

Janusz L. Wywiat

**WPROWADZENIE
DO METODY
REPREZENTACYJNEJ**

WPROWADZENIE
DO METODY
REPREZENTACYJNEJ

PODREČZNIK
MATERIAŁY DYDAKTYCZNE

WPROWADZENIE DO METODY REPREZENTACYJNEJ

Janusz L. Wywił



Katowice 2010

Komitet Redakcyjny

Krystyna Lisiecka (przewodnicząca), Anna Lebda-Wyborna (sekretarz), Halina Henzel,
Anna Kostur, Maria Michałowska, Grażyna Musiał, Irena Pyka, Marian Sołtysik,
Stanisław Stanek, Stanisław Swadźba, Janusz L. Wywiał, Teresa Żabińska

Recenzent

Czesław Bracha

Redaktor

Patrycja Keller

Skład systemem T_EX

Jacek Kmiecik

www.pretext.com.pl

© Copyright by Wydawnictwo Akademii Ekonomicznej
w Katowicach 2009

ISBN 978-83-7246-524-5

WYDAWNICTWO AKADEMII EKONOMICZNEJ
IM. KAROLA ADAMIECKIEGO W KATOWICACH
ul. 1 Maja 50, 40-287 Katowice, tel. 032 257-76-35, fax 032 257-76-43
www.ae.katowice.pl, e-mail: wydawnictwo@ae.katowice.pl

Spis treści

Wprowadzenie	7
Rozdział 1. Próba losowa i statystyki	11
1.1. Populacja i parametry jej cech	11
1.2. Definicja próby oraz jej planu i schematu losowania	13
1.2.1. Proste próby losowe	14
1.2.2. Wybrane plany losowania prób nieprostych*	23
1.3. Dane i statystyki	29
Pytania i zadania	32
Rozdział 2. Estymacja	35
2.1. Definicje podstawowe	35
2.1.1. Podejście randomizacyjne	35
2.1.2. Podejście modelowe	41
2.2. Średnia z próby prostej	45
2.2.1. Estymacja punktowa	45
2.2.2. Estymacja przedziałowa	50
2.2.3. Ocena wskaźnika struktury	51
2.2.4. Wnioskowanie modelowe	55
2.3. Statystyka Horvitz-Thompsona	57
2.3.1. Estymacja wartości globalnej	57
2.3.2. Predykcja wartości globalnej	63
2.3.3. Weryfikacja prostego modelu nadpopulacji*	68
2.4. Strategie estymacji zależne od cechy dodatkowej	71
2.4.1. Estymatory ilorazowe	71
2.4.2. Predyktor ilorazowy	75
2.4.3. Estymatory regresyjne	77
2.4.4. Predyktory regresyjne	81
2.4.5. Weryfikacja regresyjnego modelu nadpopulacji*	85
2.5. Wybrane metody oceny wariancji estymatorów złożonych*	86
Pytania i zadania	89
Rozdział 3. Popularne strategie estymacji	93
3.1. Średnia z próby systematycznej	93
3.1.1. Prosta próba systematyczna	93

3.1.2. Prawdopodobieństwa inkluzji proporcjonalne do wartości zmiennej dodatkowej*	97
3.1.3. Wnioskowanie przy modelu populacji uporządkowanej	100
3.2. Średnia z próby warstwowej	101
3.2.1. Wiadomości podstawowe	101
3.2.2. Lokalizacja proporcjonalna prób w warstwach	103
3.2.3. Warunkowa minimalizacja wariancji	105
3.2.4. Warunkowa minimalizacja kosztów obserwacji zmiennej	110
3.2.5. Warstwowanie populacji po jej wylosowaniu	111
3.2.6. Przypadek nieprostych prób składowych próby warstwowej*	112
3.2.7. Predykcja na podstawie modelu warstwowego	113
3.3. Estymacja na podstawie próby grupowej	114
3.3.1. Jednostopniowa prosta próba grupowa	115
3.3.2. Próba dwustopniowa	118
3.3.3. Predykcja na podstawie grupowego modelu nadpopulacji	122
Pytania i zadania	123

Indeks	125
---------------	-----

Bibliografia	127
---------------------	-----

Wprowadzenie

W ostatnich latach daje się zauważyć wzmożone zainteresowanie metodą reprezentacyjną wśród badaczy problemów ekonomicznych, socjologicznych, opinii publicznej itd. Tą drogą są otrzymywane szybko i stosunkowo tanio informacje o parametrach cech populacji, które są potrzebne przy podejmowaniu decyzji ekonomicznych, politycznych itd. W szczególności metoda reprezentacyjna jest coraz bardziej doceniana przez ekonomistów zajmujących się problemami marketingowymi.

W praktyce badań statystycznych dane o cechach są w populacji gromadzone na podstawie obserwacji cech obiektów dobieranych do prób. Okazuje się, że podstawowe własności takich prób odbiegają od klasycznej próby prostej będącej podstawą wnioskowania w analizach wykorzystywanych, np. w naukach przyrodniczych (biologii, antropologii itp). Dlatego te próby ogólnie rzecz biorąc są nazywane próbami nieprostymi. Właśnie wnioskowanie na podstawie prób nieprostych stanowi istotę metody reprezentacyjnej, co jednocześnie odróżnia ją od innych działów statystyki.

W Polsce mamy do czynienia z pewną tradycją prowadzenia badań reprezentacyjnych, czego dowodem są liczne prace naukowe. Warto tu podkreślić, że prekursorem w tym zakresie był J. Neyman, który w 1933 r. w swojej pracy pt. „Zarys teorii i praktyki badania struktury ludności metodą reprezentacyjną” zajmował się zagadnieniami losowania prób nieprostych i wnioskowania na ich podstawie. Znaczący wkład w rozwój i upowszechnienie metody reprezentacyjnej mają następujące prace autorów polskich: R. Zasępa (1962, 1972), Z. Pawłowski (1972), A. Czajkowski, Cz. Domański (red.), K. Pruska (1985), J. Kordos (1988), Cz. Bracha (1996, 1998), Cz. Domański, K. Pruska (2001), J. L. Wywiał (1992, 1995, 2003), W. Miszczak (2004), M. Szreder (2004), W. Miszczak, W. Ostasiewicz, J. Wawrzynek (2008) i T. Żądło (2008).

Wśród wymienionych pozycji jest stosunkowo mało dostępnych podręczników, głównie z powodu wyczerpania ich nakładów. Niniejsza praca ma dać czytelnikowi możliwie prosty obraz zagadnień, jakimi zajmuje się metoda reprezentacyjna. Dlatego też dużo miejsca poświęcono na proste, lecz dość szczegółowe przedstawienie podstaw mechanizmu uogólniania wy-

ników przetwarzania danych obserwowanych w próbie na własności całej populacji. Obok znanego randomizacyjnego podejścia w metodzie reprezentacyjnej wzięto również pod uwagę problem wnioskowania modelowego (predykcyjnego), jako że jest to w wielu sytuacjach praktycznych racjonalne podejście pozwalające pominąć klasyczny etap losowania prób, który jest zastępowany celowym wyborem próby na podstawie racjonalnych przesłanek wynikających z już dostępnej informacji o populacji. W tym przypadku umiejętne korzystanie z dodatkowych, dostępnych informacji o populacji prowadzi do wyboru wartościowych prób umożliwiających dokładne wnioskowanie statystyczne o parametrach populacji. W szczególności dotyczy to wnioskowania na podstawie danych pochodzących z rejestrów administracyjnych. Pracę tak skonstruowano, aby przedstawić możliwości oceny podstawowych parametrów w różnych sytuacjach za pomocą narzędzi dostarczanych przez podejścia zarówno randomizacyjne, jak i modelowe wnioskowania statystycznego.

W pracy znajdziemy wprowadzenie w podstawy konstrukcji planów i schematów losowania próby, przedstawiono też własności wybranych planów i schematów. Następnie wyjaśniono podstawowe własności estymatorów wybranych parametrów zmiennych z uwzględnieniem prostych przykładów wyprowadzania ich rozkładów prawdopodobieństwa. Podobnie jak to ma miejsce w podręczniku Lohr (1999), wytłumaczono podstawy wnioskowania randomizacyjnego, jak i modelowego, przechodząc stopniowo od prostych zagadnień do bardziej złożonych problemów, w których są brane pod uwagę m.in. estymatory lub predyktory ilorazowe i regresyjne. Powszechnie stosowane w praktyce procedury wnioskowania przedstawiono możliwie szczegółowo. Dotyczy to estymacji lub predykcji wartości globalnej cechy bądź jej średniej na podstawie prób warstwowych, grupowych, dwustopniowych. Zasygnalizowano również problem optymalnego ustalania liczebności prób, które mają być losowane np. z warstw, na jakie wcześniej podzielono populację. Problem racjonalizacji ustalania liczebności losowanych prób jest jednym z podstawowych problemów planowania badania reprezentacyjnego i bezpośrednio rzutuje na koszty prowadzenia takiego badania. Możliwości wykorzystania prezentowanych metod wnioskowania ilustrowano licznymi przykładami. Na końcach rozdziałów zamieszczono pytania i zadania, które mają ułatwić czytelnikowi sprawdzenie przyswajanych wiadomości i wskazać na najistotniejsze ich strony. Na końcach tytułów niektórych rozdziałów zamieszczono gwiazdki, co oznacza, że Czytelnik podczas „pierwszego czytania” może je ominąć. Te podrozdziały zawierają opis bardziej złożonych procedur wnioskowania, jakkolwiek, co należy podkreślić, bynajmniej nie oznacza to zniechęcania do ich studiowania.

W pracy ograniczono się do głównych zagadnień wnioskowania w metodzie reprezentacyjnej. Ta wiedza powinna ułatwić studiowanie bardziej zło-

zonych zagadnień, jak np. weryfikacja hipotez statystycznych lub problemy estymacji w domenach populacji. Wiadomości o wymienionych zagadnieniach można znaleźć np. w pracach Brachy (1996) i Żądły (2008).

Praca ta winna służyć osobom, które chcą poznać popularne procedury losowania prób oraz sposoby oceny na ich podstawie parametrów cech populacji. Jest więc przeznaczona dla studentów ekonomii, socjologii, pracowników urzędów statystycznych itp. Podręcznik może być przydatny podczas prowadzenia wykładów z metody reprezentacyjnej. Ponadto także będzie ważną pozycją dla osób zajmujących się analizami marketingowymi, sondażami opinii publicznej oraz innymi badaniami ankietowymi. W końcu treść podręcznika winna dać dobre podstawy metodologiczne do studiowania bardziej złożonych procedur statystycznych.

Jestem bardzo wdzięczny recenzentowi niniejszej pracy, Czesławowi Brasze, za wiele cennych uwag, które pozwoliły polepszyć jej jakość.

ROZDZIAŁ 1

Próba losowa i statystyki

Przedmiotem badań metody reprezentacyjnej jest analiza własności cech (charakterystyk) zbioru obiektów zwanego populacją. W praktyce zwykle nie interesują nas wszystkie poziomy cech przypisane poszczególnym elementom populacji, lecz tylko pewne ich funkcje, nazywane funkcjami parametrycznymi, które mają syntetycznie opisywać własności struktury wartości zmiennej w populacji. Zaliczamy do nich np.: wartość średnią i wariancję cechy, współczynnik korelacji, parametry regresji. Te wielkości są również nazywane parametrami cechy. Założmy, że obserwowane cechy przypisane poszczególnym elementom populacji są ustalone. Wtedy wnioskowanie o wartości ustalonego parametru populacji jest prowadzone na podstawie losowanej próby. Taki sposób postępowania jest nazywany randomizacyjnym lub klasycznym. W wielu praktycznych sytuacjach jest rzeczą racjonalną traktować obserwacje cechy przypisane poszczególnym elementom populacji jako wartości pewnych zmiennych losowych. Zatem przyjmujemy, że każdemu elementowi populacji jest przypisana pewna zmienna losowa. W tym przypadku trzeba wprowadzić założenia o rozkładzie prawdopodobieństwa zmiennej losowej wielowymiarowej, której realizacjami są właśnie obserwacje cechy w badanej populacji. Zbiór wszystkich możliwych realizacji tej zmiennej jest nazywany nadpopulacją w stosunku do rozważanej podczas badania populacji obiektów. Dlatego podejście modelowe jest także nazywane wnioskowaniem na podstawie modelu nadpopulacji czy superpopulacji. Dodajmy, że w pewnych sytuacjach podejście modelowe daje możliwości racjonalnego wyboru próby, jeżeli założenia dotyczące własności rozkładu prawdopodobieństwa modelującego generowanie się obserwacji wartości cechy są adekwatne do rzeczywistości.

1.1. Populacja i parametry jej cech

Prezentowane tutaj podstawowe pojęcia używane w metodzie reprezentacyjnej zebrano na podstawie przede wszystkim poniższych monografii: Brachy (1996), Cassela, Särndala i Wretmana (1977), Särndala, Swenssona, Wret-

mana (1992) oraz Tillé (2006). Wyjściowym pojęciem w badaniach statystycznych jest *populacja*.

Definicja 1.1. *Zbiór obiektów $\{u_1, \dots, u_N\}$, $N < \infty$ jest nazywany populacją (skończoną) N -elementową.*

Ważne z praktycznego punktu widzenia ma następujące określenie.

Definicja 1.2. *Mówimy, że elementy populacji $\{u_1, \dots, u_N\}$ są identyfikowalne, jeżeli mogą być wzajemnie jednoznacznie ponumerowane od 1 do N oraz gdy każdy element odpowiadający danemu numerowi jest obserwowalny.*

W związku z tym populację możemy zapisać jako zbiór numerów przypisanych jej elementom, czyli $U = \{1, 2, \dots, N\}$.

Przykład 1.1. *Populacją może być następujący zbiór elementów: $U = \{\clubsuit, \diamond, \heartsuit, \spadesuit, *\}$. Jest ona identyfikowalna, ponieważ jej elementom można wzajemnie jednoznacznie przypisać kolejne liczby naturalne tworzące zbiór $U = \{1, 2, 3, 4, 5\}$.*

Przykład 1.2. *Zbiór mieszkań na danym obszarze, np. gminy, tworzy populację mieszkań. Podobnie zbiór rejestrowanych w danym województwie sklepów detalicznych tworzy populację. Zespół szkół podstawowych znajdujących się w danym kraju może być traktowany jako populacja itd.*

Definicja 1.3. *Cechą (zmienną) w populacji nazywamy takie przyporządkowanie $y: U \rightarrow R$, że dla każdego $i = 1, \dots, N$ zachodzi, iż $y(k) = y_k$.*

Wektor obserwacji wartości zmiennej y oznaczmy przez $\mathbf{y} = [y_1 \dots y_N]$, gdzie $y_k = y(k)$. Wektor $\mathbf{y}$ jest również nazywany *parametrem populacji*.

Przykład 1.3. *Na populacji określonej w przykładzie 1.1 definiujemy następującą zmienną y : $y(\clubsuit) = 0$, $y(\diamond) = 1$, $y(\heartsuit) = 1$, $y(\spadesuit) = 2$, $y(*) = 3$. Z kolei zmienna x jest następująca: $x(\clubsuit) = 0$, $x(\diamond) = 0$, $x(\heartsuit) = 0$, $x(\spadesuit) = 1$, $x(*) = 1$. Zatem zmienna y może przyjmować wartości ze zbioru $\{0, 1, 2, 3\}$, a zmienna x wartości ze zbioru $\{0, 1\}$. Zamiast pierwotnych symboli elementów populacji można użyć ich numerów. Wówczas np. zmienną x określimy następująco: $x(1) = x_1 = 0$, $x(2) = x_2 = 0$, $x(3) = x_3 = 0$, $x(4) = x_4 = 1$, $x(5) = 1$.*

Wnioskowanie o parametrach cech można znacznie polepszyć, jeżeli są dostępne informacje dodatkowe o jednostkach populacji. W szczególności są to z góry znane badaczowi obserwacje tzw. *cech pomocniczych (wspomagających, dodatkowych)* w całej populacji. Wektor tych cech oznaczamy przez $\mathbf{x} = [x_1, \dots, x_p]$, a macierz ich wartości przez $\mathbf{x} = [x_{i,j}]$, która ma wymiary $N \times p$.

Celem badania statystycznego jest m.in. ocena wartości pewnych funkcji obserwacji zmiennych nazywanych parametrami opisowymi struktury rozkładu jej wartości w populacji bądź krótko parametrami zmiennej lub populacji. W praktyce badań reprezentacyjnych do najczęściej szacowanych parametrów opisowych należą *wartość globalna* zmiennej y i jej *średnia arytmetyczna*, które określają odpowiednio wzory:

$$y_U = \sum_{k \in U} y_k, \quad \bar{y} = \bar{y}_U = \frac{1}{N} \sum_{k \in U} y_k = \frac{y_U}{N}.$$

Zatem $y_U = N\bar{y}$. Szczególnym przypadkiem wartości średniej danej zmiennej jest *częstość względna* występowania wyróżnionego poziomu cechy w populacji, którą oznaczymy przez:

$$p = \frac{m}{N},$$

gdzie m jest *liczbą elementów populacji z cechą wyróżnioną*, która jest szczególnym przypadkiem wartości globalnej zmiennej.

Powszechnie używanym parametrem zróżnicowania wartości zmiennej jest *odchylenie standardowe*, które jest równe pierwiastkowi z *wariancji* określonej wzorem:

$$v = \frac{1}{N} \sum_{k \in U} (y_k - \bar{y})^2.$$

Wykorzystuje się także wariancję liczoną według wzoru:

$$v_* = \frac{1}{N-1} \sum_{k \in U} (y_k - \bar{y})^2 = \frac{N}{N-1} v.$$

Medianę, czyli wartość środkową zmiennej y w populacji, określa się następująco:

$$m_y = \begin{cases} y_{\frac{N+1}{2}}, & \text{gdy } N \text{ jest nieparzyste,} \\ \frac{1}{2}(y_{\frac{N}{2}} + y_{\frac{N}{2}+1}), & \text{gdy } N \text{ jest parzyste.} \end{cases}$$

Przykład 1.4. Parametry określonej w przykładzie 1.3 zmiennej y są następujące: $\bar{y} = 1.4$, $y_u = 7$, $v = 1.04$, $v_* = 1.3$, a zmiennej x są postaci: $p = 0.4$, $\tilde{y} = 7$, $x_u = 2$, $m_y = 1$, $m_x = 0$.

1.2. Definicja próby oraz jej planu i schematu losowania

Ocenę wartości parametrów populacji możemy wyliczyć, jeśli posiadamy rzetelne obserwacje zmiennej dla wszystkich elementów populacji, która zwykle

składa się z tysięcy czy nawet milionów obiektów. Wtedy obserwacja wszystkich elementów populacji jest zwykle długotrwała i kosztowna. W związku z tym trzeba zadowolić się częścią obserwacji zmiennej, czyli próbą. Co więcej, w przypadku kontroli jakości produkcji badanie własności wytworzonych detali może wiązać się nawet ze zniszczeniem ich wartości użytkowej. Próbę trzeba tak wyodrębnić, by jak najlepiej reprezentowała strukturę wartości badanej cechy w populacji. Spotyka się również inne sposoby określania kryteriów reprezentatywności próby. W szczególności postuluje się, aby błąd średni oceny najważniejszego parametru był jak najmniejszy. Wyróżnia się zwykle dwa sposoby wyodrębniania prób. Pierwszy to wybór celowy próby, a drugi to jej wybór losowy. Wybór celowy elementów do próby odbywa się na podstawie z góry znanych i racjonalnych przesłanek, które pozwalają wskazać podzbiór elementów populacji reprezentatywnych dla całej populacji. W tym przypadku wyznaczenie oszacowania błędu wnioskowania o parametrach może być złożone. Wymaga to pewnych trudnych do zweryfikowania założeń modelowych. Z kolei po to, aby była możliwość wylosowania próby, trzeba dysponować spisem elementów populacji lub przynajmniej spisem ich rozłącznych podzbiorów zwanych zespołami (jednostkami losowania). Taki spis jest nazywany *operatem losowania*. Postuluje się, aby operat losowania był kompletny, czyli aby obejmował wszystkie elementy populacji bądź jednostki losowania. Ponadto oczekuje się, aby operat losowania był aktualny oraz identyfikowalny, co polega na tym, że jego elementy i obiekty tworzące populację muszą być sobie przyporządkowane w sposób wzajemnie jednoznaczny. Dodajmy, że dobry operat losowania winien obejmować także pewne informacje dodatkowe o elementach populacji, określane zwykle przez wartości tzw. cech pomocniczych. Pozwala to m.in. na optymalne warstwowanie populacji, co w konsekwencji może dać istotne podniesienie dokładności prowadzonej estymacji parametrów populacji.

1.2.1. Proste próby losowe

Próba jest traktowana jako zbiór elementów populacji. Jest zwykle oznaczana przez s i zapisywana za pomocą zbioru $s = \{k_1, k_2, \dots, k_n\}$, gdzie k_i jest numerem elementu populacji U , czyli $k_i \in U$, $i = 1, \dots, n$. Wówczas fakt przynależności k -tego elementu populacji do próby s oznaczamy przez $k \in s$. Gdy indeks i jest numerem kolejnego losowania elementu populacji, to próbę s nazywamy uporządkowaną. Jeśli próba jest losowana zwrótnie, to numery elementów populacji mogą powtarzać się w próbie. Gdy próba jest losowana bezzwrótnie, to próba s jest podzbiorem populacji U , czyli $s \subseteq U$. Przez $\mathbf{Q}$ oznaczamy zbiór prób, które są tworzone na podstawie populacji U . Zbiór $\mathbf{Q}$ jest także nazywany *przestrzenią prób* lub *zbiorem podstawowym prób*. Symbolem $\mathbf{Q}_n$ oznaczamy próby składające się z ustalonej liczby ele-

mentów. Dodajmy, że są również rozważane próby o nieustalonej liczebności elementów.

Z punktu widzenia badania własności próby i związanych z nią innych pojęć, wygodnie jest zdefiniować próbę w następujący sposób. Niech $\{0, 1\}^N$ będzie N -krotnym iloczynem kartezjańskim zbioru $\{0, 1\}$, czyli np. $\{0, 1\}^2 = \{\{0, 0\}, \{0, 1\}, \{1, 0\}, \{1, 1\}\}$. Ponadto, niech $\mathbf{M} = \{0, 1, 2, \dots\}$ będzie zbiorem liczb naturalnych, natomiast $\mathbf{M}^N$ będzie N -krotnym iloczynem kartezjańskim zbioru $\mathbf{M}$. W szczególności więc $M^2 = \{\{0, 0\}, \{0, 1\}, \{0, 2\}, \{0, 3\}, \{0, 4\}, \dots, \{1, 0\}, \{1, 1\}, \{1, 2\}, \dots\}$.

Definicja 1.4. *Próba jest określana jako wektor*

$$s = [s_1 \dots s_N],$$

przy czym:

- a) w przypadku próby bez powtórzeń elementów populacji (losowanej bezzwrotnie), $s \in \{0, 1\}^N$ i

$$s_k = \begin{cases} 1, & \text{gdy } k\text{-ty element populacji jest w próbie,} \\ 0, & \text{gdy } k\text{-ty element populacji nie jest w próbie,} \end{cases}$$

- b) w przypadku próby z powtórzeniami elementów populacji (losowanej zwrotnie) $s \in M^N$ i

$$s_k = \begin{cases} \geq 1, & \text{gdy } k\text{-ty element populacji jest w próbie,} \\ 0, & \text{gdy } k\text{-ty element populacji nie jest w próbie.} \end{cases}$$

Przez s_k oznaczono więc krotność występowania k -ego elementu populacji w próbie. Z zapisanej definicji wynika, że próba losowana bezzwrotnie jest z formalnego punktu widzenia szczególnym przypadkiem próby losowanej zwrotnie. W szczególności wektor zerowy $s = [0 \dots 0]$ jest nazywany *próbą pustą*, natomiast wektor jednostkowy $s = [1 \dots 1]$ nazwiemy *próbą wyczerpującą* populację. Wprowadzone określeniem 1.4 pojęcie próby, jak i następne definicje zamieszczone w niniejszym paragrafie podajemy głównie za Tillé (2006).

Próbę $s = \{k: s_k = 1\}$, można traktować jako podzbiór numerów elementów populacji, które jednocześnie są numerami równych jedności elementów wektora s określonego w definicji 1.4. W przypadku próby losowanej zwrotnie s jest zbiorem par (k, s_k) , a więc $s = \{(k, s_k): s_k \geq 1\}$, gdzie k jest numerem elementu populacji U znajdującym się w próbie i jednocześnie jest numerem różnego od zera elementu wektora s określonego w definicji 1.4.

W praktyce próbę s rozumie się częściej jako podzbiór numerów elementów populacji, jakkolwiek w niektórych sytuacjach jest wygodniej definiować

ją wektorem. Dalej kontekst pracy jednoznacznie wskaże, czy zdefiniowano ją zbiorem, czy wektorem.

Liczebność (rozmiar) próby wynosi:

$$n(s) = \sum_{k \in s} s_k.$$

Przykład 1.5. Gdy populacja składa się z trzech elementów, czyli $N = 3$, to przestrzeniami prób mogą być zbiory: $\mathbf{Q} = \{[0 \ 1 \ 1], [1 \ 1 \ 0], [0 \ 0 \ 1]\}$ lub $\mathbf{Q} = \{[0 \ 0 \ 0], [1 \ 3 \ 0], [0 \ 1 \ 0], [0 \ 0 \ 4]\}$. W skład pierwszego z zapisanych zbiorów wchodzi wybrane próby losowane bezzwrotnie, a w drugiej przestrzeni znajdujemy także próby losowane zwrótnie, jak np. próba $[1 \ 3 \ 0]$, w której skład wchodzi jednokrotnie pierwszy element populacji oraz trzykrotnie drugi element populacji i rozmiar tej próby wynosi cztery, czyli $n([1 \ 3 \ 0]) = 4$.

Definicja 1.5. Przestrzeń prób $\mathbf{Q}$ jest nazywana symetryczną, gdy dla każdej próby $s \in \mathbf{Q}$, wszystkie permutacje jej elementów także należą do $\mathbf{Q}$.

Przykład 1.6. Niech populacja składa się z czterech elementów, czyli $N = 4$. Wówczas przykładem symetrycznej przestrzeni prób może być zbiór $\mathbf{Q} = \{[0 \ 0 \ 0 \ 0], [1 \ 0 \ 0 \ 0], [0 \ 1 \ 0 \ 0], [0 \ 0 \ 1 \ 0], [0 \ 0 \ 0 \ 1], [1 \ 1 \ 0 \ 0], [1 \ 0 \ 1 \ 0], [1 \ 0 \ 0 \ 1], [0 \ 1 \ 1 \ 0], [0 \ 1 \ 0 \ 1], [0 \ 0 \ 1 \ 1], [2 \ 0 \ 0 \ 0], [0 \ 2 \ 0 \ 0], [0 \ 0 \ 2 \ 0], [0 \ 0 \ 0 \ 2]\}$. Mamy więc do czynienia z przestrzenią prób losowanych zwrótnie o liczebności nie większej niż dwa elementy, czyli $n \leq 2$.

Definicja 1.6. Przestrzeń prób losowanych bezzwrotnie o ustalonej liczebności n oznaczana przez $\mathbf{Q}_n$ jest zbiorem wszystkich możliwych permutacji bez powtórzeń N -elementowego wektora $[1 \dots 1 \ 0 \dots 0]$, który zawiera n jedynek i $N - n$ zer.

Symetryczną przestrzenią prób losowanych bezzwrotnie o liczebności nie większej niż N jest suma $\mathbf{Q} = \bigcup_{n=0}^N \mathbf{Q}_n$. W szczególności gdy $N = 2$, to zbiór $\mathbf{Q} = \{[0, 0], [1, 0], [0, 1], [1, 1]\}$ jest iloczynem kartezjańskim zbiorów $\{0, 1\}$ i $\{0, 1\}$, czyli $\mathbf{Q} = \{0, 1\} \times \{0, 1\} = \{0, 1\}^2$. Uogólniając to rozumowanie dochodzimy do wniosku, że przestrzeń $\mathbf{Q}$ jest N -krotnym iloczynem kartezjańskim zbioru $\{0, 1\}$, czyli $\mathbf{Q} = \{0, 1\}^N$.

Liczba wszystkich prób tworzących przestrzeń $\mathbf{Q}_n$ wynosi $\binom{N}{n}$, natomiast liczba wszystkich prób losowanych bezzwrotnie z populacji N -elementowej jest równa 2^N .

Przykład 1.7. Gdy populacja składa się z trzech elementów, czyli $N = 3$, to symetryczną przestrzenią prób losowanych bezzwrotnie jest $\mathbf{Q} = \{0, 1\}^3 = \{[0 \ 0 \ 0], [1 \ 0 \ 0], [0 \ 1 \ 0], [0 \ 0 \ 1], [0 \ 1 \ 1], [1 \ 0 \ 1], [1 \ 1 \ 0], [1 \ 1 \ 1]\}$, a w szczególności jej podzbiór: $\mathbf{Q}_2 = \{[0 \ 1 \ 1], [1 \ 0 \ 1], [1 \ 1 \ 0]\}$ jest przestrzenią prób losowanych bezzwrotnie o liczebności $n = 2$.

Gdy próba jest losowana zwrotnie z populacji, to dany element populacji może się w niej wielokrotnie powtarzać. Niech s_k , $i = 1, \dots, N$ będzie krotnością powtarzania się k -tego elementu populacji w próbie.

Definicja 1.7. *Symetryczna przestrzeń prób losowanych zwrotnie o ustalonej liczebności n , oznaczana przez $\underline{\mathbf{Q}}_n$, jest zbiorem wszystkich możliwych prób s takich, że $\underline{s} = [\underline{s}_1, \dots, \underline{s}_N]$, $\underline{s}_i = 0, 1, \dots, n$, dla $i = 1, \dots, N$ i $\sum_{i=1}^N \underline{s}_i = n$.*

Liczba wszystkich prób tworzących przestrzeń $\underline{\mathbf{Q}}_n$ wynosi $\binom{N+n-1}{n}$. Symetryczną przestrzenią prób losowanych zwrotnie o dowolnej liczebności nie większej od m jest suma: $\underline{\mathbf{Q}}_{\leq m} = \bigcup_{n=0}^m \underline{\mathbf{Q}}_n$.

Próbę $\underline{s} = [\underline{s}_1, \dots, \underline{s}_N]$ losowaną zwrotnie można sprowadzić do próby losowanej bezzwrotnie $s = [s_1, \dots, s_N]$ za pomocą tzw. *funkcji redukującej* $s = [r(s_1), \dots, r(s_N)] = r(\underline{s})$ definowanej wyrażeniem:

$$s_k = r(s_k) = \begin{cases} 1 & \text{jeśli } \underline{s}_k > 0, \\ 0 & \text{jeśli } \underline{s}_k = 0. \end{cases} \quad (1.1)$$

Liczba różnych elementów populacji w próbie $\underline{s}$ jest nazywana *efektywną liczebnością* tej próby i jest oznaczana przez $n_e(\underline{s}) = n(s) = n(r(\underline{s}))$. Oczywiście liczebność próby losowanej bezzwrotnie jest równa jej liczebności efektywnej.

Przykład 1.8. *Przestrzeń uzyskana poprzez redukcję przestrzeni prób losowanych zwrotnie o liczebności n jest równa sumie przestrzeni prób losowanych bezzwrotnie o liczebności nie mniejszej niż 1 i nie większej niż n , czyli $r(\underline{\mathbf{Q}}_n) = \bigcup_{m=1}^n \mathbf{Q}_m$. W szczególności, gdy $N = 3$ i $n = 2$, to*

$$\begin{aligned} r(\underline{\mathbf{Q}}_2) &= r(\{[1 \ 1 \ 0], [1 \ 0 \ 1], [0 \ 1 \ 1], [2 \ 0 \ 0], [0 \ 2 \ 0], [0 \ 0 \ 2]\}) = \\ &= \{[1 \ 0 \ 0], [0 \ 1 \ 0], [0 \ 0 \ 1], [1 \ 1 \ 0], [1 \ 0 \ 1], [0 \ 1 \ 1]\} = \mathbf{Q}_1 \cup \mathbf{Q}_2 = \mathbf{Q}_{\leq 2}. \end{aligned}$$

Definicja 1.8. *Planem losowania próby s jest nazywany rozkład prawdopodobieństwa $p(s)$ określony na zbiorze podstawowym prób $\mathbf{Q}$, a więc funkcja $p(s)$, będąca prawdopodobieństwem wylosowania próby s , przyporządkowuje każdemu elementowi s ze zbioru prób $\mathbf{Q}$ liczbę z przedziału zamkniętego $[0; 1]$ i*

$$\sum_{s \in \mathbf{Q}} p(s) = 1.$$

W przypadku planu bezzwrotnego losowania próby $\mathbf{Q} \subseteq \{0, 1\}^N$.

Przykład 1.9. *Plan losowania zwrotnego próby 2-elementowej, której przestrzeń $\underline{\mathbf{Q}}_2$ określono w przykładzie 1.8 można ustalić następująco: $p([1 \ 1 \ 0]) = 0.01$, $p([1 \ 0 \ 1]) = 0.09$, $p([0 \ 1 \ 1]) = 0.5$, $p([2 \ 0 \ 0]) = 0.1$, $p([0 \ 2 \ 0]) = 0.1$, $p([0 \ 0 \ 2]) = 0.2$.*

Próbę $s = [s_1 \dots s_N]$ traktuje się również jako wartość (obserwację) N -wymiarowej zmiennej losowej $S = [S_1 \dots S_N]$, co wynika z następującego określenia (por. Tillé 2006).

Definicja 1.9. *Próbką losową jest nazywany wektor losowy S przyjmujący wartości ze zbioru $\mathbf{Q}$, którego rozkładem prawdopodobieństwa jest plan $p(s)$, czyli*

$$P(S = s) = p(s), \quad \text{dla wszystkich } s \in \mathbf{Q}.$$

Zatem rozkład prawdopodobieństwa zmiennej losowej S_k jest rozkładem brzegowym wektora S , który wyznacza następujące wyrażenie:

$$P(S_k = a) = \sum_{\{s: s_k=a\}} P(S = s).$$

Wartość oczekiwaną próby losowej określa wzór:

$$\mu = [\mu_1 \dots \mu_N] = E(S) = \sum_{s \in \mathbf{Q}} sp(s),$$

gdzie $\mu_k = E(S_k) = \sum_{s \in \mathbf{Q}} s_k p(s)$.

Macierz wariancji i kowariancji próby losowej ma postać:

$$\Sigma = [\Sigma_{k,l}] = E(S - \mu)^T (S - \mu) = \sum_{s \in \mathbf{Q}} (s - \mu)^T (s - \mu) p(s),$$

gdzie: $\Sigma_{k,l} = \mu_{k,l} - \mu_k \mu_l$, $\mu_{k,l} = E(S_k S_l) = \sum_{s \in \mathbf{Q}} s_k s_l p(s)$.

Przykład 1.10. *Dla określonego w przykładzie 1.9 planu losowania zwrotnego próby 2-elementowej mamy: $\mu_1 = E(S_1) = 0.01 + 0.09 + 0 + 0.2 + 0 + 0 = 0.3$ oraz $\mu = [0.3 \ 0.71 \ 0.99]$, $\mu_{1,1} = E(S_1^2) = 0.01 + 0.09 + 0.4 = 0.5$, $\mu_{2,2} = 0.91$, $\mu_{3,3} = 1.39$, $\mu_{1,2} = E(S_1 S_2) = 0.01$, $\mu_{1,3} = 0.09$, $\mu_{2,3} = 0.5$. Te wyliczenia prowadzą do następującej macierzy wariancji i kowariancji próby losowej.*

$$\Sigma = \begin{bmatrix} 0.4100 & -0.2030 & -0.2070 \\ -0.2030 & 0.4059 & -0.2029 \\ -0.2070 & -0.2029 & 0.4099 \end{bmatrix}$$

Definicja 1.10. *Prawdopodobieństwo inkluzji rzędu pierwszego jest prawdopodobieństwem tego, iż k -ty element populacji zostanie wylosowany do próby i jest oznaczane przez*

$$\pi_k = P(S_k > 0) = \sum_{\{s: s_k > 0\}} p(s).$$

Na podstawie wzoru (1.1) można wykazać, że $\pi_k = E(r(S_k))$. Zatem gdy próba jest losowana bezzwrotnie, to $\pi_k = E(S_k)$. Symbolem $\boldsymbol{\pi}$ oznaczamy wektor wierszowy prawdopodobieństw inkluzji rzędu pierwszego. Gdy próba jest losowana bez zwracania, to $\boldsymbol{\pi} = \boldsymbol{\mu}$.

Definicja 1.11. *Prawdopodobieństwo inkluzji rzędu drugiego jest prawdopodobieństwem tego, iż elementy populacji o numerach k oraz l , przy czym $k \neq l$, zostaną wylosowane do próby i jest oznaczane przez:*

$$\pi_{k,l} = P(S_k > 0 \text{ i } S_l > 0) = \sum_{\{s: s_k > 0 \text{ i } s_l > 0\}} p(s).$$

Ze wzoru (1.1) wynika, że $\pi_{k,l} = E(r(S_k)r(S_l))$. Niech $\boldsymbol{\Pi} = [\pi_{k,l}]$, przy czym $\pi_{k,k} = \pi_k$, $k = 1, \dots, N$.

Niech

$$\boldsymbol{\Delta} = [\Delta_{k,l}] = \boldsymbol{\Pi} - \boldsymbol{\pi}^T \boldsymbol{\pi}. \quad (1.2)$$

Zwróćmy uwagę, że $\Delta_{k,t} = \pi_{k,t} - \pi_k \pi_t$ dla $i \neq j$, natomiast $\Delta_{k,k} = \pi_k(1 - \pi_k)$. W przypadku bezzwrotnie losowanej próby

$$\boldsymbol{\Delta} = \boldsymbol{\Sigma}.$$

Gdy próby losowane bezzwrotnie mają ustaloną liczebność n , to

$$\sum_{k=1}^N \pi_k = n, \quad \sum_{k=1}^N \sum_{t=1, t \neq k}^N \pi_{k,t} = n(n-1). \quad (1.3)$$

Przykład 1.11. *Dla określonego w przykładzie 1.9 planu losowania zwrotnego próby 2-elementowej wyznaczamy następujące charakterystyki: $\pi_1 = E(r(S_1)) = 0.01 + 0.09 + 0.1 = 0.2$, $\boldsymbol{\pi} = [0.2 \ 0.61 \ 0.79]$, $\pi_{1,2} = E(r(S_1)r(S_2)) = 0.01$, $\pi_{1,3} = 0.09$, $\pi_{2,3} = 0.5$, czyli:*

$$\boldsymbol{\Pi} = \begin{bmatrix} 0.20 & 0.01 & 0.09 \\ 0.01 & 0.61 & 0.50 \\ 0.09 & 0.50 & 0.79 \end{bmatrix}.$$

Wówczas:

$$\boldsymbol{\Delta} = \begin{bmatrix} 0.1600 & -0.1120 & -0.0680 \\ -0.1120 & 0.2379 & 0.0181 \\ -0.0680 & 0.0181 & 0.1659 \end{bmatrix}.$$

Przykład 1.12. *W przykładzie 1.7 określono przestrzeń prób $\mathbf{Q}_2$ dwuelementowych losowanych bezzwrotnie z populacji trójelementowej. Definiujemy na niej plan losowania próby: $p([0 \ 1 \ 1]) = 0.5$, $p([1 \ 0 \ 1]) = 0.3$, $p([1 \ 1 \ 0]) = 0.2$. Parametry tego planu są następujące: $\boldsymbol{\pi} = [0.5 \ 0.7 \ 0.8]$.*

$$\boldsymbol{\Pi} = \begin{bmatrix} 0.5 & 0.2 & 0.3 \\ 0.2 & 0.7 & 0.5 \\ 0.3 & 0.5 & 0.8 \end{bmatrix}, \quad \boldsymbol{\Sigma} = \boldsymbol{\Delta} = \begin{bmatrix} 0.25 & -0.15 & -0.10 \\ -0.15 & 0.21 & -0.06 \\ -0.10 & -0.06 & 0.16 \end{bmatrix}.$$

Plan losowania próby określa prawdopodobieństwa, z jakimi poszczególne próby mają być wyodrębniane z populacji. Z praktycznego punktu widzenia jest rzeczą konieczną określenie sposobu doboru poszczególnych składników próby spośród elementów populacji.

Definicja 1.12. *Schemat (algorytm) losowania próby jest procedurą prowadzącą do wyselekcjonowania z populacji elementów do próby tak, aby prawdopodobieństwo wylosowania próby było równe prawdopodobieństwu określёнemu przez plan losowania próby.*

Schemat losowania próby jest nazywany *wyliczającym* (enumerative), jeżeli wszystkie możliwe próby oraz prawdopodobieństwa ich wylosowania są wyznaczone przed losowaniem próby. Wyliczające schematy losowania wymagają dużej ilości operacji numerycznych, która zwykle rośnie w sposób wybuchowy wraz ze zwiększaniem się rozmiaru populacji lub próby. Przykładowe ilości prób wchodzących w skład przestrzeni prób zamieszczono w tabeli 1.1. Wynika z niej, że są one wręcz astronomiczne już dla przypadków stosunkowo małych liczebności populacji i prób. Na przykład, gdy rozmiar populacji wzrasta 10 razy z poziomu 100 do 1000, to liczba wszystkich możliwych prób bezzwrotnie losowanych wzrasta co najmniej 10^{270} razy!

Tabela 1.1

Rozmiary symetrycznych przestrzeni prób

N :	100		1000	
n :	5	10	10	50
$Card(\mathbf{Q}_n = \binom{N}{n})$	7.528×10^7	1.731×10^{13}	2.634×10^{23}	9.460×10^{84}
$Card(\underline{\mathbf{Q}}_n) = \binom{N+n-1}{n}$	9.196×10^7	4.263×10^{13}	2.882×10^{23}	1.097×10^{86}
$Card(\mathbf{Q}) = 2^N$	1.267651×10^{30}		1.071509×10^{301}	

Chodzi więc o to, aby znaleźć efektywny algorytm losowania próby, czyli taki, który wymagałby najmniej tego typu operacji przy danym planie losowania, co powinno prowadzić do istotnej oszczędności czasu poświęcanego na wyselekcjonowanie próby z populacji.

Problemom konstrukcji algorytmów losowania próby jest poświęcona praca Tillé (2006). Tutaj przedstawimy tylko konkretne algorytmy przepisane rozważanym planom losowania próby.

Konstrukcja algorytmu wyliczającego losowania próby jest następująca:

1. Najpierw tworzy się ciąg (listę) wszystkich możliwych prób: $\mathbf{Q} = \{s_1, \dots, s_J\}$.
2. Wylicza się prawdopodobieństwa wystąpienia wszystkich prób ze zbioru $\mathbf{Q}$ zgodnie z planem losowania $P(s)$.
3. Potem, generuje się liczbę losową u o rozkładzie jednostajnym określonym na przedziale $[0, 1]$.

4. W końcu wybiera się tę próbę s_j , dla której $\sum_{i=1}^{j-1} P(s_i) \leq u < \sum_{i=1}^j P(s_i)$ przy czym, jeśli $u < P(s_1)$, to próba s_1 jest wybierana.

Rozważymy teraz własności elementarnych planów losowania próby.

Definicja 1.13. *Plan bezzwrotnego losowania próby prostej ma postać:*

$$P_{bz}(s) = \binom{N}{n}^{-1},$$

gdzie $s \in \mathbf{Q}_n$.

Parametry tego planu są następujące: $\pi_k = \frac{n}{N}$, $\pi_{k,l} = \frac{n(n-1)}{N(N-1)}$, $\Delta_{k,k} = \frac{n(N-n)}{N^2}$, $\Delta_{k,l} = -\frac{n(N-n)}{N^2(N-1)}$, przy czym $k = 1, \dots, N$, $l = 1, \dots, N$, $k \neq l$.

