wodów. Z kolei jednostkami losowania drugiego stopnia są mieszkania. Jednostki pierwszego stopnia są przy tym losowane z zastosowaniem tzw. warstwowania, a podstawę podziału na warstwy stanowi podział Polski na województwa (GUS, 2018a).

Również w Badaniu Budżetów Gospodarstw Domowych Główny Urząd Statystyczny stosuje losowanie dwustopniowe, warstwowe, z różnymi prawdopodobieństwami wyboru na pierwszym stopniu. Jednostkami losowania pierwszego stopnia są rejony statystyczne lub zespoły rejonów, a na drugim stopniu losowane są mieszkania. Szczegółowy opis procedury losowania można znaleźć w opracowaniu GUS (2017a).

Losowanie dwustopniowe jest stosowane także przez Główny Urząd Statystyczny w Europejskim Badaniu Warunków Życia Ludności (EU-SILC) z różnymi prawdopodobieństwami wyboru na pierwszym stopniu. Jednostkami pierwszego stopnia są obwody spisowe. Na drugim stopniu losowane są z kolei mieszkania. Badaniu podlegają wszystkie gospodarstwa domowe zamieszkałe w wylosowanych mieszkaniach. Warto wspomnieć, że jednostki pierwszego stopnia są przed losowaniem warstwowane. Warstwami są województwa, a wewnątrz województw jednostki pierwszego stopnia są warstwowane według klasy miejscowości. Pełen opis losowania zastosowanego przez Główny Urząd Statystyczny w badaniu EU-SILC można znaleźć w publikacji GUS (2017b).

Definicja 1.9. *Mechanizm losowania, który realizuje plan wyboru jednostek populacji do próby, nazywa się schematem losowania.*

W badaniach statystycznych na potrzeby zastosowania określonych estymatorów oraz oceny ich wariancji wykorzystuje się tzw. prawdopodobieństwa inkluzji pierwszego i drugiego rzędu. Poniższa definicja określa prawdopodobieństwo inkluzji rzędu r, tj. prawdopodobieństwo wyboru do próby s jednostek populacji $k_1, \ldots, k_r$, którego szczególnymi przypadkami są prawdopodobieństwa inkluzji pierwszego ($r = 1$) i drugiego ($r = 2$) rzędu.

Definicja 1.10. *Prawdopodobieństwem inkluzji rzędu r (ang. r-order inclusion probability) jest*:

$$\pi_{k_1,\ldots,k_r} = \sum_{s \in A(k_1,\ldots,k_r)} P(s), \tag{1.16}$$

gdzie $A(k_1, \ldots, k_r) = \{s : k_i \in s, i = 1, \ldots, r\}$.

Prawdopodobieństwa inkluzji pierwszego i drugiego rzędu dla wybranych planów losowania można wyznaczyć w dość łatwy sposób. Na przykład, dla loso-

wania prostego ze zwracaniem i bez zwracania prawdopodobieństwa te można wyrazić następującymi wzorami:

– losowanie proste ze zwracaniem:

$$\pi_k = 1 - \left(1 - \frac{1}{N}\right)^n, \tag{1.17}$$

$$\pi_{kl} = 1 - 2\left(1 - \frac{1}{N}\right)^n + \left(1 - \frac{2}{N}\right)^n, \tag{1.18}$$

– losowanie proste bez zwracania:

$$\pi_k = \frac{n}{N}, \tag{1.19}$$

$$\pi_{kl} = \frac{n(n-1)}{N(N-1)}. \tag{1.20}$$

Definicja 1.11. *Wagę wynikającą z planu losowania (ang. design weight) definiujemy jako*:

$$d_k = \frac{1}{\pi_k}. \tag{1.21}$$

Analogicznie wagę d_{kl} definiujemy jako:

$$d_{kl} = \frac{1}{\pi_{kl}}. \tag{1.22}$$

Wagi d_k oraz d_{kl} odgrywają istotną rolę w wyznaczaniu ocen punktowych estymatorów oraz ich wariancji dla różnych parametrów. Wagi d_k mają również kluczowe znaczenie w podejściu kalibracyjnym, gdyż podlegają odpowiedniej korekcie z wykorzystaniem zestawu zmiennych pomocniczych, tak aby były spełnione właściwe równania kalibracyjne. Szczegółowo proces ten opisano w rozdziale drugim niniejszej książki.

Jak wcześniej wspomniano, kluczowym elementem jest wykorzystanie próby, aby wyciągnąć wnioski dotyczące nieznanego parametru θ lub, bardziej ogólnie, pewnej funkcji g parametru θ (funkcji parametrycznej). W tym celu konstruuje się statystykę $\hat{\theta} = T(Y_1, \ldots, Y_n)$, zwaną estymatorem punktowym, w taki sposób, aby wartość tej statystyki, zwana oceną punktową, była bliska wartości parametru θ (Krzyśko, 2004, s. 67).

