wiek z przedziału 20–24 lat,} \\ 0 & \text{w przeciwnym wypadku,} \end{cases}$$

$$x_{i5} = \begin{cases} 1 & \text{jeżeli } i\text{-ta osoba ma wiek z przedziału 25–29 lat,} \\ 0 & \text{w przeciwnym wypadku,} \end{cases}$$

$$x_{i6} = \begin{cases} 1 & \text{jeżeli } i\text{-ta osoba ma wiek z przedziału 30–34 lat,} \\ 0 & \text{w przeciwnym wypadku,} \end{cases}$$

$$x_{i7} = \begin{cases} 1 & \text{jeżeli } i\text{-ta osoba ma wiek z przedziału 35–39 lat,} \\ 0 & \text{w przeciwnym wypadku,} \end{cases}$$

$$x_{i8} = \begin{cases} 1 & \text{jeżeli } i\text{-ta osoba ma wiek z przedziału 40–44 lat,} \\ 0 & \text{w przeciwnym wypadku,} \end{cases}$$

$$x_{i9} = \begin{cases} 1 & \text{jeżeli } i\text{-ta osoba ma wiek z przedziału 45–49 lat,} \\ 0 & \text{w przeciwnym wypadku,} \end{cases}$$

$$x_{i10} = \begin{cases} 1 & \text{jeżeli } i\text{-ta osoba ma wiek z przedziału 50–54 lat,} \\ 0 & \text{w przeciwnym wypadku,} \end{cases}$$

$$x_{i11} = \begin{cases} 1 & \text{jeżeli } i\text{-ta osoba ma wiek z przedziału 55–59 lat,} \\ 0 & \text{w przeciwnym wypadku.} \end{cases}$$

Ludność w wieku 15 lat i więcej według stanu cywilnego, wieku i płci w 2008 roku w powiecie chodzieskim – wartości rzeczywiste

Wyszczególnienie	Ogółem	kawaler panna	żonaty zamężna	rozwiedziony rozwiedziona	wdowiec wdowa
Ogółem	36106	11131	20739	1222	3014
15-19 lat	2535	2526	9	-	-
20-24	3612	3197	408	6	1
25-29	4229	2430	1733	61	5
30-34	3686	1021	2513	143	9
35-39	2887	481	2228	160	18
40-44	2512	334	2002	140	36
45-49	2874	332	2291	173	78
50-54	3414	282	2726	205	201
55-59	3071	184	2466	137	284
60 i więcej	7286	344	4363	197	2382
Kobiety	18550	4806	10539	696	2509
15-19 lat	1209	1200	9	-	-
20-24	1758	1449	302	6	1
25-29	2054	1004	1004	41	5
30-34	1798	363	1347	79	9
35-39	1382	169	1114	85	14
40-44	1245	106	1021	85	33
45-49	1435	106	1172	96	61
50-54	1748	100	1379	102	167
55-59	1550	84	1158	77	231
60 i więcej	4371	225	2033	125	1988

Ludność w wieku 15 lat i więcej według stanu cywilnego, wieku i płci w 2008 roku w powiecie chodzieskim – podejście 1

Wyszczególnienie	Ogółem	kawaler panna	żonaty zamężna	rozwiedziony rozwiedziona	wdowiec wdowa
Ogółem	36106	11125	20748	1215	3018
15-19 lat	2535	2526	9	-	-
20-24	3612	3192	412	6	1
25-29	4229	2434	1729	62	4
30-34	3686	1024	2512	140	10
35-39	2887	479	2231	161	16
40-44	2512	336	2001	139	37
45-49	2874	327	2297	170	81
50-54	3414	282	2730	200	203
55-59	3071	181	2465	141	283
60 i więcej	7286	344	4363	196	2384
Kobiety	18550	4797	10547	686	2520
15-19 lat	1197	1188	9	-	-
20-24	1764	1451	306	6	1
25-29	2060	1007	1007	42	4
30-34	1794	363	1348	74	10
35-39	1380	169	1112	86	13
40-44	1248	106	1023	84	34
45-49	1429	105	1167	94	63
50-54	1746	99	1383	95	169
55-59	1554	84	1159	79	232
60 i więcej	4379	226	2033	125	1995

Ludność w wieku 15 lat i więcej według stanu cywilnego, wieku i płci w 2008 roku w powiecie chodzieskim – podejście 2