Algorytm (schemat) losowania rozważanej próby prostej jest następujący:

1. Pierwszy element próby jest losowany z populacji z prawdopodobieństwem $q_1 = \frac{1}{N}$, czyli np. należy wygenerować liczbę losową u o rozkładzie jednostajnym określonym na przedziale $[0; 1]$ i wtedy jest wybierany j -ty element populacji, gdy $\frac{j-1}{N} \leq u < \frac{j}{N}$, $j = 1, \dots, N$.
2. k -ty element próby jest podobnie losowany z populacji pomniejszonej o już wcześniej wylosowane z niej $(k-1)$ elementy z prawdopodobieństwem $q_k = \frac{1}{N-k+1}$, $k = 2, \dots, n$.

Przykład 1.13. *Gdy $N = 5$ i $n = 2$, to plan bezzwrotnego losowania próby losowej prostej ma postać: $P_{bz}(s) = 0.1$, gdzie $s \in \mathbf{Q}_2$, a jego parametry określają parametry:*

$$\mathbf{\Pi} = \begin{bmatrix} 0.4 & 0.1 & 0.1 \\ 0.1 & 0.4 & 0.1 \\ 0.1 & 0.1 & 0.4 \end{bmatrix}, \quad \mathbf{\Sigma} = \mathbf{\Delta} = \begin{bmatrix} 0.24 & -0.06 & -0.06 \\ -0.06 & 0.24 & -0.06 \\ -0.06 & -0.06 & 0.24 \end{bmatrix}.$$

Algorytm losowania rozważanej próby jest następujący. Pierwszy element próby należy losować z prawdopodobieństwem $q_1 = 0.2$ z całej pięcioelementowej populacji, a drugi element wybieramy z prawdopodobieństwem $q_2 = 0.25$ z populacji czteroelementowej, czyli pomniejszonej o już wcześniej wylosowany element.

Tillé (2006, s. 54), wprowadza następujące określenie.

Definicja 1.14. *Plan zwrotnego losowania tzw. próby losowej prostej ma postać:*

$$P_z(s) = \frac{n!}{N^n \prod_{k \in U} s_k!},$$

gdzie $s \in \mathbf{Q}_n$.

Prawdopodobieństwa inkluzji tego planu określają wzory:
 $\pi_k = 1 - \left(\frac{N-1}{N}\right)^n$, $\pi_{k,l} = 1 - 2\left(\frac{N-1}{N}\right)^n + \left(\frac{N-2}{N}\right)^n$, przy czym $k = 1, \dots, N$,
 $l = 1, \dots, N$, $k \neq l$.

Schemat losowania rozważanej próby jest następujący. Każdy spośród N -elementów populacji jest losowany z prawdopodobieństwem $1/N$ do próby n -krotnie. Stąd wynika, że elementy populacji są losowane do próby niezależnie.

Przykład 1.14. Gdy $N = 4$ i $n = 2$, to plan zwrotnego losowania próby prostej ma postać: $P_z(s) = 0.1$, gdzie $s \in \underline{\mathbf{Q}}_2$. Parametry tego planu są następujące: $\boldsymbol{\mu} = [0.5 \ 0.5 \ 0.5 \ 0.5]$,

$$\boldsymbol{\Pi} = \frac{1}{16} \begin{bmatrix} 7 & 2 & 2 \\ 2 & 7 & 2 \\ 2 & 2 & 7 \end{bmatrix}, \quad \boldsymbol{\Sigma} = \frac{1}{8} \begin{bmatrix} 3 & -1 & -1 \\ -1 & 3 & -1 \\ -1 & -1 & 3 \end{bmatrix}, \quad \boldsymbol{\Delta} = \frac{1}{256} \begin{bmatrix} 63 & -17 & -17 \\ -17 & 63 & -17 \\ -17 & -17 & 63 \end{bmatrix}.$$

Po to, aby wylosować rozważaną próbę należy z całej populacji wylosować (zwrotnie) dwa elementy, przy czym za każdym razem prawdopodobieństwo wylosowania elementu populacji ma wynosić 0.25.

Definicja 1.15. Plan bezzwrotnego losowania Bernoulliego próby losowej prostej ma postać:

$$P_\pi(s) = \prod_{k=1}^N \pi^{s_k} (1 - \pi)^{1-s_k},$$

gdzie $s \in \mathbf{Q}$ i $\pi \in [0, 1]$.

Parametr π jest prawdopodobieństwem wylosowania każdego elementu populacji do próby, czyli $\pi_k = E(S_k) = \pi$ dla $k = 1, \dots, N$. Prawdopodobieństwa inkluzji rzędu drugiego są również takie same i wynoszą $\pi_{k,l} = \pi^2$ oraz $D^2(S_k) = \pi(1 - \pi)$ oraz $\text{Cov}(S_k, S_l) = 0$ dla $k \neq l$, $k = 1, \dots, N$, $l = 1, \dots, N$.

Próby Bernoulliego nie mają powtarzających się elementów populacji, a ich rozmiar jest zmienną losową, której wartość spełnia nierówności: $0 \leq n(s) \leq N$. Liczebność próby ma rozkład dwumianowy prawdopodobieństwa, który określa wzór:

$$P(n(S) = n) = \binom{N}{n} \pi^n (1 - \pi)^{1-n}, \quad (1.4)$$

przy czym $n = 0, 1, \dots, N$. Zatem, $E(n(S)) = N\pi$, $D^2(n(S)) = N\pi(1 - \pi)$.

Przykład 1.15. Sprzedano partię $N = 20\,000$ lodówek określonego typu. Prawdopodobieństwo awarii każdej z nich wynosi $\pi = 0.005$. Zbiór reklamowanych lodówek można więc traktować jako próbę Bernoulliego. W związku z tym wartość oczekiwana rozmiaru tej próby, jednocześnie będąca przeciętną liczbą reklamowanych lodówek, wynosi 100 sztuk.

1.2.2. Wybrane plany losowania prób nieprostych*

W niniejszej części przedstawimy pewne plany losowania, których prawdopodobieństwa inkluzji rzędu pierwszego nie są takie same dla elementów populacji. Zwykle będą one zależne od wartości tzw. cechy dodatkowej (pomocniczej) obserwowanej w całej populacji. Wśród nich da się wyróżnić takie, które zależą od obserwacji wartości momentów cechy dodatkowej, a inne od kwantyli tej zmiennej. Pierwszy z rozważanych planów losowania jest określany następująco (por. np. Tillé, 2006):

Definicja 1.16. *Plan Poissona bezzwrotnego losowania próby określa wyrażenie:*

$$P_P(s) = \prod_{k=1}^N \pi_k^{s_k} (1 - \pi_k)^{1-s_k},$$

przy czym $s \in \mathbf{Q}$ i $\pi_k \in [0, 1]$ dla $k = 1, \dots, N$.

Element k -ty populacji jest losowany do próby z prawdopodobieństwem π_k , $k = 1, \dots, N$, które jest jednocześnie prawdopodobieństwem inkluzji rzędu pierwszego oraz $E(S_k) = \pi_k$. Elementy populacji są losowane do próby w sposób niezależny. Zatem prawdopodobieństwo inkluzji rzędu drugiego wynosi $\pi_{k,l} = \pi_k \pi_l$ dla $k \neq l$, $k = 1, \dots, N$, $l = 1, \dots, N$. Wartość oczekiwana liczebności próby wynosi $E(n(S)) = \sum_{k=1}^N \pi_k$.

Przykład 1.16. *Niech obserwacje cechy pomocniczej będą dodatnie w całej populacji, czyli $x_k > 0$ dla $k = 1, \dots, N$. Wówczas można przyjąć, że*

$$\pi_k = \frac{x_k}{\sum_{i=1}^N x_i}, \quad k = 1, \dots, N.$$

W praktyce zmienna pomocnicza w zależności od celu badania statystycznego może być w różny sposób dobierana. Przykładowo x_k może być obserwacją wieku respondenta, który jest dostępny w różnych rejestrach administracyjnych. Wówczas prawdopodobieństwo wylosowania danej osoby do próby byłoby proporcjonalne do jej wieku. Przypuśćmy, że obserwacje wartości zmiennej pomocniczej indeksowane numerami pięcioelementowej populacji są następujące: $x_1 = 1$, $x_2 = 2$, $x_3 = 4$, $x_4 = 7$, $x_5 = 11$. Wtedy wyżej określone prawdopodobieństwa przyjmują wartości: $\pi_1 = 0.04$, $\pi_2 = 0.08$, $\pi_3 = 0.16$, $\pi_4 = 0.28$ i $\pi_5 = 0.44$.

Oznaczmy średnie i wariancje z populacji lub z próby s zmiennej dodatkowej następująco:

$$\bar{x} = \frac{1}{N} \sum_{i \in U} x_i, \quad v_*(x) = \frac{1}{N-1} \sum_{i \in U} (x_i - \bar{x})^2, \quad v(x) = \frac{N-1}{N} v_*(x). \quad (1.5)$$

$$\bar{x}_s = \frac{1}{n} \sum_{i=1}^N x_i s_i, \quad v_{*s}(x) = \frac{1}{n-1} \sum_{i=1}^N (x_i - \bar{x}_s)^2 s_i, \quad v_s(x) = \frac{n-1}{n} v_{*s}(x). \quad (1.6)$$

lub

$$\bar{x}_s = \frac{1}{n} \sum_{i \in s} x_i, \quad v_{*s}(x) = \frac{1}{n-1} \sum_{i \in s} (x_i - \bar{x}_s)^2. \quad (1.7)$$

Midzuno (1950), Lahiri (1951) oraz Sen (1953) analizowali następujący plan i schemat losowania próby.

Definicja 1.17. *Plan bezzwrotnego losowania n -elementowej próby s , proporcjonalny do średniej cechy dodatkowej z tej próby $\bar{x}_s$ ma postać:*

$$P_{LMS}(s) = \frac{1}{\binom{N}{n}} \frac{\bar{x}_s}{\bar{x}}, \quad \text{gdzie } s \in \mathbf{Q}_n.$$

Prawdopodobieństwa inkluzji rzędu pierwszego i drugiego mają postać (por. np. Chaudhuri i Vos, 1988 lub Wywiał, 1995):

$$\begin{cases} \pi_k &= \frac{n}{N} + \frac{N-n}{N(N-1)} \frac{x_k - \bar{x}}{\bar{x}}, \\ \pi_{k,l} &= \frac{n-2}{N-2} (\pi_k + \pi_l) - \frac{n(n-1)}{(N-1)(N-2)} \\ &\text{lub równoważnie} \\ \pi_{k,l} &= \frac{n(n-1)}{N(N-1)} + \frac{(n-1)(N-n)}{N(N-1)(N-2)} \frac{x_k + x_l - 2\bar{x}}{\bar{x}}, \end{cases} \quad (1.8)$$

przy czym $k \neq l$, $k = 1, \dots, N$, $l = 1, \dots, N$.

Schemat losowania próby jest następujący. Pierwszy element do próby jest losowany z prawdopodobieństwami proporcjonalnymi do wartości cechy pomocniczej, czyli:

$$p(k_1) = \frac{x_{k_1}}{\sum_{i=1}^N x_i}, \quad k = 1, \dots, N.$$

Pozostałe elementy próby tworzy losowana bezzwrotnie próba prosta o liczebności $(n-1)$.

Wywiał (1991, 1992) wyprowadza prawdopodobieństwa inkluzji dowolnego rzędu planu P_{LMS} oraz następujące odpowiednie prawdopodobieństwa

warunkowe określające schemat losowania próby. Pierwszy element jest losowany do próby z wyżej określonym prawdopodobieństwem bezwarunkowym $p(k_1)$, natomiast dla $r = 2, \dots, n$

$$p(k_r | k_{r-1}, \dots, k_1) = \frac{(N - n) \sum_{i=1}^r x_{k_i} + N(n - r)\bar{x}}{(N - 1) \left((N - n) \sum_{i=1}^{r-1} x_{k_i} + N(n - r + 1)\bar{x} \right)}.$$

Zatem jest to znacznie bardziej złożony i wolniejszy schemat losowania próby od algorytmu Midzuno i Lahiriego. Ponadto, jest to przykład na to, iż przedstawiliśmy dwa schematy losowania realizujące jeden i ten sam plan losowania próby, co jest zgodne z twierdzeniem T.V.H. Rao (1962), który wykazał, że dla każdego planu losowania próby istnieją co najmniej dwa schematy jej losowania.

Przykład 1.17. *Podobnie jak w przykładzie 1.16 niech obserwacjami wartości zmiennej pomocniczej będą następujące liczby: $x_1 = 1$, $x_2 = 2$, $x_3 = 4$, $x_4 = 7$, $x_5 = 11$, przy czym numery obserwacji zmiennej są jednocześnie numerami elementów populacji. Zatem populacja liczy $N = 5$ elementów. Na podstawie definicji 1.17 otrzymujemy następujące prawdopodobieństwa planu losowania dwuelementowej próby proporcjonalnego do średniej z próby cechy dodatkowej: $P_{LMS}(1\ 2) = 0.03$, $P_{LMS}(1\ 3) = 0.05$, $P_{LMS}(1\ 4) = 0.08$, $P_{LMS}(1\ 5) = 0.12$, $P_{LMS}(2\ 3) = 0.06$, $P_{LMS}(2\ 4) = 0.09$, $P_{LMS}(2\ 5) = 0.13$, $P_{LMS}(3\ 4) = 0.11$, $P_{LMS}(3\ 5) = 0.15$, $P_{LMS}(4\ 5) = 0.18$. Na podstawie wzoru (1.8) wyliczamy macierz prawdopodobieństw inkluzji:*

$$\mathbf{\Pi} = \begin{bmatrix} 0.28 & 0.03 & 0.05 & 0.08 & 0.12 \\ 0.03 & 0.31 & 0.06 & 0.09 & 0.13 \\ 0.05 & 0.06 & 0.37 & 0.11 & 0.15 \\ 0.08 & 0.09 & 0.11 & 0.46 & 0.18 \\ 0.12 & 0.13 & 0.15 & 0.18 & 0.58 \end{bmatrix},$$

przy czym prawdopodobieństwa inkluzji rzędu pierwszego są na głównej przekątnej macierzy $\mathbf{\Pi}$, a poza nią znajdują się prawdopodobieństwa inkluzji rzędu drugiego. Ponadto:

$$\mathbf{\Sigma} = \mathbf{\Delta} = \begin{bmatrix} 0.202 & -0.057 & -0.054 & -0.049 & -0.042 \\ -0.057 & 0.214 & -0.055 & -0.053 & -0.050 \\ -0.054 & -0.055 & 0.233 & -0.060 & -0.065 \\ -0.049 & -0.053 & -0.060 & 0.248 & -0.087 \\ -0.042 & -0.050 & -0.064 & -0.087 & 0.244 \end{bmatrix}.$$

Definicja 1.18 (Singh i Srivastava, 1980). *Plan bezzwrotnego losowania n -elementowej próby s proporcjonalny do wariancji cechy dodatkowej z tej próby v_{*s} ma postać:*

$$P_{SS}(s) = \frac{1}{\binom{N}{n}} \frac{v_{*s}(x)}{v_*(x)}, \quad \text{gdzie } s \in \mathbf{Q}_n$$

przy czym wariancję v_* i v_{*s} określają wzory (1.5)–(1.7).

Schemat losowania realizujący plan P_{SS} polega na tym, że najpierw należy wylosować z populacji parę elementów (k, l) z prawdopodobieństwami proporcjonalnymi do $(x_k - x_l)^2$. Następnie pozostałe $(n - 2)$ elementy próby dobieramy z populacji pomniejszonej o już wybrane wcześniej dwa elementy, jako próbę prostą losowaną bezzwrotnie.

Niech $z_k = \frac{x_k - \bar{x}}{\sqrt{v(x)}}$. Wtedy prawdopodobieństwa inkluzji rzędu pierwszego i drugiego mają postać:

$$\begin{cases} \pi_k &= \frac{n}{N} + \frac{N-n}{N(N-2)}(z_k^2 - 1), \\ \pi_{k,l} &= \frac{n(n-1)}{N(N-1)} + \frac{N-n}{N(N-2)} \left(1 - \frac{(N-1)(N-n-1)}{N(N-3)} \right) (z_k^2 + z_l^2 - 2) + \\ &\quad - \frac{2(N-n)(N-n-1)}{N^2(N-2)(N-3)} \left(z_k z_l - \frac{1}{N-1} \right), \end{cases} \quad (1.9)$$

przy czym $k \neq l$, $k = 1, \dots, N$, $l = 1, \dots, N$.

Prawdopodobieństwa inkluzji wyższych rzędów podaje Wywiał (1992).

Przykład 1.18. *Korzystamy z danych z przykładu 1.17 o obserwacjach wartości zmiennej pomocniczej w pięcioelementowej populacji. Na podstawie definicji 1.18 otrzymujemy następujące prawdopodobieństwa planu losowania dwuelementowej próby proporcjonalnego do wariancji z próby cechy dodatkowej: $P_{SS}(1, 2) = 0.003$, $P_{SS}(1, 3) = 0.027$, $P_{SS}(1, 4) = 0.109$, $P_{SS}(1, 5) = 0.303$, $P_{SS}(2, 3) = 0.012$, $P_{SS}(2, 4) = 0.076$, $P_{SS}(2, 5) = 0.246$, $P_{SS}(3, 4) = 0.027$, $P_{SS}(3, 5) = 0.149$, $P_{SS}(4, 5) = 0.048$. Na podstawie wzoru (1.9) wyliczamy macierz prądopodobieństw inkluzji:*

$$\mathbf{\Pi} = \begin{bmatrix} 0.442 & 0.003 & 0.027 & 0.109 & 0.303 \\ 0.003 & 0.336 & 0.012 & 0.076 & 0.246 \\ 0.027 & 0.012 & 0.216 & 0.027 & 0.149 \\ 0.109 & 0.076 & 0.027 & 0.261 & 0.048 \\ 0.303 & 0.246 & 0.149 & 0.048 & 0.745 \end{bmatrix}.$$

Zauważmy, że w rozważanym przypadku $\pi_{k,l} = p_{SS}(k,l)$ dla $k = 1, \dots, 5$, $l = 1, \dots, 5$ i $k \neq l$. Ponadto:

$$\Sigma = \Delta = \begin{bmatrix} 0.247 & -0.146 & -0.069 & -0.006 & -0.026 \\ -0.146 & 0.223 & -0.061 & -0.012 & -0.004 \\ -0.069 & -0.061 & 0.169 & -0.029 & -0.012 \\ -0.006 & -0.012 & -0.029 & 0.193 & -0.146 \\ -0.026 & -0.004 & -0.012 & -0.146 & 0.190 \end{bmatrix}.$$

Załóżmy, co nie umniejsza ogólności dalszych rozważań, że wartości cechy dodatkowej są uporządkowane od najmniejszej do największej, czyli $x_i \leq x_{i+1}$, dla $i = 1, \dots, N-1$. Niech $X_{(r)}$ będzie r -tą statystyką pozycyjną z próby prostej losowanej bezzwrotnie. Przypomnijmy, że kwantyl z próby rzędu $\alpha \in (0, 1)$ jest określany jako $q_{s,\alpha} = X_{(r)}$, gdzie $r = [n\alpha] + 1$ i $[na]$ jest częścią całkowitą liczby $n\alpha$. Zatem, $X_{(r)} = q_{s,\alpha}$, dla $(r-1)/n \leq \alpha < r/n$. Niech $\mathbf{Q}_n(r, i) = \{s: X_{(r)} = x_i\}$ będzie zbiorem wszystkich prób s , w których r -ta statystyka pozycyjna zmiennej dodatkowej jest równa x_i , gdzie $r \leq i \leq N-n+r$. Zbiory prób tworzące ciąg $(\mathbf{Q}_n(r, i), i = r, \dots, N-n+r)$ są rozłączne oraz wyczerpują przestrzeń prób o ustalonej liczebności n losowanych bezzwrotnie, oznaczanej przez $\mathbf{Q}_n$, a więc $\mathbf{Q}_n = \bigcup_{i=r}^{N-n+r} \mathbf{Q}_n(r, i)$. Wywiał (2004, 2008) zaproponował następujący plan losowania próby.

Definicja 1.19. *Plan bezzwrotnego losowania n -elementowej próby s proporcjonalny do wartości x_i statystyki pozycyjnej $X_{(r)}$ ma postać:*

$$P_q(s|r) = \frac{x_i}{a(n, r)}, \quad \text{dla } s \in \mathbf{Q}_n(r, i), \quad i = r, \dots, N-n+r,$$

gdzie:

$$a(n, r) = \sum_{j=r}^{N-n+r} \binom{j-1}{r-1} \binom{N-j}{n-r} x_j.$$

W szczególności, dla pierwszej i ostatniej ze statystyk pozycyjnych plan losowania upraszcza się do następujących wyrażeń:

$$P_q(s|1) = \frac{x_i}{\sum_{j=1}^{N-n+1} \binom{N-j}{n-1} x_j}, \quad \text{dla } s \in \mathbf{Q}_n(1, i), \quad i = 1, \dots, N-n+1, \quad (1.10)$$

$$P_q(s|n) = \frac{x_i}{\sum_{j=n}^N \binom{j-1}{n-1} x_j}, \quad \text{dla } s \in \mathbf{Q}_n(n, i), \quad i = n, \dots, N, \quad (1.11)$$

Dodajmy jeszcze, iż w szczególności, gdy $x_i = c > 0$ dla wszystkich $i = 1, \dots, N$, to plan P_q losowania próby redukuje się do planu bezzwrotnego wyboru próby prostej określonego definicją 1.13. Wyprowadzenie prawdopodobieństw inkluzji rzędu pierwszego i drugiego znajduje się w artykule

Wywiał (2008). Symulacyjną analizę dokładności strategii estymacji wykorzystującej rozważany plan losowania przeprowadza także Wywiał (2007).

Niech $U(1, \dots, i)$ będzie podpopulacją populacji U i niech s_1 będzie próbą prostą o liczebności $(r - 1)$, wylosowaną bez zwracania z podpopulacji $U(1, \dots, i)$. Podobnie, niech s_2 będzie próbą prostą o liczebności $(n - r)$ losowaną bezzwrotnie z podpopulacji $U(i + 1, \dots, N)$ populacji U . Utrzymujemy założenie, że elementy populacji są ponumerowane zgodnie z rosnącymi wartościami dodatniej cechy pomocniczej. Wówczas algorytm losowania próby zgodnie z planem P_q polega najpierw na wylosowaniu i -tego elementu populacji (któremu jest przypisana wartość x_i statystyki pozycyjnej $X_{(r)}$ na podstawie następującego rozkładu prawdopodobieństwa statystyki $X_{(r)}$:

$$p_{r,i} = \frac{\binom{i-1}{r-1} \binom{N-i}{n-r} x_i}{\sum_{j=r}^{N-n+r} \binom{j-1}{r-1} \binom{N-j}{n-r} x_j}, \quad i = r, \dots, N - n + r. \quad (1.12)$$

Następnie z podpopulacji $U(1, \dots, i)$ jest losowana bezzwrotnie próba prosta s_1 o liczebności $(r - 1)$. Potem z podpopulacji $U(i + 1, \dots, N)$ jest także losowana bezzwrotnie próba prosta s_2 o liczebności $(n - r)$. Suma tych prób z dołączonym i -tym elementem populacji $\{i\}$, czyli $s = s_1 \cup \{i\} \cup s_2$ kończy algorytm losowania próby.

Często w praktyce określa się prawdopodobieństwa inkluzji rzędu pierwszego proporcjonalnie do wartości zmiennej dodatkowej. Przykładowo zamierza się losować gospodarstwa rolne do próby z prawdopodobieństwami proporcjonalnymi do powierzchni ich użytków rolnych. Wówczas większe gospodarstwa (zwykle ważniejsze z ekonomicznego punktu widzenia) mają większą szansę dostania się do próby. Zatem należy skonstruować schemat losowania próby realizujący zadane prawdopodobieństwa inkluzji pierwszego rzędu, przy zaniedbaniu postaci funkcji planu losowania próby. Do bardziej popularnych tego rodzaju schematów losowania należy procedura zaproponowana przez Rao, Hartleya i Cochran (1962). Niech $p_i = \frac{x_i}{\sum_{j=1}^N x_j}$, $i = 1, \dots, N$, gdzie x_i jest wartością zmiennej dodatkowej przypisanej i -temu elementowi populacji U . Zakładamy, że $\sum_{i=1}^N p_i = 1$. Niech próba o liczebności n będzie bezzwrotnie losowana z populacji zgodnie z prawdopodobieństwami inkluzji $\pi_i = np_i$. To zadanie realizuje następujący algorytm. Populacja U jest dzielona losowo na n grup oznaczanych przez U_k , $k = 1, \dots, n$, przy czym każda składa się z co najmniej dwóch elementów. W szczególności Bracha (1996) rozważa podział na b grup $a + 1$ elementów oraz $n - b$ grup a elementowych, gdzie liczebność a jest częścią całkowitą ilorazu N/n . Dla każdej grupy wylicza się sumy $q_k = \sum_{i \in U_k} p_i$. W końcu z każdej grupy losuje się po jednym elemencie z prawdopodobieństwem $\frac{p_i}{q_k}$, gdzie $i \in U_k$. Tak wylosowane elementy tworzą próbę s , przy czym trzeba

zaznaczyć, że tylko w przybliżeniu prawdopodobieństwa inkluzji wchodzących w jej skład elementów są proporcjonalne do odpowiednich wartości zmiennych dodatkowych, czyli $\pi_i \approx np_i$, $i \in s$. Własności tego schematu losowania ostatnio analizowali Chaudhuri, Dihidar i Bose (2006). Po to, aby zapewnić równości $\pi_i = np_i$, $i \in s$, Hartley i Rao (1962) zaproponowali inny schemat losowania próby.

Przykład 1.19. Korzystamy z danych z przykładu 1.17 o obserwacjach wartości zmiennej pomocniczej w pięcioelementowej populacji. Na podstawie definicji 1.19 i wzoru (1.10) otrzymujemy następujące prawdopodobieństwa planu losowania dwuelementowej próby proporcjonalnego do pierwszej statystyki pozycyjnej cechy dodatkowej: $p_q(1, 2|1) = 0.04$, $p_q(1, 3|1) = 0.04$, $p_q(1, 4|1) = 0.04$, $p_q(1, 5|1) = 0.04$, $p_q(2, 3|1) = 0.08$, $p_q(2, 4|1) = 0.08$, $p_q(2, 5|1) = 0.08$, $p_q(3, 4|1) = 0.16$, $p_q(3, 5|1) = 0.16$, $p_q(4, 5|1) = 0.28$.

Wzór (1.11) daje plan losowania próby proporcjonalny do ostatniej (czyli drugiej w rozważanej próbie dwuelementowej) statystyki pozycyjnej: $p_q(1, 2|2) = 2/75$, $p_q(1, 3|2) = 4/75$, $p_q(2, 3|2) = 4/75$, $p_q(1, 4|2) = 7/75$, $p_q(2, 4|2) = 7/75$, $p_q(3, 4|2) = 7/75$, $p_q(1, 5|2) = 11/75$, $p_q(2, 5|2) = 11/75$, $p_q(3, 5|2) = 11/75$, $p_q(4, 5|2) = 11/75$.

1.3. Dane i statystyki

Przypomnijmy, że $\mathbf{y} = [y_1 \dots y_k \dots y_N]$ i $S = [S_1 \dots S_k \dots S_N]$ określają odpowiednio wektor możliwych wartości zmiennej y oraz obserwację próby losowej. Oznaczamy symbolem $*$ tzw. iloczyn Hamadarda tych wektorów, który ma postać: $\mathbf{z} = \mathbf{y} * \mathbf{s} = [y_1 s_1 \dots y_k s_k \dots y_N s_N]$, czyli $z_k = y_k s_k$, $k = 1, \dots, N$. W wyniku obserwacji elementów populacji o numerach należących do próby s otrzymujemy dane o wartościach badanej cechy. Modyfikując definicje danych Cassela, Särndala i Wretmana (1977) oraz Tillé (2006) określamy je następująco:

Definicja 1.20. Danymi o wartości zmiennej y obserwowanymi w próbie losowej S nazywamy uporządkowaną parę wektorowych zmiennych losowych $D = (\mathbf{Z}, S)$, gdzie $\mathbf{Z} = \mathbf{y} * r(S)$, przy czym $r(\cdot)$ jest określoną wzorem (1.1) funkcją redukującą dowolną próbę.

Możliwa wartość (obserwacja) danej jest oznaczana przez: $d = (\mathbf{z}, s) = (\mathbf{y} * r(s), s)$.

Zbiór (przestrzeń) wszystkich możliwych wartości danych o zmiennej y ma postać:

$$\mathbf{D} = \{d | s \in \mathbf{Q}, \mathbf{y} \in R^N\},$$

przy czym R jest zbiorem liczb rzeczywistych, natomiast R^N jest N -krotnym iloczynem kartezjańskim tego zbioru, czyli $R^N = R \times R \times \dots \times R$.

Po to, aby zdefiniować rozkład prawdopodobieństwa danych, trzeba koniecznie najpierw wprowadzić następujące określenie.

Definicja 1.21. *Mówimy, że możliwa wartość danych d jest zgodna z wektorem pewnych szczególnych wartości zmiennej y oznaczanym przez $\mathbf{y}^* = [y_1^* \dots y_k^* \dots y_N^*]$ wtedy i tylko wtedy, gdy $y_k = y_k^*$, dla takich indeksów k , że $s_k > 0$.*

Rozkład prawdopodobieństwa danych ma postać:

$$P(D = d; \mathbf{y}) = \begin{cases} p(s), & \text{jeśli } d \text{ jest zgodny z } \mathbf{y}, \\ 0 & \text{w przeciwnym przypadku.} \end{cases} \quad (1.13)$$

Przykład 1.20. *Rozważmy trójelementową populację. Jej elementom są przypisane obserwacje zmiennej y , które są wartościami wektora: $\mathbf{y} = [0 \ 5 \ 8]$. Zbiór (przestrzeń) danych dla prób losowanych bezzwrotnie ma postać: $\mathbf{D} = \{d_1 = ([0 \ 0 \ 0], [0 \ 0 \ 0]), d_2 = ([0 \ 0 \ 0], [1 \ 0 \ 0]), d_3 = ([0 \ 5 \ 0], [0 \ 1 \ 0]), d_4 = ([0 \ 0 \ 8], [0 \ 0 \ 1]), d_5 = ([0 \ 5 \ 0], [1 \ 1 \ 0]), d_6 = ([0 \ 0 \ 8], [1 \ 0 \ 1]), d_7 = ([0 \ 5 \ 8], [0 \ 1 \ 1]), d_8 = ([0 \ 5 \ 8], [1 \ 1 \ 1])\}$. Dane d_1 i d_8 są specyficznymi, ponieważ pierwsza z nich reprezentuje brak danych o zmiennej y w populacji, czyli ta próba jest pusta, a drugim przypadku mamy pełne (wyczerpujące) dane o zmiennej w populacji.*

Założmy, że plan losowania prób Bernoulliego z prawdopodobieństwem sukcesu π (por. definicję 1.15) jest określony na rozważanym planie losowania prób. Wówczas wzór (1.13) prowadzi do następującego rozkładu prawdopodobieństwa danych: $P(D = d_1) = (1 - \pi)^3$, $P(D = d_2) = P(D = d_3) = P(D = d_4) = \pi(1 - \pi)^2$, $P(D = d_5) = P(D = d_6) = P(D = d_7) = \pi^2(1 - \pi)$, $P(D = d_8) = \pi^3$. W przypadku, gdy losowana bezzwrotnie próba ma ustaloną dwuelementową liczebność, to przestrzeń danych redukuje się do zbioru $\mathbf{D}_2 = \{d_1 = ([0 \ 5 \ 0], [1 \ 1 \ 0]), d_2 = ([0 \ 0 \ 8], [1 \ 0 \ 1]), d_3 = ([0 \ 5 \ 8], [0 \ 1 \ 1])\}$.

Prawdopodobieństwo tego, iż uzyskamy dane o wartości cechy jest równe prawdopodobieństwu tego, że nie wystąpi brak danych i wynosi: $P(D \neq d_1) = 1 - P(D = d_1) = 1 - (1 - \pi)^3 = \pi(3 - 3\pi + \pi^2) = a$. Wówczas warunkowe prawdopodobieństwa wystąpienia pozostałych danych wynoszą: $P(D = d_2 | D \neq d_1) = P(D = d_3 | D \neq d_1) = P(D = d_4 | D \neq d_1) = \pi(1 - \pi)^2/a$, $P(D = d_5 | D \neq d_1) = P(D = d_6 | D \neq d_1) = P(D = d_7 | D \neq d_1) = \pi^2(1 - \pi)/a$, $P(D = d_8 | D \neq d_1) = \pi^3/a$.

Przykład 1.21. *Utrzymując założenia wprowadzone w przykładzie 1.20 losujemy tym razem bezzwrotnie próbę prostą dwuelementową. Wtedy przestrzeń prób ma postać: $\mathbf{Q}_2 = \{[1 \ 1 \ 0], [1 \ 0 \ 1], [0 \ 1 \ 1]\}$. Zbiór wszystkich możliwych danych jest następujący: $\mathbf{D} = \{d_1 = ([0 \ 5 \ 0], [1 \ 1 \ 0]), d_2 = ([0 \ 0 \ 8], [1 \ 0 \ 1]), d_3 = ([0 \ 5 \ 8], [0 \ 1 \ 1])\}$. Z definicji 1.13 wynika*

plan losowania; $p_1(s) = \frac{1}{3}$ dla $s \in S$, a rozkład prawdopodobieństwa danych ma postać: $P(D = d) = \frac{1}{3}$ dla $d \in \mathbf{D}$.

Definicja 1.22. Każda funkcja rzeczywista g_D danych D jest nazywana statystyką, czyli $g_D = g(\mathbf{y} * r(S), S)$. Wartość statystyki oznaczamy przez g_d , czyli $g_d = g(\mathbf{y} * r(s), s)$.

Stąd wynika, że statystyka g_D zależy tylko od tych elementów wektora $\mathbf{y}$, które są obserwowane w próbie losowej S . W związku z tym można używać nieco wygodniejszego oznaczenia statystyki określanego symbolem g_S , z czym można się spotkać w wielu pracach poświęconych metodzie reprezentacyjnej. Wówczas wartość statystyki g_S jest oznaczana przez g_s .

Wartość oczekiwaną i wariancję statystyki g_s pozwalają wyznaczać następujące wzory:

$$E(g_S) = \sum_{s \in \mathbf{Q}} g_s p(s),$$

$$D^2(g_S) = E(g_s - E(g_S))^2 = \sum_{s \in \mathbf{Q}} (g_s - E(g_S))^2 p(s)$$

Przykład 1.22. Kontynuując rozważania prowadzone w przykładzie 1.20 wyznaczamy rozkład prawdopodobieństwa statystyki $M = \min\{D\}$. Możliwymi wartościami tej statystyki są liczby: 0, 5 i 8. Zaobserwowanie minimalnej wartości danych jest możliwe tylko wtedy, gdy uzyskamy jakiegokolwiek dane, czyli gdy próba nie jest pusta. Wtedy:

$$\begin{aligned} P(M = 0) &= P(D = d_2 | D \neq d_1) + \\ &\quad + P(D = d_5 | D \neq d_1) + P(D = d_6 | D \neq d_1) + P(D = d_8 | D \neq d_1), \\ P(M = 0) &= (\pi(1 - \pi)^2 + 2\pi^2(1 - \pi) + \pi^3)/a = 1/(3 - 3\pi + \pi^2), \\ P(M = 5) &= P(D = d_3 | D \neq d_1) + P(D = d_7 | D \neq d_1), \\ P(M = 5) &= (\pi(1 - \pi)^2 + \pi^2(1 - \pi))/a = (1 - \pi)/(3 - 3\pi + \pi^2), \\ P(M = 8) &= P(D = d_4 | D \neq d_1) = (1 - \pi)^2/(3 - 3\pi + \pi^2). \end{aligned}$$

Reasumując mamy:

$$P(M = m) = \begin{cases} \frac{1}{3 - 3\pi + \pi^2}, & \text{dla } m = 0, \\ \frac{1 - \pi}{3 - 3\pi + \pi^2}, & \text{dla } m = 5, \\ \frac{(1 - \pi)^2}{3 - 3\pi + \pi^2}, & \text{dla } m = 8. \end{cases}$$

W szczególności, gdy $\pi = 0.5$ mamy:

$$P(M = m) = \begin{cases} \frac{4}{7} & \text{dla } m = 0, \\ \frac{2}{7} & \text{dla } m = 5, \\ \frac{1}{7} & \text{dla } m = 8. \end{cases}$$

Wartość oczekiwana i wariancja wynoszą: $E(M) = 2\frac{4}{7}$, $D^2(M) = 9\frac{5}{7}$.

Przykład 1.23. Na podstawie danych z przykładu 1.21 wyznaczamy rozkład minimalnej wartości zmiennej z próby, czyli statystyki $M = \min\{D\}$. Wówczas: $P(M = 0) = P(D = d_1) + P(D = d_2) = \frac{2}{3}$, $P(M = 5) = P(D = d_3) = \frac{1}{3}$. W końcu otrzymujemy:

$$P(M = m) = \begin{cases} \frac{2}{3} & \text{dla } m = 0, \\ \frac{1}{3} & \text{dla } m = 5. \end{cases}$$

Ponadto, $E(M) = 1\frac{2}{3}$, $D^2(M) = 5\frac{5}{9}$.

W niniejszym rozdziale przedstawiono wiadomości, które będą wykorzystywane w dalszej treści pracy. Problemy konstrukcji planów i schematów losowania prób są zwykle szeroko omawiane w monografiach poświęconych metodzie reprezentacyjnej. Na szczególną rekomendację zasługuje praca Tillé (2006), która jest jednak poświęcona tylko planom i schematom losowania prób. Ponadto, warto zwrócić uwagę na następujące pozycje: Ardilly i Tillé (2006), Brewer i Hanif (1983), Cassel, Särndall i Wretman (1977), Chaudhuri i Vos (1988), Wywiał (1992, 1995, 2003).

Pytania i zadania

1. Populacja i jej przykłady.
2. Zdefiniuj zmienną oraz podaj jej przykłady.
3. Określ próbę celową i próbę losową oraz przestrzeń prób.
4. Zdefiniuj plan losowania próby oraz jego przykłady.
5. Schemat (algorytm) losowania próby.
6. Podaj przykłady podstawowych parametrów opisowych zmiennych.
7. Określ plany i schematy losowania prób Bernoulliego i Poissona.
8. Co to jest prawdopodobieństwo inkluzji rzędu r -tego?
9. Określ plany i schematy losowania prób prostych losowanych bezzwrotnie i zwrotnie oraz ich podstawowe charakterystyki.

10. Określ plan i schemat losowania próby z prawdopodobieństwem proporcjonalnym do sumy obserwowanych w niej wartości zmiennej dodatkowej.
11. Określ plan i schemat losowania próby Singha i Srivastavy.
12. Zdefiniuj statystykę.
13. Określ plan i schemat losowania próby proporcjonalny do wartości statystyki pozycyjnej zmiennej dodatkowej.
14. Zdefiniuj dane.
15. Podaj przykład nieprostego planu losowania próby oraz realizującego go schematu losowania.
16. Określ przykłady parametrów zmiennych oraz statystyk.
17. W pięcioelementowej populacji określono zmienną y : $y_1 = 1, y_2 = 1, y_3 = 2, y_4 = 3, y_5 = 3$. Wyznacz rozkład średniej zmiennej y z losowanej: a) bezzwrotnie, b) zwrotnie próby prostej trójelementowej.
18. W czteroelementowej populacji określono zmienne y i x : $y_1 = 1, y_2 = 2, y_3 = 3, y_4 = 4$; $x_1 = 1, x_2 = 2, x_3 = 5, x_4 = 9$. Wyznacz rozkład ilorazu średniej zmiennej y przez średnią zmiennej x , przy czym obie średnie są wyznaczane z prostej próby dwuelementowej losowanej a) bezzwrotnie, b) zwrotnie.
19. W pięcioelementowej populacji określono zmienną dodatkową x : $x_1 = 1, x_2 = 2, x_3 = 3, x_4 = 4, x_5 = 5$. Wyznacz plan bezzwrotnego losowania próby trójelementowej proporcjonalnej do obserwowanej w próbie: a) sumy wartości cechy dodatkowej, b) sumy kwadratów odchyleń cechy dodatkowej od jej średniej w populacji, c) wariancji zmiennej dodatkowej z próby, d) pierwszej statystyki pozycyjnej, e) ostatniej statystyki pozycyjnej, f) mediany z próby. Wyznacz prawdopodobieństwa inkluzji rzędu pierwszego.
20. Co to jest operat losowania próby i jakie winien spełniać postulaty?
21. Wyznacz plan losowania Bernoulliego z populacji trójelementowej. Wyznacz rozkład prawdopodobieństwa z próby Bernoulliego a) średniej, b) wartości maksymalnej, c) rozstępu. Wyznacz wartości oczekiwane i wariancje tych statystyk. Rozważ szczególne przypadki, gdy $y_1 = 1, y_2 = 3, y_3 = 8$ oraz $\pi = 0.1, \pi = 0.5, \pi = 0.9$.
22. W czteroelementowej populacji określono zmienną badaną y : $y_1 = 1, y_2 = 1, y_3 = 2, y_4 = 2$ i dodatkową x : $x_1 = 1, x_2 = 2, x_3 = 3, x_4 = 4$. Wyznacz plan Poissona bezzwrotnego losowania prób, przy czym elementy populacji są losowane do próby z prawdopodobieństwami proporcjonalnymi wartości cechy dodatkowej. Wyznacz rozkład: a) średniej, b) rozstępu z próby i ich parametry.
23. W czteroelementowej populacji określono zmienną badaną y : $y_1 = 1, y_2 = 2, y_3 = 2, y_4 = 9$ i dodatkową x : $x_1 = 1, x_2 = 1, x_3 = 3, x_4 = 15$. Wyznacz plan bezzwrotnego losowania prób trójelementowych: a) prostej, b) proporcjonalnej do obserwowanej w próbie ostatniej statystyki

ki pozycyjnej. Wyznacz prawdopodobieństwa inkluzji rzędu pierwszego. Wyznacz rozkłady średnich z tak losowanych prób oraz porównaj ich wariancje.

24. Określ plany i schematy zwrotnego i bezzwrotnego losowania prób prostych oraz ich charakterystyki.

ROZDZIAŁ 2

Estymacja

2.1. Definicje podstawowe

2.1.1. Podejście randomizacyjne

Do głównych celów badania statystycznego populacji należy ocena *parametrów opisowych* cech (zmiennych). Parametr cechy populacji, oznaczany przez θ , jest funkcją parametru populacji $\mathbf{y} \in R^N$, czyli $\theta = \theta(\mathbf{y})$, która przyjmuje wartości ze zbioru Θ . Innymi słowy, $\theta: R^N \rightarrow \Theta$, $\theta = \theta(\mathbf{y})$. Zatem rozróżniamy tutaj znaczenie parametru populacji, oznaczanego wektorem $\mathbf{y}$, którego wartości są obserwacjami pewnej cechy, od pojęcia parametru (opisowego) tej cechy, która jest pewną rzeczywistą funkcją wektora $\mathbf{y}$. Parametr opisowy w sposób syntetyczny charakteryzuje własności struktury obserwacji cechy w populacji. Zwykle Θ jest przedziałem liczbowym, czyli $\Theta = [\theta_1, \theta_2]$. Używaną do oceny θ statystykę oznaczamy przez T_s . Najczęściej są brane pod uwagę parametry, które określono w pierwszym paragrafie pierwszego rozdziału, a należą do nich m.in. średnia arytmetyczna zmiennej, jej wartość globalna, liczba elementów z cechą wyróżnioną i wskaźnik struktury.