Definicja 1.12. *Estymatorem parametru θ jest statystyka $\hat{\theta} = T(Y_1, \ldots, Y_n)$. Oceną parametru θ nazywamy wartość estymatora $\hat{\theta}$.*

Naturalnym pytaniem, jakie pojawia się w tym miejscu, jest pytanie o to, jakiej statystyki powinniśmy użyć do oceny parametru θ. Punktem wyjścia do rozważań na temat jakości estymatorów jest liczbowa charakterystyka dokładności estymatora T, za którą przyjmuje się błąd średniokwadratowy. Definiuje się go w następujący sposób (Krzyśko, 2004, s. 68):

Definicja 1.13. *Błędem średniokwadratowym estymatora $\hat{\theta}$ (ang. mean squared error –* MSE*) nazywamy wyrażenie postaci* :

$$\mathrm{MSE}(\hat{\theta}) = E\Big[(\hat{\theta}-\theta)^2\Big]. \tag{1.23}$$

Pierwiastek z błędu średniokwadratowego estymatora $\hat{\theta}$ (ang. root mean squared error – RMSE*) oznaczać będziemy przez*:

$$\mathrm{RMSE}(\hat{\theta}) = \sqrt{\mathrm{MSE}(\hat{\theta})}. \tag{1.24}$$

Przyjmując błąd średniokwadratowy $\mathrm{MSE}(\hat{\theta})$ za miarę dokładności estymatora, można w klasie wszystkich estymatorów wprowadzić porządek częściowy zgodnie z poniższą definicją.

Definicja 1.14. *Estymator $\hat{\theta}_1$ jest lepszy od estymatora $\hat{\theta}_2$, jeżeli dla każdego parametru $\theta \in \Theta$*:

$$\mathrm{MSE}(\hat{\theta}_1) \leq \mathrm{MSE}(\hat{\theta}_2) \tag{1.25}$$

i chociażby dla jednej wartości θ spełniona jest nierówność ostra:

$$\mathrm{MSE}(\hat{\theta}_1) < \mathrm{MSE}(\hat{\theta}_2), \tag{1.26}$$

gdzie Θ jest przestrzenią parametrów.

Można pokazać, że najlepszy estymator w sensie powyższej definicji istnieje niezmiernie rzadko. Zwykle jest bowiem tak, że dla pewnych wartości parametru θ z dwóch estymatorów $\hat{\theta}_1$ i $\hat{\theta}_2$ lepszy jest np. estymator $\hat{\theta}_1$, a dla innych wartości parametru – estymator $\hat{\theta}_2$ (Krzyśko, 2004, s. 69).

Definicja 1.15. *Estymator $\hat{\theta} = T(Y_1,\ldots,Y_n)$ parametru θ jest nieobciążony, jeżeli dla każdego $\theta \in \Theta$ spełniony jest poniższy warunek*:

$$E(\hat{\theta}) = \theta. \tag{1.27}$$

Definicja 1.16. *Obciążeniem estymatora $\hat{\theta}$ nazywamy różnicę postaci*:

$$B(\hat{\theta}) = E(\hat{\theta}) - \theta. \tag{1.28}$$

Definicja 1.17. *Wariancją estymatora $\hat{\theta}$ nazywamy wyrażenie postaci*:

$$D^2(\hat{\theta}) = E\Big(\hat{\theta} - E(\hat{\theta})\Big)^2. \tag{1.29}$$

Można pokazać, że między błędem średniokwadratowym, wariancją i obciążeniem estymatora $\hat{\theta}$ parametru θ istnieje następująca zależność:

$$\mathrm{MSE}(\hat{\theta}) = D^2(\hat{\theta}) + B^2(\hat{\theta}), \tag{1.30}$$

która w szczególnym przypadku dla estymatora nieobciążonego prowadzi do równości $\mathrm{MSE}(\hat{\theta}) = D^2(\hat{\theta})$.

Definicja 1.18. *Średnim błędem szacunku estymatora $\hat{\theta}$ nazywamy wyrażenie postaci*:

$$D(\hat{\theta}) = \sqrt{D^2(\hat{\theta})}. \tag{1.31}$$

Definicja 1.19. *Względnym średnim błędem szacunku estymatora $\hat{\theta}$ (ang. relative estimation error –* REE*) nazywamy wyrażenie postaci*:

$$\mathrm{REE}(\hat{\theta}) = \frac{D(\hat{\theta})}{|\hat{\theta}|} \cdot 100\%. \tag{1.32}$$

Miara ta w literaturze anglojęzycznej oznaczana jest także jako CV (ang. *coefficient of variation*) i nazywana współczynnikiem zmienności. Również w publikacjach Głównego Urzędu Statystycznego, przy podawaniu wskaźników precyzji estymacji, oznaczenie to jest często używane. Zgodnie z wytycznymi, jakie Główny Urząd Statystyczny przyjmuje podczas publikowania wyników z badania reprezentacyjnego, oszacowania, dla których $CV < 10\%$, można uznać za wiarygodne. Oszacowania, dla których CV przyjmuje wartości z przedziału 10–20%, należy interpretować ze szczególną ostrożnością. Z kolei do oszacowań, dla których $CV > 20\%$, należy podchodzić sceptycznie i powinny być one publikowane tylko w postaci zagregowanej, tj. na wyższym poziomie agregacji (GUS, 2013). Wskaźnik CV można wykorzystać także do wyznaczenia odpowiedniego przedziału ufności, który z określonym prawdopodobieństwem $1 - \alpha$ pokrywa prawdziwą wartość estymowanego parametru.

W dalszej części książki dla względnego błędu szacunku estymatora będziemy jednak używać oznaczenia REE.

Definicja 1.20. *Względnym błędem średniokwadratowym estymatora $\hat{\theta}$ (ang. relative root mean squared error –* RRMSE*) nazywamy wyrażenie postaci*:

$$\mathrm{RRMSE}(\hat{\theta}) = \frac{\mathrm{RMSE}(\hat{\theta})}{\hat{\theta}}. \tag{1.33}$$

Miara ta opisuje, jaki jest udział błędu estymacji w wartości szacowanego parametru.