Wyszczególnienie	Ogółem	kawaler panna	żonaty zamężna	rozwiedziony rozwiedziona	wdowiec wdowa
Ogółem	36106	11132	20734	1222	3018
15-19 lat	2535	2526	9	-	-
20-24	3612	3200	405	6	1
25-29	4229	2423	1737	64	4
30-34	3686	1017	2518	141	10
35-39	2887	476	2235	157	19
40-44	2512	344	1991	139	38
45-49	2874	333	2289	178	75
50-54	3414	288	2721	207	199
55-59	3071	186	2464	135	287
60 i więcej	7286	339	4366	196	2385
Kobiety	18550	4805	10536	698	2511
15-19 lat	1215	1206	9	-	-
20-24	1752	1444	301	6	1
25-29	2071	1009	1016	42	4
30-34	1797	363	1346	78	10
35-39	1375	164	1112	84	14
40-44	1255	109	1027	84	36
45-49	1423	100	1166	98	58
50-54	1744	106	1366	104	169
55-59	1554	84	1160	78	231
60 i więcej	4363	218	2033	125	1987

Ludność w wieku 15 lat i więcej według stanu cywilnego, wieku i płci w 2008 roku w powiecie chodzieskim – podejście 3

Wyszczególnienie	Ogółem	kawaler panna	żonaty zamężna	rozwiedziony rozwiedziona	wdowiec wdowa
Ogółem	36106	11155	20670	1271	3010
15-19 lat	2535	2526	9	-	-
20-24	3612	3217	387	6	1
25-29	4229	2422	1732	68	6
30-34	3686	1015	2523	140	9
35-39	2887	480	2220	168	20
40-44	2512	342	1985	149	36
45-49	2874	338	2277	176	83
50-54	3414	279	2715	215	205
55-59	3071	184	2458	140	289
60 i więcej	7286	351	4363	210	2362
Kobiety	18550	4822	10490	736	2501
15-19 lat	1212	1203	9	-	-
20-24	1756	1460	288	6	1
25-29	2036	992	991	47	6
30-34	1814	365	1362	79	9
35-39	1399	174	1119	88	17
40-44	1238	108	1004	93	32
45-49	1439	103	1169	101	67
50-54	1765	98	1390	106	170
55-59	1537	85	1142	76	233
60 i więcej	4353	233	2016	140	1965

Ludność w wieku 15 lat i więcej według stanu cywilnego, wieku i płci w 2008 roku w powiecie chodzieskim – podejście 4

Wyszczególnienie	Ogółem	kawaler panna	żonaty zamężna	rozwiedziony rozwiedziona	wdowiec wdowa
Ogółem	36106	11164	20710	1231	3001
15-19 lat	2535	2527	8	-	-
20-24	3612	3202	404	6	0
25-29	4229	2452	1709	63	5
30-34	3686	1025	2512	139	10
35-39	2887	481	2224	164	19
40-44	2512	339	1997	142	34
45-49	2874	332	2289	172	81
50-54	3414	285	2730	207	192
55-59	3071	183	2474	141	274
60 i więcej	7286	339	4363	198	2386
Kobiety	18550	4823	10526	704	2497
15-19 lat	1212	1204	8	-	-
20-24	1750	1443	301	6	0
25-29	2044	1022	977	40	5
30-34	1800	358	1355	77	10
35-39	1381	171	1110	86	15
40-44	1252	109	1023	87	32
45-49	1435	109	1166	97	63
50-54	1752	103	1385	103	160
55-59	1548	82	1165	79	222
60 i więcej	4377	224	2036	128	1990

Podsumowanie wyników

- Dokonując analizy uzyskanych wyników można zauważyć, że podejście kalibracyjne daje zadowalające rezultaty bez względu na sposób generowania braków danych.
- Jest to szczególnie widoczne w sytuacji gdy z wyjściowej bazy danych usuwano w losowy sposób po 5% i 10% informacji o stanie cywilnym.
- W przypadku gdy frakcja braków wynosiła 20% można było zauważyć pewne niedoszacowania bądź przeszacowania liczby osób w poszczególnych komórkach wynikowej tablicy. Różnice jednak, w ujęciu względnym, nie przekraczały zazwyczaj 5%.
- Również w sytuacji gdy braki danych nie były generowane "równomiernie" w ramach wszystkich grup (podejście 4) wagi kalibracyjne w zadowalający sposób odtwarzały rzeczywiste struktury ludności w ramach poszczególnych domen.