Definicja 2.1. *Estymatorem parametru $\theta \in \Theta$ jest taka statystyka t_s , której wartości należą do zbioru Θ . Wartość t_s estymatora t_s jest oceną wartości parametru θ .*

Wiadomo, że wariancja w populacji może przyjmować tylko wartości nieujemne. W związku z tym np. statystyka, która może przyjmować tylko wartości ujemne, nie może być estymatorem wariancji zmiennej w populacji.

Dodajmy, że w literaturze statystycznej są rozważane inne definicje estymatora. Przykładowo przyjmuje się, że określona w definicji 2.1 statystyka t_s , jest estymatorem parametru θ jeśli znaczna część jej wartości leży nie dalej od ocenianej wartości θ niż o z góry postulowaną odległość Δ , por. np. Hellwig (1995) lub Särndal Swensson i Wretman (1992, s. 38).

Błąd estymacji jest różnicą między wartością szacowanego parametru i jego estymatorem i oznaczamy go symbolem:

$$u(t_s) = t_s - \theta$$

Definicja 2.2. Statystyka t_S jest nieobciążonym estymatorem parametru θ , jeśli dla każdego $\theta \in \Theta$

$$E(t_S) = \sum_{s \in \mathbf{Q}} t_s P(S = s) = \theta,$$

czyli gdy $E(u(t_S)) = 0$.

Gdy $E(u(t_S)) = b(t_S) \neq 0$, to parametr $b(t_S) = E(t_S) - \theta$ jest nazywany obciążeniem estymatora t_S reprezentującym przeciętny błąd systematyczny estymacji parametru θ .

Definicja 2.3. Statystyka t_S jest granicznie (asymptotycznie) nieobciążonym estymatorem parametru θ , jeśli dla każdego $\theta \in \Theta$ w przypadku losowania próby:

a) zwrotnego

$$\lim_{n \rightarrow \infty} E(t_S) = \theta,$$

b) bezzwrotnego $E(t_S) = \theta$, jeżeli $n = N$.

Względny wskaźnik obciążenia definiujemy wzorem:

$$wb(t_S) = \frac{E(t_S) - \theta}{|\theta|} 100\%. \quad (2.1)$$

Określa on, jaki procent modułu wartości szacowanego parametru stanowi obciążenie jego estymatora t_S .

Przykład 2.1. Załóżmy, że dane o zmiennej są obserwowane w n -elementowej próbie prostej losowanej zwrotnie. Na podstawie tych danych jest szacowana wariancja v określona wzorem (1.5). Do tego celu można użyć estymatora nieobciążonego v_{*S} określonego wyrażeniem (1.6) lub estymatora $v_S = \frac{n-1}{n} v_{*S}$, który jest obciążony, ponieważ $E(v_S) = \frac{n-1}{n} v < v$. Precyzyjniej rzecz biorąc estymator v_S prowadzi do systematycznego błędu oceny wariancji v , bo średnio rzecz biorąc nie doszacowuje wartości parametru v . Dodajmy jednak, że gdy liczebność próby rośnie w nieograniczony sposób, to $E(v_S) = v$, czyli statystyka v_S , jest granicznie nieobciążonym estymatorem wariancji v .

Błąd średniokwadratowy estymacji parametru θ jest wartością oczekiwaną kwadratu błędu estymacji, czyli:

$$MSE(t_S) = E(t_S - \theta)^2 \quad (2.2)$$

Błąd średniokwadratowy estymacji dekomponuje się następująco:

$$MSE(t_S) = D^2(t_S) + b^2(t_S), \quad (2.3)$$

gdzie $D^2(t_S)$ jest wariancją estymatora. Pierwiastek z błędu średniokwadratowego $MSE(t_S)$, nazywamy błędem średnim szacunku parametru θ , który określa przeciętny błąd oceny parametru θ za pomocą estymatora t_S . Pierwiastek z wariancji estymatora jest odchyleniem standardowym $D(t_S)$ estymatora określającym przeciętne odchylenie wartości estymatora od jego wartości oczekiwanej. Parametr $D(t_S)$ reprezentuje przeciętny błąd losowy estymacji, którego źródłem jest losowy wybór próby. Z kolei obciążenie estymatora reprezentuje systematyczny błąd estymacji.

Udział obciążenia estymatora w jego błędzie średnim szacunku określamy następująco:

$$ubs(t_S) = 100\% \frac{|b(t_S)|}{\sqrt{MSE(t_S)}}. \quad (2.4)$$

Gdy estymator jest nieobciążony, to błąd średniokwadratowy estymatora jest równy jego wariancji, czyli jeśli $E(t_S) = \theta$, to $MSE(t_S) = D^2(t_S)$.

Czasem użytecznym jest tzw. względny średni błąd szacunku określany przez

$$z(t_S) = \frac{\sqrt{MSE(t_S)}}{|\theta|} 100\%, \quad (2.5)$$

który wskazuje, jaki procent wartości modułu szacowanego parametru stanowi błąd średni jego szacunku.

Definicja 2.4. *Mówimy, że statystyka t_S jest zgodnym estymatorem parametru θ , jeśli w przypadku:*

- a) *zwrotnego losowania próby o liczebności n oraz każdego $\epsilon > 0$ jest prawdziwe wyrażenie:*

$$\lim_{n \rightarrow \infty} P(|t_S - \theta| \leq \epsilon) = 1,$$

- b) *bezzwrotnego losowania próby statystyka $t_S = \theta$, gdy $n = N$.*

Dodajmy, że warunek określony w przypadku b) definicji jest równoważny temu, iż $P(t_S - \theta = 0) = 1$, jeżeli $n = N$.

Z definicji granicy wynika, że dla każdego $\epsilon > 0$ i każdego $\delta > 0$ istnieje taka liczebność próby n_0 , że dla $n \geq n_0$ zachodzi $|P(|t_S - \theta| \leq \epsilon) - 1| \leq \delta$. W szczególności przy bezzwrotnym losowaniu próby $n_0 \leq N$, gdzie N jest liczebnością populacji.

W praktyce zgodność estymatora można więc rozumieć w tym sensie, że dla każdego postulowanego poziomu dokładności estymacji $\epsilon > 0$ istnieje taka liczebność próby n_0 , przy której prawdopodobieństwo nieprzekroczenia tego błędu jest dowolnie bliskie jedności.

Definicja 2.5. *Estymator t_{1S} parametru $\theta \in \Theta$ jest nie gorszy od estymatora t_{2S} tego parametru, gdy dla każdego $\theta \in \Theta$ zachodzi nierówność:*

$$MSE(t_{1S}) \leq MSE(t_{2S}).$$

Gdy dodatkowo istnieje taka wartość $\theta \in \Theta$, że powyższa nierówność staje się ostrą, to mówimy, iż estymator t_{1S} jest lepszy od estymatora t_{2S} .

Zwykle w pewnej klasie (zbiorze) estymatorów, np. nieobciążonych, poszukuje się estymatora, którego wariancja jest minimalna. Taki estymator jest nazywany efektywnym w takiej klasie. Okazuje się, że oprócz szczególnych przypadków taki estymator wartości średniej cechy w populacji skończonej i ustalonej nie istnieje. Wykazali to Godambe i Joshi (1965), porównaj również Basu (1971) oraz Chaudhuri i Stenger (2005).

Gdy nie istnieje w danej klasie estymator efektywny, to poszukuje się takiego, który jest przynajmniej dopuszczalny.

Definicja 2.6. Estymator t_{*S} parametru $\theta \in \Theta$ jest dopuszczalny w klasie $\kappa = \{t_S\}$ estymatorów tego parametru, gdy w κ nie ma estymatora lepszego od t_{*S} .

Przykład 2.2. Rozważmy czteroelementową populację. Jej elementom są przypisane obserwacje zmiennej y , które są wartościami wektora: $\mathbf{y} = [2 \ 2 \ 4, \ 5]$. Wartość średnia zmiennej y w populacji wynosi $\bar{y} = 3.25$, a wariancja $v_* = 2.25$. Bierzemy pod uwagę plan bezwrotnego losowania trójelementowej próby prostej. Zbiór (przestrzeń) danych z takich prób ma postać: $\mathbf{D} = \{d_1 = ([2 \ 2 \ 4 \ 0], [1 \ 1 \ 1 \ 0]), d_2 = ([2 \ 2 \ 0 \ 5], [1 \ 1 \ 0 \ 1]), d_3 = ([2 \ 0 \ 4 \ 5], [1 \ 0 \ 1 \ 1]), d_4 = ([0 \ 2 \ 4 \ 5], [0 \ 1 \ 1 \ 1])\}$. Wtedy rozkład prawdopodobieństwa danych ma postać: $P(D = d_i) = 0.25$ dla $i = 1, 2, 3, 4$. To już prowadzi do rozkładu prawdopodobieństwa średniej z próby: $P(\bar{y}_S = 2\frac{2}{3}) = P(D = d_1) = 0.25$, $P(\bar{y}_S = 3) = P(D = d_2) = 0.25$, $P(\bar{y}_S = 3\frac{2}{3}) = P(D = d_3) + P(D = d_4) = 0.5$. Podobnie wyznaczamy rozkład prawdopodobieństwa mediany z próby cech y , czyli jej wartości środkowej: $P(M_S = 2) = P(D = d_1) + P(D = d_2) = 0.5$, $P(M_S = 4) = P(D = d_3) + P(D = d_4) = 0.5$. W syntetyczny sposób te rozkłady zapisujemy następująco:

$$P(\bar{y}_S = y) = \begin{cases} 0.25 & \text{dla } y = 2\frac{2}{3}, \\ 0.25 & \text{dla } y = 3, \\ 0.50 & \text{dla } y = 3\frac{2}{3}, \end{cases} \quad P(M_S = y) = \begin{cases} 0.5 & \text{dla } y = 2, \\ 0.5 & \text{dla } y = 4. \end{cases}$$

Parametry rozkładu tych zmiennych są następujące: $E(\bar{y}_S) = 3.25 = \bar{y}$, $D^2(\bar{y}_S) = 0.1875 = \text{MSE}(\bar{y}_S)$, $E(M_S) = 3.0$, $D^2(M_S) = 1.0$. Średnia z próby jest więc nieobciążonym estymatorem wartości przeciętnej w populacji, natomiast mediana z próby jest obciążonym estymatorem przeciętnej w populacji. Jej obciążenie wynosi $b(M_S) = -0.25$, czyli przeciętnie rzecz biorąc mediana niedoszacowuje średniej w populacji. Błąd średniokwadratowy mediany wynosi $\text{MSE}(M_S) = 1.0625$. W związku z tym, średnia z próby jest

lepszym od mediany z próby estymatorem przeciętnej w populacji, ponieważ $MSE(\bar{y}_S) = 0.1875 < 1.0625 = MSE(M_S)$.

Przedział ufności stanowi alternatywny sposób oceny parametru populacji. Nie wchodząc w szczegóły, wyznaczenie takiego przedziału dla parametru $\theta \in \Theta$ polega na określeniu takich dwóch statystyk t_{1S} i t_{2S} , że jest spełnione wyrażenie:

$$P(t_{1S} < \theta < t_{2S}) \geq \gamma, \quad (2.6)$$

gdzie przez γ oznaczono poziom ufności, czyli prawdopodobieństwo, z jakim przedział ufności $[t_{1S}, t_{2S}]$ obejmuje wartość szacowanego parametru θ .

Precyzję estymacji przedziałowej ocenia się za pomocą połowy rozpiętości przedziału ufności, którą określa wzór:

$$\Delta = \frac{1}{2}(t_{2S} - t_{1S}) \quad (2.7)$$

Formalne metody wyznaczania końców przedziałów ufności mogą być złożone, dlatego ograniczmy się tylko do podania pewnej ogólnej zasady tworzenia przedziałów ufności dla prób o dużej liczebności. Niech t_S jest przynajmniej granicznie nieobciążonym estymatorem parametru θ , natomiast $D_S^2(t_S)$ jest zgodnym estymatorem wariancji $D^2(t_S)$. Ponadto założymy, że liczebność próby jest na tyle duża, iż rozkład prawdopodobieństwa estymatora t_S jest dostatecznie dobrze przybliżany rozkładem normalnym. Wtedy przedział ufności określa wyrażenie:

$$P(t_S - t_\gamma D_S(t_S) < \theta < t_S + t_\gamma D_S(t_S)) = \gamma \quad (2.8)$$

przy czym przez t_γ oznaczamy kwantyl rzędu $(1 + \gamma)/2$ rozkładu prawdopodobieństwa statystyki t_S .

Przykład 2.3. W $N = 500$ elementowej populacji studentów znajduje się k -osób, które pracują w trakcie wakacji. Wylosowano w sposób zwrrotny próbę prostą $n = 200$ studentów, spośród których $k_S = 160$ pracowało. Oceniamy punktowo i przedziałowo parametr k . Wiadomo, że statystyka k_S ma znany dwumianowy rozkład prawdopodobieństwa z parametrami: $E(k_S) = np$ i $D^2(k_S) = np(1 - p)$, gdzie $p = \frac{k}{N}$. W związku z tym nieobciążonym estymatorem parametru k , czyli liczby elementów (studentów) populacji z określoną cechą wyróżnioną, jest statystyka $K_S = \frac{N}{n}k_S$, ponieważ $E(K_S) = \frac{N}{n}E(k_S) = k$. Wówczas $D^2(K_S) = (\frac{N}{n})^2 D^2(k_S) = N^2 \frac{p(1-p)}{n} = \frac{k(N-k)}{n}$. Ta wariancja jest szacowana za pomocą statystyki $D_s^2(K_S) = \frac{k_S(N-k_S)}{n}$. W naszym przypadku wyliczamy, że $D_s^2(K_S) = 272$. Zatem możliwe wartości estymatora K_S odchylają się od wartości szacowanego parametru K przeciętnie o $\sqrt{272}$.

Przedział ufności dla parametu k ma postać:

$$P(K_S - t_\gamma D_s(K_S) < k < K_S + t_\gamma D_s(K_S)) = \gamma,$$

przy czym wykazuje się (por. podręczniki statystyki matematycznej), że dla liczebności próby $n \geq 100$ i $p \geq 0.01$ kwantyl t_γ można wyznaczyć z rozkładu normalnego standardowego. W naszym przypadku, gdy poziom ufności określimy liczbą $\gamma = 0.95$, to $t_\gamma = 1.96$. Wtedy realizacja wyżej określonego przedziału ufności ma postać: $[127.645, 192.325]$. Otrzymany przedział z częstością 0.95 obejmuje wartość szacowanej liczby studentów pracujących podczas wakacji. Połowa długości tego przedziału ufności wynosi $\Delta = 32.34$.

Dotąd rozważaliśmy sytuację, gdy porównywano własności co najmniej dwóch estymatorów przy tym samym planie losowania próby. W szczególności dotyczy to definicji 2.5 pozwalającej na porównywanie dokładności pary estymatorów przy tym samym planie losowania próby. Trzeba jednak liczyć się z sytuacją, gdy dany parametr populacji może być szacowany za pomocą dwóch różniących się estymatorów wyznaczanych z prób dobieranych z populacji zgodnie z dwoma planami losowania, które także różnią się od siebie. Zatem w szczególności dany estymator może być użyty do oceny parametru na podstawie danych obserwowanych w próbach dobieranych z populacji przy różnych planach losowania próby. W związku z tym, jeśli wystąpi sytuacja analizowania wielu różnych estymatorów przy również różnych planach losowania próby, uzasadnionym staje się wprowadzenie następującego określenia:

Definicja 2.7. *Strategią estymacji (losowania) dla oceny parametru θ nazywa się parę uporządkowaną $(t_S, p(s))$, gdzie t_S jest estymatorem parametru θ , natomiast $p(s)$ jest planem losowania próby.*

W związku z tym definicje 2.2–2.4 da się bezpośrednio uogólnić na pojęcia strategii odpowiednio nieobciążonej i zgodnej.

Obciążenie, błąd średniokwadratowy i wariancję strategii $(t_S, p(s))$ oznaczamy odpowiednio symbolami: $b(t_S, p(s))$, $MSE(t_S, p(s))$, $D^2(t_S, p(s))$. Te oznaczenia są jednak bardziej złożone od wcześniej używanych i nie ma potrzeby korzystania z nich w sytuacji, gdy są analizowane estymatory przy tym samym planie losowania. Wtedy, co podkreślmy, w prowadzonych rozważaniach są używane prostsze oznaczenia wprowadzone na początku niniejszego rozdziału, czyli $b(t_S)$, $MSE(t_S)$, $D^2(t_S)$.

W niniejszym punkcie ograniczyliśmy się do podania tych własności estymatorów lub strategii estymacji, które są ważne przy praktycznym ocenianiu podstawowych parametrów cech w populacji. Czytelnik może znacznie wzbogacić swoje wiadomości zwłaszcza w zakresie teoretycznych rozważań, sięgając m.in. po podręczniki: Brachy (1996), Cassela Särndala i Wretmana

(1977), Särndala, Swensona i Wretmana (1992) lub Chaudhuriego, Stengera (2005). Z kolei problemy jednoczesnej estymacji wielu wartości średnich są podejmowane przez Wywiśla (1995, 2003).

2.1.2. Podejście modelowe

Dotychczas zakładaliśmy, że wartości cech obserwowane w populacji są z góry ustalone (nielosowe). Wówczas źródłem losowości próby był przyjęty plan losowania próby. W szczególności zwykle nie mamy wątpliwości, że powierzchnia gminy czy liczba zarejestrowanych w niej samochodów to liczby ustalone, czyli nielosowe. Rozważmy jednak inną sytuację. Przypuśćmy, że celem wnioskovania jest ocena sumy plonów pszenicy gospodarstw rolnych w pewnej gminie. Wiadomo, że w wyniku obserwacji plonu pszenicy w danym gospodarstwie otrzymamy pewną ustaloną wartość, jednak jednocześnie jesteśmy świadomi, że na poziom tej wartości miały wpływ oprócz czynników głównych, jak np. gatunek sianego zboża, poziom nawożenia, także czynniki losowe, czyli m.in. te związane z warunkami atmosferycznymi, które miały miejsce podczas wegetacji tego zboża, więc w szczególności rozkłady temperatury powietrza czy opadów deszczu. Podobną sytuację mamy np. w przypadku wartości nieruchomości budowlanej, którą dla ustalenia uwagi niech będzie dom mieszkalny jednorodzinny. Jego wartość nie tylko zależy od powierzchni domu czy jego kubatury, lecz również zużytych materiałów do jego budowy, których dopuszczalne zestawy mogą się znacznie różnić. Ponadto, wartość tych materiałów jest zależna od cen, które również są wynikiem działania czynników losowych. Te przykłady skłaniają do wniosku, że w wielu przypadkach praktycznych wartości obserwowanych cech należy traktować jako wartości pewnych zmiennych losowych.

Przykład 2.4. *W pewnej urnie jest B kul białych i C kul czarnych. Zatem prawdopodobieństwo wylosowania z niej kuli białej wynosi $p = B/M$, $M = B + C$. Losujemy z tej urny w sposób zwrrotny N kul. Tą drogą otrzymujemy N -elementowy zbiór U obserwacji koloru kul. Możemy zaobserwować kulę białą albo czarną. Z racji tego, że losowanie było zwrrotne, obserwacje koloru kul są niezależne od siebie oraz prawdopodobieństwo tego, że otrzymamy obserwację kuli białej jest stałe i równe p . W wyniku obserwacji każdej wylosowanej kuli możemy stwierdzić, że wylosowano kulę białą albo czarną. Określmy następującą zmienną losową odzwierciedlającą wynik i -tego losowania. Niech $Y_i = 1$, gdy zaobserwowano kulę białą w wyniku i -tego losowania oraz $Y_i = 0$, jeśli rezultatem losowania była kula czarna. Jest więc to tzw. zmienna losowa zero-jedynkowa, a jej rozkład prawdopodobieństwa ma postać: $P(Y_i = 1) = p$ i $P(Y_i = 0) = 1 - p$. Wyniki obserwacji koloru kul otrzymywanych w kolejnych losowaniach można więc traktować*

jako wartości niezależnych zmiennych losowych będących elementami wektora $\mathbf{Y}_U = [Y_1 \ Y_2 \dots Y_i \dots Y_N]$, z których każda ma rozkład zero-jedynkowy. Łączny rozkład prawdopodobieństwa tych zmiennych zero-jedynkowych jest modelem wyników obserwacji koloru losowanych w sposób zwrrotny kul z urny U .

Przykład 2.5. Rozważmy szczególny przypadek przykładu 2.4, zakładając, że $N = 2$ i $p = 0.4$. Wówczas możliwe obserwacje wektora $\mathbf{Y}_U = [Y_1 \ Y_2]$ są następujące $\{\mathbf{y} = [y_1 \ y_2]: \mathbf{y}_1 = [0, 0]; \mathbf{y}_2 = [0, 1]; \mathbf{y}_3 = [1, 0]; \mathbf{y}_4 = [1, 1]\}$. Prawdopodobieństwo otrzymania każdej z możliwych realizacji jest stałe i równe $1/4$. Przyjmijmy teraz, że w zbiorze $Z_{\mathbf{y}}$ znajduje się po cztery z możliwych obserwacji wektora $\mathbf{Y}_U$, czyli $Z_{\mathbf{y}} = \{\mathbf{y}_1; \mathbf{y}_1; \mathbf{y}_1; \mathbf{y}_1; \mathbf{y}_2; \mathbf{y}_2; \mathbf{y}_2; \mathbf{y}_2; \mathbf{y}_3; \mathbf{y}_3; \mathbf{y}_3; \mathbf{y}_3; \mathbf{y}_4; \mathbf{y}_4; \mathbf{y}_4; \mathbf{y}_4\}$. Zatem wybierając losowo ze zbioru $Z_{\mathbf{y}}$ jeden element otrzymujemy obserwację zmiennej losowej $\mathbf{Y}_U = [Y_1 \ Y_2]$. Łatwo sprawdzamy, że prawdopodobieństwo otrzymania każdej z możliwych jej realizacji jest stałe i równe $1/4$, a każda z jej niezależnych składowych ma ten sam rozkład zero-jedynkowy z prawdopodobieństwem sukcesu $p = 0.4$. Zbiór $Z_{\mathbf{y}}$ jest więc przykładem tzw. modelu superpopulacji. Krotność występowania w nim poszczególnych obserwacji determinuje rozkład prawdopodobieństwa wektora losowego $\mathbf{Y}_U = [Y_1 \ Y_2]$ oraz jego składowych zmiennych losowych.

W ogólności zauważmy, że obserwacje wektora losowego $\mathbf{Y}_U$ są wektorami $\mathbf{y}_U$ należącymi do zbioru R^N , czyli n -krotnego iloczynu kartezjańskiego zbioru liczb rzeczywistych R . Oznaczmy przez $Z_{\mathbf{y}}$ zbiór wszystkich możliwych obserwacji $\mathbf{y}_U$. Częstość występowania poszczególnych obserwacji $\mathbf{y}_U$ w zbiorze $Z_{\mathbf{y}}$ jest wyznaczana przez funkcję prawdopodobieństwa (gęstości) $f(\mathbf{y})$ właśnie zmiennej losowej $\mathbf{Y}_U$. Reasumując przyjmuje się, że model nadpopulacji to właśnie rozkład prawdopodobieństwa wektora $\mathbf{Y}_U$. Ten rozkład zazwyczaj oznacza się przez ξ . W przykładach 2.4 i 2.5 model nadpopulacji jest łącznym rozkładem N niezależnych zmiennych losowych, z których każda ma ten sam rozkład zero-jedynkowy. Dodajmy, że próba s i odpowiadające jej dane tworzące podwektor $[Y_1 \ Y_2 \dots Y_n]$ są podstawą dalszego wnioskowania o własnościach rozkładu prawdopodobieństwa ξ wektora zmiennych losowych $\mathbf{Y}_U$.

Przypuśćmy, że teraz wybieramy bezzwrotnie próbę s o liczebności $n \leq N$ spośród obserwacji tworzących zbiór U . Probę tworzą więc wartości $y_1 \ y_2 \dots y_n$ pewnego n -elementowego podwektora $[Y_1 \ Y_2 \dots Y_n]$ wektora $\mathbf{Y}_U$. Zauważmy, że w rozważanym przypadku rozkłady wszystkich z możliwych n -elementowych podwektorów wektora $\mathbf{Y}_U$ są takie same. W związku tym nie jest rzeczą istotną, czy te podwektory są losowane z wektora $\mathbf{Y}_U$, czy też arbitralnie wskazywane, bo właśnie rozkłady prawdopodobieństwa tych podwektorów są takie same.

Przykład 2.6. W magazynie pewnej fabryki znajduje się M -wyrobów gotowych do sprzedaży. Porządek, w jakim są w nim ułożone wyroby, jest przypadkowy. Zatem populacja składa się z M -wyrobów, wśród których niech będzie $K < M$ wadliwych. Wtedy udział wyrobów wadliwych wynosi $p = K/M$. Przypuśćmy, że pewnemu odbiorcy jest sprzedawana partia $N < M$ wyrobów z tego magazynu, przy czym producent twierdzi, że wadliwość sprzedanej partii wyrobów wynosi p . Czy tak jest? Odbiorca jednak nie dowierza i z kupowanej partii N -wyrobów losuje bezzwrotnie $n < N$ elementową próbę celem sprawdzenia znajdującej się w niej liczby wyrobów wadliwych. Załóżmy jeszcze, że liczba M wyrobów w magazynie jest bardzo duża, a liczebność N wybranej spośród nich partii wyrobów jest mała. Wówczas z rachunku prawdopodobieństwa wiadomo, że wyniki losowania bezzwrotnego (wyrobów do sprzedawanej partii) są prawie niezależne. W związku z tym model generowania modelu populacji opisany w przykładzie 2.4 jest ważny dla rozważanego przypadku. Odbiorca w już wylosowanej próbie obserwuje pewien n -elementowy podwektor $[Y_1 Y_2 \dots Y_n]$ wektora $\mathbf{Y}_U = [Y_1 Y_2 \dots Y_i \dots Y_N]$. Zmienne losowe tworzące podwektor (dane) $[Y_1 Y_2 \dots Y_n]$ są niezależne oraz każda spośród zmiennych losowych będąca elementem tego podwektora ma rozkład prawdopodobieństwa zero-jedynkowy określony wzorami: $P(Y_i = 1) = p$ i $P(Y_i = 0) = 1 - p$, $i = 1, \dots, n$. Zaznaczmy, że do tego samego rozkładu dochodzimy, gdy N -elementowa partia towarów jest losowana zwrotnie z magazynu. Dalej można więc już wnioskować o parametrze za pomocą znanych podstawowych metod statystyki matematycznej. W szczególności wykorzystując znane testy statystyczne można weryfikować hipotezę statystyczną o tym, czy wadliwość partii wyrobów jest rzeczywiście równa p .

Niech celem wnioskowania jest ocena pewnej funkcji T_U wektora $\mathbf{Y}_U$. Zazwyczaj chodzi o ocenę wartości globalnej zmiennej badanej w populacji oznaczanej przez $Y_U = \sum_{i \in U} Y_i$ lub wartości średniej $\bar{Y}_U = N^{-1}Y_U$. W obu tych przypadkach parametry populacji są funkcjami zmiennych losowych, czyli również są zmiennymi losowymi. Dlatego ocena takich parametrów jest nazywana predykcją. Statystyka służąca do oceny wartości parametru T_U jest nazywana *predyktorem* będącym funkcją danych o wartościach badanej zmiennej. Zatem predyktor zależy od pewnego podciągu zmiennych losowych tworzących elementy wektora $\mathbf{Y}_U$, których wartości są już obserwowane w próbie s wylosowanej zgodnie z pewnym planem losowania $P(s)$. W próbie $s = \{i_1, \dots, i_j, \dots, i_n\}$ jest obserwowany wektor danych, który oznaczmy przez $\mathbf{Y}_s = [Y_{i_1} \dots Y_{i_j} \dots Y_{i_n}]$, przy czym przez i_j oznaczono numer elementu populacji wybranego do próby s . Zatem $\mathbf{Y}_s$ jest podwektorem wektora $\mathbf{Y}_U$. Wektor $\mathbf{Y}_s$ można również traktować jako wektor danych obserwowanych w próbie s . Predyktor oznaczany dalej symbolem T_s jest więc funkcją wektora danych $\mathbf{Y}_s$.

Uwzględnienie w procesie predykcji planu losowania próby prowadzi do określenia strategii predykcji dla parametru T_U jako pary $(T_U, P(s))$. Rozkład predyktora (strategii predykcji) zależy więc od rozkładu łącznego planu losowania próby s oraz od rozkładu obserwowanego w niej podwektora wektora $\mathbf{Y}_U$. Wartość oczekiwaną predyktora T_S wyznaczaną względem planu losowania $P(s)$ będziemy oznaczać przez $E_P(T_S)$, natomiast względem modelu nadpopulacji przez $E_\xi(T_S)$. Dalej zakładamy, że mamy do czynienia z próbami nieinformatywnymi, co oznacza, że plan losowania próby nie zależy od rozkładu prawdopodobieństwa wektora $\mathbf{Y}_U$. W związku z tym $E_\xi E_P(T_S) = E_P E_\xi(T_S)$.

Dokładność predykcji ocenia się za pomocą różnicy $Z_S = T_S - T_U$, zwanej błędem predykcji. Analizę błędu systematycznego predykcji prowadzi się na podstawie następujących definicji. Predyktor T_S jest p -nieobciążony, gdy $b_P(T_S) = 0$, gdzie $b_P(T_S) = E_P(Z_S)$. Predyktor T_S jest ξ -nieobciążony, gdy $b_\xi(T_S) = 0$, gdzie $b_\xi(T_S) = E_\xi(Z_S)$. W końcu jest on $p\xi$ -nieobciążony, gdy $b_{P\xi}(T_S) = 0$, gdzie $b_{P\xi}(T_S) = E_P E_\xi(Z_S)$. Błąd średniokwadratowy strategii predykcji $(T_S, P(s))$ definiuje wyrażenie:

$$MSE_{P\xi}(T_S, P(s)) = E_\xi E_P((T_S - T_U)^2, P(s)). \quad (2.9)$$

Przy założeniu nieinformatywności próby błąd średniokwadratowy strategii predykcji dekomponuje się następująco (por. Cassel i in., 1977, s. 94).

$$MSE_{P\xi}(T_S, P(s)) = E_P D_\xi^2(T_S) + E_P b_\xi^2(T_S) + D_\xi^2(T_U) + \\ - 2E_\xi(T_U - E_\xi(T_U))E_P(T_S - E_\xi(T_S)) \quad (2.10)$$

lub

$$MSE_{P\xi}(T_S, P(s)) = E_P D_\xi^2(T_S) + E_P b_\xi^2(T_S) + E_\xi b_P^2(T_S) + \\ - E_\xi(E_P(T_S) - E_\xi(T_U))^2. \quad (2.11)$$

Definicja 2.8. *Strategia $(T_{1S}, P(s))$ predykcji dla T_U jest nie gorsza od strategii $(T_{2S}, P(s))$, gdy zachodzi nierówność:*

$$MSE_{P\xi}(T_{1S}, P(s)) \leq MSE_{P\xi}(T_{2S}, P(s)).$$

Dalej przedstawimy kilka ważnych problemów predykcji wartości globalnej zmiennej w populacji dla podstawowych modeli generujących wartości zmiennych losowych, których obserwacje są wartościami cechy badanej w populacji U składającej się z N -elementów. Sprowadza się to do przedstawienia własności rozkładów prawdopodobieństwa wektora losowego $\mathbf{Y}_U = [Y_1 \ Y_2 \ \dots \ Y_i \ \dots \ Y_N]$. Dodajmy, że problem wnioskowania modelowego szeroko analizują Valliant, Dorfman i Royall (2000).

Przedstawiona procedura wnioskowania pozwala poprawnie oceniać parametry cech populacji, jednak pod jednym warunkiem, że przyjmowany model nadpopulacji jest prawdziwy. Gdy tak nie jest, tzn. jeśli rzeczywisty (empiryczny) łączny rozkład wektora $\mathbf{Y}_U$ różni się istotnie od modelu określonego założonym rozkładem prawdopodobieństwa ξ , to faktyczny błąd średni predykcji może być znacznie większy od oczekiwanego i co więcej, uzyskiwane oceny parametrów opisowych mogą być obciążone. Zatem występuje potrzeba weryfikacji prawdziwości założonego modelu nadpopulacji. Można tego dokonać poprzez odpowiednie formułowanie hipotez statystycznych i ich testowanie za pomocą możliwie silnych testów statystycznych. Ten problem dalej będzie szczegółowiej analizowany podczas rozważania konkretnych modeli nadpopulacji.

2.2. Średnia z próby prostej

2.2.1. Estymacja punktowa

Załóżmy, że dysponujemy danymi o zmiennej, obserwowanymi w próbie prostej. Wówczas wiadomo, że nieobciążonym estymatorem wartości średniej w populacji zarówno w przypadku zwrotnym, jak i bezzwrotnym jest średnia z próby, czyli

$$\bar{y}_S = \frac{1}{n} \sum_{k \in S} y_k = \frac{1}{n} \sum_{k=1}^N y_k S_k. \quad (2.12)$$

Wariancję średniej z próby prostej losowanej bezzwrotnie, czyli strategii $(\bar{y}_S, P_{bz}(s))$, określa wzór:

$$D^2(\bar{y}_S, P_{bz}(s)) = \frac{N-n}{Nn} v_*(y), \quad (2.13)$$

przy czym wariancję z populacji v_* określa wzór (1.5).

Gdy próba prosta jest losowana zwrotnie, to

$$D^2(\bar{y}_S, P_z(s)) = \frac{1}{n} v(y), \quad (2.14)$$

przy czym wariancję z populacji v określa wzór (1.5).

W badaniach reprezentacyjnych do porównywania wariancji dowolnej nieobciążonej strategii estymacji $(t_S, P(s))$ z wariancją średniej z próby prostej losowanej bezzwrotnie używa się tzw. współczynnika efektywności względnej, który ma postać:

$$def(t_S, P(s)) = \frac{D^2(t_S, P(s))}{D^2(\bar{y}_S, P_{bz}(s))}. \quad (2.15)$$

W przypadku analizowanych teraz strategii estymacji dla $n \geq 2$ otrzymujemy:

$$\text{def}(\bar{y}_S, P_z(s)) = \frac{D^2(\bar{y}_S, P_z(s))}{D^2(\bar{y}_S, P_{bz}(s))} = \frac{N-1}{N-n} > 1.$$

Zatem średnia z próby prostej losowanej zwrotnie jest mniej precyzyjnym estymatorem przeciętnej w populacji niż średnia z próby prostej losowanej bezzwrotnie.

Ocenę wariancji średniej z próby prostej otrzymujemy zastępując we wzorach (2.13) i (2.14) wariancje v_* i v statystyką:

$$v_{*S} = \frac{1}{n-1} \sum_{k \in S} (y_k - \bar{y}_S)^2. \quad (2.16)$$

Nieobciążone estymatory wariancji określonych wzorami (2.13) i (2.14) wyznaczają odpowiednio wzory:

$$D_S^2(\bar{y}_S, P_{bz}(s)) = \frac{N-n}{Nn} v_{*S}(y), \quad D_S^2(\bar{y}_S, P_z(s)) = \frac{1}{n} v_{*S}(y). \quad (2.17)$$

Estymatorem wartości globalnej zmiennej y jest statystyka:

$$\tilde{y}_S = N\bar{y}_S. \quad (2.18)$$

Na podstawie wzorów (2.13) i (2.14) łatwo wyprowadzamy wariancje estymatora, która w przypadku bezzwrotnego losowania próby prostej ma postać:

$$D^2(\tilde{y}_S, P_{bz}(s)) = \frac{N(N-n)}{n} v_*(y), \quad (2.19)$$

a w przypadku losowania zwrotnego:

$$D^2(\tilde{y}_S, P_z(s)) = \frac{N^2}{n} v(y). \quad (2.20)$$

W celu wyznaczenia oceny tych wariancji należy występujące w określających je wzorach wariancje z populacji v_* i v zastąpić ich nieobciążonym estymatorem zdefiniowanym wzorem (2.16).

Podobnie jak w przypadku oceny średniej w populacji strategia estymacji $(\tilde{y}_s, P_{bz}(s))$ jest precyzyjniejsza od $(\tilde{y}_s, P_z(s))$.

Niech d będzie dopuszczalnym poziomem błędu średniego szacunku wartości w populacji, czyli $D^2(\bar{y}_S, P_{bz}(s)) \leq d$. Stąd i z równania (2.13) wynika, że niezbędna liczebność próby prostej losowanej bezzwrotnie spełniająca określony postulat dokładności jest wyznaczana przez wyrażenie:

$$\underline{n} = [n_d] + 1, \quad n_d = \left(\frac{d^2}{v_*(y)} + \frac{1}{N} \right)^{-1}, \quad (2.21)$$

przy czym przez $[a]$ oznaczono część całkowitą z liczby a .

W przypadku próby prostej losowanej zwrotnie mamy:

$$\underline{n} = [n_d] + 1, \quad n_d = \frac{d^2}{v(y)}.$$

W praktyce rozsądnym może się okazać nałożenie ograniczenia na moduł współczynnika zmienności estymatora, czyli względnego błędu średniego szacunku, który określa wyrażenie:

$$z(\bar{y}_S, P_{bz}(s)) = 100\% \frac{D(\bar{y}_S, P_{bz}(s))}{|\bar{y}|}.$$

Zatem, jeśli postulowany dopuszczalny względny średni błąd szacunku jest na poziomie e , to niezbędną liczebność próby wyznaczamy z nierówności $z(\bar{y}_S, P_{bz}(s)) \leq e$. Wtedy potrzebną liczebność próby określa wyrażenie:

$$\underline{n} = [n_e] + 1, \quad n_e = \left(\frac{e^2}{z_*^2(y)} + \frac{1}{N} \right)^{-1}, \quad (2.22)$$

gdzie $z_*(y) = \frac{\sqrt{v_*(y)}}{|\bar{y}|} 100\%$.

W praktyce zwykle nie są znane wartości wariancji badanej cechy lub jej współczynnika zmienności. Dlatego występującą we wzorze (2.21) wariancję $v_*(y)$ należy zastąpić jej nieobciążonym estymatorem wyznaczonym ze wstępnej próby prostej o liczebności n , który określa wzór (2.16). Wówczas otrzymujemy następującą cenę niezbędnej liczebności próby.

$$\underline{n}_S = [n_{d,S}] + 1, \quad n_{d,S} = \left(\frac{d^2}{v_{*S}(y)} + \frac{1}{N} \right)^{-1}. \quad (2.23)$$

W przypadku konieczności korzystania ze wzoru (2.22) w sytuacji, gdy nie znamy parametru $z_*(y)$, należy go oszacować za pomocą statystyki $z_{*S}(y) = \frac{\sqrt{v_{*S}(y)}}{|\bar{y}_S|} 100\%$. Następnie podstawiamy ją odpowiednio do wzoru (2.22), co prowadzi do wyrażenia:

$$\underline{n}_S = [n_{e,S}] + 1, \quad n_{e,S} = \left(\frac{e^2}{z_{*S}^2(y)} + \frac{1}{N} \right)^{-1}. \quad (2.24)$$

Przykład 2.7. Na podstawie danych o dochodach z podatków z opłat lokalnych 15 gmin woj. katowickiego w 1995 r., pochodzących z Rocznika statystycznego gmin województwa katowickiego (1996), wyliczono następujące parametry: średnia 12 124,80 zł i odchylenie standardowe 7622,00 zł. Z kolei parametry dokładnych rozkładów średniej z próby prostej losowanej bezzwrotnie o liczebnościach 2–14 zamieszczono w tabeli 2.1. Oprócz błędu średnie-

go szacunku $D = D(\tilde{y}_s, P_{bz}(s))$ oraz względnego średniego błędu szacunku $z = z(\tilde{y}_s, P_{bz}(s))$ znajdujemy w niej również kwantyle $q_{0.05}$ i $q_{0.95}$ rozkładu średniej z próby odpowiednio rzędów 0.05 i 0.95, a także ich różnicę q . W naszym przypadku kwantyl $q_{0.05}$ to taka wartość średniej z próby, że co najmniej 5% możliwych wartości średnich obserwowanych w próbach jest mniejszych lub równych liczbie $q_{0.05}$ i jednocześnie co najmniej 95% możliwych wartości średnich obserwowanych w próbach jest większych lub równych liczbie $q_{0.05}$. Podobnie interpretujemy kwantyl $q_{0.95}$. Ponadto, wyliczono prawdopodobieństwo tego, iż średnia z próby odchyli się od szacowanej średniej w populacji nie więcej niż o wartość 1000, czyli $p = P(|\bar{y}_S - \bar{y}| \leq 1000)$.

Tabela 2.1

Parametry rozkładów średnich z prób (losowanych bezzwrotnie) dochodów gmin

n	D	z %	$q_{0.05}$	$q_{0.95}$	q	p
1	7622	62	3416	22737	19321	0.0667
2	5042	42	4301	20987	16686	0.1714
3	3940	33	5698	18663	12965	0.1736
5	2784	23	7629	16761	9132	0.2784
8	1841	15	9102	15192	6090	0.6973
10	1392	12	9807	14373	4566	0.7529
12	984	8	10490	13732	3242	0.8440
14	526	4	11366	12747	1381	1.0000

Z analizy tabeli 2.1 wynika, że wraz ze wzrostem liczebności próby dokładność estymacji rośnie, ponieważ średni błąd szacunku maleje. Zatem wartości estymatora średniej w populacji skupiają się coraz ściślej wokół szacowanej wartości średniej w populacji. świadczą o tym również obserwacja kwantyli $q_{0.05}$ i $q_{0.95}$, których wartości sukcesywnie przybliżają się do szacowanej wartości średniej w populacji wraz ze wzrostem liczebności próby. Tym samym wraz ze wzrostem liczebności próby co najmniej 90% wszystkich obserwacji średniej z próby znajduje się w coraz to węższym otoczeniu szacowanej przeciętnej w populacji, którego końcami są właśnie wymienione kwantyle. W końcu z analizy ostatniej kolumny tabeli 2.1 wynika, że prawdopodobieństwo tego, iż wartość modułu odchylenia średniej z próby od przeciętnej w populacji nie więcej niż o liczbę 1000 rośnie wraz ze wzrostem liczebności próby. Przeprowadzona analiza świadczy więc o tym, że rozkład średniej z próby skupia się coraz ściślej wokół szacowanej średniej z populacji, gdy rozmiar próby rośnie.

Przykład 2.8. *Przypuśćmy, że celem badania reprezentacyjnego jest ocena średniej wydatków na dojazd do uczelni studentów tworzących $N = 3000$ populację. W wylosowanej bezzwrotnie próbie prostej o liczebności $n = 15$ ocena średniej i wariancji wydatków wynoszą odpowiednio $\bar{y}_s = 20$ zł i $v_{*s} = 400$. Zadanie polega na odpowiedzi na pytania: jak duża winna być próba, aby średnią wydatków na dojazd oszacować: a) z błędem średnim szacunku nie większym od 2 zł, b) względnym błędem średnim szacunku nie większym od 2%.*

Korzystając ze wzorów (2.23) i (2.24) wyliczamy, że po to, aby błąd średni szacunku nie przekroczył poziomu $d = 2$, należy wylosować próbę o liczebności $n \geq \underline{n}_s = 97$. Z kolei, gdy żąda się, aby względny błąd średni szacunku nie przekroczył poziomu $e = 2\%$, to trzeba wylosować próbę o liczebności $n \geq \underline{n}_s = 50$. W związku z tym, jeśli do już dobranej próby prostej dolo-
sujemy bezzwrotnie 82 elementy populacji, to uzyskamy próbę o liczebności $n = 97$ elementów, co wystarczy, aby zostały spełnione obydwie wysunięte postulaty dokładności estymacji punktowej średniej w populacji.