1.3. Estymator Horvitza-Thompsona wartości globalnej

W badaniach próbkowych prowadzonych przez krajowe urzędy statystyczne na całym świecie procesowi estymacji podlega wiele różnych parametrów. Jak wcześniej wspomniano, najczęściej są to wartość globalna, średnia, kwantyle czy bardziej złożone parametry takie jak iloraz dwóch wartości globalnych. Parametry te mogą się odnosić na przykład do liczby osób bezrobotnych, pracujących czy biernych zawodowo (wartość globalna), mediany bądź średniego przychodu przedsiębiorstw.

Estymacja parametrów odbywa się na podstawie próby odpowiednio wylosowanej ze skończonej populacji, pobranej zgodnie z określonym planem losowania. Istnieje możliwość uzyskania informacji na temat wybranych parametrów również ze spisów czy rejestrów administracyjnych, które w statystyce publicznej odgrywają szczególną rolę. Ze względu na incydentalny charakter spisów oraz ich ograniczony zakres informacyjny, badania próbkowe są ważnym źródłem danych na temat sytuacji społeczno-gospodarczej w kraju. Jednak spisy oraz rejestry administracyjne mogą stanowić bogate zasoby zmiennych pomocniczych, które odgrywają istotną rolę w procesie estymacji, również w kontekście rozważanego w pracy podejścia kalibracyjnego. Równie ważne są inne źródła, takie jak internet czy tzw. big data (na przykład dane pochodzące z Facebooka, Twittera, od operatorów telefonii komórkowej itd.), które w coraz większym zakresie wykorzystywane są przez krajowe urzędy statystyczne (Beręsewicz i Szymkowiak, 2015; Daas i Puts, 2014; Daas i in., 2015; Szreder, 2015a; Zeelenberg, 2016). Mimo że w statystyce publicznej są one w coraz większym stopniu wykorzystywane jako źródło cennych informacji, również w kontekście podejścia kalibracyjnego (Beręsewicz i Szymkowiak, 2018), nie są przedmiotem głębszych rozważań w niniejszej pracy.

W dalszej części książki analizie poddano problem estymacji wartości globalnej, która odgrywa kluczową rolę wśród wszystkich potencjalnych parametrów.

Opisano estymator Horvitza-Thompsona oraz uogólniony estymator regresyjny, który (jak pokazano w rozdziale drugim) jest szczególnym przypadkiem estymatora kalibracyjnego. Ujmując zagadnienie bardziej formalnie, załóżmy, że przedmiotem estymacji jest wartość globalna w populacji U zmiennej Y określona wzorem (por. wzór (1.1)):

$$\tau_Y = \sum_{k=1}^{N} y_k, \tag{1.34}$$

gdzie N oznacza liczebność populacji, a y_k to wartość badanej cechy Y dla k-tej jednostki populacji, przy czym $k = 1,\ldots,N$. Zakładamy, że z populacji $U = \{1,\ldots,N\}$ losujemy zgodnie z określonym planem losowania próbę s o liczebności n. Niech $\pi_k > 0$ oznacza prawdopodobieństwo inkluzji pierwszego rzędu (por. wzór (1.16)), a d_k wagę wynikającą z planu losowania próby zdefiniowaną jako $d_k = 1/\pi_k$ (por. wzór (1.21)). Przy przyjętych powyżej oznaczeniach wartość globalną (1.34) zmiennej Y można wyrazić alternatywnie jako $\tau_Y = \sum_{k\in U} y_k$ lub w skrócie $\tau_Y = \sum_U y_k$. Wartość globalna we wzorze (1.34) może się odnosić zarówno do zmiennej Y ciągłej (na przykład przychód przedsiębiorstwa), jak i do zmiennych dyskretnych. Na przykład, jeśli przyjmiemy, że Y jest zmienną dychotomiczną, dla której:

$$y_k = \begin{cases} 1, & \text{jeżeli osoba jest pracująca,} \\ 0, & \text{jeżeli osoba jest bezrobotna,} \end{cases}$$

gdzie $k = 1,\ldots,N$, to wartość globalna (1.34) odnosić się będzie do liczby osób pracujących.

W badaniach statystycznych prowadzonych przez krajowe urzędy statystyczne do szacowania wartości globalnej wyrażonej wzorem (1.34) wykorzystuje się często tzw. estymator Horvitza-Thompsona (1952) określony wzorem:

$$\hat{\tau}_{\text{HT}} = \sum_{k=1}^{n} d_k y_k = \sum_{k\in s} d_k y_k. \tag{1.35}$$

Niewątpliwą zaletą tego estymatora jest jego prostota, a także to, że jest on nieobciążony, tj. $E(\hat{\tau}_{\text{HT}}) = \tau_Y$. Wynika to z następujących przekształceń:

$$E(\hat{\tau}_{\text{HT}}) = E\left(\sum_{k\in U} \pi_k^{-1} y_k I(k\in s)\right) = \sum_{k\in U} \pi_k^{-1} y_k E(I(k\in s)) =$$
$$= \sum_{k\in U} \pi_k^{-1} y_k \pi_k = \tau_Y.$$