Kalibracja w NSP 2011 – ujęcie problemu

- Narodowy Spis Powszechny Ludności i Mieszkań 2011 – metoda mieszana
- Wykorzystanie danych pochodzących ze źródeł administracyjnych a także danych zbieranych od ludności w ramach przeprowadzonego na szeroką skalę badania reprezentacyjnego
- Uogólnianie wyników badania reprezentacyjnego – kalibracja wag
- Konieczność odpowiedniego zintegrowania i zachowania spójności pomiędzy wynikami badania reprezentacyjnego z danymi pochodzącymi z rejestrów

CALMAR

- na potrzeby wyznaczenia wag kalibracyjnych w NSP 2011 wykorzystano makro CALMAR,
- makro jest napisane w języku 4GL w środowisku SAS i służy do wyznaczania wag kalibracyjnych,
- jest makrem napisanym na potrzeby prac francuskiego urzędu statystycznego (stąd dokumentacja techniczna jest dostępna tylko w języku francuskim),
- oferuje 4 sposoby wyznaczania wag kalibracyjnych w zależności od przyjętej postaci funkcji G.

CALMAR

CALMAR, jak stwierdzono powyżej, oferuje cztery sposoby wyznaczania wag kalibracyjnych, w zależności od postaci funkcji G:

- podejście liniowe

$$G(x) = \frac{1}{2}(x-1)^2, \tag{39}$$

- raking ratio

$$G(x) = x(\log x - 1) + 1, \tag{40}$$

- podejście logitowe

$$G(x) = \left[(x-L)\,log\frac{x-L}{1-L} + (U-x)\,log\frac{U-x}{U-1}\right]\frac{1}{A}, \tag{41}$$

gdzie:

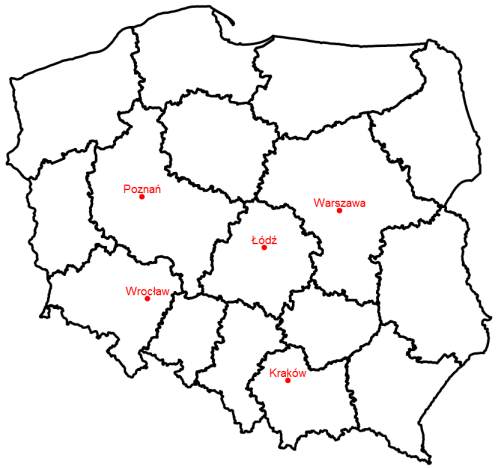
$$A = \frac{U-L}{(1-L)(U-1)}, \tag{42}$$

- podejście liniowe z warunkami ograniczającymi

$$G(x) = \frac{1}{2}(x-1)^2, \quad L \leq \frac{w_i}{d_i} \leq U. \tag{43}$$

- płeć
- wiek
- miejsce zamieszkania - poziom powiatu z wyodrębnieniem części miejskiej i wiejskiej

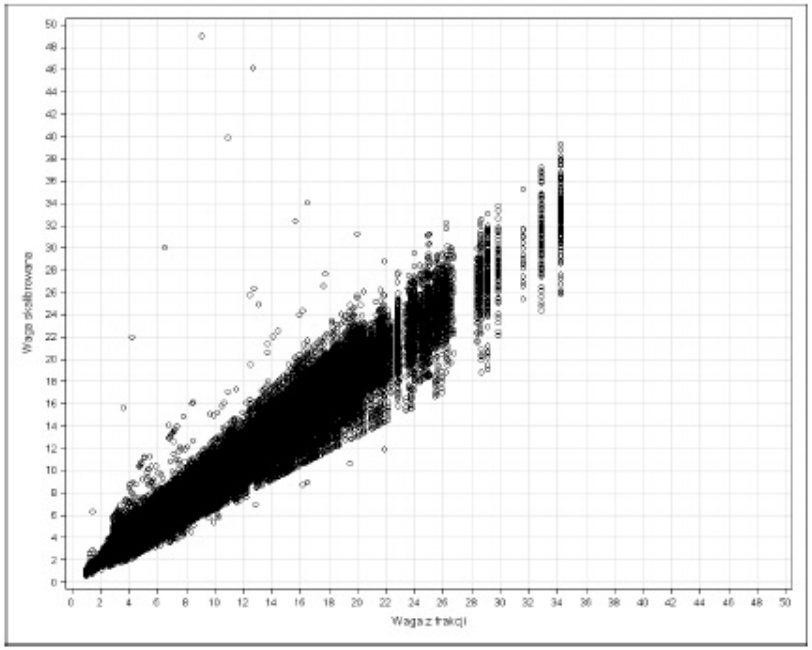
- **Województwo: płeć x miejsce zamieszkania x pojedyncze roczniki wieku (0,1,...,83,84,85+)**
- **Powiaty: płeć x miejsce zamieszkania x grupy wieku (0–4,5–9,...,80–84,85+)**
- **Największe miasta: płeć x pojedyncze roczniki wieku (0,1,...,83,84,85+ lub 100+ dla Warszawy)**