Przykład 2.9. *Na podstawie danych analizowanych w przykładzie 2.7 wyznaczamy dokładne rozkłady prawdopodobieństwa dwóch estymatorów mediany z próby prostej losowanej bezzwrotnie przy wybranych liczebnościach próby. Mediana wpływów z podatków w populacji wymienionych gmin wynosi 9584,20 zł, co oznacza, że co najmniej połowa gmin ma wpływy z podatków nie mniejsze od tej kwoty oraz co najmniej połowa gmin ma wpływy z podatków nie większe od podanej wartości. Niech $y_{(i)}$ będzie i -tą statystyką porządkową obserwowanej zmiennej y w próbie S , a więc $y_{(1)} \leq y_{(2)} \leq \dots \leq y_{(n)}$. Wartość i -tej statystyki porządkowej jest z kolei i -tą co do wielkości wartością zmiennej y obserwowaną w próbie S . Wtedy znany estymator mediany jest określany jak następuje:*

$$M_{0S} = \begin{cases} y_{([n/2]+1)} & \text{dla } n \text{ nieparzystych,} \\ \frac{1}{2} (y_{(n/2)} + y_{(n/2+1)}) & \text{dla } n \text{ parzystych.} \end{cases} \quad (2.25)$$

Drugi estymator mediany, której własności studiował m.in. Fisz (1967), ma postać:

$$M_{1S} = y_{([n/2]+1)}. \quad (2.26)$$

Charakterystyki rozkładów statystyk M_{0S} i M_{1S} prezentuje tabela 2.2, w której m.in. zamieszczono względny wskaźnik obciążenia estymacji wb zdefiniowany wzorem (2.1) oraz współczynnik ubs będący udziałem kwadratu obciążenia estymacji w błędzie średniokwadratowym, który wyjaśnia wzór (2.4). Ponadto podano iloraz błędów średniokwadratowych $ibsr = 100\% \text{MSE}(M_{1S}) / \text{MSE}(M_{0S})$. Dokładność ocen mediany porównywano tylko dla parzystej liczebności próby, ponieważ wtedy tylko różnią się od siebie wartości obu estymatorów mediany.

Tabela 2.2

Parametry rozkładów median z prób dochodów gmin

n	$\sqrt{MSE(M_{0S})}$	$ibsr$	$wb(M_{0S})$	$wb(M_{1S})$	$ubs(M_{0S})$	$ubs(M_{1S})$
2	5645	100	-27	-19	20	12
8	3987	84	-15	-7	2	3
10	3324	80	-12	-7	2	4
12	2402	72	-10	-6	2	7
14	1019	1	-7	-1	46	1

Analiza drugiej i trzeciej z kolei kolumny tablicy 2.2 prowadzi do wniosku, że w badanej piętnastoelementowej populacji estymator mediany oznaczony przez M_{1S} jest dokładniejszy od M_{0S} . Podobnie, statystyka M_{1S} jest mniej obciążonym od M_{0S} estymatorem mediany w populacji. W przypadku obu estymatorów udział kwadratu obciążenia w błędzie średniokwadratowym estymacji jest mały (nie przekracza 7%) z wyjątkiem przypadku, gdy rozmiar próby jest bardzo mały ($n = 2$) albo bardzo duży ($n = 14$).

2.2.2. Estymacja przedziałowa

Przyjmijmy, że próba prosta jest losowana bezzwrotnie. Wówczas przedział ufności dla wartości przeciętnej w populacji, konstruowany zgodnie z zasadą określoną wyrażeniem (2.8), ma znaną postać:

$$P\left(\bar{y}_S - z_\gamma \sqrt{\frac{N-n}{Nn}} v_{*S} \leq \bar{y} \leq \bar{y}_S + z_\gamma \sqrt{\frac{N-n}{Nn}} v_{*S}\right) = \gamma, \quad (2.27)$$

przy czym v_{*S} określa wzór (2.16), a z_γ jest kwantylem rozkładu normalnego standardowego rzędu $(1 + \gamma)/2$, czyli $\varphi(z_\gamma) = (1 + \gamma)/2$, przy czym $\varphi(\cdot)$ jest dystrybuantą rozkładu normalnego standardowego. Określony przedział ufności zaleca się używać do oceny przeciętnej w populacji tylko w przypadku, gdy liczebność próby prostej jest duża, a przynajmniej równa stu elementom. Dokładność przedziału ufności określa wyrażenie:

$$\Delta = z_\gamma \sqrt{\frac{N-n}{Nn}} v_{*S}. \quad (2.28)$$

Zatem połowa rozpiętości przedziału ufności rośnie wraz ze wzrostem poziomu ufności γ lub ze spadkiem liczebności próby, lecz oczywiście przy ustalonej wartości wariancji badanej zmiennej. Załóżmy, że dopuszczalny błąd estymacji przedziałowej wynosi δ . Wtedy z nierówności $\delta \geq \Delta$ otrzymujemy w przybliżeniu niezbędną liczebność próby potrzebną do wyznaczenia

przedziału ufności o rozpiętości nie większej od δ , który z prawdopodobieństwem γ winien obejmować wartość przeciętną zmiennej w populacji. Zatem ta liczebność wynosi:

$$n_s = \left\lceil \left(\frac{\delta}{z_\gamma v_{*S}} + \frac{1}{N} \right)^{-1} \right\rceil + 1. \quad (2.29)$$

Przedział ufności dla wartości globalnej jest wyznaczany na podstawie wzoru:

$$P \left(N\bar{y}_S - z_\gamma \sqrt{\frac{N(N-n)}{n} v_{*S}} \leq y_U \leq N\bar{y}_S + z_\gamma \sqrt{\frac{N(N-n)}{n} v_{*S}} \right) = \gamma. \quad (2.30)$$

Przykład 2.10. W bezzwrotnie losowanej próbie prostej liczącej 120 studentów uczelni, która ma 5500 studentów, obserwowano w ustalonym semestrze ich wydatki na książki. Okazało się, że średnia z próby tych wydatków wynosi 104 zł, a nieobciążona ocena wariancji tych wydatków jest równa 400. Przy współczynniku ufności $\gamma = 0.95$ mamy $z_\gamma = 1.96$ oraz przedział ufności: [100.46; 107.54]. Ocena precyzji estymacji średniej za pomocą otrzymanego przedziału wynosi $\Delta = 3.54$.

Traktując próbę, w której obserwowano dane jako wstępną, wyznaczamy niezbędną liczebność próby, której wylosowanie pozwoli na wyznaczenie przedziału ufności dla przeciętnej o długości nieprzekraczającej poziomu 4 zł, przy tym samym poziomie współczynnika ufności. Na podstawie wzoru (2.29) wyliczamy, że niezbędna liczebność próby wynosi 360 studentów, a zatem należy jeszcze dołosować 240 studentów, aby spełnić postulaty dotyczące precyzji i wiarygodności estymacji przedziałowej.

2.2.3. Ocena wskaźnika struktury

Udział elementów z cechą wyróżnioną, inaczej wskaźnik struktury, jest parametrem szczególnym w stosunku do wartości średniej. Podobnie jak liczba elementów z cechą wyróżnioną jest szczególnym parametrem w stosunku do wartości globalnej zmiennej w populacji. Niech M będzie liczbą elementów z cechą wyróżnioną w populacji o liczebności N . Wtedy wskaźnik struktury oznaczmy przez

$$p = \frac{M}{N}. \quad (2.31)$$

Estymatorem parametru p jest częstość względna z próby:

$$p_S = \frac{m_S}{n}, \quad (2.32)$$

gdzie realizacja zmiennej losowej p_S jest częstością elementów z cechą wyróżnioną obserwowaną w próbie s o liczebności n , która jest realizacją prostej próby losowej S . Jeśli próba jest losowana bezzwrotnie, to statystyka p_S ma znany rozkład hipergeometryczny, który określa następująca funkcja prawdopodobieństwa:

$$P\left(p_S = \frac{m_s}{n}\right) = \frac{\binom{M}{m} \binom{N-M}{n-m}}{\binom{N}{n}} = \frac{\binom{Np}{m} \binom{N(1-p)}{n-m}}{\binom{N}{n}}. \quad (2.33)$$

Wartość oczekiwaną i wariancję statystyki p_S prezentują wzory:

$$E(p_S) = p, \quad D^2(p_S) = \frac{N-n}{(N-1)} \frac{p(1-p)}{n}. \quad (2.34)$$

Nieobciążonym estymatorem wariancji zmiennej losowej p_S jest statystyka:

$$D_S^2(p_S) = \frac{N-n}{(N-1)} \frac{p_S(1-p_S)}{n}. \quad (2.35)$$

Gdy próba prosta jest losowana zwrotnie, to estymator p_S ma znany rozkład dwumianowy postaci:

$$P\left(p_S = \frac{m}{n}\right) = \binom{N}{n} p^m (1-p)^{n-m}. \quad (2.36)$$

W tym przypadku:

$$E(p_S) = p, \quad D^2(p_S) = \frac{p(1-p)}{n}. \quad (2.37)$$

Zatem zarówno przy bezzwrotnym, jak i zwrotnym wariancie losowania próby prostej statystyka p_S daje nieobciążone i zgodne oceny parametru p . Nieobciążonym estymatorem wariancji zmiennej losowej p_S w przypadku losowania zwrotnego jest statystyka:

$$D_S^2(p_S) = \frac{p_S(1-p_S)}{n}. \quad (2.38)$$

Załóżmy, że dopuszczalny poziom średniego błędu szacunku wynosi d . Wówczas w przypadku losowania bezzwrotnego na podstawie nierówności $D(p_S) \leq d$ wyliczamy, że:

$$n \geq n_d = \left\lceil \frac{N}{\frac{d^2(N-1)}{p(1-p)} + 1} \right\rceil + 1, \quad (2.39)$$

gdzie przypomnijmy, przez $[x]$ oznaczono funkcję wyznaczającą część całkowitą liczby x . Nieznana wartość parametru p wyznacza się na podstawie informacji dodatkowych lub ocenia na podstawie próbki wstępnej lub przyjmuje się jego oszacowanie otrzymane we wcześniejszych badaniach statystycznych. Gdy te rozwiązania nie są możliwe, to niezbędną liczebność próby ustala się na poziomie maksymalnym $n_{\max}$, który jest osiągany dla $p = 0.5$ i wynosi:

$$n_{\max} = \left[\frac{N}{4d^2(N-1)+1} \right] + 1. \quad (2.40)$$

Gdy $N \rightarrow \infty$ bądź jest stosowany zwrotny wariant losowania próby, to

$$n_d = \left[\frac{p(1-p)}{d^2} \right] + 1, \quad n_{\max} = \left[\frac{1}{4d^2} \right] + 1. \quad (2.41)$$

Dla bezzwrotnego i zwrotnego wariantu losowania próby względny średni błąd szacunku, czyli $z(p_S) = D(p_S)/p$ określają odpowiednio wzory:

$$z(p_S) = \sqrt{\frac{N-n}{Nn} \frac{1-p}{np}}, \quad z(p_S) = \sqrt{\frac{1-p}{np}}. \quad (2.42)$$

Stąd i z nierówności $z(p_S) \leq e$ wyliczamy niezbędną liczebność próby, którą dla bezzwrotnego i zwrotnego wariantu losowania próby, określają wzory:

$$n_e = \left[\frac{N(1-p)}{1-p+e^2p(N-1)} \right] + 1, \quad n_e = \left[\frac{1-p}{e^2p} \right] + 1. \quad (2.43)$$

Stąd wynika, że niezbędną liczebność próby rośnie wraz ze spadkiem wartości parametru e .

Estymacja przedziałowa wskaźnika struktury jest zwykle prowadzona na podstawie dużych prób o liczebności równej co najmniej sto elementów. Przedział ufności dla wskaźnika struktury w populacji, konstruowany zgodnie z zasadą określoną wyrażeniem (2.8), ma w przypadku bezzwrotnego losowania próby następującą znaną postać:

$$P \left(p_S - z_\gamma \sqrt{\frac{N-n}{N} \frac{p_S(1-p_S)}{n}} < p < p_S + z_\gamma \sqrt{\frac{N-n}{N} \frac{p_S(1-p_S)}{n}} \right) = \gamma, \quad (2.44)$$

a w przypadku zwrotnie losowanej próby prostej mamy:

$$P \left(p_S - z_\gamma \sqrt{\frac{p_S(1-p_S)}{n}} < p < p_S + z_\gamma \sqrt{\frac{p_S(1-p_S)}{N}} \right) = \gamma, \quad (2.45)$$

gdzie kwantyl z_γ jest wyznaczany na podstawie dystrybucji rozkładu normalnego standardowego $\varphi(z)$, tak aby $\varphi(z_\gamma) = 0.5(1+\gamma)$.

Dokładność estymacji przedziałowej określana połową długości przedziału ufności prezentują odpowiednio następujące wzory dla wariantu bezzwrotnego i zwrotnego losowania próby:

$$\Delta = z_\gamma \sqrt{\frac{N-n}{N} \frac{p_S(1-p_S)}{n}}, \quad \Delta = z_\gamma \sqrt{\frac{p_S(1-p_S)}{n}}. \quad (2.46)$$

Przyjmując na poziomie δ dopuszczalny błąd estymacji przedziałowej, wyznaczamy ocenę niezbędnej liczebności próby potrzebnej do oceny wskaźnika struktury z dokładnością δ i wiarygodnością (ufnością) γ za pomocą wzoru:

$$n_\delta = \left\lceil \frac{z_\gamma^2 N p_S(1-p_S)}{(N-1)\delta^2 + z_\gamma^2 p_S(1-p_S)} \right\rceil + 1. \quad (2.47)$$

Gdy p_S nie jest znane, to przyjmujemy jego najmniej korzystną wartość $p_S = \frac{1}{2}$ i wtedy:

$$n_\delta = \left\lceil \frac{N z_\gamma^2}{4(N-1)\delta^2 + z_\gamma^2} \right\rceil + 1. \quad (2.48)$$

Jeśli $n \rightarrow \infty$ lub jest stosowany zwrotny wariant losowania próby, to

$$n_\delta = \left\lceil \frac{z_\gamma^2 p_S(1-p_S)}{\delta^2} \right\rceil + 1, \quad (2.49)$$

$$n_\delta = \left\lceil \frac{z_\gamma^2}{4\delta^2} \right\rceil + 1, \quad \text{gdy } p_S = \frac{1}{2}. \quad (2.50)$$

Przykład 2.11. W bezzwrotnie losowanej próbie prostej liczącej 110 mieszkań miasta, w którym jest 4000 mieszkań, obserwowano w ustalonym roku, że 50 mieszkań ma centralne ogrzewanie. Ocena z próby udziału mieszkań z centralnym ogrzewaniem wynosi 0.455, a oceny błędu średniego szacunku oraz względnego średniego błędu szacunku wynoszą odpowiednio 0.047 i 10.3%. Przy współczynniku ufności $\gamma = 0.95$ przedział ufności dla udziału mieszkań z centralnym ogrzewaniem ma postać: [0.363; 0.546]. Zatem precyzja estymacji przedziałowej wynosi $\delta = 0.092$. Przyjmujemy, że w populacji udział mieszkań z centralnym ogrzewaniem wynosi 0.5, czyli przyjmuje najmniej korzystną wartość z punktu widzenia precyzji estymacji. Wtedy wyznaczamy tak niezbędną liczebność próby, aby z ufnością $\gamma = 0.95$ przedział obejmował nieznaną wartość z precyzją $\delta = 0.05$. Na podstawie wzoru (2.48) otrzymujemy, że należy spodziewać się spełnienia postawionych postulatów, gdy liczebność próby osiągnie przynajmniej poziom 351 elementów. Zatem, należy jeszcze dołosować w sposób bezzwrotny do próby 241 mieszkań.

Problem estymacji przedziałowej wskaźnika struktury w małych próbach jest rozwiązywany przez Chunga i DeLury'ego (1950). W przypadku umiarkowanie dużych prób problem ten analizowano w pracy Krysińskiego i in. (1989).

2.2.4. Wnioskowanie modelowe

Model prosty jest określany następująco. Niech składowe wektora $\mathbf{Y}_U$ spełniają założenia:

$$\begin{cases} E_{\xi}(Y_i) = \mu, & D_{\xi}^2(Y_i) = \sigma^2 & \text{dla } i = 1, \dots, N, \\ \text{Cov } \xi(Y_i, Y_t) = 0 & \text{dla } i \neq t, & t = 1, \dots, N, \quad i = 1, \dots, N. \end{cases} \quad (2.51)$$

W świetle wprowadzonych założeń populacja jest jednorodna w tym sensie, że wszystkim elementom populacji przypisano zmienne losowe (nieskorelowane), które mają tę samą nadzieję matematyczną i tę samą wariancję.

Rozważmy problem predykcji wartości globalnej Y_U za pomocą predyktora liniowego $\tilde{Y}_S$ przy pewnym planie losowania $P(s)$, przy czym:

$$\tilde{Y}_{\gamma S} = \sum_{i=1}^N S_i \gamma_i Y_i, \quad (2.52)$$

$S_i = 1$, gdy i -ty element populacji został wybrany do próby, natomiast w przeciwnym przypadku $S_i = 0$, $i = 1, \dots, N$. Wtedy wiadomo, że $E_P(S_i) = \pi_i$. Zakładamy, że próba ma ustaloną efektywną liczebność na poziomie n , czyli $\sum_{i=1}^N S_i = n$. Przez γ_i , $i = 1, \dots, N$ oznaczono nielosowe współczynniki (wagi).

Predyktor $\tilde{Y}_{\gamma S}$ będzie ξ -nieobciążony, gdy $E_{\xi}(\tilde{Y}_{\gamma S} - Y_U) = 0$, co zachodzi wtedy, jeśli:

$$\sum_{i=1}^N S_i \gamma_i = N. \quad (2.53)$$

Predyktor ten będzie $p\xi$ -nieobciążony, jeśli $E_P E_{\xi}(\tilde{Y}_{\gamma S} - Y_U) = 0$, co zachodzi wówczas, gdy:

$$\sum_{i=1}^N \pi_i \gamma_i = N. \quad (2.54)$$

Przy założeniu, że $\mu = 0$ wyprowadzenie ξ -błędu średniokwadratowego ma postać:

$$\begin{aligned}
E_{\xi}(\tilde{Y}_{\gamma S} - Y_U)^2 &= E_{\xi} \left(\sum_{i=1}^N (S_i \gamma_i - 1) Y_i \right)^2 = \\
&= \sum_{i=1}^N (S_i \gamma_i - 1)^2 E_{\xi}(Y_i^2) + \sum_{i=1}^N \sum_{j=1, j \neq i}^N (S_i \gamma_i - 1) Y_i (S_j \gamma_j - 1) E_{\xi}(Y_i Y_j) = \\
&= \sum_{i=1}^N (S_i \gamma_i - 1)^2 D_{\xi}^2(Y_i^2) + \sum_{i=1}^N \sum_{j=1, j \neq i}^N (S_i \gamma_i - 1) Y_i (S_j \gamma_j - 1) \text{Cov}_{\xi}(Y_i Y_j) = \\
&= \sigma^2 \sum_{i=1}^N (S_i \gamma_i - 1)^2. \quad (2.55)
\end{aligned}$$

Stąd i ze wzoru (2.53) już wynika następujące wyprowadzenie p - ξ błędu średniokwadratowego:

$$\begin{aligned}
E_P E_{\xi}(\tilde{Y}_{\gamma S} - Y_U)^2 &= \sigma^2 \sum_{i=1}^N E_P (S_i \gamma_i - 1)^2 = \sigma^2 \sum_{i=1}^N E_P (S_i \gamma_i^2 - 2S_i \gamma_i + 1) = \\
&= \sigma^2 \left(\sum_{i=1}^N \pi_i \gamma_i^2 - 2 \sum_{i=1}^N \pi_i \gamma_i + N \right) = \sigma^2 \left(\sum_{i=1}^N \pi_i \gamma_i^2 - N \right). \quad (2.56)
\end{aligned}$$

Zadanie polega na takim wyznaczeniu wag γ_i oraz prawdopodobieństw π_i , aby błąd średniokwadratowy był jak najmniejszy przy założeniu nieobciążoności predyktora, co jest równoważne warunkowi danemu wzorem (2.53). Ponadto pamiętamy, iż z własności planu losowania próby wynika, że: $\sum_{i=1}^N \pi_i = n$ i $\sum_{i=1}^N S_i = n$. Funkcja Lagrange'a ma postać:

$$L = \sigma^2 \left(\sum_{i=1}^N \pi_i \gamma_i^2 - N \right) + \lambda_1 \left(N - \sum_{i=1}^N \pi_i \gamma_i \right) + \lambda_2 \left(n - \sum_{i=1}^N \pi_i \right).$$

Pochodne względem tej funkcji względem γ_i są postaci: $\frac{\partial L}{\partial \gamma_i} = 2\pi_i \gamma_i - \pi_i \lambda_1$, $i = 1, \dots, N$. Przyrównując je do zera i sumując tak otrzymane równania stronami dochodzimy do rozwiązania: $\gamma_i = \frac{N}{n}$ dla $i = 1, \dots, N$. Zatem określony wzorem (2.52) predyktor przyjmuje postać:

$$\tilde{Y}_S = \frac{N}{n} \sum_{i=1}^N S_i Y_i = \frac{N}{n} \sum_{i \in S} Y_i = N \bar{Y}_S. \quad (2.57)$$

Wyrażenie (2.56) prowadzi do następującego wzoru na p - ξ błąd średniokwadratowy:

$$\begin{aligned}
E_P E_{\xi}(\tilde{Y}_{\gamma S} - Y_U)^2 &= \sigma^2 \left(\frac{N^2}{n^2} \sum_{i=1}^N \pi_i - N \right) = \sigma^2 \left(\frac{N^2}{n} - N \right). \\
MSE_{p,\xi}(\tilde{Y}_{\gamma S}) &= \frac{N(N-n)}{n} \sigma^2. \quad (2.58)
\end{aligned}$$

Stąd już ewidentnie wynika, że $p\xi$ błąd średniokwadratowy predyktora $\tilde{Y}_{\gamma S}$ nie zależy od przyjętego planu losowania próby. W związku z tym $p\xi$ błąd średniokwadratowy predyktora jest taki sam przy dowolnym schemacie losowania próby. Można więc brać również celowe plany losowania próby pod uwagę, byleby elementy populacji nie powtarzały się w próbie o ustalonym rozmiarze. W szczególności do próby można wybrać n -pierwszych z kolei elementów populacji. Wówczas, formalnie rzecz biorąc, $S_i = 1$, $\pi_i = 1$ dla $i = 1, \dots, n$, natomiast $S_i = 0$, $\pi_i = 0$ dla $i = n + 1, \dots, N$. W sytuacji celowego planu losowania próby nie ma już sensu rozważać p nieobciążoności predyktora.

ξ -nieobciążony estymator wariancji określonego wzorem (2.57) predyktora $\tilde{Y}_S$ ma postać:

$$MSE_S(\tilde{Y}_S) = \frac{N(N-n)}{n} \hat{V}_S^2 \quad (2.59)$$

gdzie:

$$\hat{V}_S^2 = \frac{1}{n-1} \sum_{i \in S} (Y_i - \bar{Y}_S)^2, \quad \bar{Y}_S = \frac{1}{n} \sum_{i \in S} Y_i$$

Przedstawione wyniki dotyczące predykcji są oczywiście ważne, gdy opisany model nadpopulacji jest adekwatny do rzeczywistości, tzn. gdy założony rozkład prawdopodobieństwa ξ istotnie nie odbiega od rzeczywistego. Przedstawiona procedura predykcji wartości globalnej może jednak być obciążona znacznie większym błędem średnim od oczekiwanego, jeśli np. zakładana jednorodność modelu określonego wzorem (2.51) nie jest prawdziwa. Konkretniej chodzi tu o to, że obserwowany w próbie rozkład wartości cechy nie pozwala dalej utrzymywać postulatu, że wartości oczekiwane zmiennych losowych Y_k , $k = 1, \dots, N$, tworzących nadpopulację są takie same lub że wariancje tych zmiennych przyjmują tę samą wartość. Rozstrzygnięciem tego problemu będziemy zajmować się m.in. w punkcie 2.3.3.

2.3. Statystyka Horvitz-Thompsona

2.3.1. Estymacja wartości globalnej

W niniejszym rozdziale przedstawimy problem estymacji wartości globalnej zmiennej y w populacji na podstawie próby losowanej z różnymi prawdopodobieństwami doboru do niej elementów, które są zwykle funkcją wartości zmiennych pomocniczych obserwowanych w całej populacji. Horvitz i Thompson (1952) zaproponowali estymator wartości globalnej $y_U = y$ w populacji ustalonej, który ma postać:

$$y_{HTS} = \sum_{k=1}^N \frac{y_k S_k}{\pi_k}. \quad (2.60)$$

Gdy $\pi_k > 0$ dla wszystkich $k = 1, \dots, N$, statystyka y_{HTS} jest nieobciążonym estymatorem parametru y_U , a jej wariancja jest postaci:

$$D^2(y_{HTS}) = \sum_{k=1}^N \left(\frac{y_k}{\pi_k} \right)^2 \pi_k (1 - \pi_k) + \sum_{k=1}^N \sum_{i=1, i \neq k}^N \frac{y_k y_i}{\pi_k \pi_i} (\pi_{k,i} - \pi_k \pi_i) \quad (2.61)$$

lub

$$D^2(y_{HTS}) = \sum_{k=1}^N \frac{y_k^2}{\pi_k} + \sum_{k=1}^N \sum_{i=1, i \neq k}^N \frac{y_k y_i}{\pi_k \pi_i} \pi_{k,i} - y^2. \quad (2.62)$$

Gdy efektywna liczebność próby jest ustalona, to wariancję określa wzór (por. Yates i Grundy, 1953):

$$D^2(y_{HTS}) = \sum_{k=1}^N \sum_{i=1, i \neq k}^N \left(\frac{y_k}{\pi_k} - \frac{y_i}{\pi_i} \right)^2 (\pi_{k,i} - \pi_k \pi_i). \quad (2.63)$$

Do estymacji parametru $D^2(y_{HTS})$ jest używana statystyka:

$$D_S^2(y_{HTS}) = \sum_{k=1}^N \left(\frac{y_k}{\pi_k} \right)^2 (1 - \pi_k) S_k + \sum_{k=1}^N \sum_{i=1, i \neq k}^N \frac{y_k y_i}{\pi_k \pi_i} \frac{\pi_{k,i} - \pi_k \pi_i}{\pi_{k,i}} S_k S_i, \quad (2.64)$$

która jest nieobciążonym estymatorem wariancji $D^2(y_{HTS})$. Warto zwrócić uwagę na to, że estymator $D_S^2(y_{HTS})$ w pewnych sytuacjach może przyjmować wartości ujemne z prawdopodobieństwem większym od zera. Jeśli $\pi_{k,i} - \pi_k \pi_i \geq 0$ dla każdego $i = 1, \dots, N$, $k = 1, \dots, N$, $i \neq k$, to następujący estymator Sena-Yatesa-Grundy'ego przyjmuje tylko wartości nieujemne i ma postać:

$$\hat{D}_S^2(y_{HTS}) = \sum_{k=1}^N \sum_{i=1, i \neq k}^N \left(\frac{y_k}{\pi_k} - \frac{y_i}{\pi_i} \right)^2 \frac{\pi_{k,i} - \pi_k \pi_i}{\pi_{k,i}} S_k S_i. \quad (2.65)$$

Na podstawie statystyki y_{HTS} w sposób natychmiastowy otrzymujemy następujący estymator wartości średniej $\bar{y}$

$$\bar{y}_{HTS} = \frac{1}{N} \bar{y}_{HTS} = \frac{1}{N} \sum_{k=1}^N \frac{y_k S_k}{\pi_k}. \quad (2.66)$$

Wariancję tej statystyki oraz jej estymatory otrzymujemy prosto w następujący sposób:

$$D^2(\bar{y}_{HTS}) = \frac{1}{N^2} D^2(y_{HTS}),$$

$$D_S^2(\bar{y}_{HTS}) = \frac{1}{N^2} D_S^2(y_{HTS}), \quad \hat{D}_S^2(\bar{y}_{HTS}) = \frac{1}{N^2} \hat{D}_S^2(y_{HTS}).$$

Przykład 2.12. Podstawienie $y_k = 1$ dla $k = 1, \dots, N$, we wzorze (2.60) prowadzi do następującego estymatora liczebności populacji N :

$$N_{HTS} = \sum_{k=1}^N \frac{S_k}{\pi_k}. \quad (2.67)$$

Wariancję tej statystyki oraz jej estymatory otrzymujemy prosto ze wzorów (2.62) i (2.64), a określają je wzory:

$$D^2(N_{HTS}) = \left(\sum_{k=1}^N \frac{1}{\pi_k} + \sum_{k=1}^N \sum_{i=1, i \neq k}^N \frac{\pi_{k,i}}{\pi_k \pi_i} \right) - N^2, \quad (2.68)$$

$$D_S^2(N_{HTS}) = \sum_{k=1}^N \frac{1 - \pi_k}{\pi_k^2} S_k + \sum_{k=1}^N \sum_{i=1, i \neq k}^N \left(\frac{1}{\pi_k \pi_i} - \frac{1}{\pi_{k,i}} \right) S_k S_i. \quad (2.69)$$

Teraz przeprowadzamy analizę wpływu losowania próby z dowolnymi prawdopodobieństwami inkluzji pierwszego rzędu na wartość oczekiwaną zwykłej średniej z próby. Niech $\bar{y}_S$ będzie określoną wzorem (2.12) średnią z próby losowanej bezwrotnie. Wtedy:

$$\begin{aligned} E(\bar{y}_S) &= \frac{1}{n} E \left(\sum_{k=1}^N y_k S_k \right) = \frac{1}{n} \sum_{k=1}^N y_k \pi_k, \\ E(\bar{y}_S) &= \bar{y} - \bar{y} \left(1 - \frac{N}{n} \bar{\pi} \right) + \frac{N}{n} v(\pi) v(y) \rho(\pi, y), \end{aligned} \quad (2.70)$$

gdzie:

$$\begin{aligned} \bar{\pi} &= \frac{1}{N} \sum_{k=1}^N \pi_k, & v(\pi) &= \frac{1}{N} \sum_{k=1}^N (\pi_k - \bar{\pi})^2, \\ v(y) &= \frac{1}{N} \sum_{k=1}^N (y_k - \bar{y})^2, & \rho(\pi, y) &= \frac{\sum_{k=1}^N (\pi_k - \bar{\pi})(y_k - \bar{y})}{N \sqrt{v(\pi) v(y)}}, \end{aligned}$$

przy czym $\rho(\pi, y)$ jest współczynnikiem korelacji liniowej między zmiennymi y i π . Ze wzoru (2.70) wynika, że zwykła średnia z próby może dać obciążone oceny przeciętnej w populacji, gdy próba jest losowana z różnymi prawdopodobieństwami inkluzji rzędu pierwszego. Zwróćmy uwagę na to, że nawet gdy prawdopodobieństwa inkluzji nie są skorelowane z wartościami badanej cechy (czyli $\rho(\pi, y) = 0$), to średnia z próby może być znacznie obciążonym estymatorem przeciętnej w populacji, gdy $\bar{\pi} \neq \frac{n}{N}$. To winno przekonać praktyków, że trzeba stosować estymator Horvitz-Thompsona, gdy mają do czynienia z próbami nieprostymi.

Częstym celem estymacji jest wskaźnik struktury elementów populacji z cechą wyróżnioną, czyli parametr $p = \frac{M}{N}$, gdzie M jest liczbą elementów z cechą wyróżnioną w populacji. Niech $z_k = 1$, gdy k -ty element populacji posiada wyróżnioną cechę, natomiast $z_k = 0$, gdy jej nie posiada. Wtedy populację $U = \{u_k\}$ można rozłożyć na takie dwa rozłączne podzbiory U_0 i U_1 , że $U = U_0 \cup U_1$ i $U_0 = \{u_i : z_i = 0, i = 1, \dots, N\}$ oraz $U_1 = \{u_i : z_i = 1, i = 1, \dots, N\}$. Wtedy parametr p możemy zapisać następująco:

$$p = \frac{1}{N} \sum_{k=1}^N z_k.$$

Następująca statystyka będąca szczególnym przypadkiem estymatora Horvitz-Thompsona, danego wzorem (2.60), jest estymatorem parametru p .

$$p_{HTS} = \frac{1}{N} \sum_{k=1}^N \frac{z_k S_k}{\pi_k} = \frac{1}{N} \sum_{k \in U_1} \frac{S_k}{\pi_k}. \quad (2.71)$$

Daje ona nieobciążone oceny wskaźnika struktury p , gdy $\pi_k > 0$ dla $k = 1, \dots, N$. Wariancję estymatora $\bar{p}_{HTS}$ otrzymujemy ze wzoru (2.62), który redukuje się do postaci:

$$D^2(p_{HTS}) = \frac{1}{N^2} \left(\sum_{k \in U_1} \frac{1}{\pi_k} + \sum_{k \in U_1} \sum_{i \in U_1, k \neq i} \frac{\pi_{k,i}}{\pi_k \pi_i} \right) - p^2. \quad (2.72)$$

Nieobciążony estymator zapisanej wariancji otrzymujemy ze wzoru (2.64), i ma on postać:

$$D_S^2(p_{HTS}) = \frac{1}{N^2} \sum_{k \in U_1} \frac{(1 - \pi_k) S_k}{\pi_k^2} + \frac{1}{N^2} \sum_{k \in U_1} \sum_{i \in U_1, i \neq k} \left(\frac{1}{\pi_k \pi_i} - \frac{1}{\pi_{k,i}} \right) S_k S_i. \quad (2.73)$$

W szczególności, jeśli bezzwrotnie losowana próba S o liczebności n jest próbą prostą, to estymator p_{HTS} redukuje się do zwykłej częstości p_S względnej z próby, którą określa wzór (2.32), a jego wariancję i jej estymator odpowiednio wzory (2.34) i (2.35).

Można wykazać, że wpływ losowania próby z różnymi prawdopodobieństwami na wartość oczekiwaną estymatora p_S określa wzór:

$$E(p_S) = \frac{1}{n} \sum_{k \in U_1} \pi_k = p - p \left(1 - \frac{N}{n} \bar{\pi}_1 \right), \quad (2.74)$$

gdzie $\bar{\pi}_1 = \frac{1}{M} \sum_{k \in U_1} \pi_k$ jest średnią arytmetyczną prawdopodobieństw inkluzji (rzędu pierwszego) przypisanych elementom populacji z cechą wyróżnioną. W tym więc przypadku również należy używać estymator Horvitz-Thompsona zamiast częstości względnej z próby.

Przykład 2.13. Elementom populacji $U = \{1, 2, 3, 4, 5\}$ są przyporządkowane następujące wartości zmiennej y : $y(1) = 8$, $y(2) = 1$, $y(3) = 3$, $y(4) = 1$, $y(5) = 3$. Średnia w populacji wynosi $\bar{y} = 3.2$, a wariancja $v_* = 8.2$. Do oceny średniej użyto przeciętnej z próby oraz estymatora Horvitz-Thompsona, który określa wzór. Pierwszy plan bezzwrotnego losowania próby trójelementowej określają wyrażenia:

$$P_1(1, 2, 3) = P_1(1, 2, 5) = P_1(1, 3, 4) = P_1(1, 4, 5) = \frac{1}{8},$$

$$P_1(1, 2, 4) = P_1(2, 3, 4) = P_1(2, 3, 5) = P_1(3, 4, 5) = P_1(1, 3, 5) = P_1(2, 4, 5) = \frac{1}{12}.$$

Prawdopodobieństwa inkluzji pierwszego i drugiego rzędu są następujące:

$$\begin{aligned} \pi_1^{(1)} &= \frac{2}{3}, & \pi_2^{(1)} &= \pi_3^{(1)} = \pi_4^{(1)} = \pi_5^{(1)} = \frac{7}{12}, & \pi_{1,2}^{(1)} &= \pi_{1,3}^{(1)} = \pi_{1,4}^{(1)} = \pi_{1,5}^{(1)} = \frac{1}{3}, \\ \pi_{2,3}^{(1)} &= \pi_{2,5}^{(1)} = \pi_{3,4}^{(1)} = \pi_{4,5}^{(1)} = \frac{7}{24}, & \pi_{2,4}^{(1)} &= \pi_{3,5}^{(1)} = \frac{1}{4}. \end{aligned}$$

Drugi plan losowania próby prezentują wzory:

$$P_2(1, 2, 3) = P_2(1, 2, 5) = P_2(1, 3, 4) = P_2(1, 4, 5) = \frac{1}{16}, \quad P_2(1, 2, 4) = \frac{1}{8},$$

$$P_2(2, 3, 4) = P_2(2, 3, 5) = P_2(3, 4, 5) = P_2(1, 3, 5) = P_2(2, 4, 5) = \frac{1}{8},$$

$$\pi_1^{(2)} = \frac{13}{35}, \quad \pi_2^{(2)} = \pi_3^{(2)} = \pi_4^{(2)} = \pi_5^{(2)} = \frac{23}{35}, \quad \pi_{2,4}^{(2)} = \pi_{3,5}^{(2)} = \frac{33}{70}$$

$$\pi_{1,2}^{(2)} = \pi_{1,3}^{(2)} = \pi_{1,4}^{(2)} = \pi_{1,5}^{(2)} = \frac{13}{70}, \quad \pi_{2,3}^{(2)} = \pi_{2,5}^{(2)} = \pi_{3,4}^{(2)} = \pi_{4,5}^{(2)} = \frac{23}{70}.$$

Na podstawie wzorów (2.61) lub (2.62), lub (2.63) wyliczamy wariancje strategii losowania, a wynoszą one: $D^2(\bar{y}_{HTS}, P_1(s)) = 0,7706$, $D^2(\bar{y}_{HTS}, P_2(s)) = 3.3691$. Z kolei wariancja średniej z próby prostej losowanej bezzwrotnie wyliczana na podstawie wzoru (2.13) wynosi $D^2(\bar{y}_s, P_{bz}(s)) = 1.0933$, przy czym plan losowania $P_{bz}(s)$ określa definicja 1.13. Wariancja $D^2(\bar{y}_{HTS}, P_1(s))$ stanowi 70.5% wariancji średniej z próby prostej, a z kolei ta stanowi 32.5% wariancji strategii $D^2(\bar{y}_{HTS}, P_2(s))$. Zatem spośród trzech rozważanych strategii losowania najlepszą jest strategia $(\bar{y}_{HTS}, P_1(s))$, ponieważ daje najprecyzyjniejsze oceny wartości średniej w populacji.

Przykład 2.14. Definicją 1.16 określono plan losowania próby Poissona, który oznaczmy symbolem $P_P(s)$. W tym przypadku estymator wartości globalnej Horvitz-Thompsona oraz jego wariancje określają odpowiednio wzory:

$$y_{HTS} = \sum_{k=1}^N \frac{y_k S_k}{\pi_k}, \quad D^2(y_{HTS}, P_P(s)) = \sum_{k=1}^N \left(\frac{y_k}{\pi_k} \right)^2 \pi_k (1 - \pi_k). \quad (2.75)$$

Ze wzoru (2.64) wynika następujący nieobciążony estymator wariancji:

$$D_S^2(y_{HTS}, P_P(s)) = \sum_{k=1}^N \left(\frac{y_k}{\pi_k} \right)^2 (1 - \pi_k) S_k. \quad (2.76)$$

Przykład 2.15. Definicją 1.15 określono plan losowania próby Bernoulliego, który oznaczmy przez $P_B(s)$. Plan ten jest oczywiście szczególnym przypadkiem planu losowania Poissona. Zakładamy, że prawdopodobieństwo π wylosowania k -tego elementu populacji do próby jest znane. Wtedy estymator wartości globalnej Horvitz-Thompsona oraz jego wariancję określają odpowiednio wzory:

$$y_{HTS} = \frac{1}{\pi} \sum_{k=1}^N y_k S_k, \quad D^2(y_{HTS}, P_B(s)) = \frac{1 - \pi}{\pi} \sum_{k=1}^N y_k^2. \quad (2.77)$$

Nieobciążony estymator wariancji ma postać:

$$D_S^2(y_{HTS}, P_B(s)) = \frac{1 - \pi}{\pi^2} \sum_{k=1}^N y_k^2 S_k. \quad (2.78)$$

Jeśli jest znana liczebność populacji N , to na podstawie powyższych wzorów otrzymujemy łatwo następujący estymator wartości średniej $\bar{y}$ zmiennej badanej $\bar{y}_{HTS} = \frac{1}{N} y_{HTS}$. Zatem

$$\bar{y}_{HTS} = \frac{1}{N\pi} \sum_{k=1}^N y_k S_k. \quad (2.79)$$

Statystyka $\bar{y}_{HTS}$ jest nieobciążonym estymatorem średniej w populacji, a jej wariancja wynosi $D^2(\bar{y}_{HTS}) = N^{-2} D^2(y_{HTS})$, gdzie $D^2(y_{HTS})$ określa wzór (2.77).

W szczególności pierwszy wzór wyrażenia (2.77) pozwala określić następujący nieobciążony estymator liczebności populacji N .

$$N_{HTS} = \frac{1}{\pi} \sum_{k=1}^N S_k. \quad (2.80)$$

Wyrażenia (2.77) i (2.78) prowadzą do następującej wariancji statystyki N_{HTS} i jej nieobciążonego estymatora:

$$D^2(N_{HTS}, P_B(s)) = \frac{1 - \pi}{\pi} N, \quad (2.81)$$

$$D_S^2(N_{HTS}, P_B(s)) = \frac{1 - \pi}{\pi^2} n_S, \quad \text{gdzie } n_S = \sum_{k=1}^N S_k. \quad (2.82)$$

Przykład 2.16. Dalej rozważamy określony definicją 1.15 plan losowania próby Bernoulliego, lecz teraz zakładamy, że prawdopodobieństwo π wylosowania k -tego elementu populacji do próby nie jest znane. Wówczas z własności rozkładu Bernoulliego wynika, że statystyka

$$\pi_S = \frac{n_S}{N} = \frac{\sum_{k=1}^N S_k}{N} \quad (2.83)$$

jest nieobciążonym estymatorem prawdopodobieństwa π , a jej wariancja ma oczywiście postać:

$$D^2(\pi_S) = \frac{\pi(1-\pi)}{N}. \quad (2.84)$$

Zastępując we wzorze (2.77) prawdopodobieństwo π przez jego estymator π_S mamy nowy estymator wartości średniej $\bar{y}$:

$$\hat{y}_S = N \frac{\sum_{k=1}^N y_k S_k}{\sum_{k=1}^N S_k} = N \frac{\sum_{k=1}^N y_k S_k}{n_S} = N \frac{y_s}{n_S}. \quad (2.85)$$

Można wykazać, że statystyka $\hat{y}_s$ jest w przybliżeniu nieobciążonym estymatorem wartości globalnej w populacji, a jej błąd średniokwadratowy w przybliżeniu określa wzór:

$$MSE(\hat{y}_{HTS}) = \frac{1-\pi}{\pi} v(y)_*. \quad (2.86)$$

Wyprowadzenie wzoru przedstawiono w przykładzie 2.21.