ROZDZIAŁ 2

Teoretyczne podstawy podejścia kalibracyjnego

2.1. Wprowadzenie

W badaniach reprezentacyjnych prowadzonych przez krajowe urzędy statystyczne przedmiotem estymacji są różnego rodzaju parametry: wartość globalna, średnia czy kwantyle. Spośród wymienionych parametrów populacji generalnej zdecydowanie najczęściej jest to wartość globalna. Dotyczy to również wartości globalnej w odpowiednio zdefiniowanych przekrojach o charakterze przestrzennym (województwo czy powiat), a także wynikających z uwzględnienia dodatkowych cech – najczęściej demograficznych (jak płeć, miejsce zamieszkania, grupy wieku czy wykształcenie). Wynika to z tego, że w publikacjach i różnego rodzaju raportach badawczych najczęściej przedstawia się odpowiednie tabele z liczebnościami. Liczebności te odnoszą się do oszacowanych wartości globalnych z wykorzystaniem odpowiednio skonstruowanych wag kalibracyjnych. Przykłady takich tabel można znaleźć w publikacjach odnoszących się do Badania Budżetów Gospodarstw Domowych (GUS, 2017a), Europejskiego Badania Warunków Życia Ludności (GUS, 2017b) czy Badania Aktywności Ekonomicznej Ludności (GUS, 2018a).

W poprzednim rozdziale wskazano, że kalibracja jest techniką estymacji powszechnie wykorzystywaną w badaniach statystycznych. Zwrócono uwagę na szerokie spektrum zastosowań tej metody w praktyce.

W tym rozdziale podejście kalibracyjne zostało sformalizowane od strony matematycznej. Uwaga została skoncentrowana na konstruowaniu estymatorów kalibracyjnych wartości globalnej, których budowa polega na znalezieniu minimum dla odpowiednio dobranej funkcji odległości, przy jednoczesnej konieczności spełnienia tzw. równań kalibracyjnych[8]. Rozważania dotyczyć będą nie tylko najczęściej wykorzystywanej w praktyce tzw. liniowej funkcji kalibracyjnej, która prowadzi do uzyskania estymatora typu GREG wartości globalnej, ale również innych rzadziej stosowanych funkcji. Omówiono wady i zalety poszczególnych funkcji odległości. Przedstawiono również estymatory wariancji rozważanych estymatorów kalibracyjnych zarówno w populacji generalnej, jak i w ujęciu domen, na które podzielona może być populacja.

W odpowiednio zaprojektowanym badaniu symulacyjnym, z wykorzystaniem rzeczywistych danych pochodzących z Badania Budżetów Gospodarstw Domowych, dokonano oceny najważniejszych własności estymatorów kalibracyjnych wartości globalnej dla różnych funkcji odległości. Dyskusji poddano również analizę wag kalibracyjnych, która w tego typu badaniach jest dość często pomijana[9].

2.2. Estymatory kalibracyjne wartości globalnej – podejście minimalizujące funkcję odległości

Załóżmy, że celem badania jest oszacowanie wartości globalnej cechy statystycznej Y w populacji U danej wzorem (1.1). Z populacji tej pobieramy n-elementową próbę s zgodnie z określonym planem jej losowania. Zakładamy, że $d_k = \frac{1}{\pi_k}$, $k = 1,\ldots,N$, jest odpowiednią wagą (por. wzór (1.21)), gdzie π_k oznacza prawdopodobieństwo inkluzji pierwszego rzędu (por. wzór (1.16)). W praktyce badań

[8] W książce podejście to jest określane mianem podejścia minimalizującego funkcję odległości. W rozdziale trzecim pokazano, że nie jest to jedyna metoda konstruowania estymatorów kalibracyjnych wartości globalnej.

[9] Najczęściej dokonuje się oceny oszacowań wartości globalnej, co zdaniem autora jest dużym uproszczeniem i może prowadzić do błędów w estymacji. Pożądane własności estymatora kalibracyjnego wartości globalnej, takie jak małe obciążenie czy wariancja dla jednej zmiennej Y, przy ustalonym wektorze wag, niekoniecznie musi się przełożyć na inne zmienne uwzględnione w badaniu. Z tego punktu widzenia konieczna jest ocena wag kalibracyjnych uwzględniająca występowanie wag ujemnych lub odstających. Jest to szczególnie istotne, gdy estymacji dokonujemy nie tylko na poziomie całej populacji, ale również w dodatkowych przekrojach.

krajowych urzędów statystycznych w procesie estymacji wartości globalnej (1.1) wykorzystuje się bardzo często wspomniany już estymator Horvitza-Thompsona, który wyraża się wzorem (1.35).

Zdarza się jednak, że wagi d_k wynikające z planu losowania próby nie odtwarzają znanych wartości globalnych w odniesieniu do niektórych kluczowych zmiennych. Przyjmijmy, że wektor:

$$\sum_{k\in U} \mathbf{x}_k = \left(\sum_{k\in U} x_{k1}, \ldots, \sum_{k\in U} x_{kJ} \right)^{\mathrm{T}} \tag{2.1}$$

jest wektorem wartości globalnych wszystkich zmiennych pomocniczych. Oznacza to, że co najmniej dla jednej zmiennej $j = 1, \ldots, J$ nie jest spełniony poniższy warunek:

$$\sum_{k\in s} d_k x_{kj} = \sum_{k\in U} x_{kj}, \tag{2.2}$$

gdzie x_{kj} oznacza wartość j-tej zmiennej pomocniczej dla k-tej jednostki badania oraz $\sum_{k\in U} x_{kj}$ jest wartością globalną tej zmiennej. Wartości globalne takich zmiennych są znane zazwyczaj ze spisów powszechnych czy rejestrów administracyjnych. Ich przykładem mogą być informacje na temat liczby ludności w przekroju płci, klasy miejsca zamieszkania czy odpowiednich grup wieku. Wartości globalne zmiennych pomocniczych są wykorzystywane w procesie kalibracji wag d_k. Po jej zastosowaniu do wszystkich zmiennych pomocniczych nowe wagi – tzw. wagi kalibracyjne w_k – odtwarzają znane ich wartości globalne (2.1) dokładnie. Kalibracja wag jest zatem niezbędna w celu spełnienia często postulowanego w badaniach statystycznych wymogu zgodności. Oznacza to, że oszacowania tych samych zmiennych w różnych badaniach powinny dawać te same rezultaty. Co więcej, można oczekiwać, że jeśli zmienna Y jest skorelowana ze zmiennymi pomocniczymi, to również warunek ten w przybliżeniu powinien być spełniony i dla tej zmiennej.