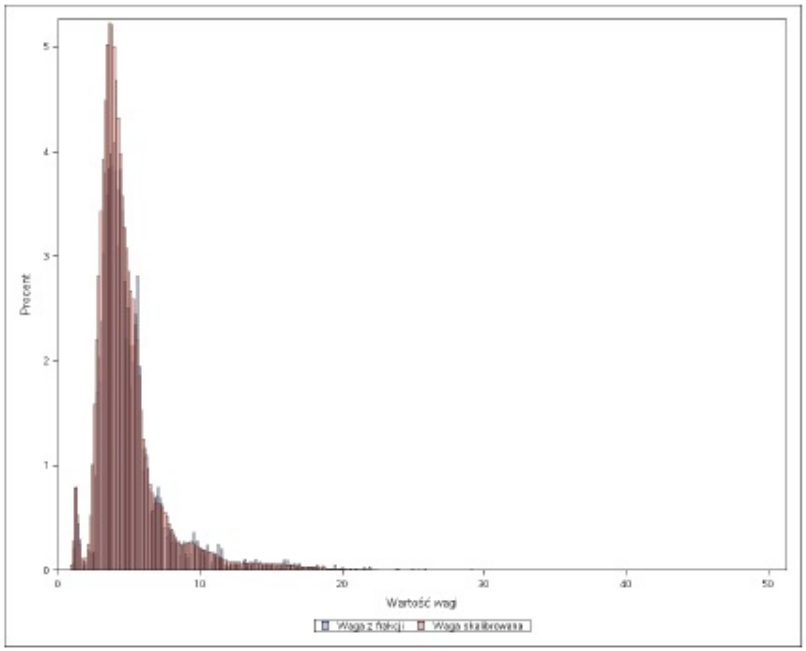


	Część miejska/ Część wiejska	Płeć	Grupy wieku	Pojedyncze roczniki wieku	Pojedyncze roczniki wieku
	1,2	1,2	0-4, 5-9,..., 80-84, 85+	0, 1, . . . ,83, 84; 85+	0, 1, . . . ,98 99, 100+
Polska	1	1	1	1	0
Województwa	1	1	1	1	0
Powiaty (bez 5 największych miast)	1	1	1	0	0
4 największe miasta	x	1	1	1	0
Warszawa	x	1	1	1	1
Dzielnice Warszawy	x	1	1	1	0
Delegatury 4 największych miast	x	1	1	1	0

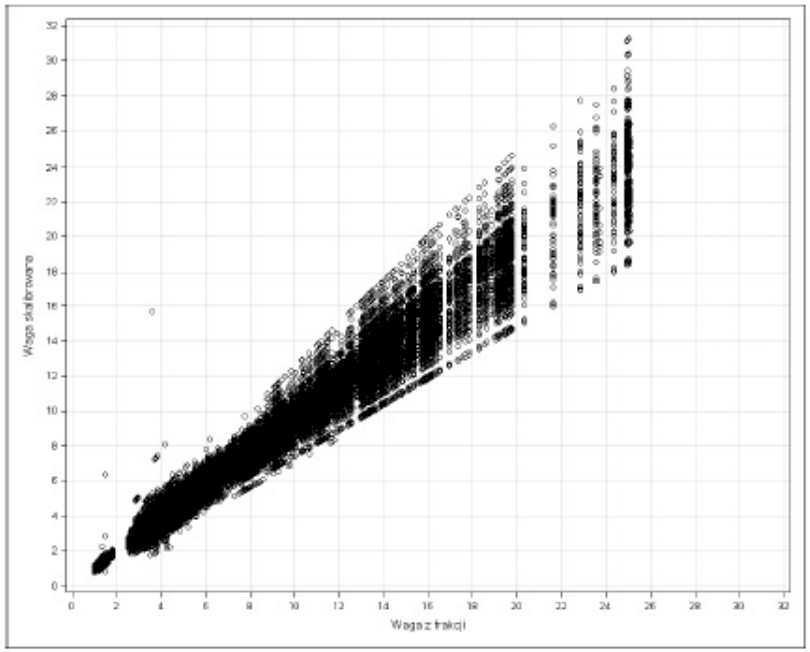
- **Legenda:** 1–kalibracja możliwa, 0–kalibracja niemożliwa, x–przekrój nieadekwatny