2.3.2. Predykcja wartości globalnej

Rozważmy następujący model nadpopulacji:

$$\begin{cases} E_\xi(Y_i) = \mu_i, & D_\xi^2(Y_i) = \sigma_i^2, \quad i = 1, \dots, N, \\ \text{Cov}_\xi(Y_i, Y_j) = 0, & i = 1, \dots, N, \quad j = 1, \dots, N, \quad i \neq j. \end{cases} \quad (2.87)$$

Naszym celem jest predykcja wartości globalnej Y_U na podstawie danych obserwowanych w próbie o ustalonej liczebności n . Najpierw rozważmy zwykły predyktor postaci:

$$\tilde{Y}_S = N \bar{Y}_S, \quad \bar{Y}_S = \frac{1}{n} \sum_{k \in S} Y_k = \frac{1}{n} Y_S. \quad (2.88)$$

Gdy próbę dobierzemy w sposób arbitralny, to da ona ξ -obciążone oceny wartości globalnej Y_U . Dla wykazania tego faktu wystarczy założyć, że próba s składa się z ustalonych z góry elementów populacji. Wtedy otrzymujemy:

$$E(\tilde{Y}_s) = \frac{N}{n} \sum_{i \in s} \mu_i = N \bar{\mu}_s \neq N \bar{\mu} = \sum_{i \in U} \mu_i = \mu. \quad (2.89)$$

ξ -obciążenie predyktora jest następujące:

$$b_\xi(\tilde{Y}_s) = E(\tilde{Y}_s) - E(Y_U) = N(\bar{\mu}_s - \bar{\mu}) = (N - n)(\bar{\mu}_s - \bar{\mu}_{U-s}), \quad (2.90)$$

gdzie: $\bar{\mu}_{U-s} = \frac{1}{N-n} \sum_{i \in U-s} \mu_i$. Rozstęp możliwych wartości oczekiwanych obserwowanych zmiennych oznaczamy przez $\tau = \mu_{\max} - \mu_{\min}$, przy czym $\mu_{\max} = \max_{i=1, \dots, N} \{\mu_i\}$, $\mu_{\min} = \min_{i=1, \dots, N} \{\mu_i\}$. Wtedy otrzymujemy:

$$|b_\xi(\tilde{Y}_s)| \leq (N - n) \tau. \quad (2.91)$$

Stąd wynika, że maksymalny poziom obciążenia maleje wraz ze wzrostem liczebności próby, a znika dopiero wtedy, gdy $n = N$. Wariancja predyktora ma postać:

$$D_\xi^2(Y_s) = \frac{N^2}{n^2} \sum_{i \in s} D^2(Y_i) = \frac{N^2}{n} \bar{\sigma}_s^2, \quad \bar{\sigma}_s^2 = \frac{1}{n} \sum_{i \in s} \sigma_i^2. \quad (2.92)$$

Zatem błąd średniokwadratowy wynosi:

$$MSE_\xi(\tilde{Y}_s) = \frac{N^2}{n} \bar{\sigma}_s^2 + b_\xi^2(\tilde{Y}_s) \leq \frac{N^2}{n} \sigma_{\max}^2 + (N - n)^2 \tau^2, \quad (2.93)$$

gdzie $\sigma_{\max}^2 = \max_{i=1, \dots, N} \{\sigma_i^2\}$.

Dodajmy, że predyktor $\tilde{Y}_s$ byłby nieobciążony, gdyby $E_\xi(Y_i) = \mu$, dla $i = 1, \dots, N$.

Rozważmy teraz ocenę Y_U za pomocą strategii predykcji $(\tilde{Y}_S, P_{bz})$, gdzie P_{bz} jest planem bezzwrotnego losowania próby prostej. Ta strategia daje P -nieobciążone oraz $P\xi$ -nieobciążone oceny Y_U . Jej błąd średniokwadratowy wynosi:

$$MSE_{\xi P}(\tilde{Y}_S, P_{bz}) = E_\xi D^2(\tilde{Y}_S, P_{bz}) = \frac{N(N - n)}{n} (\bar{\sigma}^2 + v_*(\mu)), \quad (2.94)$$

gdzie:

$$\bar{\sigma}^2 = \frac{1}{N} \sum_{i=1}^N \sigma_i^2, \quad v_*(\mu) = \frac{1}{N - 1} \sum_{i=1}^N (\mu_i - \bar{\mu})^2.$$

Zatem randomizacja wyboru próby prowadzi do nieobciążoności predykcji wartości globalnej. Ma to duże znaczenie praktyczne pozwalające na racjonalny wybór jednego z dwóch rozważanych teraz predyktorów. Najpierw wyobraźmy sobie rzadziej występującą sytuację, gdy wartości oczekiwane obserwowanych zmiennych są takie same, czyli $\mu_i = \mu$ dla $i = 1, \dots, N$. Wówczas oczywiście predyktor $\tilde{Y}_S$ jest nieobciążony. Jeśli nieobciążoność predykcji jest najważniejszym celem naszego wnioskowania, to wówczas nie ma potrzeby sięgania po strategię $(\tilde{Y}_S, P_{bz})$, bo to tylko podniesie koszty

wnioskowania. Chodzi tutaj o wydatki (czasem znaczne) na przygotowanie operatu losowania, który jest niezbędny do przeprowadzenia losowania próby.

Naturalne jest teraz pytanie, czy po to, aby dokładnie ocenić wartość Y_U , należy użyć strategii predykcji $(\tilde{Y}_S, P_{bz})$, czy po prostu predyktora $\tilde{Y}_{s_a}$ z próby s_a , która nie jest losowana, lecz wybrana w sposób arbitralny lub celowy (czyli merytorycznie uzasadniony). W drugim przypadku mamy do czynienia z planem losowania, który można nazwać osobliwym, bo $P_a(s) = 1$, gdy $S = s_a$, natomiast $P_a(S) = 0$, gdy $S \neq s_a$. Wprawdzie formalnie tego typu strategię należałoby oznaczyć przez $(\tilde{Y}_S, P_a)$, lecz użycie w tym symbolu predyktora $\tilde{Y}_{s_a}$ nie powinno prowadzić do nieporozumień. Rozważmy teraz następującą różnicę błędów średniokwadratowych:

$$\begin{aligned} \Delta &= MSE_{\xi P}(\tilde{Y}_S, P_{bz}) - MSE_{\xi P_a}(\tilde{Y}_S, P_a) = E_{\xi} D^2(\tilde{Y}_S, P_{bz}) - MSE_{\xi}(\tilde{Y}_{s_a}) = \\ &= \frac{N(N-n)\bar{\sigma}_{U-s}^2 - (N^2 + n^2 - nN)\bar{\sigma}_s^2 + N(N-n)v_*(\mu)}{n} - N^2(\bar{\mu}_s - \bar{\mu})^2, \end{aligned} \quad (2.95)$$

gdzie: $\bar{\sigma}_{U-s}^2 = \frac{1}{N-n} \sum_{i \in U-s} \sigma_i^2$. Stąd wynika, że $\Delta \geq 0$, jeśli

$$\left(1 - \frac{n}{N}\right) \bar{\sigma}_{U-s}^2 - \left(1 + \left(\frac{n}{N}\right)^2 - \frac{n}{N}\right) \bar{\sigma}_s^2 + \left(1 - \frac{n}{N}\right) v_*(\mu) \geq n(\bar{\mu}_s - \bar{\mu})^2. \quad (2.96)$$

Gdy liczebność populacji jest znacznie większa od liczebności próby, to $\Delta \geq 0$, jeśli

$$\bar{\sigma}_{U-s}^2 - \bar{\sigma}_s^2 + v_*(\mu) \geq n(\bar{\mu}_s - \bar{\mu})^2. \quad (2.97)$$

Stąd wynika, że należałoby nie dobierać próby w sposób arbitralny, lecz celowy i to taki, aby wartości wariancji zmiennych obserwowanych w tej próbie były możliwie małe oraz by obciążenie predykcji było także małe. Wówczas predyktor $\tilde{Y}_s$ z takiej próby celowej byłby dokładniejszym predyktorem od strategii $(\tilde{Y}_S, P_{bz})$. Określonych postulatów jednak nie można urzeczywistnić bez dodatkowych informacji o rozkładach zmiennych określanych przez model nadpopulacji.

Teraz rozpatrzmy szczególną sytuację w stosunku do analizowanej. Założmy, że $\mu_k = \mu$, dla $k = 1, \dots, N$. Wtedy ze wzoru (2.96) wynika, że $\Delta \geq 0$, jeżeli

$$\left(1 - \frac{n}{N}\right) \bar{\sigma}_{U-s}^2 \geq \left(1 + \left(\frac{n}{N}\right)^2 - \frac{n}{N}\right) \bar{\sigma}_s^2, \quad (2.98)$$

a dla bardzo dużych liczebności populacji (w stosunku do rozmiaru próby) mamy:

$$\bar{\sigma}_{U-s}^2 \geq \bar{\sigma}_s^2.$$

Stąd, podobnie jak to miało miejsce w przypadku analizy nierówności (2.96) i (2.97), należy tak dobrać próbę, aby wariancje zmiennych losowych obserwowanych w próbie były możliwie małe.

Następny szczególny przypadek występuje przy założeniu, że $\sigma_k = \sigma$, dla $k = 1, \dots, N$. Wtedy ze wzoru (2.96) wynika, że $\Delta \geq 0$, jeśli

$$-\frac{1}{N}\sigma^2 + \left(1 - \frac{n}{N}\right)v_*(\mu) \geq n(\bar{\mu}_s - \bar{\mu})^2, \quad (2.99)$$

a dla bardzo dużych liczebności populacji (w stosunku do rozmiaru próby) mamy:

$$v_*(\mu) \geq n(\bar{\mu}_s - \bar{\mu})^2.$$

Na wariancję $v_*(\mu)$ nie mamy wpływu, bo odzwierciedla rzeczywistość. Zatem, jeśli próba celowa zostałaby tak dobrana, aby predyktor $\tilde{Y}_s$ oceniał wartość globalną w populacji Y_U z małym obciążeniem, to $\tilde{Y}_s$ może dać dokładniejsze oceny Y_U niż strategia $(\tilde{Y}_S, P_{bz})$.

Teraz rozważmy najsilniejsze założenia, gdy $\mu_i = \mu$ oraz $\sigma_i = \sigma$ dla każdego $i = 1, \dots, N$. Wtedy

$$\Delta = -\frac{1}{N}\sigma^2. \quad (2.100)$$

Zatem w tym przypadku strategia $(\tilde{Y}_S, P_{bz})$ jest dokładniejsza od predyktora $\tilde{Y}_s$ dla arbitralnie dobieranej próby. Jednak przy dostatecznie dużej liczebności populacji te dwie metody predykcji są prawie tak samo dokładne.

Rozważmy jeszcze problem występowania tzw. wartości oddalonych (inaczej odstających czy nietypowych). W tym celu w populacji U wyróżniamy dwie niepuste podpopulacje U_1 i U_2 , przy czym $U = U_1 \cup U_2$. Niech próba s bezzwrotnie dobierana dzieli się na dwie podpróby $s_1 \subseteq U_1$ i $s_2 \subseteq U_2$, przy czym $s = s_1 \cup s_2$. Liczebności podpopulacji U_1 , U_2 i prób s_1 , s_2 oznaczamy odpowiednio symbolami N_1 , N_2 i n_1 , n_2 , przy czym $N = N_1 + N_2$, $n = n_1 + n_2$. Ponadto niech $W_1 = \frac{N_1}{N}$, $W_2 = \frac{N_2}{N}$, $w_1 = \frac{n_1}{n}$, $w_2 = \frac{n_2}{n}$. Wtedy obciążenie predyktora $\tilde{Y}_s$ ma postać:

$$b_\xi(\bar{Y}_s) = \frac{N - n}{n} (n_1 \bar{\mu}_{s_1} + n_2 \bar{\mu}_{s_2}) + (N_1 - n_1) \bar{\mu}_{U_1 - s_1} + (N_2 - n_2) \bar{\mu}_{U_2 - s_2}, \quad (2.101)$$

gdzie

$$\bar{\mu}_{s_i} = \frac{1}{n_i} \sum_{k \in s_i} \mu_k, \quad \bar{\mu}_{U_i - s_i} = \frac{1}{N_i - n_i} \sum_{k \in U_i - s_i} \mu_k, \quad i = 1, 2.$$

Dodatkowo zakładając, że $\bar{\mu}_i = \bar{\mu}_{s_i} = \bar{\mu}_{U_i - s_i}$, $i = 1, 2$, mamy

$$b_\xi(\bar{Y}_s) = N((w_1 - W_1)\bar{\mu}_1 + (w_2 - W_2)\bar{\mu}_2) = N(w_1 - W_1)(\bar{\mu}_1 - \bar{\mu}_2). \quad (2.102)$$

Zatem predyktor będzie nieobciążony, gdy $\bar{\mu}_1 = \bar{\mu}_2$.

Teraz załóżmy, że w próbie $s_1 = U_1$, czyli $n_1 = N_1$, są obserwowane tylko wartości oddalone, natomiast pozostała podpopulacja jest jednorodna ze względu na wartości oczekiwane zmiennych, czyli $\mu_k = \bar{\mu}_2$ dla $k \in U_2$. Wtedy ze wzoru (2.102) wynika, że

$$b_{\xi 1}(\bar{Y}_s) = n_1 \frac{N - n}{n} (\bar{\mu}_1 - \bar{\mu}_2). \quad (2.103)$$

Podobnie jak poprzednio załóżmy, że w podpopulacji U_1 znajdują się tylko wartości oddalone, natomiast w próbie s ich nie ma. Zatem $s = s_2 \subseteq U_2$. Wówczas również zakładając, iż $\mu_k = \bar{\mu}_2$ dla $k \in U_2$, na podstawie wzoru (2.102) mamy:

$$b_{\xi 2}(\bar{Y}_s) = N_1 (\bar{\mu}_2 - \bar{\mu}_1). \quad (2.104)$$

Zakładając, że mamy tylko jedną wartość oddaloną, otrzymujemy

$$b_{\xi 1}(\bar{Y}_s) = \frac{N - n}{n} (\bar{\mu}_1 - \bar{\mu}_2), \quad b_{\xi 2}(\bar{Y}_s) = \bar{\mu}_2 - \bar{\mu}_1. \quad (2.105)$$

Wtedy

$$|b_{\xi 1}(\bar{Y}_s)| - |b_{\xi 2}(\bar{Y}_s)| = \frac{N - 2n}{n} |\bar{\mu}_2 - \bar{\mu}_1|. \quad (2.106)$$

Stąd wynika, że jeśli $n < N/2$ to obciążenie predykcji jest większe, gdy wartość oddalona jest w próbie niż w przypadku, gdy jest poza nią. Takiej sytuacji należy w praktyce oczekiwać, bo zwykle liczebność próby jest znacznie mniejsza od rozmiaru populacji. W związku z tym w przypadku celowego doboru próby, należy pomijać przy liczeniu wartości predyktora występującą w próbie wartość oddaloną. To czy ona występuje w próbie można sprawdzić za pomocą odpowiednich testów statystycznych. Wspomnijmy, że najprostszą procedurą identyfikowania ewentualnego występowania wartości oddalonych jest znana zasada tzw. trzech sigm.

Na zakończenie przedstawiamy pewne optymalne strategie predykcji dla rozważanego modelu nadpopulacji określonego wyrażeniem (2.87). Rozważmy liniowy i $p\xi$ -nieobciążony predyktor wartości globalnej Y_U postaci

$$Y_{lS} = \sum_{i \in S} l_{iS} Y_i + l_0,$$

gdzie l_{iS} , $i \in S$ oraz l_0 są pewnymi nielosowymi współczynnikami. Niech w szczególności

$$Y_{\mu S} = \sum_{k \in S} \frac{Y_k - \mu_k}{\pi_k} + \mu, \quad (2.107)$$

wówczas (por. Chaudhuri i Stenger, 2005) dla dowolnego planu losowania próby $P(S)$, takiego że $\pi_k > 0$ dla $k = 1, \dots, N$, strategia predykcji

$(Y_{\mu S}, P(S))$, jest nie mniej precyzyjna od nieobciążonej strategii $(Y_{lS}, P(S))$, czyli:

$$E_{\xi} D^2(Y_{lS}, P(S)) \geq E_{\xi} D^2(Y_{\mu S}, P(S)) = \sum_{k=1}^N \sigma_k^2 \left(\frac{1}{\pi_k} - 1 \right). \quad (2.108)$$

Szczególną postacią określonego wyżej liniowego $p\xi$ -nieobciążonego predyktora Y_{lS} jest jednorodny, liniowy i $p\xi$ -nieobciążony predyktor postaci: $Y_{jIS} = \sum_{i \in S} l_{iS} Y_i$, Przez $P_{\mu}(S)$ oznaczamy plan losowania próby, którego prawdopodobieństwa inkluzji rzędu pierwszego określa wzór:

$$\pi_k = n \frac{\mu_k}{\sum_{i=1}^N \mu_i}, \quad k = 1, \dots, N. \quad (2.109)$$

Szczególnym przypadkiem predyktora liniowego jest statystyka Horvitz-Thompsona:

$$Y_{HTS} = \sum_{k \in S} \frac{Y_k}{\pi_k}. \quad (2.110)$$

Godambe (1955) wykazuje, że w klasie $p\xi$ -nieobciążonych i jednorodnych strategii predykcji wartości globalnej Y_U oznaczonych przez $(Y_{jIS}, P_{\mu}(S))$ najlepszą jest strategia $(Y_{HTS}, P_{\mu}(S))$, czyli:

$$E_{\xi} D^2(Y_{HTS}, P_{\mu}(S)) \leq E_{\xi} D^2(Y_{jIS}, P_{\mu}(S)). \quad (2.111)$$

Trzeba jeszcze dodać, że z praktycznego punktu widzenia strategia predykcji $(Y_{jIS}, P_{\mu}(S))$ nie ma znaczenia, ponieważ m.in. nie jest możliwe wyznaczenie prawdopodobieństw inkluzji określonych wzorem (2.109), jako że nadzieje matematyczne μ_k , $k \in U$, nie są znane z założenia. Można je oceniać w sytuacji, gdy np. istnieje zależność regresyjna między nimi a wartościami zmiennych pomocniczych. Ten problem będzie szerzej rozważany w dalszej części pracy.

2.3.3. Weryfikacja prostego modelu nadpopulacji*

Wcześniej już rozważano problem predykcji wartości globalnej w sytuacji, gdy nie są spełnione założenia modelu nadpopulacji. Zaniedbanie tego stanu rzeczy może doprowadzić np. nawet do znacznego obciążenia predykcji wartości globalnej. Wtedy, jak to pokazano na przykładach w uprzednim punkcie, randomizacja wnioskowania polegająca na losowaniu w odpowiedni sposób próby zamiast na jej celowym doborze może uchronić nas przed obciążeniem predykcji. Drugi sposób postępowania, który pozostawia możliwość wnioskowania na podstawie odpowiednio konstruowanej próby celowej

już wcześniej m.in. w punkcie 2.2.4 był sygnalizowany, a polega on na weryfikacji założeń modelu nadpopulacji. Zajmiemy się tu problemem adaptacji znanych testów statystycznych do weryfikacji prawdziwości założeń modelu nadpopulacji. Innymi słowy będziemy analizowali przydatność wybranych testów statystycznych do weryfikacji hipotezy, że rozkład z próby zmiennej jest zgodny z rozkładem prawdopodobieństwa ξ , który określa właśnie model nadpopulacji.

Niech przedmiotem hipotezy sprawdzanej, określanej symbolem H_0 , będzie model jednorodny populacji określony wzorem (2.51). Wyrażenie (2.87) określa założenia konkurencyjnego modelu nadpopulacji (w stosunku do modelu jednorodnego populacji), które są przedmiotem hipotezy H_1 alternatywnej w stosunku do H_0 . Postawione hipotezy możemy jedynie weryfikować na podstawie obserwowanych w próbie s wartości zmiennej. Dodatkowo zakładając, że występujące w rozważanych modelach nadpopulacji zmienne losowe są niezależne oraz że każda z nich ma rozkład normalny, sformułowane hipotezy zapisujemy następująco:

$$\begin{aligned} H_0: \mathbf{Y}_s^T &\sim N(\mu \mathbf{J}_n, \sigma^2 \mathbf{I}_n) \\ H_1: \mathbf{Y}_s^T &\sim N(\boldsymbol{\mu}_s, \sigma^2 \text{diag } \boldsymbol{\delta}_s), \quad \boldsymbol{\mu}_s \neq \mu \mathbf{J}_n \text{ lub } \boldsymbol{\delta}_s \neq \mathbf{J}_n. \end{aligned}$$

gdzie: $E(\mathbf{Y}_s) = \boldsymbol{\mu}_s = [\mu_1 \dots \mu_n]^T$ i $\boldsymbol{\delta}_s^T = [\delta_1 \delta_2 \dots \delta_n]$, $\delta_i > 0$, $i = 1, \dots, n$. Symbolem $\mathbf{I}_n$ oznaczono macierz jednostkową stopnia n , natomiast $\mathbf{J}_n$ jest n -elementowym wektorem kolumnowym składającym się z samych jedynek. Hipoteza sprawdzana H_0 określa jednorodność rozkładu prawdopodobieństwa wektora $\mathbf{Y}_s$ w tym sensie, że wartości oczekiwane jego składowych są takie same oraz że ich wariancje są również te same. Z kolei hipoteza alternatywna H_1 głosi, iż elementy wektora $\boldsymbol{\mu}_s$ lub wektora $\boldsymbol{\delta}_s$ nie są takie same. Dodajmy jeszcze, iż można określić hipotezy alternatywne bardziej restrykcyjnie. W sytuacji, gdy wiadomo, że rozkłady badanych zmiennych mają tę samą wariancję, to hipoteza alternatywna ma postać:

$$H'_1: \mathbf{Y}_s^T \sim N(\boldsymbol{\mu}_s, \sigma^2 \mathbf{I}_n), \quad \boldsymbol{\mu}_s \neq \mu \mathbf{J}_n. \quad (2.112)$$

Z kolei, gdy wiadomo, iż elementy wektora wartości oczekiwanych są takie same, czyli $\boldsymbol{\mu}_s = \mu \mathbf{J}_n$ to

$$H''_1: \mathbf{Y}_s^T \sim N(\mu \mathbf{J}_n, \sigma^2 \text{diag } \boldsymbol{\delta}_s), \quad \boldsymbol{\delta}_s \neq \mathbf{J}_n. \quad (2.113)$$

Do weryfikacji postawionych hipotez można m.in. użyć testu Tjetiena i Moore'a (1972). Po to, aby wyznaczyć jego sprawdzian należy najpierw zmienne $\{Y_1, \dots, Y_n\}$ obserwowane w próbie s tak uporządkować w ciąg statystyk $(Y^{(1)}, Y^{(2)}, \dots, Y^{(n)})$, aby spełnione były nierówności: $|Y^{(1)} - \bar{Y}| \leq$

$|Y^{(2)} - \bar{Y}| \leq \dots \leq |Y^{(m)} - \bar{Y}|$. Wówczas statystyka testowa Tietjena-Moore'a ma postać:

$$T_k = \frac{1}{nS^2} \sum_{i=1}^{m-k} \left(Y^{(i)} - \bar{Y}_k \right)^2, \quad (2.114)$$

gdzie

$$\bar{Y}_k = \frac{1}{m-k} \sum_{i=1}^m Y^{(i)}, \quad S^2 = \frac{1}{m} \sum_{i=1}^m (Y_i - \bar{Y})^2, \quad \bar{Y} = \frac{1}{m} \sum_{i=1}^m Y_i.$$

Istotnie małe wartości sprawdzianu T_k świadczą przeciwko hipotezie H_0 , a na korzyść H'_1 . Wartości krytyczne testu można znaleźć w książce Smaliaka i Titarienko (1980) dla $k = 1(1)10$ i $m = 3(1)20(5)50$.

Z punktu widzenia stosowania niżej proponowanego testu wygodniej będzie przekształcić wektor $\mathbf{Y}_s$ za pomocą liniowej transformacji w wektor $\mathbf{Z}$, który składa się z $m = n - 1$ elementów oraz $\mathbf{Z} = \mathbf{0}$, gdy hipoteza H_0 jest prawdziwa. To przekształcenie prowadzi do następującej redefinicji hipotez H_0 i H_1 .

$$H_0: \mathbf{Z}^T \sim N(\mathbf{0}_m, \sigma^2 \mathbf{I}_m), \quad (2.115)$$

$$H_1: \mathbf{Z}^T \sim N(\boldsymbol{\mu}, \sigma^2 \text{diag } \boldsymbol{\delta}), \quad \boldsymbol{\mu} \neq \mathbf{0}_m, \text{ lub } \boldsymbol{\delta} \neq \mathbf{J}_m. \quad (2.116)$$

Wywiał (1988) zaproponował następującą statystykę testową.

$$L_m = -\frac{1}{m} \sum_{i=1}^m \ln(Z_i^2) + \ln\left(\sum_{i=1}^m Z_i^2\right). \quad (2.117)$$

Duże wartości sprawdzianu L_m świadczą przeciwko hipotezie H_0 , czyli obszar krytyczny testu jest prawostronny. Niech $\eta_r(X) = E(X - E(X))^r$, $r = 1, 2, 3, \dots$, jest momentem centralnym rzędu r zmiennej losowej X . Gdy hipoteza H_0 jest prawdziwa, to momenty statystyki L_m określają następujące przybliżone wzory (dokładne wzory na te momenty podaje Wywiał, 1988).

$$\begin{aligned} E(L_m) &\approx \ln(2m) + 0,5 + \frac{1}{m}, \\ D^2(L_m) &\approx \frac{2,935}{m} - \frac{2}{m^2}, \\ \eta_3(L_m) &\approx \frac{12,828}{m^2} - \frac{8}{m^3}, \\ \eta_4(L_m) &\approx 3D^2(L_m) + \frac{81,409}{m^3} - \frac{48}{m^4}. \end{aligned}$$

Standardową postać statystyki oznaczmy przez $Q_m = \frac{L_m - E(L_m)}{D(L_m)}$. Gdy hipoteza H_0 jest prawdziwa, to sprawdzian Q_m ma granicznie rozkład normalny standardowy. Ta zbieżność jest jednak wolna. Dlatego w celu aproksymacji kwantyla q_p rzędu p rozkładu standardowego statystyki L_m dla kilkunastu elementowych prób można użyć następującego tzw. przybliżenia Cornish-Fishera:

$$q_p = u_p + \frac{1}{6}\beta_1(u_p^2 - 1) + \frac{1}{24}(\beta_2 - 3)(u_p^3 - 3u_p),$$

gdzie u_p jest p -tym kwantylem rzędu p rozkładu normalnego standardowego, natomiast

$$\beta_1 = \frac{\eta_3(L_m)}{D^3(L_m)}, \quad \beta_2 = \frac{\eta_4(L_m)}{D^4(L_m)}$$

są odpowiednio współczynnikami asymetrii i kurtozy rozkładu zmiennej losowej L_m . Zatem dla dostatecznie licznej próby i poziomu istotności $\alpha = 1 - p$ wartość krytyczna testu wynosi $q_{1-\alpha}$. Wówczas hipotezę H_0 odrzucamy, gdy wartość standardowa sprawdzianu testu nie jest mniejsza od wartości krytycznej, czyli gdy $q \geq q_{1-\alpha}$.

Test Tietjena-Moore'a jest mocniejszy od testu wykorzystującego statystykę L_n w przypadku występowania tylko wartości oddalonych. Podkreślimy jeszcze, że obydwa testy dają tylko racjonalne podstawy do odrzucenia hipotezy H_0 , wówczas oznacza to, iż rozkład zmiennych tworzących próbę jest niejednorodny, przy czym wniosek ten nie jest prawdziwy z prawdopodobieństwem małym, bo nieprzekraczającym poziomu istotności. Gdy test nie odrzuca hipotezy H_0 o jednorodności próby, bynajmniej nie oznacza to, że rozkład zmiennych losowych znajdujących się poza próbą jest jednorodny.

Dodajmy jeszcze, że problem testowania jednorodności próby występuje bardzo często w literaturze statystycznej. Tym zagadnieniem zajmowali się m.in. Galdfeld i Quandt (1965), Grubbs (1950), Račkauskas i Suquet (2006).

2.4. Strategie estymacji zależne od cechy dodatkowej

W niniejszym podrozdziale koncentrujemy uwagę na własnościach znanych estymatorów parametrów populacji wykorzystujących dostępne informacje o wartościach tzw. zmiennych dodatkowych. Ich użycie zwykle prowadzi do podniesienia dokładności estymacji, jeśli cechy dodatkowe są dostatecznie silnie skorelowane ze zmienną badaną.

2.4.1. Estymatory ilorazowe

Załóżmy, że wartość średnia $\bar{x}$ w populacji cechy dodatkowej jest znana. Dysponujemy obserwacjami zmiennych y i x w próbie prostej losowanej

bezzwrotnie o liczebności n , której plan losowania określa definicja 1.13. Wartości średnie z próby prostej zmiennych badanej i dodatkowej oznaczmy odpowiednio przez $\bar{x}_S$ i $\bar{y}_S$. Wówczas estymator ilorazowy wartości globalnej y_U ma postać:

$$y_{IS} = \frac{\bar{y}_S}{\bar{x}_S} x. \quad (2.118)$$

Wariancję statystyki y_{IS} z dokładnością do składnika rzędu $O(n^{-1})$ określają wzory:

$$D^2(y_{IS}, P_{bz}(s)) = \frac{N(N-n)}{n} v_{*I}(y), \quad (2.119)$$

gdzie:

$$v_{*I}(y) = \frac{1}{N-1} \sum_{k=1}^N (y_k - hx_k)^2, \quad h = \frac{\bar{y}}{\bar{x}} \quad (2.120)$$

lub

$$D^2(y_{IS}, P_{bz}(s)) = \frac{N(N-n)}{n} v_*(y) \left(1 + \left(\frac{z(x)}{z(y)} \right)^2 - 2 \frac{z(x)}{z(y)} \rho(x, y) \right), \quad (2.121)$$

gdzie przez $z(x)$, $z(y)$ i $\rho(x, y)$ oznaczono odpowiednio współczynniki zmienności cech x , y i współczynnik korelacji liniowej między tymi zmiennymi, czyli:

$$z(x) = \frac{\sqrt{v_*(x)}}{\bar{x}}, \quad z(y) = \frac{\sqrt{v_*(y)}}{\bar{y}}, \quad \rho(x, y) = \frac{\sum_{k=1}^N (x_k - \bar{x})(y_k - \bar{y})}{(N-1)\sqrt{v_*(x)v_*(y)}}.$$

Estymator y_{IS} daje obciążone oceny średniej, ponieważ (por. np. Cochran, 1963, s. 161 i Konijn, 1973, s. 100):

$$o_I = E(y_{IS}, P_{bz}(s)) - \bar{y} = \frac{N-n}{Nn} \sqrt{v_*(y)} z(x) \left(\frac{z(x)}{z(y)} - \rho(x, y) \right). \quad (2.122)$$

Jeśli współczynnik korelacji $\rho(x, y) > 0$, to obciążenie estymatora wyznaczone względem jego odchylenia standardowego spełnia nierówność (por. np. Cochran, 1963, s. 162 i Konijn 1973, s. 102):

$$\frac{|o_I|}{D(y_{IS}, P_{bz}(s))} \leq z(x). \quad (2.123)$$

Z powyższego oraz ze wzoru określającego wariancję estymatora y_{IS} wynika, że cechę dodatkową należy tak dobrać, aby jej współczynnik zmienności był możliwie mały oraz aby była ona silnie i dodatnio skorelowana ze zmienną badaną.

Wariancję estymatora y_{IS} można oceniać za pomocą dwóch statystyk:

$$D_S^2(y_{IS}, P_{bz}(s)) = \frac{N(N-n)}{n} v_{*IS}(y), \quad (2.124)$$

gdzie:

$$v_{*IS}(y) = \frac{1}{n-1} \sum_{k=1}^N (y_k - h_S x_k)^2 S_k, \quad h_S = \frac{\bar{y}_S}{\bar{x}_S} \quad (2.125)$$

lub

$$D_S^2(y_{IS}, P_{bz}(s)) = \frac{N(N-n)}{n} \left(v_{*S}(y) + h_S^2 v_{*S}(x) - 2h_S v_{*S}(x, y) \right). \quad (2.126)$$

gdzie:

$$v_{*S}(x, y) = \frac{1}{n-1} \sum_{k=1}^N (x_k - \bar{x})(y_k - \bar{y}) S_k \quad (2.127)$$

oraz $v_{*S}(y) = v_{*S}(y, y)$, $v_{*S}(x) = v_{*S}(x, x)$. Z dokładnością do składnika rzędu $O(n^{-1})$ precyzję estymatora ilorazowego z próby prostej y_{IS} względem strategii $(\tilde{y}_S, P_{bz}(s))$, gdzie $\tilde{y}_S = N\bar{y}_S$, pozwala analizować następujący wzór:

$$def(y_{IS}, \tilde{y}_S) = \frac{D^2(y_{IS}, P_{bz}(s))}{D^2(\tilde{y}_S, P_{bz}(s))} = 1 + \frac{z(x)}{z(y)} \left(\frac{z(x)}{z(y)} - 2\rho(x, y) \right). \quad (2.128)$$

Stąd wynika, że estymator ilorazowy z próby prostej jest lepszy od strategii $(\tilde{y}_S, P_{bz}(s))$, gdy:

$$\rho(x, y) > \frac{1}{2} \frac{z(x)}{z(y)}. \quad (2.129)$$

Warunkiem koniecznym do spełnienia powyższej nierówności jest, aby jej prawa strona była mniejsza od liczby jeden, co jest równoważne nierówności $z(x) < 2z(y)$.

Przykład 2.17. Niech celem badania jest ocena wartości globalnych następujących cech 96 gmin województwa katowickiego: y_1 – dochody z podatków i opłat lokalnych w 1995 r., y_2 – dochody z podatku rolnego w 1995 r. (źródło danych o tych zmiennych podano w przykładzie 2.7). Dysponujemy również wartościami średnimi z populacji dwóch cech dodatkowych. Wartość średnia cechy x_1 – liczba ludności w gminie (popul) wynosi 40883.5 osób, natomiast wartość średnia zmiennej x_2 – powierzchnia użytków rolnych w gminie (podrol) wynosi: 3376.9 ha. Na podstawie próby prostej o liczebności 20 gmin wyliczono wartości średnie zmiennych: y_1 , y_2 , x_1 i x_2 , które odpowiednio wynoszą: 11770.8, 104.7, 71269.1 i 3683.2. Współczynniki korelacji z próby między zmiennymi wchodzącymi w skład par: (y_1, x_1)

$i(y_2, x_2)$ wynoszą odpowiednio 0.930 i 0.877. Współczynniki zmienności z próby zmiennych: y_1 , y_2 , x_1 i x_2 wynoszą odpowiednio: 1.013, 1.163, 1.126, 0.720. Zatem $r(y_1, x_1) = 0.930 > 0.556 = 0.5z(x_1)/z(y_1)$ oraz $r(y_2, x_2) = 0.877 > 0.31 = 0.5z(x_2)/z(y_2)$. Te nierówności sugerują więc użycie estymatora ilorazowego do oceny wartości średnich cech y_1 , y_2 w populacji. Na podstawie wzoru (2.118) wnioskujemy, że wartości estymatora y_{IS} dla cech y_1 i y_2 wynoszą odpowiednio: 649620.8 i 5616.0. W końcu na podstawie wzoru (2.126) wyliczamy oceny błędu, które wynoszą 41625.6 i 566.4 odpowiednio w przypadku cechy y_1 i y_2 , a względne błędy średnie szacunku wynoszą: 6.4% i 6.2%.

Rozważany estymator ilorazowy z próby prostej (losowanej bezzwrotnie) jest tylko asymptotycznie nieobciążonym estymatorem wartości globalnej w populacji. Okazuje się, że gdy zamiast losowania prostego zastosujemy schemat losowania realizującego plan losowania próby proporcjonalny do średniej z próby cechy dodatkowej, to estymator ilorazowy daje nieobciążone oceny globalnej w populacji. Zatem $E(y_{IS}, P_{LM}(s)) = y_U$, gdzie plan losowania $p_{LMS}(s)$ określa definicja 1.17.

Wraz ze wzrostem liczebności populacji, prawdopodobieństwa inkluzji rzędu pierwszego i drugiego planu losowania $p_{LM}(s)$ są prawie równe odpowiednim prawdopodobieństwom inkluzji planu losowania bezzwrotnego próby prostej, co wynika z definicji 1.13, 1.17 i ze wzoru (1.8). To prowadzi do tego, iż wraz z nieograniczonym wzrostem liczebności populacji wariancje obu strategii $(y_{IS}, P_{LM}(s))$ i $(y_S, P_b(s))$ są sobie równe. Z badań symulacyjnych prowadzonych m.in. przez Wywiśla (2000) wynika, że już przy stosunkowo małych liczebnościach populacji, rzędu kilkuset, wariancje obu strategii są prawie sobie równe. W związku z tym w praktyce strategię $(y_{IS}, P_{LM}(s))$ należy stosować w przypadku niezbyt licznych populacji.

Rzeczą naturalną jest, aby w przypadku dowolnego nieprostego schematu losowania próby do estymacji wartości globalnej wykorzystać iloraz statystyk typu Horvitz-Thompsona. Wiadomo, że estymator Horvitz-Thompsona daje nieobciążone oceny wartości globalnej w populacji (gdy wszystkie prawdopodobieństwa inkluzji rzędu pierwszego są dodatnie). Rozważmy następujący estymator ilorazowy:

$$y_{HTS} = \frac{y_{HTS}}{x_{HTS}} x. \quad (2.130)$$

Statystyka y_{HTS} daje obciążone oceny przeciętnej w populacji, lecz obciążenie to zwykle szybko maleje wraz ze wzrostem rozmiaru próby. Przybliżony wzór na jego wariancję jest postaci:

$$D^2(y_{HTS}) = D^2(y_{HTS}) + h^2 D^2(x_{HTS}) - 2h \text{Cov}(x_{HTS}, y_{HTS}), \quad (2.131)$$

gdzie wariancje $D^2(x_{HTS})$ i $D^2(y_{HTS})$ są wyznaczane na podstawie wzorów (2.61)–(2.63), natomiast kowariancję określa wzór:

$$\text{Cov}(x_{HTS}, y_{HTS}) = \sum_{k=1}^N \frac{x_k y_k}{\pi_k} (1 - \pi_k) + \sum_{k=1}^N \sum_{i=1, i \neq k}^N \frac{x_k y_i}{\pi_k \pi_i} (\pi_{k,i} - \pi_k \pi_i). \quad (2.132)$$

Wariancja estymatora Horvitz-Thompsona jest szacowana m.in. za pomocą estymatora określonego wzorem (2.64), natomiast kowariancja za pomocą wzoru:

$$\begin{aligned} \text{Cov}_S(x_{HTS}, y_{HTS}) = \\ = \sum_{k=1}^N \frac{x_k y_k}{\pi_k^2} (1 - \pi_k) S_k + \sum_{k=1}^N \sum_{i=1, i \neq k}^N \frac{x_k y_i}{\pi_k \pi_i} \frac{\pi_{k,i} - \pi_k \pi_i}{\pi_{k,i}} S_k S_i, \end{aligned} \quad (2.133)$$

który jest nieobciążonym estymatorem kowariancji $\text{Cov}(y_{HTS}, x_{HTS})$. Zastępując we wzorze (2.131) wariancje i kowariancje statystyki H-T ich estymatorami otrzymujemy estymator wariancji statystyki y_{HTS} .

W literaturze można znaleźć dużą ilość prac poświęconych strategii ilorazowej estymacji oraz różnym ich modyfikacjom. W szczególności Wywiał (2004, 2008) proponuje następujący estymator typu ilorazowego:

$$\hat{y}_{IS} = \bar{y}_S \frac{E(X_{(r)})}{X_{(r)}},$$

gdzie $E(X_{(r)})$ jest wartością oczekiwaną r -tej statystyki pozycyjnej $X_{(r)}$ zmiennej dodatkowej. Ten estymator daje nieobciążone oceny średniej w populacji, gdy próba jest losowana zgodnie z planem losowania proporcjonalnym do kwantyla z próby cechy dodatkowej, określonego definicją 1.19.

2.4.2. Predyktor ilorazowy

Model nadpopulacji określa wyrażenie:

$$\begin{cases} Y_k = \beta x_k + \xi_k, & k = 1, \dots, N, \\ E(\xi_k) = 0, & D^2(\xi_k) = \sigma^2 x_k, & k = 1, \dots, N, \\ \text{Cov}(\xi_k, \xi_t) = 0, & k \neq t, & k = 1, \dots, N, \quad t = 1, \dots, N. \end{cases} \quad (2.134)$$

Zatem składniki losowe mają wariancje proporcjonalne do odpowiednich wartości zmiennej dodatkowej.

Nieobciążonym predyktorem wartości globalnej w populacji Y_U jest statystyka:

$$Y_{I,s} = \frac{\bar{x}}{\bar{x}_s} \tilde{Y}, \quad (2.135)$$

gdzie:

$$\tilde{Y}_s = N\bar{Y}_s, \quad \bar{Y}_s = \frac{1}{n}Y_s, \quad Y_s = \sum_{k \in s} Y_k.$$

Ten predyktor (zwany ilorazowym) ma najmniejszy błąd średniokwadratowy, który przybiera postać:

$$MSE_{\xi}(Y_{I,s}) = \sigma^2 \frac{N(N-n)}{n} \frac{\bar{x}\bar{x}_{U-s}}{\bar{x}_s} = \sigma^2 N\bar{x} \left(\frac{N\bar{x}}{n\bar{x}_s} - 1 \right). \quad (2.136)$$

Dodajmy, że $MSE_{\xi}(Y_{I,s}) = D_{\xi}^2(Y_{I,s} - Y_U)$, czyli błąd średniokwadratowy predyktora $Y_{I,s}$ jest równy wariancji błędu predykcji $D_{\xi}^2(Y_{I,s} - Y_U)$, gdy predyktor $Y_{I,s}$ jest nieobciążony.

Ze wzoru (2.136) wynika, że błąd średniokwadratowy predyktora $Y_{I,s}$ osiąga minimalną wartość, gdy próba s_* jest celowo tak dobrana, że w jej skład wchodzi te elementy populacji, którym przypisano n -największych wartości zmiennej dodatkowej. Ponadto,

$$MSE_{\xi}(Y_{I,s}, P(s_*)) \leq MSE_{\xi}(Y_{I,s}, P(s)),$$

gdzie $P(s)$ jest dowolnym planem losowania próby różnym od planu losowania próby celowej $P(s_*)$. Zatem losowanie jakiejkolwiek próby (co wiąże się z dodatkowymi kosztami) nie przyniesie zysku na dokładności estymacji.

Dodajmy, że predyktor ilorazowy da się zapisać następująco:

$$Y_{I,s} = Y_s + \hat{Y}_{U-s}, \quad (2.137)$$

gdzie

$$\hat{Y}_{U-s} = \sum_{k \in U-s} \hat{Y}_{ks}, \quad \hat{Y}_{ks} = \beta_s x_k, \quad \beta_s = \frac{1}{n} \sum_{k \in s} \frac{Y_k}{x_k},$$

przy czym β_s jest nieobciążonym estymatorem współczynnika β . Statystyka $\hat{Y}_{U-s}$ jest nieobciążonym predyktorem wartości globalnej zmiennej poza próbą, czyli wielkości $Y_{U-s} = \sum_{k \in U-s} Y_k$.

W przypadku rozważanego tutaj zdefiniowanego wyrażeniem (2.134) modelu określony wzorem (2.93) błąd średniokwadratowy predykcji wartości globalnej Y_U za pomocą statystyki $\tilde{Y}_s$ przyjmuje postać:

$$MSE_{\xi}(\tilde{Y}_s) = \frac{N^2 \bar{x}_s \sigma^2}{n} + b_{\xi}(\tilde{Y}_s)^2,$$

gdzie $b_{\xi}(\tilde{Y}_s) = \beta(\bar{x}_s - \bar{x})$ jest obciążeniem predykcji.