Poniżej opiszemy proces kalibracji oraz sposób konstrukcji takich wag dla estymatora kalibracyjnego wartości globalnej zmiennej Y, aby dla każdej zmiennej pomocniczej odtwarzane były znane wartości globalne (2.1). Załóżmy w dalszym ciągu, że $\mathbf{d} = (d_1, \ldots, d_n)^{\mathrm{T}}$ jest wektorem wag wynikającym z planu losowania próby, a $\mathbf{w} = (w_1, \ldots, w_n)^{\mathrm{T}}$ poszukiwanym wektorem końcowych wag kalibracyjnych.

Zgodnie z ideą zaproponowaną w pracy Deville'a i Särndala (1992) estymator kalibracyjny wartości globalnej (1.1) jest postaci:

$$\hat{\tau}_{\text{CAL}} = \sum_{k \in s} w_k y_k, \tag{2.3}$$

gdzie wagi kalibracyjne w_k są rozwiązaniem następującego zadania optymalizacyjnego:

- (W1) – minimalizacja funkcji odległości:

$$D(\mathbf{d}, \mathbf{w}) = \sum_{k \in s} \frac{d_k}{q_k} G\left(\frac{w_k}{d_k}\right) \rightarrow \min, \tag{2.4}$$

- (W2) – równania kalibracyjne:

$$\sum_{k \in s} w_k \mathbf{x}_k = \sum_{k \in U} \mathbf{x}_k. \tag{2.5}$$

Pierwszy warunek (W1) orzeka, że wagi kalibracyjne w_k powinny być w taki sposób wyznaczone, aby były możliwie bliskie – w sensie przyjętej funkcji odległości $D(\cdot)$ – wag d_k wynikających z planu losowania próby. Funkcja $G(\cdot)$ mierzy odległość między ilorazem wag w_k/d_k a 1 i jest wykorzystywana w konstruowaniu funkcji odległości $D(\cdot)$, dla której szuka się lokalnego minimum warunkowego. W tym miejscu należy podkreślić, że funkcja $D(\cdot)$ nie jest metryką. Na ogół nie jest bowiem spełniony warunek symetryczności. Co więcej, w wypadku tej funkcji odległości trudno w ogóle mówić o tzw. nierówności trójkąta. Nie bierze się bowiem pod uwagę dodatkowego – trzeciego wektora $\mathbf{z}$ w celu sprawdzenia prawdziwości nierówności $D(\mathbf{d}, \mathbf{w}) \leq D(\mathbf{d}, \mathbf{z}) + D(\mathbf{z}, \mathbf{w})$. Czynnik q_k stanowi dodatkową wagę, która w zależności od jej postaci może prowadzić do uzyskania różnych estymatorów kalibracyjnych. Podobnie jak w wypadku uogólnionego estymatora regresyjnego typu GREG, często przyjmuje się, że $q_k = 1$, a w wypadku jednej zmiennej objaśniającej x_k można przykładowo założyć, że $q_k = 1/x_k$. Z kolei warunek (W2) stanowi istotę teorii kalibracji i orzeka, że wagi powinny być tak dobrane, aby po ich zastosowaniu dla wszystkich zmiennych pomocniczych można było odtworzyć ich znane wartości globalne. Jeśli ten warunek zostanie spełniony, to wykorzystanie wag kalibracyjnych w_k do innych zmiennych Y w badaniu powinno się przyczynić do lepszego oszacowania wartości globalnych tych zmiennych w populacji i do poprawienia precyzji.

Dla niektórych funkcji odległości $D(\cdot)$ może się zdarzyć, że wyznaczone wagi kalibracyjne w_k będą przyjmowały wartości ujemne. Jest to sprzeczne z definicją wagi, która stanowi odwrotność prawdopodobieństwa inkluzji pierwszego rzędu i powinna przyjmować wartość $w_k \geq 1$ dla każdego k. Może się również zdarzyć sytuacja, że wagi są dodatnie i większe od 1, ale przyjmują wartości ekstre-

malne, tj. znacznie odbiegają od wag d_k wynikających z planu losowania próby. Na ogół obserwuje się to, gdy dla danego przekroju w próbie istnieje niewielka liczba jednostek, podczas gdy w odpowiadającym mu przekroju w badaniu pełnym (na przykład spisie) liczba jednostek jest znacznie większa. Taki przypadek jest zatem szczególnie niebezpieczny, gdy estymatory kalibracyjne wykorzystywane są do szacowania wartości globalnych w domenach. W wypadku małej liczby jednostek w próbie reprezentujących daną domenę oszacowana wartość globalna ze względu na znaczne zniekształcenie wagi oryginalnej wpływa na istotne przeszacowanie nieznanej wartości prawdziwej. Z tego powodu w procesie wyznaczania wag kalibracyjnych w niektórych przypadkach wprowadza się dodatkowy warunek ograniczający na wagi, a dokładniej na iloraz wag w_k i d_k:

- (W3) – warunki ograniczające:

$$L \leq \frac{w_k}{d_k} \leq U, \qquad \text{gdzie: } 0 \leq L \leq 1 \leq U,\ k = 1, \ldots, n. \tag{2.6}$$

Przy wyborze funkcji $G(\cdot)$ w procesie wyznaczania wag kalibracyjnych w_k istnieje pewna dowolność. Funkcja $G(\cdot)$ powinna jednak spełniać między innymi następujące własności matematyczne: $G(\cdot)$ jest ściśle wypukła i dwukrotnie różniczkowalna, $G(\cdot) \geq 0$, $G(1) = 0$, $G'(1) = 0$ oraz $G''(1) = 1$.

Do wyznaczania wag kalibracyjnych w_k szczególnie przydatna jest znajomość funkcji $F(\cdot)$ odwrotnej do pierwszej pochodnej funkcji $G(\cdot)$, tj.:

$$F(\cdot) = G'^{-1}(\cdot), \tag{2.7}$$

przy czym $F(0) = 1$. Jest to tzw. funkcja kalibracyjna (ang. *calibration function*). Wynika to ze sposobu poszukiwania rozwiązań zadania minimalizacji opisanego w warunku (W1), przy jednoczesnym spełnieniu równań kalibracyjnych (W2). Rozwiązania poszukać można bowiem, wykorzystując metodę czynników nieoznaczonych Lagrange'a służącą do znajdowania ekstremum warunkowego funkcji różniczkowalnej. Odpowiednia funkcja Lagrange'a wyraża się wzorem:

$$\mathcal{L}(\mathbf{w}, \boldsymbol{\lambda}; \mathbf{d}) = \sum_{k \in s} \frac{d_k}{q_k} G\left(\frac{w_k}{d_k}\right) - \boldsymbol{\lambda}^{\mathrm{T}} \left(\sum_{k \in s} w_k \mathbf{x}_k - \sum_{k \in U} \mathbf{x}_k \right), \tag{2.8}$$

gdzie $\boldsymbol{\lambda}^{\mathrm{T}} = (\lambda_1, \ldots, \lambda_J)^{\mathrm{T}}$ jest tzw. wektorem mnożników (czynników nieoznaczonych) Lagrange'a. Równania Lagrange'a dla $k = 1, \ldots, n$ wyrażają się zatem wzorem:

$$\frac{\partial \mathcal{L}}{\partial w_k} = \frac{d_k}{q_k} \cdot \frac{\partial}{\partial w_k} G\left(\frac{w_k}{d_k}\right) - \boldsymbol{\lambda}^{\mathrm{T}} \mathbf{x}_k = 0. \tag{2.9}$$

Z powyższego równania i z twierdzenia o pochodnej funkcji złożonej wynika, że:

$$\frac{1}{q_k} G'\left(\frac{w_k}{d_k}\right) - \boldsymbol{\lambda}^{\mathrm{T}} \mathbf{x}_k = 0. \tag{2.10}$$

Przyjmując, że $G'(\cdot) = \psi(\cdot)$ oraz że $F(\cdot) = \psi^{-1}(\cdot)$ jest funkcją odwrotną względem pierwszej pochodnej funkcji $G(\cdot)$, otrzymujemy, że:

$$\psi\left(\frac{w_k}{d_k}\right) = q_k \boldsymbol{\lambda}^{\mathrm{T}} \mathbf{x}_k = q_k \mathbf{x}_k^{\mathrm{T}} \boldsymbol{\lambda} \Rightarrow \frac{w_k}{d_k} = \psi^{-1}\left(q_k \mathbf{x}_k^{\mathrm{T}} \boldsymbol{\lambda}\right) = F\left(q_k \mathbf{x}_k^{\mathrm{T}} \boldsymbol{\lambda}\right). \tag{2.11}$$

Ostatecznie wagi kalibracyjne można przedstawić w następującej postaci:

$$w_k = d_k F\left(q_k \mathbf{x}_k^{\mathrm{T}} \boldsymbol{\lambda}\right) = d_k g_k, \tag{2.12}$$

gdzie $g_k = F\left(q_k \mathbf{x}_k^{\mathrm{T}} \boldsymbol{\lambda}\right)$ to tzw. mnożniki wagowe (kalibracyjne). Do wyznaczenia wag kalibracyjnych (2.12) w dalszym ciągu niezbędna jest znajomość wektora mnożników Lagrange'a $\boldsymbol{\lambda}$, który można znaleźć, rozwiązując odpowiednie równanie kalibracyjne – por. równanie (2.5):

$$\sum_{k \in s} d_k F\left(q_k \mathbf{x}_k^{\mathrm{T}} \boldsymbol{\lambda}\right) \mathbf{x}_k = \sum_{k \in U} \mathbf{x}_k. \tag{2.13}$$