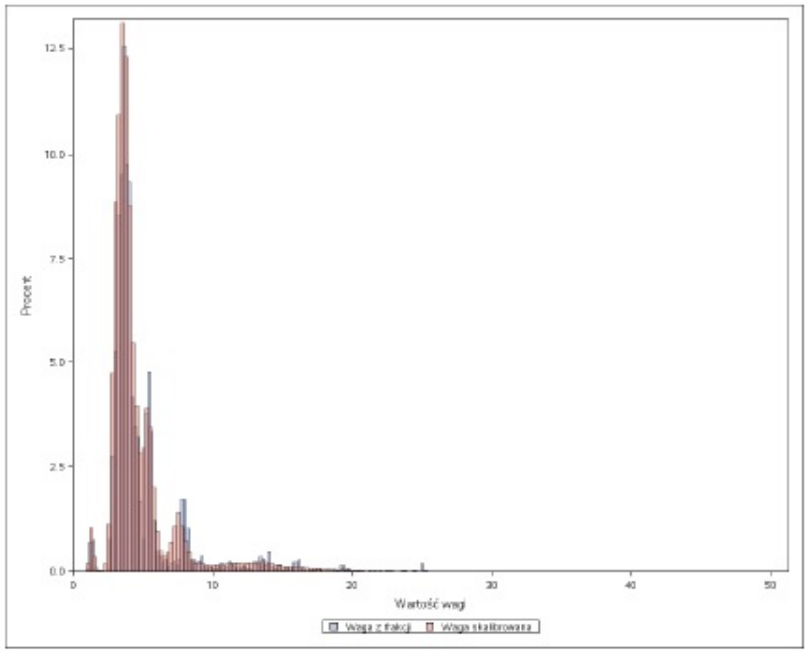
Korelogram i histogram rozkładu wag wynikających ze schematu losowania próby i kalibracyjnych – Polska