Obciążenie predyktora można więc zlikwidować poprzez taki wybór próby celowej $s_{\#}$, aby $\bar{x}_{s_{\#}} = \bar{x}$. Wtedy

$$MSE_{\xi}(\tilde{Y}_{s_{\#}}) = \frac{N^2 \bar{x} \sigma^2}{n}.$$

Wówczas ze wzoru (2.136) wynika, że dla $\bar{x}_s \geq \bar{x}$ mamy:

$$\frac{MSE_{\xi}(Y_{I,s})}{MSE_{\xi}(\tilde{Y}_{s\#})} = \frac{\bar{x}}{\bar{x}_s} - \frac{n}{N} \leq 1 - \frac{n}{N} = \frac{MSE_{\xi}(Y_{I,s\#})}{MSE_{\xi}(\tilde{Y}_{s\#})}.$$

Zatem dla próby celowej s tak wybieranej, że $\bar{x}_s \geq \bar{x}$ predyktor ilorazowy jest nie mniej precyzyjny niż statystyka $\tilde{Y}_{s\#}$ z próby celowej $s\#$.

Rozważmy teraz nieco ogólniejszy model nadpopulacji:

$$\begin{cases} Y_k = \beta_0 + \beta_1 x_k + \xi_k, & k = 1, \dots, N, \\ E(\xi_k) = 0, & D^2(\xi_k) = \sigma^2 x_k, & k = 1, \dots, N, \\ \text{Cov}(\xi_k, \xi_t) = 0, & k \neq t, & k = 1, \dots, N, \quad t = 1, \dots, N. \end{cases} \quad (2.138)$$

Model ten jest nazywany regresyjnym, a jego szczególnym przypadkiem jest model określony wzorem (2.134). Gdy model (2.138) jest prawdziwy, to predyktor $Y_{I,s}$ jest już ξ -obciążony, ponieważ

$$o_{\xi} = E_{\xi}(Y_{I,s} - Y_U) = \left(\frac{\bar{x}}{\bar{x}_s} - 1 \right) N\beta_0.$$

W tym przypadku $MSE_{\xi}(Y_{I,s}) = D_{\xi}^2(Y_{I,s}) + o_{\xi}^2$, gdzie wariancję określa wzór (2.136). Występujące tutaj obciążenie jest spowodowane niewłaściwym wyborem predyktora w stosunku do modelu nadpopulacji. To obciążenie znika, gdy celowo wybierzemy wcześniej określoną próbę $s\#$. Ten zabieg kalibracyjny powoduje, że predyktor ilorazowy $Y_{I,s\#}$ jest nieobciążony zarówno w przypadku modelu regresyjnego określonego wzorem (2.134), jak i jego uogólnienia zdefiniowanego wyrażeniem 2.138.

Drugim sposobem likwidacji obciążenia predykcji wartości globalnej jest losowanie próby. Z poprzedniego paragrafu wynika, że strategię predykcji wartości globalnej $(Y_{I,s}, P_{bz}(s))$, $(Y_{I,s}, P_{LMS}(s))$ są ξP -nieobciążone, przy czym $P_{bz}(s)$ jest zwykłym planem bezzwrotnego losowania próby prostej, a $P_{LMS}(s)$ planem losowania Lahiriego-Midzuno-Sena, który określa definicja (1.17). Zatem losowanie próby niejako uodparnia wnioskowanie na odchylenia od założonego modelu.

2.4.3. Estymatory regresyjne

Celem estymacji jest dalej ocena wartości globalnej w populacji. Estymator regresyjny jest szczególnym przypadkiem estymatora różnicowego, który ma postać:

$$\tilde{y}_d s = \tilde{y}_s - c(\tilde{x}_s - x_U), \quad (2.139)$$

gdzie c jest pewną stałą, natomiast

$$\tilde{y}_s = \frac{N}{n} \sum_{k \in S} y_k, \quad \tilde{x}_s = \frac{N}{n} \sum_{k \in S} x_k.$$

Statystyka $\tilde{y}_S$ jest nieobciążonym estymatorem średniej w populacji, którego wariancja jest następująca:

$$D^2(\tilde{y}_{dS}, P_{bz}(s)) = \frac{N(N-n)}{n} \left(v_*(y) - 2c\sqrt{v_*(x)v_*(y)}\rho(x, y) + c^2v_*(x) \right). \quad (2.140)$$

Wtedy:

$$\begin{aligned} D^2(\tilde{y}_{dS}, P_{bz}(s)) - D^2(\tilde{y}_S, P_{bz}(s)) &= \\ &= \frac{N(N-n)}{n} \left(c^2v_*(x) - 2c\sqrt{v_*(x)v_*(y)}\rho(x, y) \right). \end{aligned} \quad (2.141)$$

Zatem przy planie losowania bezzwrotnego próby prostej estymator różnicowy $\tilde{y}_{dS}$ z próby prostej jest precyzyjniejszy od strategii $(\tilde{y}_S, P_{bz}(s))$, gdy $2\rho(x, y)\sqrt{\frac{v_*(y)}{v_*(x)}} < c < 0$ i $\rho(x, y) < 0$ albo $0 < c < 2\rho(x, y)\sqrt{\frac{v_*(y)}{v_*(x)}}$ i $\rho(x, y) > 0$.

Analiza własności estymatora różnicowego znacznie uprości się, jeżeli założymy, że $|c| = 1$. Wówczas po to, by oceniać wartość globalną w populacji, trzeba dodatkowo znać wartość średnią zmiennej wspomagającej z populacji i z próby. Wtedy $D^2(\tilde{y}_{dS}, P_{bz}(s)) < D^2(\tilde{y}_S, P_{bz}(s))$, gdy $c = 1$ i $\rho(x, y) > \frac{1}{2}\sqrt{\frac{v_*(x)}{v_*(y)}}$, bądź $c = -1$ i $\rho(x, y) < -\frac{1}{2}\sqrt{\frac{v_*(x)}{v_*(y)}}$.

Wnioskujemy więc, że gdy $|b| = 1$, to estymator różnicowy należy stosować zamiast strategii $(\tilde{y}_S, P_{bz}(s))$ wtedy, gdy cechy x i y są silnie skorelowane oraz wariancja cechy pomocniczej jest mała w stosunku do wariancji zmiennej badanej.

Przykład 2.18. Zadanie polega na ocenie wartości globalnych następujących cech: y_1 – dochody z podatków i opłat lokalnych w 1995 r. (podłok), y_2 – dochody z podatku rolnego w 1995 r. w populacji liczącej 96 gmin (podrol). Dysponujemy również wartościami średnimi z populacji dwóch cech dodatkowych: x_1 – liczba ludności w gminie (popul) oraz x_2 – powierzchnia użytków rolnych w gminie (uzrol). Wartości średnie zmiennych pomocniczych w populacji oraz wybrane parametry z próby wszystkich analizowanych teraz cech znajdziemy w przykładzie 2.17. Dodajmy, że odchylenia standardowe tych zmiennych z próby 20-elementowej wynoszą 11917.4, 121.8, 80213.4 i 2653.4 odpowiednio dla podłok, podrol, popul i uzrol. Wyliczamy, że: $\rho(x_1, y_1) = 0.93 < 3.365 = \frac{1}{2}\sqrt{\frac{v_{*s}(x_1)}{v_{*s}(y_1)}}$, co oznacza, iż estymator różnicowy dla $b = 1$ jest mniej precyzyjny od średniej z próby prostej losowanej bezzwrotnie. Z kolei $\rho(x_2, y_2) = 0.877 > 0.360 = \frac{1}{2}\sqrt{\frac{v_{*s}(x_2)}{v_{*s}(y_2)}}$, co wskazywałoby na to, że estymator różnicowy (dla $b = 1$) średniego podatku rolne-

go będzie precyzyjniejszy od średniej z próby prostej, a jego wartość wynosi $\tilde{y}_{ds} = 104.7 + (3683.2 - 3376.9) = 10656$.

Można łatwo wykazać, że określona wzorem (2.140) wariancja estymatora różnicowego osiąga wartość minimalną, gdy $c = b$, gdzie:

$$b = \rho(x, y) \sqrt{\frac{v_*(y)}{v_*(x)}} = \frac{v_*(x, y)}{v_*^2(x)}, \quad (2.142)$$

gdzie $v_*(x, y) = \frac{1}{N-1} \sum_{k=1}^N (x_k - \bar{x})(y_k - \bar{y})$ jest kowariancją zmiennych x i y . Parametr b jest współczynnikiem regresji liniowej zmiennej y względem x . Wtedy podstawiając prawą stronę równania (2.142) do wzoru (2.139) otrzymujemy tzw. estymator regresyjny, który ma postać:

$$y_{rS} = \tilde{y}_S - b(\tilde{x}_S - x_U). \quad (2.143)$$

Zatem, jeżeli próba prosta jest losowana bezzwrotnie, to estymator y_{rS} jest efektywny w klasie estymatorów różnicowych. Statystyka y_{rS} jest nieobciążonym estymatorem średniej w populacji, a jej wariancja jest następująca:

$$D^2(y_{rS}, P_{bz}(s)) = \frac{N(N-n)}{n} v_*(y) (1 - \rho^2(x, y)). \quad (2.144)$$

Stąd wynika, że

$$def(y_{rS}, P_{bz}(s)) = \frac{D^2(y_{rS}, P_{bz}(s))}{D^2(\tilde{y}_S, P_{bz}(s))} = 1 - \rho^2(x, y). \quad (2.145)$$

Zatem wnioskujemy, że dla próby prostej losowanej bezzwrotnie estymator y_{rS} jest nie mniej efektywny niż strategia $(\tilde{y}_S, P_{bz}(s))$, a jego precyzja rośnie wraz ze wzrostem zależności liniowej między zmiennymi y i x mierzonej zwykłym współczynnikiem korelacji. Dodajmy jeszcze, że estymator regresyjny jest nie mniej efektywny od ilorazowego w próbie prostej, co można wykazać na podstawie wzorów (2.144) i (2.121).

Przypuśćmy teraz, że statystykowi jest znana tylko wartość średnia cechy dodatkowej w populacji. Pozostaje jeszcze problem zdobycia informacji o parametrze b . Szacujemy go na podstawie obserwacji cech y i x w losowanej bezzwrotnie próbie prostej o liczebności n , za pomocą statystyki:

$$b_S = \frac{c_{*S}(x, y)}{v_{*S}^2(x)}, \quad (2.146)$$

gdzie przez $v_{*S}(x)$ i $c_{*S}(x, y)$ określa wzór (2.127). Zastępując we wzorze (2.143) parametr b przez b_S mamy:

$$\tilde{y}_{rS} = \tilde{y}_S - b_S(\tilde{x}_S - x_U). \quad (2.147)$$

Można wykazać, że statystyka $\tilde{y}_{rs}$ jest asymptotycznie nieobciążonym estymatorem wartości globalnej w populacji (przybliżony wzór określający to obciążenie można znaleźć w pracy Wywiśla, 1992), a jej wariancję w przybliżeniu określa wzór (2.144). Estymator tej wariancji określa wzór:

$$D_S^2(\tilde{y}_{rs}) = \frac{N(N-n)}{n} v_{*S}(y) \left(1 - \rho_S^2(x, y)\right), \quad (2.148)$$

gdzie współczynnik korelacji z próby określa wzór:

$$\rho_S^2(x, y) = \frac{v_{*S}(x, y)}{\sqrt{v_{*S}(x) v_{*S}(y)}}.$$

Przez $v_{*S}(x, y)$, $v_{*S}(x)$ i $v_{*S}(y)$ oznaczono nieobciążone oceny z próby prostej odpowiednio kowariancji i wariancji zmiennych x i y , określone wzorem (2.127).

Dodajmy jeszcze, że wykorzystanie określonego definicją 1.18 planu losowania próby Singha i Srivastavy w połączeniu ze zwykłym estymatorem regresyjnym określonym wzorem (2.147) daje nieobciążoną strategię estymacji średniej w populacji.

Przykład 2.19. Rozważmy ponownie dane z przykładów 2.7 i 2.17. Podobnie jak tam celem estymacji są wartości średnie następujących cech: y_1 – dochody z podatków i opłat lokalnych w 1995 r. (podlok), y_2 – dochody z podatku rolnego w 1995 r. (podrol). Przypomnijmy również, że wartości średnie z populacji dwóch cech dodatkowych: x_1 – liczba ludności w gminie (popul) oraz x_2 – powierzchnia użytków rolnych w gminie (uzrol) znajdziemy w przykładzie 2.17. Dodajmy, że odchylenia standardowe tych zmiennych z próby 20-elementowej wynoszą 11 917.4, 121.8, 80 213.4 i 2653.4 odpowiednio dla podlok, podrol, popul i uzrol. Do oceny wartości średnich użyjemy teraz estymatorów regresyjnych. Średnią zmiennej y_1 szacujemy na podstawie liniowej funkcji regresji z próby tej zmiennej względem zmiennej x_1 , która ma postać:

$$y_1 = 0.138(x_1 - 71\,269.1) + 11\,770.8 + u_1, \quad i = 1, \dots, 20.$$

Z kolei funkcja regresji zmiennej y_2 względem x_2 ma postać:

$$y_2 = 0.04(x_2 - 3683.2) + 104.7 + u_2, \quad i = 1, \dots, 20.$$

Ze wzoru (2.147) oraz z powyższych wyrażeń wynika, że $\tilde{y}_{1,rs} = 9615\,964$ i $\tilde{y}_{2,rs} = 9692.5$ są ocenami odpowiednio wartości globalnych y_{1U} i y_{2U} . Korzystając ze wzoru (2.148) wyliczamy oceny błędów średnich szacunku obu estymatorów $D_s(\tilde{y}_{1,rs}) = 96871.5$ i $D_s(\tilde{y}_{2,rs}) = 9611.6$. W końcu oceny względnych średnich błędów szacunku wynoszą odpowiednio 5.5% i 12.5%.

Estymator regresyjny wykorzystujący statystyki H-T ma postać:

$$\tilde{y}_{rHTS} = \tilde{y}_{HTS} - b_{rHTS}(\tilde{x}_{rHTS} - x_U), \quad (2.149)$$

gdzie:

$$b_{rHTS} = \frac{v_{HTS}(x, y)}{v_{HTS}(x)}, \quad (2.150)$$

$$v_{HTS}(x, y) = \frac{1}{n-1} \sum_{k=1}^N (x_k - \bar{x}_{HTS})(y_k - \bar{y}_{HTS}) \frac{S_k}{\pi_k} \quad v_{HTS}(x) = v_{HTS}(x, x).$$

Statystyka $\hat{y}_{rHTS}$ daje obciążone oceny wartości globalnej w populacji, a jej wariancja ma w przybliżeniu postać:

$$D^2(\tilde{y}_{rHTS}) = D^2(\tilde{y}_{HTS}) - 2b \text{Cov}(\tilde{x}_{rHTS}, \tilde{y}_{rHTS}) + b^2 D^2(\tilde{y}_{HTS}). \quad (2.151)$$

Zastępując we wzorze (2.151) wariancje i kowariancje statystyk typu Horvitz-Thompsona ich estymatorami określonymi odpowiednio wzorami (2.64) i (2.133) otrzymujemy estymator wariancji statystyki $\hat{y}_{rHTS}$.

Oprócz przedstawionych estymatorów w literaturze proponuje się wiele innych strategii estymacji wykorzystujących dane o cechach dodatkowych. W szczególności rozważa się tzw. estymator iloczynowy, który w pewnym sensie jest komplementarnym do estymatora ilorazowego, ponieważ zaleca się go stosować w przypadku ujemnie skorelowanych zmiennych badanej i pomocniczej. Analizuje się również wykorzystanie różnych funkcji estymatorów ilorazowych, regresyjnych i iloczynowych do oceny średniej w populacji. W praktyce badań metodą reprezentacyjną nie zawsze można dysponować informacją o wartości średniej cechy pomocniczej w populacji. Wtedy do estymacji średniej w populacji można użyć np. estymatorów ilorazowych lub regresyjnych z tzw. próby podwójnej. Naszkicowane zagadnienia oraz m.in. przypadek, gdy mamy dane o wielu cechach dodatkowych są bardzo szeroko rozważane w literaturze, a w szczególności przez następujących autorów Bracha (1978, 1982, 1983), Cochran (1963), Greń (1969, 1970), Konijn (1962, 1973), Murthy (1977), Tripathi (1973) i Wywiał (1992, 1995).

2.4.4. Predyktory regresyjne

Niech nielosowa macierz $\mathbf{X}_U$ ma wymiary $N \times (k+1)$. Pierwsze k -kolumny tej macierzy to wartości tzw. zmiennych pomocniczych (dodatkowych), a ostatnia kolumna składa się z samych jedynek. Zakłada się, że wartości zmiennych pomocniczych są z góry znane i obserwowane w całej populacji oraz że odzwierciedlają wpływ pewnych głównych czynników na kształtowanie się wartości zmiennej badanej. Przez $\mathbf{x}_i$ oznaczamy i -ty wiersz macierzy $\mathbf{X}_U$, który jest przypisany i -temu elementowi populacji $i = 1, \dots, N$. Założenia tzw. modelu regresyjnego są następujące.

$$\begin{cases} \mathbf{Y}_U = \mathbf{X}_U \boldsymbol{\beta} + \boldsymbol{\xi}_U \\ E(\boldsymbol{\xi}) = \mathbf{0}, \quad \mathbf{V}(\boldsymbol{\xi}_U) = E(\boldsymbol{\xi}_U \boldsymbol{\xi}_U^T), \end{cases} \quad (2.152)$$

gdzie: $\mathbf{Y}_U$ i $\boldsymbol{\xi}_U$ są odpowiednio N -elementowymi wektorami kolumnowymi odpowiednio zmiennych badanych i składników losowych, natomiast $\mathbf{V}(\boldsymbol{\xi}_U)$ jest dodatnio określoną macierzą stopnia N wariancji i kowariancji składników losowych. $\boldsymbol{\beta}^T = [\beta_1 \ \beta_2 \ \dots \ \beta_{k+1}]$ jest wektorem $(k+1)$ nieznanymi parametrów funkcji regresji. Przez $\boldsymbol{\xi}_U$ oznaczono tzw. wektor składników losowych modelu, odzwierciedlających wpływ przypadkowych przyczyn na wartości zmiennej badanej. Zapisane wyrażenia odzwierciedlają zakładaną zależność liniową między wartością oczekiwaną wektora zmiennej badanej a zmiennymi dodatkowymi, czyli $E(\mathbf{Y}_U) = \mathbf{X}_U \boldsymbol{\beta}$.

W niniejszym paragrafie będziemy zakładać, że

$$\mathbf{V}(\boldsymbol{\xi}_U) = \sigma^2 \mathbf{d}_U. \quad (2.153)$$

Z wprowadzonego założenia wynika, że zmienne tworzące wektor $\mathbf{Y}_U$ nie są skorelowane i $\sigma^2 d_i = D^2(\xi_i) = E(\xi_i)^2$, $i = 1, \dots, N$, przy czym d_i jest i -tym elementem macierzy diagonalnej $\mathbf{d}_U$. Wówczas mówi się, iż rozkład prawdopodobieństwa składników losowych jest heteroscedastyczny. Ten rozkład jest homoscedastyczny, gdy $d_i = 1$ dla $i = 1, \dots, N$.

Gdy przyjmiemy, że macierz $\mathbf{X}_U$ redukuje się do kolumny jednostkowej (czyli wektora, którego wszystkie elementy są równe liczbie jeden) oraz $\mathbf{d}_U = \mathbf{I}_U$, to model dany wyrażeniem (2.152) upraszcza się do modelu prostego określonego wyrażeniem (2.51).

Zapiszmy równanie macierzowe znajdujące się w wyrażeniu (2.152) w postaci blokowej:

$$\mathbf{Y}_U = \begin{bmatrix} \mathbf{Y}_s \\ \mathbf{Y}_{U-s} \end{bmatrix} = \begin{bmatrix} \mathbf{X}_s \\ \mathbf{X}_{U-s} \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \boldsymbol{\xi}_s \\ \boldsymbol{\xi}_{U-s} \end{bmatrix},$$

gdzie: $\mathbf{Y}_s$, $\boldsymbol{\xi}_s$ mają wymiary $n \times 1$, $\mathbf{Y}_{U-s}$, $\boldsymbol{\xi}_{U-s}$ mają wymiary $(N-n) \times 1$, a $\mathbf{X}_s$ ma wymiar $n \times (k+1)$, $\mathbf{X}_{U-s}$ ma wymiar $(N-n) \times (k+1)$. Niech $\mathbf{J}_N$ jest wektorem kolumnowym składającym się z N jedynek.

Celem wnioskowania jest ocena wartości globalnej, która w zapisie macierzowym ma postać: $Y_U = \mathbf{J}_N^T \mathbf{Y}_U$. Do jej oceny można użyć następującego predyktora:

$$\hat{Y}_s = \sum_{i \in s} Y_i + \sum_{i \in U-s} \hat{Y}_i, \quad (2.154)$$

gdzie $\hat{Y}_i$ jest prognozą wartości zmiennej losowej Y_i wyznaczaną według wzoru:

$$\hat{Y}_i = \mathbf{x}_i \mathbf{B}_s, \quad i \in U-s, \quad (2.155)$$

przy czym $\mathbf{B}_s$ jest estymatorem metody najmniejszych kwadratów wektora parametrów β , a określa go wzór:

$$\mathbf{B}_s = (\mathbf{X}_s^T \mathbf{d}_U^{-1} \mathbf{X}_s)^{-1} \mathbf{X}_s^T \mathbf{d}_U^{-1} \mathbf{Y}_s. \quad (2.156)$$

Używając wygodnego zapisu macierzowego predyktor $\hat{Y}_s$ można zapisać:

$$\hat{Y}_s = \mathbf{J}_n^T \mathbf{Y}_s + \mathbf{J}_{N-n}^T \hat{\mathbf{Y}}_{U-s}, \quad (2.157)$$

gdzie

$$\hat{\mathbf{Y}}_{U-s} = \mathbf{X}_{U-s} \mathbf{B}_s. \quad (2.158)$$

Błąd predykcji można zapisać następująco:

$$Z_s = Y_U - \hat{Y}_s = \sum_{i \in U-s} Z_{i,s}, \quad \text{gdzie } Z_{i,s} = Y_i - \hat{Y}_i.$$

Gdy spełnione są założenia modelu określone wzorami (2.152) i (2.153), to $\hat{Y}_s$ daje nieobciążone oceny wartości Y_U , czyli $b_\xi(\hat{Y}_s) = E(\hat{Y}_s - Y_U) = 0$. Właśnie przy założeniu, że $b_\xi(\hat{Y}_s) = 0$ błąd średniokwadratowy predykcji ma postać:

$$MSE_\xi(\hat{Y}_s) = D^2(Z_s) = \sum_{i \in U-s} D^2(Z_{i,s}) + \sum_{i,j \in U-s, i \neq j} \text{Cov}(Z_{i,s}, Z_{j,s}), \quad (2.159)$$

gdzie:

$$D^2(Z_{i,s}) = \sigma^2 \left(d_i + \mathbf{x}_i \left(\mathbf{X}_s^T \mathbf{d}_U^{-1} \mathbf{X}_s \right)^{-1} \mathbf{x}_i^T \right),$$

$$\text{Cov}(Z_{i,s}, Z_{j,s}) = \sigma^2 \mathbf{x}_i \left(\mathbf{X}_s^T \mathbf{d}_U^{-1} \mathbf{X}_s \right)^{-1} \mathbf{x}_j^T,$$

przy czym $\text{Cov}(Z_{i,s}, Z_{j,s})$ jest kowariancją błędów predykcji wartości zmiennych Y_i, Y_j , wariancja $D^2(Z_{i,s}) = \text{Cov}(Z_{i,s}, Z_{i,s})$. W zapisie macierzowym wariancja predykcji wartości globalnej Y_U przyjmuje postać:

$$\begin{aligned} MSE_\xi(\hat{Y}_s) &= \\ &= D^2(Z_s) = \sigma^2 \sum_{i \in U-s} d_i + \sigma^2 \mathbf{J}_{N-n}^T \mathbf{X}_{U-s} \left(\mathbf{X}_s^T \mathbf{d}_U^{-1} \mathbf{X}_s \right)^{-1} \mathbf{X}_{U-s}^T \mathbf{J}_{N-n}. \end{aligned} \quad (2.160)$$

Estymator $D_s^2(Z_s)$ wariancji predykcji $D^2(Z_s)$ otrzymujemy poprzez zastąpienie we wzorze (2.160) wariancji σ^2 jej estymatorem postaci:

$$\begin{aligned} \hat{V}_s^2 &= \frac{1}{n-k-1} \sum_{i \in s} (Y_i - \hat{Y}_i)^2 = \\ &= \frac{1}{n-k-1} \mathbf{Y}_s^T \left(\mathbf{I}_n - \mathbf{X}_s \left(\mathbf{X}_s^T \mathbf{d}_U^{-1} \mathbf{X}_s \right)^{-1} \mathbf{X}_s^T \right) \mathbf{Y}_s. \end{aligned} \quad (2.161)$$

Estymator $D_s^2(Z_s)$ daje nieobciążone oceny wariancji $D^2(Z_s)$ ponieważ $\hat{V}_s^2$ jest nieobciążonym estymatorem parametru σ^2 . To zachodzi oczywiście tylko wtedy, gdy model określony wzorem (2.152) jest prawdziwy.

Oceną wartości globalnej Y_U w sytuacji, gdy występuje skorelowanie składników losowych (czyli ich macierz wariancji i kowariancji $\mathbf{V}(\boldsymbol{\xi}_U)$ nie jest diagonalna) zajmują się m.in. Chaudhuri, Stenger (2005).

Przypuśćmy, że prawdziwy jest model regresyjny nadpopulacji, natomiast stosowany jest predyktor $\tilde{Y}_s$ określony wzorem (2.57), który jest poprawny wówczas, gdy jest prawdziwy model prosty. Wtedy ξ -wartość oczekiwana błędu predykcji ma postać:

$$E_\xi(\hat{Y}_s - Y_U) = b_\xi(\hat{Y}_s) = (N - n)(\bar{\mathbf{x}}_s - \bar{\mathbf{x}}_{U-s})\boldsymbol{\beta} = N(\bar{\mathbf{x}}_s - \bar{\mathbf{x}}_U)\boldsymbol{\beta}, \quad (2.162)$$

przy czym $\bar{\mathbf{x}}_s$, $\bar{\mathbf{x}}_U$, i $\bar{\mathbf{x}}_{U-s}$ są wektorami wierszowymi średnich zmiennych dodatkowych odpowiednio z próby, populacji i populacji pomniejszonej o elementy z próby. Predyktor $\tilde{Y}_s$ da nieobciążone oceny Y_U wtedy, gdy zajdzie równość $\bar{\mathbf{x}}_s = \bar{\mathbf{x}}_U$ lub $\bar{\mathbf{x}}_s = \bar{\mathbf{x}}_{U-s}$. To oznacza, że należy tak dobrać próbę s aby jedna z tych równości zachodziła. Zatem, podobnie jak to ma miejsce w przypadku predyktora $\hat{Y}_s$, trzeba znać wartości cech dodatkowych w całej populacji. Gdy próba s jest tak dobrana, aby była spełniona równość $\bar{\mathbf{x}}_s = \bar{\mathbf{x}}_U$, to mówi się, że jest ona skalibrowana właśnie ze względu na tę równość.

Ważny szczególny przypadek modelu regresyjnego wszechstronnie analizowanego przez Royalla (1970 i 1971) dotyczy sytuacji, gdy mamy do czynienia z jedną zmienną dodatkową. Wtedy wyrażenie (2.152) można uprościć do postaci:

$$\begin{cases} Y_k = \beta_1 x_k + \beta_2 + \xi_k, & k = 1, \dots, N, \\ E(\xi_k) = 0, & D^2(\xi_k) = \sigma^2, & k = 1, \dots, N, \\ \text{Cov}(\xi_k, \xi_t) = 0, & k \neq t, & k = 1, \dots, N, \quad t = 1, \dots, N. \end{cases} \quad (2.163)$$

Wówczas najlepszym liniowym i nieobciążonym predyktorem wartości globalnej jest statystyka:

$$\hat{Y}_{1,s} = Y_s + NB_s(\bar{x} - \bar{x}_s), \quad (2.164)$$

gdzie

$$B_s = \frac{\sum_{i \in s} (x_i - \bar{x}_s) Y_i}{\sum_{i \in s} (x_i - \bar{x}_s)^2}.$$

Jego błąd średniokwadratowy ma postać:

$$MSE_\xi(\hat{Y}_{1,s}) = \sigma^2 N^2 \left(\frac{1}{n} + \frac{(\bar{x}_s - \bar{x})^2}{\sum_{i \in s} (x_i - \bar{x}_s)^2} \right). \quad (2.165)$$

Stąd wynika, że błąd średniokwadratowy predyktora $\hat{Y}_{1,s}$ osiąga minimalną wartość, gdy próba s jest celowo tak dobrana, że $\bar{x}_s = \bar{x}$. Wówczas predyktor $\hat{Y}_{1,s}$ redukuje się do predyktora $\hat{Y}_s$ określonego wzorem (2.57).

2.4.5. Weryfikacja regresyjnego modelu nadpopulacji*

Problem weryfikacji prawdziwości modelu regresji określonego wzorem (2.152) można rozważać z punktu widzenia stabilności (jednorodności) wartości współczynników regresji. Przypuśćmy, że są uzasadnione racjonalne przesłanki, iż w dwóch rozłącznych i niepustych podzbiorach U_1 i U_2 , przy czym $U_1 \cup U_2 = U$, wektor parametrów modelu linowego regresji jest różny. Po to, by bardziej szczegółowo analizować ten problem, wprowadźmy następujące oznaczenia oraz dodatkowe założenia. Niech wektor składników losowych modelu (2.152) oznaczony wektorem ξ_U ma sferyczny N -wymiarowy rozkład normalny. Teraz zakładamy jeszcze, $\mathbf{Y}_{U_h} = \mathbf{X}_h \beta_{U_h} + \xi_{U_h}$, $h = 1, 2$, przy czym $\mathbf{Y}_U^T = [\mathbf{Y}_{U_1}^T \ \mathbf{Y}_{U_2}^T]$, $\mathbf{X}_U^T = [\mathbf{X}_{U_1}^T \ \mathbf{X}_{U_2}^T]$. Weryfikowana hipoteza ma postać: $H_0: \beta_{U_1} = \beta_{U_2}$, przeciwko alternatywnej: $H_0: \beta_{U_1} \neq \beta_{U_2}$. Teraz założymy, iż z każdej podpopulacji została wybrana próba, $s_h \subseteq U_h$, $h = 1, 2$ i $s = s_1 \cup s_2$. Niech $\hat{V}_{s_1}^2$ będzie uzyskaną metodą najmniejszych kwadratów oceną wariancji resztowej modelu $\mathbf{Y}_{U_1} = \mathbf{X}_h \beta_{U_1} + \xi_{U_1}$, przy czym $n_1 > k+1$. Wtedy do weryfikacji hipotezy H_0 można użyć tzw. predykcyjnego testu Chowa (1995), którego sprawdzian ma postać:

$$T_C = \frac{(n - k - 1)\hat{V}_s^2 - (n_1 - k - 1)\hat{V}_{s_1}^2}{n_2 \hat{V}_{s_1}^2}, \quad (2.166)$$

przy czym $\hat{V}_s^2$ jest określoną wzorem (2.161) wariancją resztową modelu $\mathbf{Y}_U = \mathbf{x}\beta_U + \xi_U$ dla $\mathbf{d}_U = \mathbf{I}_N$, natomiast

$$\hat{V}_{s_1}^2 = \frac{1}{n_1 - k - 1} \mathbf{Y}_{s_1}^T \left(\mathbf{I}_{n_1} - \mathbf{X}_{s_1} \left(\mathbf{X}_{s_1}^T \mathbf{X}_{s_1} \right)^{-1} \mathbf{X}_{s_1}^T \right) \mathbf{Y}_{s_1}. \quad (2.167)$$

Gdy hipoteza H_0 jest prawdziwa, to statystyka testowa T_C ma rozkład F Snedecora z $n_1 - k$ oraz n_2 stopniami swobody. Obszar krytyczny testu jest prawostronny, czyli istotnie duże wartości tej statystyki świadczą przeciwko hipotezie H_0 . Prezentowany test daje tylko racjonalne podstawy do odrzucenia hipotezy H_0 . Bowiem wówczas oznacza to, iż model regresji w próbie nie jest prawdziwy, przy czym wniosek ten jest fałszywy z prawdopodobieństwem nieprzekraczającym poziomu istotności testu, który z natury rzeczy jest mały.

Dodajmy jeszcze, że problem testowania stabilności modelu regresji występuje bardzo często w literaturze ekonometryczno-statystycznej. Tym zagadnieniem zajmowali się m.in. Chow G.C. (1960, 1995), Galdfeld i Quandt (1965), Maddala (2006), Račkauskas (2006) oraz Wilson (1978).

2.5. Wybrane metody oceny wariancji estymatorów złożonych*

W niniejszym podrozdziale naszkicujemy problem oceny wariancji estymatorów parametrów będących m.in. funkcjami momentów z populacji. Dla uproszczenia prezentacji tych metod będziemy zwykle zakładać, że jest losowana próba prosta. Powszechnie jest znaną metoda wyznaczania błędu średniokwadratowego oceny parametru populacji otrzymana na podstawie rozwinięcia w szereg Taylora jego estymatora. W ogólnym zarysie ta metoda służy m.in. do oceny funkcji momentów z populacji, którą oznaczymy przez $\theta = \theta(\mu_1, \dots, \mu_L)$, gdzie przez μ_k , oznaczono moment rzędu k zmiennej w populacji. Do oceny parametru θ używamy funkcji $\theta_s = \theta(m_{1s}, \dots, m_{Ls})$, gdzie przez m_{kS} oznaczono moment z próby. Funkcja T_s jest rozwijana w szereg Taylora z dokładnością do dwóch jego pierwszych składników w otoczeniu punktu ($m_{1S} = \mu_1, \dots, m_{LS} = \mu_L$) w następujący sposób:

$$\theta(m_{1S}, \dots, m_{LS}) \approx \theta(\mu_1, \dots, \mu_L) + \sum_{k=1}^L \theta_k(m_{kS} - \mu_k), \quad (2.168)$$

przy czym przez $\theta_k = \theta_k(\mu_1, \dots, \mu_L)$ oznaczono pochodną cząstkową funkcji $\theta(\mu_1, \dots, \mu_L)$ względem jej k -tego argumentu wyznaczoną w punkcie ($m_{1S} = \mu_1, \dots, m_{LS} = \mu_L$). Wówczas można wykazać, że:

$$E(\theta(m_{1S}, \dots, m_{LS})) \approx \theta(\mu_1, \dots, \mu_L), \quad (2.169)$$

$$MSE(\theta(m_{1S}, \dots, m_{LS})) \approx \sum_{k=1}^L \sum_{t=1}^L \theta_k \theta_t \text{Cov}(m_{kS}, m_{tS}), \quad (2.170)$$

gdzie przez $\text{Cov}(m_{kS}, m_{tS})$ oznaczono kowariancję momentów z próby m_{kS} i m_{tS} . Z kolei do estymacji błędu średniokwadratowego $MSE(\theta(m_{1S}, \dots, m_{LS}))$ używa się następującej statystyki:

$$MSE_S(\theta(m_{1S}, \dots, m_{LS})) \approx \sum_{k=1}^L \sum_{t=1}^L \theta_{kS} \theta_{tS} \text{Cov}_S(m_{kS}, m_{tS}), \quad (2.171)$$

przy czym θ_{kS} oraz $\text{Cov}_S(m_{kS}, m_{tS})$ są odpowiednio estymatorami pochodnej cząstkowej θ_k oraz kowariancji $\text{Cov}(\cdot, \cdot)$. Wyprowadzony przybliżony wzór pozwala m.in. wyznaczać błędy średniokwadratowe funkcji statystyk typu Horvitz-Thompsona, a w szczególności wcześniej określonych estymatorów ilorazowych lub regresyjnych.

Przykład 2.20. Niech próba prosta będzie losowana w sposób bezzwrotny. Wtedy pierwsze pochodne cząstkowe estymatora ilorazowego określonego

wzorem (2.118) względem średnich $\bar{x}_s$ i $\bar{y}_s$ w punkcie $(\bar{x}_s = \bar{x}, \bar{y}_s = \bar{y})$ określając odpowiednio wzory: $-\frac{\bar{y}}{\bar{x}}$ i 1. Stąd i na podstawie wzoru (2.170) oraz z faktu, że $\text{Cov}(\bar{x}_s, \bar{y}_s) = \frac{N-n}{Nn}v_*(x, y)$ wyprowadzamy już błąd średniokwadratowy estymatora ilorazowego, który jednocześnie jest dla dostatecznie dużej liczebności próby wariancją tej statystyki określoną wzorem (2.121), ponieważ estymator ilorazowy jest granicznie nieobciążony. Podobnie, można wyprowadzić określoną wzorem (2.144) wariancję estymatora regresyjnego z próby prostej.

Przykład 2.21. Wyprowadzamy wzór (2.86) na błąd średniokwadratowy estymatora $\hat{y}_s$ określonego wzorem (2.85). Można wykazać, że $E(n_S) = \pi N$, $E(y_S) = \pi y_u$, $D^2(n_S) = N\pi(1 - \pi)$, $D^2(y_S) = \pi(1 - \pi) \sum_{k \in U} y_k^2$, $\text{Cov}(n_S, y_S) = y_U$. Stąd i na podstawie wzoru (2.170) mamy:

$$MSE(\hat{y}_S) = \frac{N^2}{E^2(n_S)} D^2(y_S) + \frac{N^2 E^2(y_S)}{E^4(n_S)} D^2(n_S) - 2 \frac{N^2 E(y_S)}{E^3(n_S)} \text{Cov}(y_S, n_S).$$

Po odpowiednich uproszczeniach otrzymanego wyrażenia otrzymujemy już wzór (2.86).

Następna metoda oceny wariancji estymatora $\theta_S = \theta(m_{1S}, \dots, m_{LS})$ parametru θ jest nazywana *metodą losowych podprób*. Przypuśćmy, że z populacji jest losowana seria prób prostych $s_1, \dots, s_q$. Ich suma daje próbę s . Na podstawie każdej podpróby o liczebności n_i jest wyznaczana statystyka θ_{S_i} , przy czym jej postać analityczna jest taka sama jak statystyki θ_S . Niech:

$$\bar{\theta}_S = \frac{1}{q} \sum_{i=1}^q \theta_{S_i}. \quad (2.172)$$

Zapisana statystyka również może być traktowana jako estymator parametru θ . Wariancję estymatora θ_S ocenia się za pomocą następujących dwóch statystyk:

$$D_{1S}^2(\theta_S) = \frac{1}{q(q-1)} \sum_{i=1}^q (\theta_{S_i} - \theta_S)^2 \quad (2.173)$$

lub

$$D_{2S}^2(\theta_S) = \frac{1}{q(q-1)} \sum_{i=1}^q (\theta_{S_i} - \bar{\theta}_S)^2, \quad (2.174)$$

przy czym $D_{2S}^2(\theta_S) \leq D_{1S}^2(\theta_S)$.

Kolejny sposób oceny wariancji estymatora θ_S jest nazywany *techniką jackknife*. Upraszczając jej prezentację założmy, że θ_{s-e} , $e = 1, \dots, n$, są ocenami parametru θ wyznaczanymi na podstawie tego samego wzoru co

statystyka θ_S , lecz na podstawie próby s pomniejszonej o e -tą daną obserwowaną w próbie s o liczebności n . Wtedy wyznacza się tzw. pseudowartości:

$$\hat{\theta}_{s,e} = n\theta_s - (n-1)\theta_{s-e}. \quad (2.175)$$

Do estymacji parametru θ można użyć następującej statystyki:

$$\hat{\theta}_S = \frac{1}{n} \sum_{e=1}^n \hat{\theta}_{S,e}. \quad (2.176)$$

Wykazuje się, że dla prób prostych losowanych zwrotnie obciążenie estymatora $\hat{\theta}_s$ jest nie większe od obciążenia estymatora θ_S . Wariancję estymatora θ_S ocenia się za pomocą następujących statystyk:

$$D_{3S}^2(\theta_S) = \frac{1}{n(n-1)} \sum_{e=1}^n (\hat{\theta}_{S,e} - \theta_S)^2 \quad (2.177)$$

lub

$$D_{4S}^2(\theta_S) = \frac{1}{n(n-1)} \sum_{e=1}^n (\hat{\theta}_{S,e} - \hat{\theta}_S)^2. \quad (2.178)$$

Ostatni z prezentowanych sposobów oceny wariancji estymatora θ_s jest nazywany *metodą bootstrap*. Załóżmy, że k -ty element populacji jest dobie-rany do próby S z prawdopodobieństwem π_k , $k = 1, \dots, N$, z populacji U . Wówczas tworzy się (por. np. Särndal i in., 1992) sztuczną populację U^* o liczebności N^* . W praktyce U^* składa się z powtarzających się elementów już wylosowanej próby s , czyli $U^* = \bigcup_{k \in s} U_k^*$, przy czym k -ty element próby s powtarza się N_k^* -krotnie w zbiorze U_k^* . Ponadto niech $y_i^* = y_k$ oraz $\pi_i^* = \pi_k$ dla $i \in U_k^*$, $k \in s$. Można przyjąć, że krotność N_k^* występowania k -tego elementu próby w sztucznej populacji U^* jest tak określona by $N_k^* = \frac{c}{\pi_k}$, przy czym stała c jest tak dobrana, aby wszystkie liczebności N_k^* , $k \in s$ były liczbami naturalnymi (lub w przybliżeniu naturalnymi). Następnie, używając identycznego schematu losowania jak to miało w przypadku doboru próby s z populacji U , losujemy niezależnie n -elementowe próbki S_{*t} , $t = 1, \dots, a$ ze sztucznej populacji U^* . Potem wyliczamy wartości statystyk $\theta_{S_{*t}}$, $t = 1, \dots, a$ według tego samego wzoru analitycznego co estymator θ_S . Wtedy do oceny wariancji estymatora θ_S wykorzystuje się statystykę:

$$D_{5S}^2(\theta_S) = \frac{1}{a-1} \sum_{t=1}^a (\theta_{S_{*t}} - \bar{\theta}_{S*})^2, \quad \bar{\theta}_{S*} = \frac{1}{a} \sum_{t=1}^a \theta_{S_{*t}}. \quad (2.179)$$

Przykład 2.22. Wiadomo, że określona wzorem $V_S = \frac{1}{n} \sum_{i \in S} (Y_i - \bar{Y}_S)^2$ wariancja z próby prostej jest obciążonym estymatorem wariancji σ^2 . Można wykazać, że estymator typu jackknife mający postać $\hat{V}_S = \frac{1}{n-1} \sum_{i \in S} (Y_i - \bar{Y}_S)^2$ jest już nieobciążonym estymatorem parametru σ^2 .

Dodajmy jeszcze, że pewne uogólnienia i więcej szczegółów na temat prezentowanych tu metod wyznaczania wariancji statystyk znajdziemy m.in. w pracach Rao (1982), Särndala i in. (1992) oraz Woltera (1985).