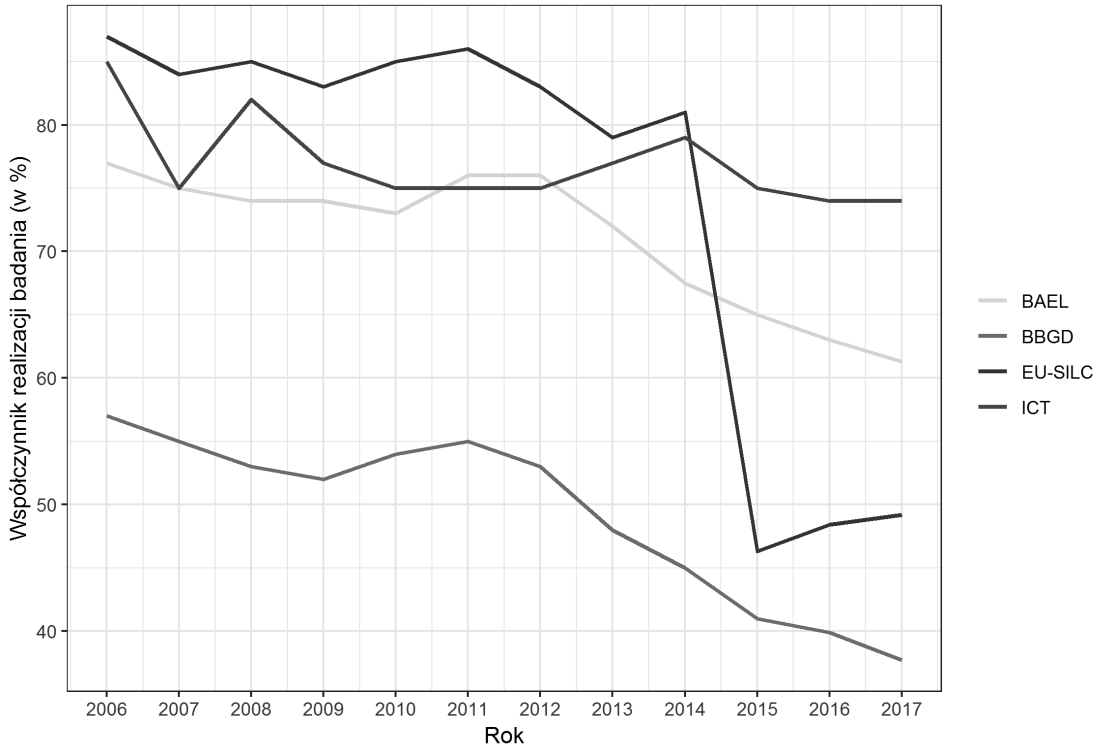
Finalnie estymator kalibracyjny wartości globalnej można przedstawić w postaci:

$$\hat{\tau}_{\mathrm{CAL}} = \sum_{k \in s} d_k F\left(q_k \mathbf{x}_k^{\mathrm{T}} \boldsymbol{\lambda}\right) y_k = \sum_{k \in s} w_k y_k. \tag{2.14}$$

Wyznaczenie wektora $\boldsymbol{\lambda}$ czynników nieoznaczonych Lagrange'a, a w konsekwencji wag kalibracyjnych (2.12), zależy od postaci funkcji $G(\cdot)$. W wypadku niektórych funkcji istnieje możliwość jawnego przedstawienia wag kalibracyjnych w postaci odpowiedniego wzoru. W innych przypadkach wymagane jest z kolei zastosowanie podejścia iteracyjnego. Przegląd wybranych funkcji $G(\cdot)$ można znaleźć m.in. w pracach Deville'a i Särndala (1992) czy Szymkowiaka (2013)[10]. Szczególnie ważna jest funkcja $G(x) = \frac{1}{2}(x-1)^2$, dla której wektor wag kalibracyjnych można wyrazić za pomocą jawnego wzoru. Mówi o tym poniższe twierdzenie.

Twierdzenie 2.1 (Deville i Särndal, 1992)**.** *Rozwiązaniem zadania minimalizacji* (2.4) *przy warunku* (2.5) *dla funkcji* $G(x) = \frac{1}{2}(x-1)^2$ *jest wektor wag*

[10] Przedstawiono je również wraz z odpowiadającymi im funkcjami kalibracyjnymi $F(\cdot)$ w tabeli 2.1.



Rysunek 5.1. Wskaźniki realizacji dla wybranych badań reprezentacyjnych

Źródło: na podstawie danych GUS.

Szczególnie niepokojący jest przypadek Badania Budżetów Gospodarstw Domowych, dla którego wskaźnik realizacji jest poniżej progu 40%. Nieco korzystniej sytuacja wygląda w wypadku Badania Aktywności Ekonomicznej Ludności i przedsiębiorstw, jednak i tutaj obserwuje się trend spadkowy wskaźnika kompletności. Podobne trendy występują w badaniach realizowanych przez inne kraje należące do Unii Europejskiej (Eurostat, 2015). Różne są przyczyny niepodejmowania udziału w badaniach przez gospodarstwa domowe. Wyjaśnieniu determinant występowania braków danych w badaniach gospodarstw domowych realizowanych w Polsce poświęcona jest praca Rószkiewicz (2015). Z kolei Cobben (2009) omawia ten problem na przykładzie badań prowadzonych przez holenderski urząd statystyczny. Warto jednak zaznaczyć, że literatura przedmiotu poświęcona zagadnieniu braków odpowiedzi w badaniach gospodarstw jest bardzo obszerna. Dotyczy to zarówno przyczyn niepodejmowania udziału w badaniu, jak i metod korygowania finalnych wyników (Fricker i Tourangeau, 2010; Groves, 2006; Groves i Couper, 2012).