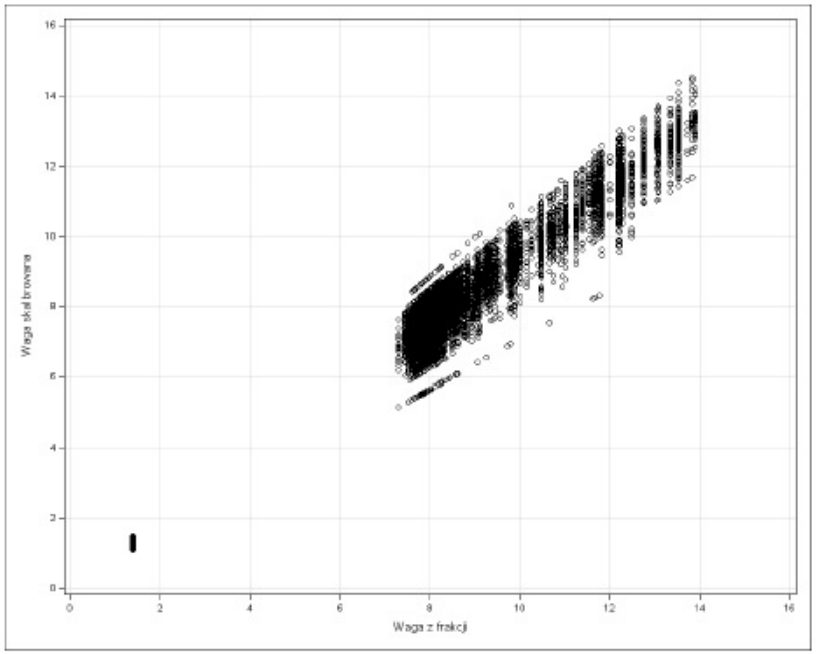


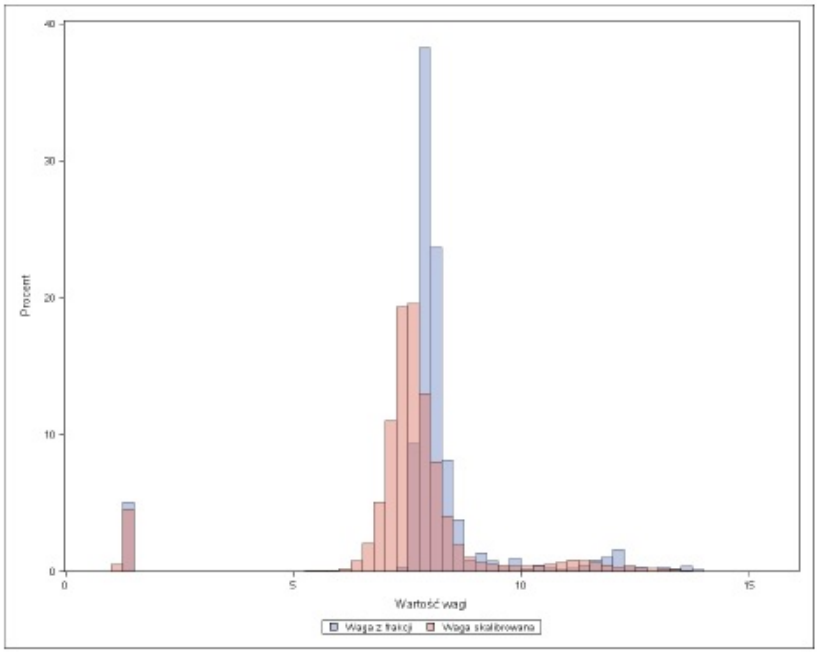
Korelogram i histogram rozkładu wag wynikających ze schematu losowania próby i kalibracyjnych – wielkopolskie





Korelogram i histogram rozkładu wag wynikających ze schematu losowania próby i kalibracyjnych – powiat poznański





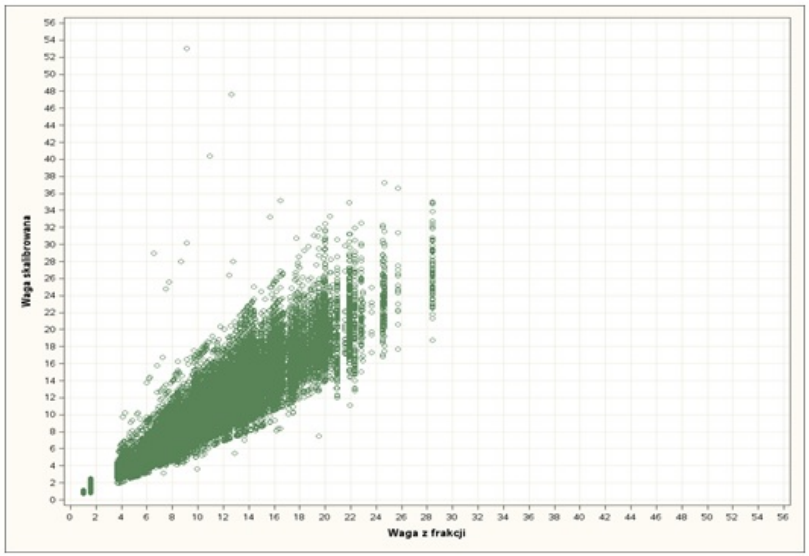
Statystyki opisowe wag – województwo wielkopolskie

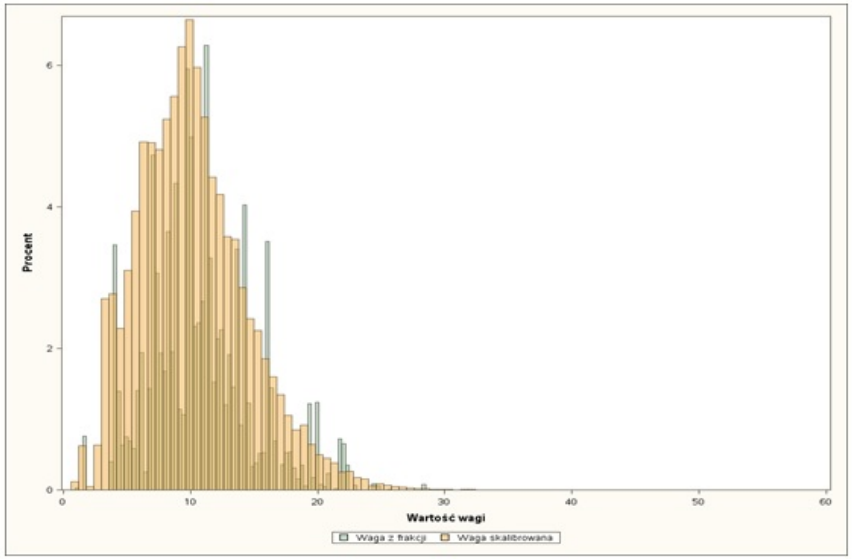
Waga	Min	Q1	Q2	Q3	Max	Suma	Średnia	Sx	Vx	K	As
Waga z frakcji skorygowana	1	3,39	3,88	5,21	25,08	3486223,3	4,78444	2,78	58,07	12,6	3,187
Waga skalibrowana na ludność rezydującą	0,6	3,33	3,82	5	36,31	3405808	4,67408	2,73	58,39	14,8	3,371
Iloraz wagi skalibrowanej i wagi z frakcji	0,5	0,93	0,98	1,02	4,351	713909,56	0,97976	0,08	7,976	12,7	0,677