Pytania i zadania

1. Podaj definicję estymatora.
2. Jakie są podstawowe własności, które powinien posiadać dobry estymator?
3. Określ dekompozycję błędu średniokwadratowego estymacji.
4. Porównaj wariancję średnią z prób prostych losowanych zwrotnie i bezzwrotnie.
5. W pięcioelementowej populacji określono zmienne y i x : $y(1) = 1$, $y(2) = 1$, $y(3) = 2$, $y(4) = 3$, $y(5) = 5$; $x(1) = 1$, $x(2) = 2$, $x(3) = 3$, $x(4) = 4$, $x(5) = 5$. Wyznacz rozkłady średniej i mediany zmiennej y z losowanej bezzwrotnie próby prostej trójelementowej. Porównaj dokładność oceny wartości przeciętnej za pomocą obu wymienionych estymatorów.
6. Korzystając z danych z poprzedniego zadania wyznacz rozkład statystyki Horvitz-Thompsona jako estymatora wartości średniej zmiennej y . Zakładamy, że plan bezzwrotnego losowania próby trójelementowej jest proporcjonalny do obserwowanej w niej sumy wartości cechy pomocniczej. Wyznacz wariancję tego estymatora.
7. Wyprowadź wzór na niezbędną liczebność próby prostej losowanej bezzwrotnie, gdy jest ustalony dopuszczalny: a) błąd średni, b) względny średni błąd szacunku przeciętnej w populacji.
8. Wyznacz niezbędną liczebność próby prostej losowanej zwrotnie, gdy względny średni błąd szacunku wskaźnika struktury ma nie przekraczać dopuszczalnego poziomu e .
9. Wyznacz przedział ufności dla wskaźnika struktury w populacji w przypadku zwrotnego losowania próby prostej.
10. Wyznacz tak niezbędną liczebność próby prostej losowanej zwrotnie, aby z ustaloną ufnością i precyzją przedział ufności objął wartość szacowanego wskaźnika struktury.
11. Wyznacz niezbędną liczebność próby prostej, tak aby przedział ufności objął udział kobiet zamieszkujących pewne miasteczko liczące 3000 mieszkańców z ufnością 0.95 i precyzją 0.05.
12. Na podstawie próby prostej losowanej bezzwrotnie o rozmiarze 120 wyliczono, że udział rolników ubezpieczających się dodatkowo wynosi 0.0833. Wyznacz tak niezbędną liczebność próby prostej losowanej bezzwrotnie,

- aby względny średni błąd szacunku wskaźnika struktury nie przekroczył 2%.
13. Od czego i jak zależy długość przedziału ufności dla wartości średniej w populacji?
 14. Od czego i jak zależy precyzja estymacji przedziałowej wartości wskaźnika struktury w populacji?
 15. Wylosowano bezzwrotnie próbę prostą składającą się ze 150 mieszkań pewnej gminy, w której jest 4900 mieszkań. Na podstawie tej próby wyliczono, że średnie zużycie wody w ustalonym roku wynosi 14 m^3 , a nieobciążona ocena wariancji wynosi 36. Ocenic punktowo i przedziałowo wartość przeciętnego zużycia wody w mieszkaniach gminy w ustalonym roku. Niech poziom ufności wynosi 0.99. Na podstawie danych ocen tak niezbędną liczebność próby, aby względny średni błąd szacunku przeciętnego zużycia nie przekroczył poziomu 5% lub postulowana dokładność przedziału ufności nie była większa od 2 m^3 .
 16. Jak należałoby wyznaczać prawdopodobieństwa inkluzji w zależności od wartości cechy dodatkowej, aby wariancja estymatora Horvitz-Thompsona przyjmowała jak najmniejszą wartość?
 17. Przypuśćmy, że próba była losowana z różnymi prawdopodobieństwami inkluzji rzędu pierwszego. Z obserwowanych w takiej próbie wartości cechy wyznaczono wartość średnią w celu estymacji wartości przeciętnej w populacji. Czy ta średnia wyznaczana z próby nieprostej jest obciążonym estymatorem wartości przeciętnej w populacji? Jeśli tak, to od czego to obciążenie zależy?
 18. Przypuśćmy, że próba była losowana z różnymi prawdopodobieństwami inkluzji rzędu pierwszego. Z obserwowanych w takiej próbie danych wyznaczono częstość względną w celu estymacji udziału elementów z cechą wyróżnioną w populacji. Od czego i jak zależy ewentualne obciążenie tego estymatora?
 19. Wyprowadź wzór na wariancję estymatora regresyjnego.
 20. Jak należy wybierać cechę pomocniczą, aby obciążenie i wariancja estymatora ilorazowego były jak najmniejsze?
 21. Wyjaśnij konstrukcje estymatorów ilorazowego, regresyjnego i różnicowego.
 22. Kiedy estymator ilorazowy jest dokładniejszy od średniej z próby prostej?
 23. Od czego i jak zależy obciążenie predyktora ilorazowego?
 24. Jak należy dobrać cechę pomocniczą oraz parametr b estymatora różnicowego, aby jego wariancja była jak najmniejsza?
 25. Określ warunki, przy których estymator różnicowy nie jest gorszy od średniej z próby prostej.
 26. Wyznacz przedział ufności dla wartości średniej w populacji na podstawie estymatorów: a) ilorazowego, b) różnicowego, c) regresyjnego.

27. Jak należy wybierać cechę pomocniczą, aby wariancja estymatora regresyjnego była jak najmniejsza?
28. Porównaj dokładność estymatora regresyjnego z estymatorem ilorazowym.
29. Porównaj dokładność następujących estymatorów wartości przeciętnej w populacji wyznaczanych na podstawie próby prostej losowanej bezzwrotnie: średniej z próby prostej, estymatora ilorazowego i estymatora regresyjnego.
30. Wyjaśnij metodę oceny wariancji estymatora z wykorzystaniem jego rozwinięcia w szereg Taylora.
31. Niech próba prosta będzie losowana w sposób bezzwrotny. Wyznacz wariancję ilorazu średnich $h_S = \frac{\bar{y}_S}{\bar{x}_S}$.
32. Opisz istotę estymacji wariancji metodą jackknife.
33. Na czym polega ocena wariancji estymatora metodą podprób?
34. Wyjaśnij istotę konstrukcji regresyjnego modelu nadpopulacji.
35. Opisz wpływ wartości oddalonych na błąd średniokwadratowy predykcji.
36. Jak należy wybrać próbę, aby predyktor ilorazowy przyjmował jak najmniejszą wartość?
37. Zanalizuj wpływ założeń modelu nadpopulacji na obciążenie i błąd średniokwadratowy predykcji.
38. Określ zależność błędu średniokwadratowego predyktora regresyjnego od wyboru próby.
39. Podaj własność wybranych modeli nadpopulacji.
40. Zdefiniuj nadpopulację.
41. Określ zalety i wady zrandomizowanych strategii predykcji.

ROZDZIAŁ 3

Popularne strategie estymacji

3.1. Średnia z próby systematycznej

3.1.1. Prosta próba systematyczna

Rozważymy tutaj najprostszy przypadek systematycznego losowania próby, gdy jej liczebność n jest liczbą naturalną otrzymaną z ilorazu N/q , gdzie parametr q jest nazywany odstępem (interwałem) losowania systematycznego. Schemat losowania próby s polega na tym, że spośród q -pierwszych elementów populacji losuje się jeden z nich, a następnie wybiera się te elementy, których numery są odległe od numeru pierwszego wylosowanego elementu o krotności liczby q . Powiedzmy, że wylosowano i_e -ty element populacji spośród q -pierwszych elementów. Wówczas do próby s_e wejdą elementy o numerach $i_e, i_{e+q}, i_{e+2q}, \dots, i_{e+q}, \dots, i_{e+(n-1)q}$ lub w zapisie bardziej syntetycznym $s_e = (i_{e+eq}; e = 0, 1, \dots, n-1)$. Zatem przestrzeń wszystkich możliwych do wylosowania prób składa się z q -prób tworzących zbiór $\mathbf{Q} = \{s_1, \dots, s_q\}$. Wtedy plan losowania prostej próby systematycznej można określić wzorem:

$$P_{sys}(s) = \frac{1}{q} \quad \text{dla } s \in \mathbf{Q}. \quad (3.1)$$

Prawdopodobieństwa inkluzji rzędu pierwszego i drugiego są takie same i wynoszą:

$$\pi_i = \frac{1}{q} \quad \text{dla } i = 1, \dots, N, \quad (3.2)$$

$$\pi_{i,j} = \begin{cases} \frac{1}{q} & \text{dla } i \neq j, i \in s, j \in s, \\ 0 & \text{dla } i \neq j, i \notin s, \text{ lub } j \notin s. \end{cases} \quad (3.3)$$

Algorytm schematu losowania próby prostej systematycznej o liczebności n , z interwałem losowania q jest następujący. Należy wygenerować liczbę losową u z rozkładu jednostajnego prawdopodobieństwa określonego na przedziale $[0; 1]$. Próbę wyznacza ciąg numerów elementów w populacji: $s = (i_1, i_1 + q, \dots, i_1 + (n-1)q)$, przy czym element populacji o numerze i_1

wyznacza nierówność:

$$\frac{i_1 - 1}{q} < u \leq \frac{i_1}{q}. \quad (3.4)$$

Przykład 3.1. Gdy $N = 12$, $n = 3$. Wtedy interwał losowania systematycznego prostego wynosi $q = \frac{N}{n} = 4$. Niech liczba losowa o rozkładzie jednostajnym na przedziale $[0, 1]$ wynosi $u = 0.48$. Wówczas powyższy algorytm prowadzi do wyboru próby $s = (2, 6, 10)$, ponieważ dla pierwszego elementu próby, będącego jednocześnie elementem populacji o numerze $i_1 = 2$ zachodzi nierówność:

$$\frac{1}{q} = 0.25 < u = 0.48 < 0.5 = \frac{2}{q}.$$

Do estymacji wartości globalnej w populacji używa się następującej strategii estymacji: $(\tilde{y}_S, P_{sys}(s))$, gdzie:

$$\tilde{y}_S = \frac{N}{n} \sum_{i \in S} y_i. \quad (3.5)$$

Daje ona nieobciążone oceny wartości średniej w populacji, a jej wariancję określa wzór:

$$D^2(\tilde{y}_S, P_{sys}(s)) = \frac{1}{q} \sum_{e=1}^q (\bar{y}_{s_e} - \bar{y})^2, \quad (3.6)$$

gdzie:

$$\bar{y}_{s_e} = \frac{1}{n} \sum_{i \in s_e} y_i = \frac{1}{n} \sum_{e=1}^n y_{i_e + (e-1)q}.$$

Po to, aby porównać precyzję średniej z próby prostej, czyli strategii losowania $(\tilde{y}_S, P_{bz}(s))$ i strategii $(\tilde{y}_S, P_{sys}(s))$, wprowadza się współczynnik jednorodności, który określa wzór (por. Särndal, Swensson i Wretman, 1992):

$$\delta = 1 - \frac{v_{*w}}{v_*}, \quad -\frac{1}{n} \leq \delta \leq 1, \quad (3.7)$$

gdzie:

$$v_* = \frac{1}{N-1} \sum_{i=1}^N (y_i - \bar{y})^2,$$

natomiast v_{*w} jest tzw. przeciętną wariancją wewnątrzgrupową i określają ją wzory:

$$v_{*w} = \frac{1}{q} \sum_{e=1}^q v_{*e}, \quad v_{*e} = \frac{1}{n-1} \sum_{i \in s_e} (y_i - \bar{y}_{s_e})^2. \quad (3.8)$$

Wówczas względny współczynnik efektywności ma postać:

$$deff = \frac{D^2(\tilde{y}_S, P_{sys}(s))}{D^2(\tilde{y}_S, P_0(s))} = 1 + \frac{N(n-1)}{N-n}\delta. \quad (3.9)$$

Stąd więc wynika, że średnia z próby systematycznej jest bardziej precyzyjnym estymatorem przeciętnej w populacji niż średnia z próby prostej, gdy współczynnik jednorodności $\delta < 0$. Oznacza to, że jeśli statystyk może wpływać na uporządkowanie elementów populacji, to winien to tak uczynić, aby przeciętna wariancja wewnątrzgrupowa była większa od wariancji badanej cechy w populacji.

Sposobami oceny wariancji średniej z próby systematycznej zajmował się m.in. Bracha (1996), który rozważał również pewne uogólnienia schematu losowania systematycznego. Analizował sytuację, gdy iloraz N/n nie jest liczbą całkowitą, ponadto opisał przypadek losowania wielu elementów spośród q -pierwszych elementów populacji, co pozwala m.in na ocenę wariancji strategii $(\tilde{y}_S, P_{sys}(s))$.

Przykład 3.2. *Rozważmy 16-elementową populację, z której ma być losowana systematycznie próba 4-elementowa. Zatem interwał losowania wynosi $q = 4$, czyli są możliwe do wylosowania cztery próby. Rozważmy trzy przypadkowe przyporządkowania wartości tego samego zbioru liczb elementom populacji, a dokładniej numerom elementów populacji. To jest równoważne z określeniem trzech różnych zmiennych y_A , y_B i y_C . Wartości tych zmiennych prezentuje tablica 3.1.*

Tabela 3.1

Rozkład wartości zmiennych y_A , y_B , y_C w populacji

k	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
s_e	s_1	s_2	s_3	s_4	s_1	s_2	s_3	s_4	s_1	s_2	s_3	s_4	s_1	s_2	s_3	s_4
y_A	2	4	6	8	2	4	6	8	2	4	6	8	2	4	6	8
y_B	2	2	2	2	4	4	4	4	6	6	6	6	8	8	8	8
Y_C	6	2	4	4	8	6	2	2	4	8	6	8	8	4	2	6

W pierwszym wierszu tej tablicy podano kolejne numery elementów populacji, a w drugim przynależność każdego z nich do jednej z możliwych czterech prób. W kolejnych trzech wierszach podano wartości zmiennych. Wyliczamy, że wartość średnia w populacji każdej z trzech zmiennych wynosi $\bar{y} = 5$, a wariancja $v_* = 5\frac{1}{3}$. Wariancje wewnątrzgrupowe dla rozkładów zmiennych y_A , y_B , i y_C odpowiednio wynoszą $v_{B*w} = 6\frac{2}{3}$, $v_{A*w} = 0$ i $v_{C*w} = 6.5$. Z kolei współczynniki jednorodności wartości cechy dla wyróżnionych rozkładów zmiennych wynoszą $\delta_B = -0.25$, $\delta_A = 1$, $\delta_3 = -\frac{7}{32} = -0.219$.

W przypadku rozważanych rozkładów zmiennych y_A , y_B , i y_C mamy: $\text{deff}_A = D^2(\tilde{y}_S, P_{\text{sys},A}(s)) = 5$, $\text{deff}_B = D^2(\tilde{y}_S, P_{\text{sys},B}(s)) = 0$, $\text{deff}_C = D^2(\tilde{y}_S, P_{\text{sys},C}(s)) = 0.875$, ponieważ $D^2(\tilde{y}_S, P_0(s)) = 1$.

Najbardziej precyzyjna jest ocena wartości globalnej w populacji dla rozkładu zmiennej y_B , ponieważ prowadzi on do najmniejszej wariancji $D^2(\tilde{y}_S, P_{\text{sys},B}(s)) = \text{deff}_B = 0$. Wówczas wszystkie wariancje wewnątrzgrupowe są takie same i przyjmują największą możliwą wartość, która jest równa średniej wariancji wewnątrzgrupowej wynoszącej $v_{B*W} = 6\frac{2}{3}$. Z tabeli 3.1 wynika, że w tym przypadku częstości występowania wartości cechy w populacji i w możliwych do wylosowania próbach są takie same, czyli w pewnym sensie z tego punktu widzenia każda próba jest miniaturą populacji i średnie z tych wszystkich prób są równe średniej w populacji. Zatem wariancja tych możliwych średnich musi być równa zeru, czyli $D^2(\tilde{y}_S, P_{\text{sys},B}(s)) = 0$.

W przypadku zmiennej y_A mamy do czynienia z odwrotną sytuacją, bo wariancja strategii $(\tilde{y}_S, P_{\text{sys},A}(s))$ jest maksymalna i $D^2(\tilde{y}_S, P_{\text{sys},A}(s)) = \text{deff}_A = 5$. Zatem w tym przypadku strategia $(\tilde{y}_S, P_{\text{sys},A}(s))$ jest pięciokrotnie mniej precyzyjna od strategii $(\tilde{y}_S, P_{bz}(s))$. Mamy tu do czynienia z sytuacją, gdy stopień zróżnicowania wartości zmiennej wewnątrz prób (grup) jest najmniejszy, bo wariancja wewnątrzgrupowa wynosi $v_{A*W} = 0$. Spowodowane to jest tym (por. tabela 3.1), że w potencjalnie możliwych do wylosowania próbach powtarzają się te same wartości cechy. Wówczas współczynnik niejednorodności przyjmuje największą możliwą wartość $\delta_A = 1$.

Przypadek zmiennej y_C jest w pewnym sensie pośredni w stosunku do dwóch poprzednio rozważanych. W tym przypadku współczynnik jednorodności $\delta_3 = -0.219$, czyli mamy do czynienia z dużym, chociaż nie największym, zróżnicowaniem wartości cechy w grupach (czyli możliwych do wylosowania próbach). Prowadzi to do bardziej precyzyjnej estymacji wartości globalnej w populacji za pomocą strategii $(\tilde{y}_S, P_{\text{sys},C}(s))$ niż by to miało miejsce przy estymacji z wykorzystaniem strategii $(\tilde{y}_S, P_{bz}(s))$, ponieważ współczynnik efektywności względnej wynosi $\text{deff}_C = 0.875 = D^2(\tilde{y}_S, P_{\text{sys}}(s))$. Zatem wariancja strategii $(\tilde{y}_S, P_{\text{sys},C}(s))$ stanowi 87.5% wartości wariancji strategii $(\tilde{y}_S, P_{bz}(s))$.

Zasygnalizujemy jeszcze przypadek występowania wahań cyklicznych wartości oczekiwanych ciągu zmiennych obserwowanych w populacji uporządkowanej oraz gdy interwał losowania próby systematycznej jest równy krotności długości cyklu wahań. Wówczas strategia $(\tilde{y}_S, P_{\text{sys}}(s))$ daje P -nieobciążoną ocenę wartości globalnej, lecz zwykle dla każdej możliwej próby losowanej zgodnie z planem systematycznym estymator $\tilde{y}_s$ istotnie może różnić się od wartości globalnej. Ilustruje to następujący przykład.

Przykład 3.3. Niech nadpopulacja jest następującym ciągiem wartości $(a, b, a, b, \dots, b, a, b)$, gdzie $a \neq b$. Załóżmy jeszcze, że liczebność populacji

jest parzysta, co prowadzi do wniosku, że $\bar{y} = (a+b)/2$ oraz $y = N(a+b)/2$. Jeśli założymy, że interwał losowania systematycznego jest liczbą parzystą, to w każdej możliwej do wylosowania próbie będą obserwowane albo same wartości równe liczbie a , albo liczbie b . Wówczas, dla każdej próby możliwe są jedynie dwie wartości średniej z próby $\bar{y}_s = a$, albo $\bar{y}_s = b$. Zatem każda z tych dwóch możliwych ocen nie będzie równa średniej w populacji, jakkolwiek strategia $(\bar{y}_s, P_{sys}(s))$ jest P -nieobciążona dla $\bar{y}$. Podobny wniosek oczywiście dotyczy również ocen wartości globalnej za pomocą strategii $(\tilde{y}_s, P_{sys}(s))$.

3.1.2. Prawdopodobieństwa inkluzji proporcjonalne do wartości zmiennej dodatkowej*

Zakładamy, że $\pi_k \propto x_k$, $k = 1, \dots, N$. Wtedy zgodnie z własnością $\sum_{k=1}^N \pi_k = n$, należałoby przyjąć, że:

$$\pi_k^{(0)} = \frac{nx_k}{\sum_{i=1}^N x_i}, \quad k = 1, \dots, N. \quad (3.10)$$

Może się jednak zdarzyć, że dla pewnego e wartość x_e jest na tyle duża, iż $\pi_e > 1$. Wówczas przyjmuje się (por. np. Bracha, 1998 lub Tillé, 2006), że $\pi_e = 1$. Oznacza to, że e -ty element populacji będzie celowo dobierany do próby, a prawdopodobieństwa doboru pozostałych elementów należy skorygować. Bardziej szczegółowo ten problem opisujemy następująco. Wyznaczamy taki wskaźnik h spośród $k = 1, \dots, N$, że

$$\pi_h^{(0)} = \text{maximum}_{k \in U} \{ \pi_k^{(0)} \},$$

gdzie, przypomnijmy, przez U oznaczono populację, a dokładnie zbiór numerów elementów populacji. Jeśli $\pi_h^{(0)} \leq 1$, to przyjmujemy, że $\pi_k = \pi_k^{(0)}$, dla $k = 1, \dots, N$, to już koniec algorytmu wyznaczania wartości prawdopodobieństw inkluzji.

W przypadku, gdy $\pi_h^{(0)} > 1$, to przyjmujemy, że $\pi_h = 1$, i wyznaczamy następujące ilorazy:

$$\pi_k^{(1)} = \frac{(n-1)x_k}{\sum_{i=1}^N x_i - x_h}, \quad k = 1, \dots, N; \quad k \neq h.$$

Znów wyznaczamy wskaźnik g tak, aby

$$\pi_g^{(1)} = \text{maximum}_{k \in \{U-h\}} \{ \pi_k^{(1)} \}.$$

Jeśli $\pi_k^{(1)} \leq 1$, to przyjmujemy, że $\pi_h = 1$, $\pi_k = \pi_k^{(1)}$ dla $k = 1, \dots, N$, przy czym $k \neq h$. To kończy algorytm wyznaczania wartości prawdopodobieństw inkluzji.

Jeśli $\pi_g^{(1)} > 1$, to przyjmujemy, że $\pi_h = 1$, $\pi_g = 1$, i ponownie wyznaczamy odpowiednie ilorazy.

Uogólniając przedstawione rozumowanie przypuścimy, że mamy do czynienia z e -tym etapem algorytmu, przy czym $e = 1, \dots, r < n$, $\pi_{k_i} = 1$ dla $i = 1, \dots, e - 1$ oraz

$$\pi_k^{(e)} = \frac{(n - e)x_k}{\sum_{i=1}^N x_i - \sum_{h=1}^e x_{k_h}}, \quad k \in \{U - \{k_1, \dots, k_e\}\}.$$

Wyznaczamy wskaźnik k_{e+1} tak, aby

$$\pi_{k_{e+1}}^{(e)} = \text{maximum}_{k \in \{U - \{k_1, \dots, k_e\}\}} \left\{ \pi_k^{(e)} \right\}.$$

Jeśli $\pi_{k_{e+1}}^{(e)} \leq 1$, to przyjmujemy, że $\pi_{k_i} = 1$ dla $i = 1, \dots, e$ oraz $\pi_k = \pi_{k_i}^{(e)}$ dla $k \in \{U - \{k_1, \dots, k_e\}\}$. Wtedy algorytm wyznaczania prawdopodobieństw inkluzji zatrzymujemy.

W przeciwnym przypadku, gdy $\pi_{k_{e+1}}^{(e)} > 1$, to kontynuujemy algorytm wyznaczania tych prawdopodobieństw, przyjmując, że $\pi_{k_{e+1}} = 1$ i wyliczamy:

$$\pi_k^{(e+1)} = \frac{(n - e - 1)x_k}{\sum_{i=1}^N x_i - \sum_{h=1}^{e+1} x_{k_h}}, \quad k \in \{U - \{k_1, \dots, k_{e+1}\}\}.$$

Przykład 3.4. *Populacja składa się z $N = 6$ elementów, a próba z $n = 2$ elementów. Obserwacje zmiennej pomocniczej są następujące: $x_1 = 6$, $x_2 = 4$, $x_3 = 100$, $x_4 = 44$, $x_5 = 10$, $x_6 = 6$. Korzystając ze wzoru (3.10) mamy: $\pi_3^{(0)} = \frac{200}{170} > 1$. Przyjmujemy, że $\pi_3 = 1$, i wyliczamy, że: $\pi_1 = \pi_6 = \frac{3}{35}$, $\pi_2 = \frac{2}{35}$, $\pi_4 = \frac{22}{35}$ i $\pi_5 = \frac{5}{35}$. Można sprawdzić, że $\sum_{k=1}^6 \pi_k = 2$.*

Przykład 3.5. *Populacja liczy $N = 6$ elementów, a próba $n = 3$ elementy. Obserwacje zmiennej pomocniczej są następujące: $x_1 = 6$, $x_2 = 4$, $x_3 = 8$, $x_4 = 42$, $x_5 = 2$, $x_6 = 8$. Korzystając ze wzorów (3.10) mamy: $\pi_4^{(0)} = \frac{126}{70} > 1$. Przyjmujemy, że $\pi_4 = 1$, i wyliczamy, że: $\pi_1 = \frac{3}{7}$, $\pi_2 = \frac{2}{7}$, $\pi_3 = \pi_6 = \frac{4}{7}$ i $\pi_5 = \frac{1}{7}$.*

Hartley i Rao (1962) zaproponowali następujący schemat losowania próby z zadanymi prawdopodobieństwami inkluzji rzędu pierwszego. Niech $\pi_k = np_k$, $p_k > 0$, $k = 1, \dots, N$, $\sum_{k=1}^N p_k = 1$. Oznaczmy próbę jako ciąg kolejno losowanych do niej elementów postaci: $s = (i_1, i_2, \dots, i_n)$. Najpierw wyliczamy sumę:

$$V_e = \sum_{k=1}^e \pi_k,$$

przy czym $V_0 = 0$ i $V_N = n$. Następnie jest generowana liczba losowa o rozkładzie jednostajnym z przedziału $(0; 1)$. Element populacji z kolei o numerze i_1 należy dobrać do próby, czyli $i_1 \in s$, gdy zachodzi następująca nierówność:

$$V_{i_1-1} < u \leq V_{i_1}.$$

Drugim elementem próby jest element populacji indeksowany przez i_2 , przy czym musi być spełniona nierówność:

$$V_{i_2-1} < u + 1 \leq V_{i_2}.$$

W ogólności, element $i_z \in s$, $z = 1, \dots, n$ wtedy i tylko wtedy, gdy:

$$V_{i_z-1} < u + z - 1 \leq V_{i_z}. \quad (3.11)$$

Podane przez Rao i Hartleya przybliżone wzory dla prawdopodobieństw inkluzji rzędu drugiego są następujące (por. Bracha, 1996).

$$\begin{aligned} \pi_{k,t} = n(n-1)p_k p_t \left(1 + p_k + p_t - A + 2 \left(p_k^2 + p_t^2 + p_k p_t (1 - B) \right) \right) + \\ - 3n(n-1)p_k p_t A (p_k + p_t - A) + O(N^{-4}), \end{aligned} \quad (3.12)$$

gdzie

$$A = \sum_{k=1}^N p_k^2, \quad B = \sum_{k=1}^N p_k^3.$$

Przedstawiony algorytm wyznaczania elementów tworzących próbę dotyczy również określonego w uprzednim podrozdziale schematu losowania prostej próby systematycznej.

Dodajmy jeszcze, że opisany algorytm prowadzi do losowania próby z pominięciem planu losowania próby, który, jak przekonaaliśmy się, nie jest niezbędny przy tworzeniu schematu losowania próby. Ta sytuacja jest zgodna ze znanym faktem, że dla ustalonych prawdopodobieństw inkluzji rzędu pierwszego istnieje co najmniej jeden plan losowania próby. Tym zagadnieniem zajmowali się m.in. Henzel (1986), Gabler i Scheigkoffer (1990) oraz Singha (1973).

Przykład 3.6. Dla danych z przykładu 3.4 losujemy próbę systematyczną dwuelementową. Niech wygenerowana wartość zmiennej losowej o rozkładzie jednostajnym na przedziale $(0; 1)$ wynosi 0.48. Wówczas korzystając ze wzoru (3.11) mamy:

$$\begin{aligned} i_1 = 3, \quad \text{ponieważ} \quad V_2 = \frac{5}{35} < u = 0.48 < 1 \frac{5}{35} = V_3, \\ i_1 = 4, \quad \text{ponieważ} \quad V_3 = 1 \frac{5}{35} < u + 1 = 1.481 < 1 \frac{27}{35} = V_4. \end{aligned}$$

Zatem próba składa się z elementów nr 3 i nr 4, czyli $s = \{3, 4\}$.

Przykład 3.7. Dla danych z przykładu 3.5 losujemy próbę systematyczną trójelementową. Niech wygenerowana wartość zmiennej losowej o rozkładzie jednostajnym na przedziale $(0; 1)$ wynosi 0.21. Wówczas korzystając ze wzoru (3.11) otrzymujemy próbę: $s = \{1, 3, 4\}$, ponieważ:

$$\begin{aligned} i_1 &= 1, \quad \text{bo} \quad V_0 = 0 < u = 0.21 < \frac{3}{7} = V_1, \\ i_1 &= 3, \quad \text{bo} \quad V_2 = \frac{5}{7} < u + 1 = 1.21 < 1\frac{2}{7} = V_3, \\ i_1 &= 4, \quad \text{bo} \quad V_3 = 1\frac{2}{7} < u + 2 = 2.21 < 2\frac{2}{7} = V_4. \end{aligned}$$

Przykład 3.8. Niech ciąg prawdopodobieństw inkluzji rzędu pierwszego dla próby trójelementowej ma postać: $\pi_1 = 0.2$, $\pi_2 = 0.8$, $\pi_3 = 0.05$, $\pi_4 = 0.3$, $\pi_5 = 0.5$, $\pi_6 = 0.15$, $\pi_7 = 0.1$, $\pi_8 = 0.3$, $\pi_9 = 0.25$, $\pi_{10} = 0.35$. Wygenerowana wartość zmiennej losowej o rozkładzie jednostajnym na przedziale $(0; 1)$ wynosi 0.8. Wówczas korzystając ze wzoru (3.11) mamy:

$$\begin{aligned} i_1 &= 2, \quad \text{bo} \quad V_1 = 0.2 < u = 0.8 < 1.0 = V_2, \\ i_2 &= 5, \quad \text{bo} \quad V_4 = 1.35 < u + 1 = 1.8 < 1.85 = V_5, \\ i_3 &= 10, \quad \text{bo} \quad V_9 = 2.65 < u + 2 = 2.8 < 3.0 = V_{10}. \end{aligned}$$

Stąd wynika, że próba ma postać: $s = \{2, 5, 10\}$.

3.1.3. Wnioskowanie przy modelu populacji uporządkowanej

Założmy, że rozkłady prawdopodobieństwa ciągu zmiennych losowych $Y_1, \dots, Y_t, \dots, Y_N$ mogą zależeć od numerów indeksów. Wówczas taki model będziemy nazywali *modelem nadpopulacji uporządkowanej*. W szczególności indeks oznaczony tutaj literą t może oznaczać numer kolejnego okresu czasu lub momentu czasu. Z kolei wartość zmiennej losowej Y_t może oznaczać stan zapasu w magazynie na t -ty moment, cenę akcji w t -tym notowaniu giełdowym, poziom zużycia energii elektrycznej w ciągu t -tego okresu, liczbę wyprodukowanych żarówek w ciągu t -ego dnia, wartość sprzedaży wody mineralnej w ciągu t -ego miesiąca, wartość sprzedaży leków w ciągu t -ej godziny, poziom zasolenia wody oceanu na t -ej głębokości, poziom temperatury skorupy ziemskiej na t -ej głębokości, stopień skażenia terenu w t -ej odległości od jego źródła itd.

Rozważmy następujący prosty model trendu liniowego:

$$\begin{cases} Y_t = \beta_1 t + \beta_0 + \xi_t, \\ E_\xi(\xi_t) = 0, \quad D_\xi^2(\xi_t) = \sigma^2, \\ \text{Cov}_\xi(\xi_t, \xi_l) = 0, \quad t = 1, \dots, N, \quad l = 1, \dots, N, \quad t \neq l. \end{cases} \quad (3.13)$$

To jest oczywiście szczególny przypadek modelu regresyjnego analizowanego w podrozdziale 2.4.4. W celu oceny wartości globalnej można użyć predyktora $\hat{Y}_s$ określonego wzorem (2.164), który jednak daje ξ -obciążone oceny wartości globalnej. Można łatwo pokazać, że predyktor $\hat{Y}_{s_*}$ jest ξ -nieobciążony dla wartości globalnej, gdy próba s_* jest tak celowo dobierana, aby $\bar{y}_{s_*} = \frac{N+1}{2}$. Z kolei zrandomizowane strategie $(\hat{Y}_s, P_{bz}(s))$ oraz $(\hat{Y}_s, P_{sys}(s))$ dają już P - ξ -nieobciążone oceny wartości globalnej. Dodajmy jeszcze, że gdy występują wahania sezonowe wartości oczekiwanych ciągu zmiennych tworzących model uporządkowany nadpopulacji oraz gdy interwał losowania próby systematycznej jest równy krotności długości cyklu wahań, to strategia $(\hat{Y}_s, P_{sys}(s))$ daje P - ξ -nieobciążoną ocenę wartości globalnej, lecz zwykle dla każdej możliwej próby losowanej zgodnie z planem systematycznym predyktor $\hat{Y}_s$ jest już ξ -obciążony dla wartości globalnej, por. przykład 3.3.

Dodajmy jeszcze, że prezentowane w punkcie 2.4.5 metody weryfikacji prawdziwości modelu regresyjnego można łatwo adoptować do potrzeby testowania liniowości trendu. Ponadto można wykorzystać test proponowany przez Wywiśla (1989), który jest z kolei adaptacją jego testu prezentowanego w punkcie 2.3.3.

3.2. Średnia z próby warstwowej

3.2.1. Wiadomości podstawowe

Prezentowana tutaj strategia estymacji jest jedną z najczęściej używanych w praktyce. Zakładamy, że populacja U jest podzielona na niepuste i rozłączne podzbiory U_h , $h = 1, \dots, H$, zwane warstwami oraz $U = \bigcup_{h=1}^H U_h$. Liczbę elementów populacji tworzących h -tą warstwę oznaczmy przez $0 < N_h < N$, przy czym $N = \sum_{h=1}^H N_h$. Niech $w_h = N_h N^{-1}$. Przez s_h oznaczmy próbę prostą o liczebności $1 \leq n_h \leq N_h$ losowaną bezzwrotnie z h -tej warstwy. Próbę składającą się z prób prostych losowanych bezzwrotnie z poszczególnych warstw oznaczamy przez $s = \{s_1, s_2 \dots s_H\}$. Tak tworzoną próbę będziemy nazywać prostą próbą warstwową losowaną bezzwrotnie lub krótko próbą warstwową z uwagi na jej powszechne stosowanie w praktyce. Jej plan losowania ma postać:

$$P_w(s) = \prod_{h=1}^H \binom{N_h}{n_h}^{-1} \quad s \in \mathbf{Q}_w, \quad (3.14)$$

gdzie:

$$\mathbf{Q}_w = \mathbf{Q}_1 \times \dots \times \mathbf{Q}_h \times \dots \times \mathbf{Q}_H,$$

przy czym $\mathbf{Q}_h$ jest przestrzenią próby s_h losowanej z warstwy U_h , $h = 1, \dots, H$.

Prawdopodobieństwa inkluzji określa wyrażenie:

$$\begin{cases} \pi_k = \frac{n_h}{N_h}, & \text{dla } k \in U_h \\ \pi_{k,l} = \frac{n_h}{N_h} \frac{n_t}{N_t}, & \text{dla } k \in U_h, \quad l \in U_t, \\ \pi_{k,l} = \frac{n_h(n_h-1)}{N_h(N_h-1)}, & \text{dla } k, l \in U_h, \end{cases} \quad (3.15)$$

przy czym $k = 1, \dots, N$, $l = 1, \dots, N$, $l \neq k$, $h = 1, \dots, H$, $t = 1, \dots, H$, $h \neq t$. Stąd wynika, że prawdopodobieństwa inkluzji rzędu pierwszego nie zawsze są takie same, a więc próba warstwowa nie jest próbą prostą, mimo że jej składowe podpróby losowane z poszczególnych warstw są próbami prostymi.

Wprowadźmy oznaczenia:

$$\bar{y}_h = \frac{1}{N_h} \sum_{k \in U_h} y_k, \quad v_{*h} = \frac{1}{N_h - 1} \sum_{k \in U_h} (y_k - \bar{y}_h)^2. \quad (3.16)$$

Nieobciążonymi estymatorami parametrów $\bar{y}_h$ i v_{*h} są odpowiednio statystyki: $\bar{y}_{S_h}$ i v_{S_h} gdzie:

$$\bar{y}_{S_h} = \frac{1}{n_h} \sum_{k \in S_h} y_k, \quad v_{*S_h} = \frac{1}{n_h - 1} \sum_{k \in S_h} (y_k - \bar{y}_{S_h})^2. \quad (3.17)$$

Można łatwo pokazać, że średnia i wartość globalna cechy w populacji są odpowiednio następującymi funkcjami średnich zmiennych w poszczególnych warstwach:

$$\bar{y} = \sum_{h=1}^H w_h \bar{y}_h, \quad y_U = \sum_{h=1}^H N_h \bar{y}_h = N \bar{y}.$$

Znanym w metodzie reprezentacyjnej nieobciążonym estymatorem wartości globalnej zmiennej w populacji jest statystyka:

$$\tilde{y}_{wS} = \sum_{h=1}^H N_h \bar{y}_{S_h}. \quad (3.18)$$

Wariancję tego estymatora określa wzór:

$$D^2(\tilde{y}_{wS}, P_w) = \sum_{h=1}^H \frac{N_h(N_h - n_h)}{n_h} v_{*h}, \quad (3.19)$$

ponieważ

$$D^2(\tilde{y}_{wS}, P_w) = \sum_{h=1}^H w_h^2 D^2(\tilde{y}_{S_h}, P_w) \quad \text{ i } \quad D^2(\tilde{y}_{S_h}, P_w) = \frac{N_h(N_h - n_h)}{n_h} v_{*h}.$$

Wariancja $D^2(\tilde{y}_{S_h}, P_w)$ służy do oceny precyzji estymacji wartości globalnej cechy w h -tej warstwie.

Nieobciążony estymator wariancji otrzymujemy zastępując we wzorze (3.19) parametr v_{*h} przez statystykę, którą określa wyrażenie (3.17). Wtedy mamy:

$$D_S^2(\tilde{y}_{wS}, P_w) = \sum_{h=1}^H \frac{N_h(N_h - n_h)}{n_h} v_{*S_h}. \quad (3.20)$$

Przykład 3.9. Niech celem badania statystycznego jest estymacja wartości średnich dwóch cech gmin województwa katowickiego. Pierwsza z nich oznaczana przez y to dochody z podatków i opłat lokalnych w 1995 r., a druga to dochody z podatku rolnego w 1995 r., którą oznaczamy przez g . Populację gmin podzielono na trzy warstwy, którymi są gminy: miejskie, miejsko-wiejskie oraz wiejskie. Liczą one odpowiednio: 38, 15 oraz 43 gminy. Wylosowano trzy próby proste: spośród gmin miejskich 5 gmin (Bieruń, Imielin, Mysłowice, Pszów, Żory), natomiast spośród gmin miejsko-wiejskich 4 gminy miejskie (Brzeszcze, Pilicę, Pszczynę i Trzebinę) oraz 6 gmin wiejskich (Kobiór, Krzanowice, Lubonię, Pilichowice, Świerklaniec i Tworóg). Dane o tych gminach znajdziemy w Roczniku statystycznym gmin województwa katowickiego (1996). Wyliczone oceny wartości średnich cech za pomocą wzoru (3.18) wynoszą: $\tilde{y}_{ws} = 441\,628.8$ zł i $\tilde{g}_{ws} = 12\,182.4$ zł. Wyznaczone na podstawie wzoru (3.20) oceny błędów średnich szacunku wynoszą $D_s(\tilde{y}_{ws}, P_w) = 3138.2$ i $D_s(\tilde{g}_{ws}, P_w) = 196.7$. Z kolei oceny względnych średnich błędów szacunku wynoszą odpowiednio $z(\tilde{y}_{ws}, P_w) = 100\% \frac{D_s(\tilde{y}_{ws}, P_w)}{\tilde{y}_{ws}} = 23.3\%$ i $z(\tilde{g}_{ws}, P_w) = 3.3\%$. Precyzja oceny wartości globalnej dochodów z podatków lokalnych nie jest dopuszczalna, bo względny błąd szacunku przekracza 5%, co przyjmuje się zwykle w praktyce za górną granicę względnego błędu estymacji, którą można akceptować.

Dodajmy, że po to, aby oszacować wartość średnią cechy wystarczy estymator $\tilde{y}_{ws}$ podzielić przez liczebność populacji, wówczas wariancję tak otrzymanego estymatora średniej oraz jej ocenę otrzymujemy dzieląc prawe strony równań (3.19) i (3.20) przez kwadrat liczebności populacji.

3.2.2. Lokalizacja proporcjonalna prób w warstwach

Ważnym problemem przy estymacji parametrów populacji na podstawie próby warstwowej jest zagadnienie ustalania liczebności prób, jakie mają być losowane z poszczególnych warstw. Najprościej byłoby ustalić ich rozmiary w sposób równomierny, co oznacza, że z każdej warstwy losowanoby tę samą ilość elementów. Takie podejście nie jest jednak zalecane, bo istnieją inne metody wyznaczania liczebności tych prób, prowadzące do precyzyjniejszej estymacji wartości globalnej w populacji.

Oznaczmy przez n sumę liczebności prób losowanych ze wszystkich warstw, czyli $n = \sum_{h=1}^H n_h$. Wtedy liczebność $n_{\#h}$ elementów losowanych z h -tej warstwy jest proporcjonalna do frakcji ilości elementów populacji w h -tej warstwie, gdy:

$$n_{\#h} = nw_h, \quad h = 1, \dots, H. \quad (3.21)$$

Plan losowania prób prostych z tak ustalonymi liczebnościami elementów losowanych z poszczególnych warstw oznaczamy przez P_p . Wtedy na podstawie wzoru (3.19) wnioskujemy, że wariancja strategii estymacji $(\tilde{y}_{wS}, P_p)$ w przypadku bezzwrotnego losowania próby przyjmuje postać:

$$D^2(\tilde{y}_{wS}, P_p) = \frac{N(N-n)}{n} \sum_{h=1}^H w_h v_{*h}. \quad (3.22)$$

Do jej estymacji używa się statystyki

$$D_S^2(\tilde{y}_{wS}, P_p) = \frac{N(N-n)}{n} \sum_{h=1}^H w_h v_{*S_h}. \quad (3.23)$$

Niech $\Delta = D^2(\tilde{y}_S, P_0) - D^2(\tilde{y}_{wS}, P_p)$. Można wykazać, że w przypadku, gdy wszystkie liczebności warstw są dostatecznie duże, to

$$\Delta = \frac{N(N-n)}{n} v_m \geq 0, \quad (3.24)$$

gdzie parametr

$$v_m = \sum_{h=1}^H (\bar{y}_h - \bar{y})^2 w_h \geq 0 \quad (3.25)$$

jest wariancją średnich z warstw, również nazywany wariancją międzygrupową (międzywarstwową). Stąd wynika, że jeśli przeciętna zmiennej w populacji będzie szacowana za pomocą strategii $(\tilde{y}_{wS}, P_p)$ oraz istnieje możliwość wyróżniania warstw w populacji, to należy to tak czynić, aby stopień zróżnicowania średnich zmiennej w warstwach był możliwie duży, czyli aby wariancja v_m była jak największa.

Przykład 3.10. Podobnie jak w przykładzie 3.9 celem badania jest ocena wartości globalnej cechy y oraz g . Przypuśćmy, że środki, jakimi dysponujemy, pozwalają na wylosowanie i obserwację 15 gmin. Stosując proporcjonalny wariant ustalania liczebności prób i wykorzystując dane z treści przykładu 3.9 oraz wzór (3.21) wyliczamy, że z warstw gmin miejskich, miejsko-wiejskich i wiejskich należy wylosować odpowiednio 6, 2 i 7 gmin. Na podstawie tabeli 3.2 oraz wzorów (2.19) i (3.22) wyliczamy: $D(\tilde{y}_{wS}, P_p) =$

210539, $D(\tilde{y}_S, P_0) = 247197$ oraz $D(\tilde{g}_{wS}, P_p) = 3150$, $D(\tilde{g}_S, P_0) = 3256$. Następnie wyznaczamy względne współczynniki efektywności. W przypadku zmiennej y ten współczynnik wynosi $\text{def}(\tilde{g}_{wS}, P_p) = 0.725$, a w przypadku cechy g jest równy $\text{def}(\tilde{g}_{wS}, P_p) = 0.936$. Zatem precyzja estymacji wartości globalnej na podstawie próby warstwowej jest większa niż na podstawie próby prostej, jakkolwiek zysk na precyzji w przypadku oceny przeciętnej cechy g jest nikły.