Paradoksalnie, problem braków odpowiedzi dotyczy nie tylko badań reprezentacyjnych mających charakter dobrowolny, ale również badań pełnych, takich jak spisy czy sprawozdawczość statystyczna, w których udział jednostek jest czę-

sto obligatoryjny (Gołata, 2018; Hora, 2009). Na przykład, w badaniach towarzyszących spisowi, takich jak badanie dzietności kobiet w Narodowym Spisie Powszechnym 1970 czy 1988, braki odpowiedzi były na poziomie 30%. Z kolei w Narodowym Spisie Powszechnym Ludności i Mieszkań 2011 około 1,5 mln osób odmówiło odpowiedzi na temat niepełnosprawności (GUS, 2012). Również w rejestrach administracyjnych, które na potrzeby statystyki publicznej przekształcane są w rejestry statystyczne, braki danych stanowią poważne źródło błędów (Groen, 2012; Laitila, A. Wallgren i B. Wallgren, 2011; 2014). Na przykład, w rejestrze PESEL, będącym głównym źródłem informacji na temat wieku czy płci mieszkańców Polski, stan cywilny jest zmienną, dla której obserwuje się spory odsetek braków danych.

Mówiąc o brakach danych jako najważniejszym źródle błędów nielosowych, należy wyróżnić całkowite (ang. *unit nonresponse*) oraz częściowe (ang. *item nonresponse*) braki odpowiedzi (Särndal i Lundström, 2005; Yan i Curtin, 2010). Omówione rodzaje braków ukazuje w syntetyczny sposób tabela 5.1.

Tabela 5.1. Rodzaje braków danych w hipotetycznym badaniu statystycznym

Jednostka	Zmienne identyfikacyjne			Zmienne w kwestionariuszu		
	1	2	3	1	2	3
1	x	x	x	x	x	x
2	x	x	x	x	x	x
3	x	x	x	.	.	x
4	x	x	x	x	x	.
5	x	x	x	.	x	x
⋮	⋮	⋮	⋮	⋮	⋮	⋮
$n-1$	x	x	x	.	.	.
n	x	x	x	.	.	.

Objaśnienia: x – istniejące informacje; . – brak danych.

Źródło: na podstawie Särndal i Lundström (2005).

Z pierwszą sytuacją mamy do czynienia, gdy znane są jedynie dane identyfikacyjne jednostki (najczęściej adresowe), która powinna wziąć, a nie bierze udziału w badaniu, na przykład na skutek odmowy, utrudnionego kontaktu, rozpadu gospodarstwa domowego czy nieobecności podczas badania. Z kolei z drugą sytuacją mamy do czynienia, gdy jednostka bierze udział w badaniu, ale nie udziela odpowiedzi na niektóre zadane pytania. Najczęściej związane jest to z drażliwością tych pytań lub obawą, że udzielone odpowiedzi zostaną wykorzystane prze-

ciwko respondentowi (na przykład pytanie o dochody, kwestie etniczne, obyczajowe itp.). Przyczyny niewzięcia udziału w badaniu bądź udzielenia odpowiedzi tylko na niektóre pytania kwestionariusza mogą mieć charakter obiektywny (choroba, podeszły wiek, nieobecność w mieszkaniu w trakcie przeprowadzenia badania ankietowego czy zmiana miejsca zamieszkania) bądź subiektywny (niechęć do badania czy brak czasu).

Należy podkreślić, że bez względu na rodzaj braków danych i przyczyny ich powstania, ich występowanie w badaniu jest zazwyczaj źródłem wielu problemów, zwłaszcza w procesie estymacji. Wynika to z tego, że respondenci różnią się na ogół od nierespondentów ze względu na pewne kluczowe cechy[41]. Powoduje to w konsekwencji powstanie błędu systematycznego. Braki danych, stanowiące główną kategorię błędów nielosowych, mają ponadto wpływ na (Lundström i Särndal, 1999; Manski, 2016; Peytchev, 2013; Toepoel i Schonlau, 2017):

- efektywną liczebność badanej próby bądź populacji, przez co zwiększa się wariancja wykorzystywanych estymatorów, tj. zmniejsza precyzja oszacowań,
- obciążenie uzyskanych wyników – oszacowane parametry znacznie odbiegają od ich „prawdziwych" wartości, a wyznaczone przedziały ufności koncentrują się wokół niewłaściwych wartości,
- zniekształcenie rozkładów analizowanych cech, przez co estymacja parametrów jest utrudniona,
- zmniejszenie zaufania do wyników badania przez końcowych odbiorców danych statystycznych.

W praktyce badań statystycznych stosuje się wiele różnego rodzaju metod, które mają zapobiegać występowaniu braków danych lub wpływać na zwiększenie wskaźnika realizacji badania oraz poprawę procesu estymacji (Brick, 2013). Mają one zastosowanie zarówno na etapie zbierania danych (stosowanie bodźców materialnych), jak i ich opracowywania (techniki korygujące wyniki). Metody te można podzielić na trzy grupy:

- **Techniki prewencyjne** – które polegają przede wszystkim na zapobieganiu występowaniu w badaniu braków odpowiedzi zarówno całkowitych, jak i częściowych. Mają one zniwelować sceptycyzm oraz niechęć respondenta do wzięcia udziału w badaniu, a także promować pozytywne do niego nastawienie. Działania prewencyjne mogą również obejmować odpowiednie przeszkolenie ankieterów czy właściwe przygotowanie kwestionariu-

[41] W literaturze anglojęzycznej funkcjonuje termin *nonrespondent* na określenie jednostki, która nie bierze udziału w badaniu. W polskiej terminologii w zasadzie brak odpowiedniego tłumaczenia. W pracy na określenie takiej jednostki przyjmować będziemy termin „nierespondent".