Statystyki opisowe wag – powiat poznański

Waga	Min	Q1	Q2	Q3	Max	Suma	Średnia	Sx	Vx	K	As
Waga z frakcji skorygowana	1,392	7,835	7,969	8,222	13,894	348324,570	7,991	1,866	23,349	6,777	-1,458
Waga skalibrowana na ludność rezydującą	1,094	7,261	7,588	8,005	14,657	329826,000	7,567	1,818	24,024	5,965	-1,265
Iloraz wagi skalibrowanej i wagi z frakcji	0,693	0,915	0,946	0,973	1,082	41280,390	0,947	0,051	5,364	1,098	-0,344

Korelogram rozkładu wag wynikających ze schematu losowania próby i kalibracyjnych – Warszawa



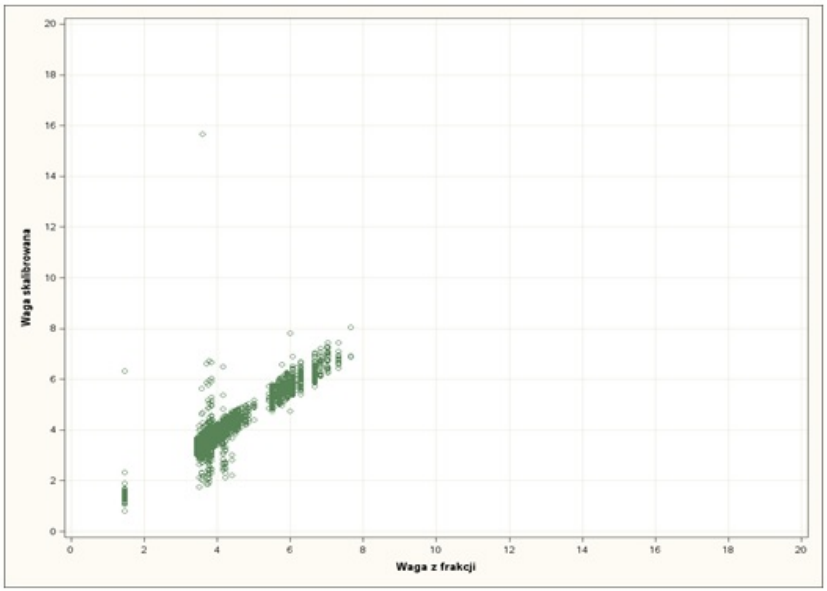


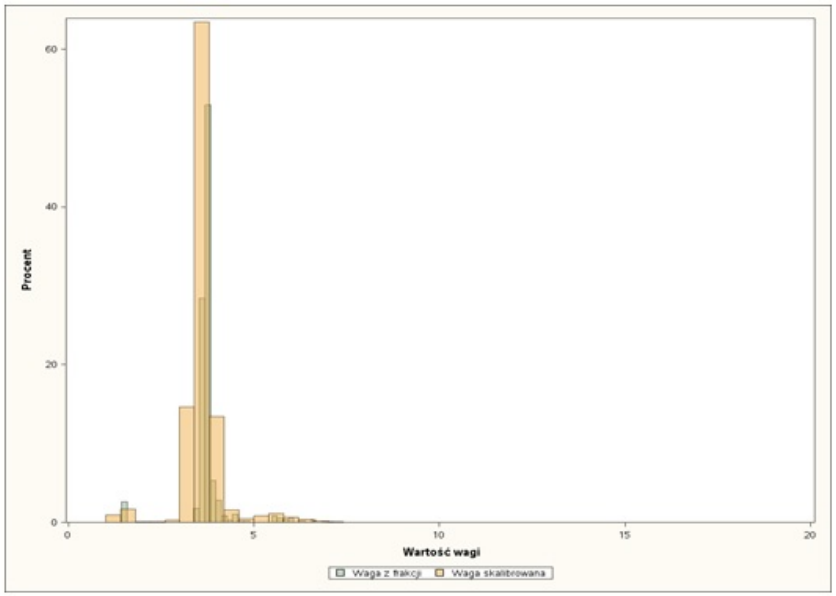
Statystyki opisowe wag – Warszawa

Waga	Min	Q1	Q2	Q3	Max	Suma	Średnia	Sx	Vx	K	As
Waga z frakcji skorygowana	1	8,07	10,5	13,3	28,39	1787796,5	10,8408	4,1	37,84	0,47	0,57
Waga skalibrowana na ludność rezydującą	0,71	7,07	9,79	12,7	53	1678349	10,1771	4,3	42,29	0,8	0,676
Iloraz wagi skalibrowanej i wagi z frakcji	0,36	0,83	0,92	1,02	5,817	153748,37	0,93229	0,15	16,27	13,4	1,19

- Jak pokazuje analiza, korelogram wag dla Warszawy w przypadku ogólnym nie ma charakteru jednolitej „smugi" - występują w nim punkty leżące z „dala" od niej.
- Oznacza to, że po kalibracji wagi zmieniły znacznie swoją wartość i odbiegają od wag wejściowych d_i wynikających ze schematu losowania próby.
- W takich sytuacjach należy zachować szczególną uwagę w procesie estymacji, zwłaszcza gdyby wagi ekstremalne występowały na dużą skalę.
- Dla Warszawy można jednak było zaobserwować, że wartości ekstremalnych wśród wag kalibracyjnych nie ma zbyt dużo. Co więcej dotyczą one starszych roczników wieku.

Korelogram i histogram rozkładu wag wynikających ze schematu losowania próby i kalibracyjnych – powiat kaliski





Statystyki opisowe wag – powiat kaliski

Waga	Min	Q1	Q2	Q3	Max	Suma	Średnia	Sx	Vx	K	As
Waga z frakcji skorygowana	1,47	3,67	3,71	3,77	7,644	84619,8	3,75387	0,58	15,5	13,3	0,682
Waga skalibrowana na ludność rezydującą	0,82	3,46	3,58	3,74	15,67	81621	3,62084	0,6	16,58	17,6	1,026
Iloraz wagi skalibrowanej i wagi z frakcji	0,5	0,94	0,96	0,99	4,351	21746,99	0,96473	0,06	6,421	782	14,6

- Występowanie wag skrajnych można było zauważyć przede wszystkim w niektórych grupach wieku oraz rocznikach.
- Dotyczy to przykładowo grupy wieku 15–17 lat czy samych 17-latków w powiecie kaliskim.
- Wyjściowa waga d_i dla pewnego mężczyzny w grupie wiekowej 15–17 lat wynosiła 3,6 podczas gdy po kalibracji jej wartość wzrosła do poziomu 15,67.
- Podobnie w powiecie tym w badanej grupie wieku mężczyzn znalazł się 17-latek, który miał wagę wejściową 1,47 podczas gdy po kalibracji jego waga wynosiła 6,34.

Podsumowanie

- Kalibracja umożliwiła dopasowanie struktur z badania reprezentacyjnego do znanych wartości globalnych z rejestrów administracyjnych.
- Analiza wag kalibracyjnych we wszystkich badanych przekrojach umożliwiła kompleksową ich ocenę.
- Wyznaczone wagi kalibracyjne mogą stanowić postawę uogólniania wyników z wykorzystaniem danych pochodzących z badania reprezentacyjnego.
- Szczególną uwagę podczas uogólniania wyników należy zwrócić w starszych grupach wiekowych (85+) czy w starszych rocznikach wieku (85,86,...) oraz dla grup wieku 15–17 czy poszczególnych roczników z tej grupy.
- W przypadku uogólniania wyników na dość dużym stopniu agregacji (np. całe województwo, czy nawet powiat) wpływ ekstremalnych wag kalibracyjnych będzie pomijalny.
- Ich znaczenie może jednak wzrastać w sytuacji, gdy estymacja wyników z wykorzystaniem wag kalibracyjnych odbywać się będzie dla szczegółowo zdefiniowanych przekrojów czy domen.

Literatura


Särndal C-E., Lundström S. (2005), „*Estimation in Surveys with Nonresponse*", John Wiley & Sons, Ltd.


Deville J-C., Särndal C-E. (1992), „*Calibration Estimators in Survey Sampling*", Journal of the American Statistical Association, Vol. 87, 376–382.


Särndal C-E. (2007), „*The Calibration Approach in Survey Theory and Practice*", Survey Methodology, Vol. 33, No. 2, 99–119.

Dziękuję za uwagę!