Tabela 3.2

Odchylenia standardowe cech w warstwach i w populacji

Warstwy	Wariancje	
gminy:	y	g
miejskie	14446	81
miejsko-wiejskie	3023	158
wiejskie	1814	168
populacja	10857	143

3.2.3. Warunkowa minimalizacja wariancji

W sposób racjonalny problemem optymalnej lokalizacji prób w warstwach początkowo zajmowali się Tschuprov (1923) i Neyman (1933, 1937). Otrzymane przez nich wyniki dotyczą optymalizacji liczebności prób w przypadku estymacji wartości globalnej pojedynczej cechy w populacji. Zaproponowali oni, by przy ograniczonej sumie liczebności wszystkich prób losowanych z warstw tak wyznaczyć liczebności tych prób, aby wariancja estymatora osiągnęła minimum. Ponadto sugeruje się, że ten sposób otrzymywania liczebności jest także użyteczny w przypadku jednoczesnej estymacji średnich wielu cech. Proponuje, aby optymalizować liczebności prób na podstawie cechy najsilniej skorelowanej z pozostałymi. Zaleca się również, aby ta cecha z punktu widzenia celu badania była najważniejsza.

Koszty obserwacji zmiennych zwykle traktuje się jako funkcję liniową liczebności prób w poszczególnych warstwach postaci:

$$k(n_1, n_2, \dots, n_H) = k(n_1, \dots, n_H) = \sum_{h=1}^H k_h n_h, \quad (3.26)$$

gdzie k_h jest kosztem jednostkowym obserwacji cech w h -tej warstwie. Koszty stałe obserwacji pomijamy, ponieważ ich uwzględnienie nie ma istotnego wpływu na otrzymywane rozwiązania formułowanych dalej zadań optymalizacyjnych. Dopuszczalny koszt zmienny obserwacji cech oznaczamy przez K .

Rozważmy wariancję średniej z prostej próby warstwowej jako funkcję liczebności prób postaci: $D^2(\tilde{y}_{wS}, P_w) = f_1(n_1, n_2, \dots, n_H)$, gdzie

$$f_1(n_1, n_2, \dots, n_H) = \sum_{h=1}^H \frac{N_h^2 v_{*h}}{n_h} - \sum_{h=1}^H N_h v_{*h}. \quad (3.27)$$

Nasze zadanie polega na takim wyznaczeniu wektora liczebności $\mathbf{n} = [n_1 \dots n_H]$, aby przy ustalonych kosztach zmiennych K funkcja ryzyka osiągała minimum, co syntetycznie zapisujemy wyrażeniem:

$$\begin{cases} f_1(n_1, n_2, \dots, n_H) = \text{minimum} \\ k(n_1, n_2, \dots, n_H) \leq K, \quad 2 \leq n_h \leq N_h, h = 1, \dots, H. \end{cases} \quad (3.28)$$

Ogólniejsza postać tego zadania jest rozwiązywana przez Hughesa i Rao (1979), porównaj również prace Wywiśla (1992, 1995). Tutaj podamy rozwiązanie zadania przy następujących upraszczających je założeniach. Po pierwsze, niech $K \geq 2 \sum_{h=1}^H k_h$, co oznacza, że dopuszczalny koszt obserwacji K umożliwia wylosowanie co najmniej dwóch elementów populacji z każdej warstwy. Po drugie, niech każda z liczebności prób będzie mniejsza od ilości elementów warstwy, z której jest losowana. Warunek ten jest spełniony, jeśli $N_h \rightarrow \infty$ dla każdego $h = 1, \dots, H$ lub gdy

$$N_h \geq n_0 = \frac{1}{k_0} \left(K - 2 \sum_{t=1}^H k_t + k_0 \right) \quad \text{dla } h = 1, \dots, H, \quad (3.29)$$

przy czym $k_0 = \min_{h=1, \dots, H} \{k_h\}$. Prawa strona nierówności określa maksymalną licznosc próby, jaką można uzyskać, gdyby losowano próbę z warstwy, w której jest minimalny koszt obserwacji danych, a z pozostałych warstw tylko po dwa elementy do odpowiednich prób. Te założenia znacznie upraszczają rozwiązywanie postawionego zadania. Ponadto będzie ono często w praktyce badań statystycznych spełnione, gdyż zwykle warstwy są bardzo liczne, a koszty dopuszczalne małe. Zatem przy założeniu, że $n_h \leq N_h$ dla każdego $h = 1, \dots, H$, rozwiązanie optymalne ma postać:

$$\underline{n}_h = c_0 w_h \sqrt{\frac{v_{*h}}{k_h}}, \quad \text{dla } h = 1, \dots, H, \quad (3.30)$$

gdzie:

$$c_0 = \frac{K}{\sum_{h=1}^H w_h \sqrt{v_{*h} k_h}}.$$

Stąd wniosek, że optymalna liczebność próby losowanej z h -tej warstwy jest wprost proporcjonalna do iloczynu frakcji w_h i odchylenia standardowego cechy w h -tej warstwie, a odwrotnie proporcjonalna do pierwiastka z kosztu jednostkowego obserwacji w niej zmiennej.

W szczególności niech koszty jednostkowe obserwacji cech w każdej warstwie będą takie same. Wtedy występujące w zadaniu (3.28) ograniczenie narzucone na koszty redukuje się do postaci:

$$\sum_{h=1}^H n_h \leq n. \quad (3.31)$$

Wówczas rozwiązanie (3.30) uprasza się do postaci:

$$\underline{n}_h = c_0 w_h \sqrt{v_{*h}}, \quad \text{dla } h = 1, \dots, H, \quad (3.32)$$

gdzie:

$$c_0 = \frac{n}{\sum_{h=1}^H w_h \sqrt{v_{*h}}}.$$

Wariancja estymatora $\tilde{y}_{wS}$ dla optymalnych liczebności określonych wzorem (3.32) ma postać:

$$D^2(\tilde{y}_{wS}, P_{opt}) = \frac{1}{n} \left(\sum_{h=1}^H N_h d_h \right)^2 - \sum_{h=1}^H N_h v_{*h},$$

gdzie: $d_h = \sqrt{v_{*h}}$ jest odchyleniem standardowym badanej zmiennej w h -tej próbie S_h , natomiast P_{opt} jest optymalną wersją planu losowania prostej próby warstwowej, którą wyznaczają wzory (3.14) i (3.32). Różnica między wariancjami estymatora $\tilde{y}_{ws}$ dla proporcjonalnego wariantu ustalania liczebności prób i optymalnego wariantu ustalania liczebności prób przy ustalonej liczebności próby warstwowej na poziomie n ma postać (por. Zasepa, 1962 i Pawłowski, 1972):

$$D^2(\tilde{y}_{wS}, P_p) - D^2(\tilde{y}_{wS}, P_{opt}) = \frac{N}{n} \sum_{h=1}^H (d_h - \bar{d})^2 N_h, \quad \bar{d} = \frac{1}{N} \sum_{h=1}^H d_h.$$

Zatem użycie optymalnego wariantu losowania prób zamiast wariantu proporcjonalnego daje tym większy zysk na dokładności estymacji wartości globalnej im stopień zróżnicowania odchyłeń standardowych zmiennej badanej między warstwami jest większy. W związku z tym, gdy wariancje zmiennej we wszystkich warstwach są takie same, czyli $v_{*h} = v_*$ dla każdego $h = 1, \dots, H$, to $\underline{n} = n w_h$. W tym więc przypadku optymalny wariant ustalania liczebności prób redukuje się do proporcjonalnego sposobu ich wyznaczania.

Przykład 3.11. Tak jak w przykładzie 3.9 celem badania jest ocena wartości średnich y – dochody z podatków i opłat lokalnych, oraz g – dochody z podatku rolnego. Środki, jakimi dysponujemy, pozwalają na wylosowanie i obserwację 15 gmin. Stosując optymalny wariant ustalania liczebności prób dany wzorem (3.32) i wykorzystując dane z tabeli 3.2 wyliczamy, że z punktu widzenia warunkowej minimalizacji wariancji estymatora wartości średniej zmiennej y należy wylosować odpowiednio 12.3, 1.0 i 1.7 gmin z warstw odpowiednio miejskich, miejsko-wiejskich i wiejskich. Korygujemy te wartości do liczb całkowitych jednocześnie zwiększając liczebność próby losowanej z gmin miejsko-wiejskich do dwóch elementów, co pozwoli już na ocenę wariancji zmiennej w tej warstwie. Ostatecznie więc należy z warstw gmin miejskich, miejsko-wiejskich i wiejskich wylosować odpowiednio: 11, 2 i 2 gminy. Podobnie wyznaczamy optymalne liczebności prób, lecz na podstawie wariancji cechy g . W tym przypadku należy z warstw gmin miejskich, miejsko-wiejskich i wiejskich wylosować odpowiednio: 10, 2 i 3 gminy. Zatem ta lokalizacja nieznacznie różni się od wyżej otrzymanej lokalizacji wyznaczonej na podstawie wariancji zmiennej y .

Przykład 3.12. Przeprowadzamy analizę dokładności ocen wartości średnich cech małych firm. Dane o działalności małych firm pochodzą z badania reprezentacyjnego przeprowadzonego przez Wojewódzki Urząd Statystyczny w Katowicach w 1993 r. Niniejszą analizę ograniczono do cech, których symbole niżej wyjaśniono. Badana populacja składała się z 167094 firm. Wylosowana próba liczyła 6656 firm, co stanowi 4% liczby wszystkich elementów populacji. y_1 – liczba właścicieli i współwłaścicieli, y_2 – uczniowie, y_3 – osoby, z którymi zawarto umowę agencyjno-prowizyjną, y_4 – osoby zatrudnione na podstawie umowy o pracę, y_5 – razem zatrudnieni, y_6 – liczba osób wykonująca prace na umowę zlecenie, y_7 – liczba zatrudnionych w przeliczeniu na etaty, y_8 – kwota wynagrodzeń brutto, y_9 – największe wynagrodzenie brutto. Celem analizy jest ocena dokładności estymacji wartości średnich wymienionych zmiennych. Z metodologicznego punktu widzenia gromadzenie danych odbywało się poprzez bezzwrotne losowanie prób prostych z warstw, przy czym warstwy były tworzone na podstawie cech określających rodzaj działalności firmy zgodny z Europejską Klasyfikacją Danych (EKD). Wykorzystując określone wzorami (3.18) i (3.20) estymatory wartości średnich i ich błędów szacunku wyliczono na podstawie próby warstwowej oceny średnich badanych cech oraz rząd ich odchylen od nieznanych średnich w populacji. Oceny wartości średnich i błędów ich estymacji znajdujemy w tablicy 3.3. W jej trzeciej kolumnie zapisano ocenę wariancji estymatora średniej, z kolei w następnej zapisano jej pierwiastek, który jest oceną przeciętnego błędu szacunku średniej. Przykładowo, średnia wynagrodzeń brutto (zmienna y_8) wynosi 2060.45 zł i odchyła się ta wartość od nieznannej szacowanej wartości

średniej wynagrodzeń brutto przeciętnie o 137.73zł in plus bądź in minus. Błąd ten stanowi 6.68% oceny wartości średniej. Należy uznać, że błąd ten jest zbyt duży, zwykle bowiem przyjmuje się poziom dopuszczalnego błędu nie powinien być większy od 5%.

Tabela 3.3

Oceny dokładności estymacji średnich cech przedsiębiorstw

Zmienna	y1	y2	y3	y4	y5	y6	y7	y8	y9
średnia	1.40	0.14	0.04	0.78	2.36	0.17	0.74	2060	914
błąd śr. szac.	0.016	0.016	0.013	0.042	0.054	0.019	0.040	137	44

Oceny parametrów zamieszczonych w tabeli 3.3 były wyliczone przy założeniu, że ich rozmiary bynajmniej nie były proporcjonalne do liczebności warstw, co zaleca się jako najprostszy zabieg prowadzący zwykle do podniesienia dokładności estymacji. Wyliczono więc oceny błędów szacunku średniej przy założeniu, że liczebności losowanych prób byłyby proporcjonalne do rozmiarów warstw. Wyniki obliczeń zamieszczono w tabeli 3.4. Ponadto znajdujemy w niej oceny błędów estymacji średnich przy założeniu, że liczebności prób losowanych z warstw były wyznaczane w sposób optymalny na podstawie wzoru (3.32). Przyjęto, że optymalne liczebności prób wyznaczamy w stosunku do ocen odchyłeń standardowych w warstwach wynagrodzeń brutto w firmach (zmienna y8).

Tabela 3.4

Względne średnie błędy szacunku (wyrażone w %) dla liczebności faktycznych oraz wyznaczanych proporcjonalnie i optymalnie

Zmienna	y1	y2	y3	y4	y5	y6	y7	y8	y9
liczebności faktyczne	1.1	11.8	30.1	5.4	2.3	10.8	5.5	6.7	4.9
liczebności proporcjonalne	0.6	5.8	15.9	3.2	1.3	8.6	3.3	4.1	2.5
liczebności optymalne	0.6	6.8	16.1	2.0	1.2	10.2	2.9	3.5	2.6

Z tabel 3.3 i 3.4 wynika m.in., że ocena względnego błędu estymacji średnich wynagrodzeń brutto jest już niższa od dopuszczalnego poziomu 5%, bo wynosi 4.1% dla proporcjonalnie wyznaczanych liczebności prób, a 3.5% dla wyznaczanych optymalnie.

Z analizy danych w tabelach 3.3 i 3.4 wynika, że zarówno wariant proporcjonalny, jak i optymalny ustalania liczebności warstw prowadzi do znacznie większej dokładności estymacji przeciętnych badanych zmiennych niż to miało miejsce przy rzeczywistych rozmiarach losowanych prób. W przypadku optymalnego sposobu wyznaczania rozmiarów prób otrzymujemy podobne wyniki.

3.2.4. Warunkowa minimalizacja kosztów obserwacji zmiennej

Czasami w praktyce żąda się, aby dokładność estymacji była na z góry określonym poziomie. Wówczas formuluje się problem minimalizacji kosztów obserwacji cech przy ustalonym poziomie dopuszczalnym wariancji estymatora wartości globalnej za pomocą wyrażenia:

$$\begin{cases} k(n_1, n_2, \dots, n_H) = \text{minimum} \\ f_1(n_1, n_2, \dots, n_H) \leq f_0, \quad 2 \leq n_h \leq N_h, \quad h = 1, 2, \dots, H, \end{cases} \quad (3.33)$$

przy czym f_0 jest dopuszczalnym poziomem wariancji oznaczonej przez $f_1(n_1, n_2, \dots, n_H)$, którą określa wzór (3.27). Funkcję kosztów $k(n_1, n_2, \dots, n_H)$ określa wzór (3.26), a pozostałe oznaczenia wyjaśniono przy rozważaniu zadania optymalizacyjnego określonego wzorem (3.28). Rozwiązanie zadania podają Hughes i Rao (1979). Tutaj przedstawimy jego uproszczoną postać ważną przy założeniu, że $n_h \leq N_h$ dla każdego $h = 1, \dots, H$. Wtedy rozwiązanie optymalne określa wzór (3.30), przy czym:

$$c_0 = \frac{\sum_{h=1}^H w_h \sqrt{v_{*h} k_h}}{f_0 + \sum_{h=1}^H w_h \frac{\sqrt{v_{*h}}}{N_h}}.$$

Zatem optymalna liczebność próby losowanej z h -tej warstwy jest wprost proporcjonalna do iloczynu frakcji występowania elementów populacji w h -tej warstwie przemnożonej przez odchylenie standardowe zmiennej w h -tej warstwie, a odwrotnie proporcjonalna do pierwiastka z kosztu jednostkowego obserwacji cech w tej warstwie. W szczególności, gdy koszty jednostkowe obserwacji cechy w warstwach są takie same, to funkcja kryterium zadania (3.33) upraszcza się do postaci:

$$k(n_1, n_2, \dots, n_H) = \sum_{h=1}^H n_h = \text{minimum}.$$

Wtedy rozwiązanie zadania upraszcza się do postaci określonej wzorem (3.32), przy czym występujący w nim parametr c_0 ma teraz postać:

$$c_0 = \frac{\sum_{h=1}^H w_h \sqrt{v_{*h}}}{f_0 + \sum_{h=1}^H w_h \frac{\sqrt{v_{*h}}}{N_h}}.$$

Przykład 3.13. *Tak jak w przykładzie 3.9 celem badania jest ocena wartości średnich następujących cech: y – dochody z podatków i opłat lokalnych w 1995 r., g – dochody z podatku rolnego w 1995 r. Teraz postulujemy, aby błędy średnie szacunku estymacji przeciętnych cech y i g nie przekraczały odpowiednio poziomów: 1100 zł i 20 zł. Przyjmując, że koszty jednostkowe*

obserwacji wartości zmiennej we wszystkich warstwach są takie same, stosujemy optymalny wariant ustalania liczebności wynikający z określonego wzorem (3.33) zadania optymalizacyjnego, którego funkcją celu jest suma liczebności prób losowanych z warstw. Przy ustalonej dopuszczalnej wartości wariancji zmiennej y wykorzystując dane z tabeli 3.2 wyliczamy, że z warstw gmin miejskich, miejsko-wiejskich i wiejskich należy wylosować odpowiednio 33,1, 2,7 i 4,7 gmin. Po to, by spełnić narzucony warunek na dokładność estymacji zaokrąglamy z nadmiarem te wartości do liczb całkowitych. Ostatecznie więc należy z warstw gmin miejskich, miejsko-wiejskich i wiejskich wylosować odpowiednio: 34, 3 i 5 gmin. Stąd wynika, że w sumie należy wylosować 42 gminy.

Minimalizacja sumy liczebności prób przy założonej dokładności oceny wartości średniej cechy g prowadzi do następującego wyniku. Z warstw gmin miejskich, miejsko-wiejskich i wiejskich należy wylosować odpowiednio: 16, 3 i 5 gmin. Suma liczebności prób wynosi więc 24. Otrzymane liczebności prób są odpowiednio majoryzowane przez liczebności prób otrzymane wyżej przy optymalizacji względem cechy y . Zatem wybierając liczebności prób 34, 3 i 5 spełnimy obydwa postulaty precyzji estymacji.

W praktyce badań reprezentacyjnych mamy często do czynienia z jednoczesną estymacją wektora średnich lub wartości globalnych. Wówczas rozwiązuje się zadanie polegające na takim ustaleniu liczebności prób, aby funkcja kosztów obserwacji zmiennych osiągnęła minimum przy z góry ustalonych poziomach błędów dopuszczalnych szacunku parametrów poszczególnych cech. Zagadnienie to sformułował Dalenius (1957) i rozwiązał dla przypadku losowania zwrotnego prób z dwóch warstw podczas estymacji średnich dwóch zmiennych. W przypadku ogólnym, lecz dla zwrotnego wariantu losowania prób problem ten rozwiązywali m.in.: Greń (1963, 1966), Hartley (1965), Huddleston, Claypool i Hocking (1970), Jaganathan (1965, 1965a), Kish (1961, 1965), Kokan (1963), Yates (1960), a w sposób wyczerpujący dla zwrotnego i bezzwrotnego wariantu losowania próby zadanie to rozwiązali Kokan i Khan (1967). Szeroko tym problemem zajmował się również Wywiał (1992, 1995, 2003). Zasygnalizujemy jeszcze, że Ghosh (1963) zagadnienie wyboru liczebności prób losowych z warstw traktuje jako grę statystyczną.

3.2.5. Warstwowanie populacji po jej wylosowaniu

W badaniach praktycznych często mamy do czynienia z sytuacją, że dopiero po wylosowaniu elementu do próby i zaobserwowaniu przypisanej mu odpowiedniej wartości zmiennej dodatkowej jest możliwa identyfikacja przynależności elementu do odpowiedniej warstwy. W tym wypadku obserwowana zmienna dodatkowa jest nazywana właśnie warstwującą. Zatem próba jest

warstwowana po jej wylosowaniu. W wyniku tego dostajemy podpróby elementów przynależnych do odpowiednich warstw. Zakłada się, że z góry są znane frakcje elementów populacji przynależnych do odpowiednich warstw. Ponadto jeszcze przyjmijmy, że próba prosta o liczebności n jest losowana bezzwrotnie. Jej plan losowania oznaczono przez P_{bz} . Wówczas nieobciążonym estymatorem wartości globalnej w populacji $\bar{y}$ jest następująca statystyka:

$$\hat{y}_{wS} = \sum_{h=1}^H N_h \bar{y}_S, \quad (3.34)$$

przy czym oznaczenia wyjaśniono wcześniej w podrozdziale 3.2.1. W tym przypadku $S = S_1 \cup \dots \cup S_H$, gdzie S_h jest podpróbą o liczebności n_h należącą do h -tej warstwy wyróżnioną w próbie S po jej wylosowaniu. Gdy $n_h > 0$ dla $h = 1, \dots, H$, to wariancję określa wzór:

$$\begin{aligned} D^2(\hat{y}_{wS}, P_{bz}) &= \frac{N(N-n)}{n} \left(\sum_{h=1}^H w_h v_{*h} + \frac{1}{n} \sum_{h=1}^H (1-w_h) v_{*h} \right) + O(n^{-3}) = \\ &= D^2(\tilde{y}_{wS}, P_p) + \frac{N(N-n)}{n^2} \frac{1}{n} \sum_{h=1}^H (1-w_h) v_{*h} + O(n^{-3}) = \\ &= D^2(\tilde{y}_{wS}, P_p) + O(n^{-2}), \end{aligned} \quad (3.35)$$

przy czym $D^2(\tilde{y}_{wS}, P_p)$ określa wzór (3.22).

3.2.6. Przypadek nieprostych prób składowych próby warstwowej*

Przypuśćmy, że próby składowe próby warstwowej nie są proste. Oznaczmy przez P_h plan losowania próby z h -tej warstwy, a prawdopodobieństwa inkluzji przez $\pi_k^{(h)} > 0$, $\pi_{k,t}^{(h)}$, $h = 1, \dots, H$, $k = 1, \dots, N$, $t = 1, \dots, N$, $k \neq t$. Wówczas nieobciążonym estymatorem wartości globalnej w populacji jest następująca kombinacja liniowa statystyk Horvitz-Thompsona:

$$y_{wHTS} = \sum_{h=1}^H w_h y_{HTS_h}, \quad (3.36)$$

której wariancja ma postać:

$$D^2(y_{wHTS}, P_*) = \sum_{h=1}^H w_h^2 D^2(y_{HTS_h}, P_h), \quad (3.37)$$

przy czym $P_* = \prod_{h=1}^H P_h$, gdzie P_h jest planem losowania podpróby S_h z h -tej warstwy. Statystykę y_{HTS_h} oraz jej wariancję pozwalają wyznaczyć odpowiednio wzory (2.60)–(2.63). Nieobciążony estymator wariancji

strategii (y_{wHTS}, P_*) wyznaczamy zastępując we wzorze (3.37) parametry $D^2(y_{HTS_h}, P_h)$ ich nieobciążonymi estymatorami, które wyznaczają odpowiednio wzory (2.64) lub (2.65).

3.2.7. Predykcja na podstawie modelu warstwowego

Model warstwowy ma odzwierciedlać niejednorodność rozkładu wartości zmiennej w populacji. Syntetycznie model warstwowy opisujemy wyrażeniem:

$$\begin{cases} E_{\xi}(Y_i) = \mu_h, & D_{\xi}^2(Y_i) = \sigma_h^2 & \text{dla } i \in U_h, h = 1, \dots, H, \\ \text{Cov } \xi(Y_i, Y_t) = 0 & & \text{dla } i \neq t, t = 1, \dots, N, i = 1, \dots, N. \end{cases} \quad (3.38)$$

Model ten jest szczególnym przypadkiem modelu regresyjnego dla wielu zmiennych objaśniających określonego wyrażeniem (2.152), w którym przyjmujemy, że elementy macierzy $\mathbf{X}_U$ o wymiarach $N \times H$ są równe zero lub jeden, przy czym w każdym wierszu tej macierzy jest jedna jedynka, a pozostałe elementy są równe zero. Dany element h -tej kolumny tej macierzy jest równy jeden, gdy odpowiadający temu elementowi obiekt należy do h -tej warstwy. Wówczas elementy wektora β są równe oczekiwanym wartościom zmiennych w odpowiednich warstwach, czyli $\beta^T = [\mu_1 \mu_2 \dots \mu_H]$. Ponadto macierz wariancji i kowariancji $\mathbf{V}$ jest diagonalna, każdy z jej N_1 -pierwszych elementów, jest równy σ_1^2 . Elementy o numerach od $\sum_{i=1}^{h-1} N_i + 1$ do $\sum_{i=1}^h N_i$ są równe σ_h^2 , $h = 1, \dots, H - 1$, czyli $\mathbf{V} = \text{diag}[\sigma_1^2 \sigma_1^2 \dots \sigma_1^2 \sigma_2^2 \dots \sigma_h^2 \dots \sigma_h^2 \sigma_H^2 \sigma_H^2]$.

W analizowanym przypadku rozkłady zmiennych w poszczególnych warstwach są jednorodne w tym sensie, że zmienne należące do danej warstwy mają tę samą wariancję i tę samą wartość oczekiwaną. Określony model jest nazywany warstwowym. Niech s_h będzie próbą o liczebności n_h wybraną spośród elementów warstwy U_h , przy czym $1 < n_h \leq N_h$, $h = 1, \dots, H$. W związku z tym próba $s = \bigcup_{h=1}^H s_h$ oraz podpróby są rozłączne, czyli $s_h \cap s_t = \emptyset$, dla $h \neq t$, $h, t = 1, \dots, H$. Niech

$$\bar{Y}_{U_h} = \frac{1}{N_h} \sum_{i \in U_h} Y_i, \quad \bar{Y}_{s_h} = \frac{1}{n_h} \sum_{i \in s_h} Y_i.$$

Nieobciążonym predyktorem wartości globalnej jest następująca statystyka:

$$\hat{Y}_s = \sum_{h=1}^H N_h \bar{Y}_{s_h}. \quad (3.39)$$

Jest ona ξ -nieobciążonym predyktorem wartości globalnej Y_U . Jej ξ -błąd średniokwadratowy ma postać:

$$MSE_{\xi}(\hat{Y}_s) = E_{\xi}(\hat{Y}_s - Y_U)^2 = \sum_{h=1}^H \frac{N_h(N_h - n_h)}{n_h} \sigma_h^2, \quad (3.40)$$

a jego nieobciążony estymator ma postać:

$$MSE_s(\hat{Y}_s) = \sum_{h=1}^H \frac{N_h(N_h - n_h)}{n_h} \hat{V}_{s_h}^2, \quad (3.41)$$

gdzie:

$$\hat{V}_{s_h}^2 = \frac{1}{n_h - 1} \sum_{i \in s_h} (Y_i - \bar{Y}_{s_h})^2, \quad \bar{Y}_{s_h} = \frac{1}{n} \sum_{i \in s} Y_i.$$

Dodajmy, że otrzymane wyniki są również ważne w przypadku rozważanym w podrozdziale 3.2.5, gdy próba s jest warstwowana (czyli rozdzielana na podpróby s_h , $h = 1, \dots, H$) po jej wyodrębnieniu, pod warunkiem, iż liczebności warstw są z góry znane. Ponadto, każda z podprób s_h , $h = 1, \dots, H$, musi zawierać co najmniej jeden element, bo w przeciwnym przypadku predyktor $\hat{Y}_s$ może być obciążony.

Na zakończenie rozważmy ważny problem dotyczący trafności wyboru modelu. Przypuśćmy, że jest prawdziwy model warstwowy określony założeniami (3.38). Z kolei statystyk ignoruje możliwość występowania niejednorodności wartości oczekiwanej zmiennej badanej między warstwami i przyjmuje model określony wyrażeniem (2.51), a co za tym idzie użyje predyktora $\tilde{Y}_s$ określonego wzorem (2.57) do oceny wartości globalnej Y_U . Wówczas predyktor będzie dawał ξ -obciążone oceny Y_U , ponieważ można wykazać że:

$$E(\tilde{Y}_s - Y_U) = \sum_{h=1}^H \left(N \frac{n_h}{n} - N_h \right) \mu_h \neq 0.$$

Stąd wynika, że nietrafnie określony model prowadzi do błędu systematycznego oceny wartości globalnej.

Z drugiej jednak strony załóżmy, że próba S jest prostą losowaną bezzwrotnie. Wówczas strategia predykcji $(\tilde{Y}_S, P_{bz}(S))$ jest p - ξ -nieobciążona. Zatem losowanie próby prostej w tym przypadku prowadzi do odporności predyktora $\tilde{Y}_S$ na ewentualne występowanie odchyłeń od założeń modelu prostego w kierunku tych określających model warstwowy.

Weryfikację jednorodności rozkładu zmiennych losowych odpowiadających określonej warstwie można przeprowadzić za pomocą np. testów rozważanych w punkcie 2.3.3.

3.3. Estymacja na podstawie próby grupowej

Losowanie grupowe zwykle jest stosowane wtedy, gdy nie jest możliwe posiadanie dokładnego operatu losowania, a jedynie spisu grup (zespołów) elementów tworzących populację. Te zespoły są zazwyczaj zdeterminowane

w sposób naturalny, jak np. zbiór gmin traktowanych jako grupy gospodarstw domowych lub klasy uczniów, grupy studentów.

3.3.1. Jednostopniowa prosta próba grupowa

Przyjmujemy, że N -elementowa populacja U jest podzielona na G rozłącznych i niepustych zbiorów U_p o liczebności N_p , $p = 1, \dots, G$, w taki sposób, że $U = U_1 \cup \dots \cup U_G$, $N = \sum_{p=1}^G N_p$. Te zbiory są właśnie nazywane *grupami* lub *zespołami* elementów populacji.

Przyjmujemy, że g -elementowa próba S jest dobierana za pomocą znanego schematu losowania prostej próby grupowej realizującej plan losowania:

$$P_g(s) = \binom{G}{g}^{-1}. \quad (3.42)$$

Schemat losowania takiej próby różni się od schematu losowania próby prostej tylko tym, że zamiast poszczególnych elementów populacji są losowane do niej zespoły tych elementów. Dodajmy, że opisany wcześniej plan losowania systematycznego może być traktowany jako szczególny przypadek planu losowania grupowego, przy założeniu, że do próby jest losowana tylko jedna grupa. Ponadto, gdy założymy, że grupy są jednoelementowe, to otrzymujemy plan losowania próby prostej.

Wprowadzamy następujące oznaczenia: przez y_k oznaczamy k -tą obserwację zmiennej. Sumę wartości zmiennej w p -tej grupie określa wzór:

$$z_p = \sum_{k \in U_p} y_k,$$

a wartość średnią w p -tej grupie przez

$$\bar{y}_p = \frac{z_p}{N_p}.$$

Średnia suma wartości zmiennej przypadająca na grupę:

$$\bar{z} = \frac{1}{G} \sum_{p=1}^G z_p,$$

średnia wartość zmiennej w populacji:

$$\bar{y} = \frac{1}{N} \sum_{p=1}^G z_p = \frac{1}{N} \sum_{p=1}^G \sum_{k \in U_p} y_k.$$

Wariancja cechy w populacji:

$$v_* = \frac{1}{N-1} \sum_{p=1}^G \sum_{k \in U_p} (y_k - \bar{y})^2.$$

Wariancja sum wartości cechy w grupach:

$$v_*(z) = \frac{1}{G-1} \sum_{p=1}^G (z_k - \bar{z})^2. \quad (3.43)$$

Teraz zakładamy, że grupy są równoliczne i każda z nich składa się z M elementów, czyli rozmiar populacji $N = GM$. Wtedy wartość globalna w populacji jest oceniana za pomocą estymatora:

$$y_{gS} = \frac{G}{g} \sum_{p \in S} \sum_{k \in U_p} y_k = \frac{G}{g} \sum_{p \in S} z_p. \quad (3.44)$$

Statystyka y_{gS} daje nieobciążone oceny wartości globalnej y_U , a jej wariancja ma postać:

$$D^2(y_{gS}) = \frac{G(G-g)}{g} v_*(z). \quad (3.45)$$

Nieobciążony estymator tej wariancji ma postać:

$$D_S^2(y_{gS}, P_g) = \frac{G(G-g)}{g} v_s(z), \quad (3.46)$$

gdzie

$$v_s(z) = \frac{1}{g-1} \sum_{p \in S} (z_p - \bar{z}_S)^2, \quad \bar{z}_S = \frac{1}{g} \sum_{p \in S} z_p. \quad (3.47)$$

Przykład 3.14. *Populację 96 gmin byłego województwa katowickiego podzielono na równoliczne grupy, z których każda składa się z 4 gmin. Następnie wylosowano w sposób prosty i bezzwrotny następujące grupy gmin: {Bytom, Chorzów, Czeladź, Dąbrowa Górnicza}, {Ruda Śląska, Rybnik, Rydułtowy, Siemianowice Śląskie}, {Sławków, Sosnowiec, Świętochłowice, Tarnowskie Góry}, {Czechowice-Dziedzice, Czerwionka-Leszczyny, Kuźnia Raciborska, Libiąż}. Wartości estymatora określonego wzorem (3.44) wynoszą 9614 127.2 i 9652.8 odpowiednio dla zmiennych podlok (wpływy z podatków lokalnych) i podrol (wpływy z podatku rolnego). Na podstawie wzoru (3.47) wyliczamy wariancje z próby, które wynoszą: 764 310 948.8 i 7485.0 odpowiednio dla cech podlok i podrol. Wyliczone na podstawie wzoru (3.46) oceny błędów średnich szacunku wartości globalnych w populacji odpowiednio dla cech podlok i podrol wynoszą odpowiednio: 963154.7 i 969.9. W końcu oceny względnych błędów szacunku wynoszą 22.3% i 18.1% odpowiednio dla zmiennych: podlok i podrol. Zatem to oznacza, że oceny przeciętnych w populacji nie są dostatecznie dokładne, jeśli przyjelibyśmy dopuszczalny błąd względny na poziomie 5%.*

Określona wcześniej wariancję $v_*(z)$ można zapisać w następującej postaci:

$$v_*(z) = \frac{N-1}{G-1} v_* \left(1 + (M-1)r^{(w)} \right), \quad (3.48)$$

gdzie:

$$r^{(w)} = \frac{c^{(w)}}{v_*} \quad (3.49)$$

jest nazywany *współczynnikiem korelacji wewnątrzgrupowej*, przy czym

$$c^{(w)} = \frac{2}{(N-1)(M-1)} \sum_{p=1}^G \sum_{k \in U_p} \sum_{t \in U_p, t \neq k} (y_k - \bar{y})(y_t - \bar{y}).$$

Współczynnik $r^{(w)}$ przyjmuje wartości z przedziału $[-\frac{1}{M-1}; 1]$.

Na podstawie wzorów (3.45) i (3.48) otrzymujemy następującą postać wzoru na wariancję strategii estymacji wartości globalnej:

$$D^2(y_{gS}, P_g) = \frac{G(N-1)(G-g)}{(G-1)g} v_* \left(1 + (M-1)r^{(w)} \right). \quad (3.50)$$

Współczynnik korelacji wewnątrzgrupowej można zapisać następująco (por np. Konijn 1973, s. 225–227; Zasępa 1972, s. 311–312):

$$r^{(w)} = \frac{v^{(m)} - \frac{1}{M}v^{(w)}}{v} \quad (3.51)$$

lub

$$r^{(w)} = 1 - \frac{v^{(w)}}{v}, \quad (3.52)$$

lub

$$r^{(w)} = \frac{1}{M-1} \left(M \frac{v^{(m)}}{v} - 1 \right), \quad (3.53)$$

przy czym $v = \frac{N-1}{N} v_*$, natomiast przez $v^{(m)}$ oznaczono tzw. wariancję międzygrupową:

$$v^{(m)} = \frac{1}{G} \sum_{p=1}^G (\bar{y}_p - \bar{y})^2, \quad (3.54)$$

a przez $v^{(w)}$ oznaczono tzw. wariancję wewnątrzgrupową:

$$v^{(w)} = \frac{1}{G(M-1)} \sum_{p=1}^G \sum_{k \in U_p} (y_k - \bar{y}_p)^2 = \frac{1}{G} \sum_{p=1}^G v_{*p}, \quad (3.55)$$

gdzie

$$v_{*p} = \frac{1}{M-1} \sum_{k \in U_p} (y_k - \bar{y}_p)^2. \quad (3.56)$$

Wariancja wewnątrzgrupowa jest więc średnią wariancji cechy w grupach. Zatem określa ona przeciętny stopień zróżnicowania zmiennej wewnątrz grup, na jakie populacja została podzielona.

Wariancję $D^2(\tilde{y}_S, P_{bz})$ określa wzór (2.13). Dla $n = Mg$ i dostatecznie dużej liczby grup na podstawie (3.50) mamy:

$$e = \frac{D^2(\tilde{y}_{gS}, P_g)}{D^2(\tilde{y}_S, P_{bz})} = 1 + (M - 1)r^{(w)}. \quad (3.57)$$

Jeśli ilość równolicznych grup jest duża i współczynnik korelacji wewnątrzgrupowej $r^{(w)}$ jest niedodatni (ujemny), to strategia $(\tilde{y}_{gS}, P_g)$ jest nie gorszą (lepszą) od strategii $(\tilde{y}_S, P_{bz})$. W związku z tym, o ile jest to możliwe, należy tak wyróżniać rozłączne grupy w populacji, by jak najsilniej były zróżnicowane obserwacje cech wewnątrz tych grup, bo wówczas współczynnik korelacji wewnątrzgrupowej będzie ujemny.

Możliwość tworzenia grup równolicznych pojawia się w niektórych badaniach praktycznych, np. dotyczących badań reprezentacyjnych w rolnictwie i środowisku naturalnym. Wobec ograniczonych kosztów przeznaczonych na badania trzeba zdecydować, jak dużo grup mamy losować i o jakiej liczebności. Ponadto pojawia się potrzeba określenia związku między liczebnością grupy i wariancją wewnątrzgrupową bądź współczynnikiem korelacji wewnątrzgrupowej. Tym problemem zajmowali się początkowo Jessen (1942), Mahalanobis (1944), Hedricks (1944), który na podstawie badań empirycznych wywnioskowali m.in., że wariancja wewnątrzgrupowa $v^{(w)}$, dana wzorem (3.55), jest niemalejącą funkcją liczebności grupy. Dla odpowiednio sformułowanej nieliniowej funkcji kosztów obserwacji cech w próbie grupowej rozważają dwa zadania optymalizacyjne. Pierwsze polega na takim ustaleniu liczebności g i M , aby wariancja estymatora osiągnęła minimum przy ustalonych kosztach dopuszczalnych badania statystycznego. Drugie zadanie ma na celu znalezienie takich liczebności g i M , które zminimalizują funkcję kosztów przy ustalonej dokładności estymacji określonej dopuszczalnym poziomem wariancji. Więcej szczegółów na ten temat można znaleźć np. w pracach Cochra (1963) i Wywiła (1995).

Pozostaje jeszcze problem optymalnego ustalania przynależności elementów populacji do wyróżnianych grup. Zadanie to podejmował Wywił (1992, 1993, 1995), formułując odpowiednie kryteria wyróżniania równolicznych grup na podstawie z góry dostępnych danych o wartościach cech dodatkowych.

3.3.2. Próba dwustopniowa

Utrzymujemy w mocy oznaczenia wprowadzone w uprzednim podrozdziale, lecz teraz dodatkowo zakładamy, że próba S jest dobierana za pomocą

dwustopniowego schematu losowania. Pierwszy stopień losowania polega na bezzwrotnym losowaniu g grup, przy czym zakłada się, że każda z grup jest losowana z tym samym prawdopodobieństwem jej wyboru do próby. Następnie z każdej tak wybranej grupy U_p , jest losowana bezzwrotnie próbka prosta S_p o liczebności n_p . Plan tak losowanej próby określa wzór:

$$P_d = \left(\binom{G}{g} \right)^{-1} \prod_{p=1}^G \left(\binom{N_p}{n_p} \right)^{-1}.$$

Niech

$$\hat{y}_{gS} = \frac{G}{g} \sum_{p \in S} N_p \bar{y}_{S_p} = G \hat{z}_S, \quad (3.58)$$

gdzie

$$\hat{z}_S = \frac{1}{g} \sum_{p \in S} \hat{z}_{S_p}, \quad \hat{z}_{S_p} = N_p \bar{y}_{S_p}, \quad \bar{y}_{S_p} = \frac{1}{n_p} \sum_{k \in S_p} y_k. \quad (3.59)$$

Statystyka $\hat{z}_{S_p}$ jest nieobciążonym estymatorem wartości globalnej cechy w p -tej grupie. Strategia $(\hat{y}_{gS}, P_d)$ daje nieobciążone oceny przeciętnej w populacji, a jej wariancja ma postać:

$$D^2(\hat{y}_{gS}, P_d) = \frac{G(G-g)}{g} v_*(z) + \frac{G}{g} \sum_{p=1}^G \frac{N_p(N_p - n_p)}{n_p} v_{*p}, \quad (3.60)$$

przy czym wariancję $v_*(z)$ określa wzór (3.43), a wariancję v_{*p} wyrażenie:

$$v_{*p} = \frac{1}{N_p - 1} \sum_{k \in U_p} (y_k - \bar{y}_p)^2, \quad \bar{y}_p = \frac{1}{N_p} \sum_{k \in U_p} y_k. \quad (3.61)$$

Nieobciążony stymator wariancji $D^2(\bar{y}_{gS}, P_d)$ ma postać:

$$D_S^2(\hat{y}_{gS}, P_d) = \frac{G(G-g)}{g} v_{*S}(z) + \frac{G}{g} \sum_{p=1}^G \frac{N_p(N_p - n_p)}{n_p} v_{*S_p}, \quad (3.62)$$

przy czym:

$$\begin{aligned} v_{*S}(z) &= \frac{1}{g-1} \sum_{p \in S} (\hat{z}_{S_p} - \hat{z}_S)^2, \\ v_{*S_p} &= \frac{1}{n_p - 1} \sum_{k \in S_p} (y_k - \bar{y}_{S_p})^2. \end{aligned} \quad (3.63)$$

Nieobciążony estymator średniej w populacji ma postać:

$$\bar{y}_{gS} = \frac{\hat{y}_S}{N}. \quad (3.64)$$

Wówczas: $D^2(\bar{y}_{gS}, P_d) = N^{-2} D^2(\hat{y}_{gS}, P_d)$, $D_S^2(\bar{y}_{gS}, P_d) = N^{-2} D_S^2(\hat{y}_{gS}, P_d)$.

Przykład 3.15. Rozważaną w przykładzie 3.9 populację gmin województwa katowickiego podzielono na grupy, które nie są równoliczne. Średnia liczebności grupy wynosi 4 gminy. Następnie wylosowano w sposób prosty i bezwrotny $g = 5$ grup gmin spośród $G = 24$. Z kolei z każdej z tych grup wylosowano trzy gminy, tak że ostatecznie w próbie znajdują się następujące gminy: {Dąbrowa Górnicza, Imielin, Jaworzno, Łędziny, Miasteczko Śląskie, Mikołów, Pyskowice, Racibórz, Ruda Śląska, Wojkowice, Zawiercie, Żory, Babice, Borowniki, Bojszowy}. Na podstawie wzorów (3.58) i (3.64) wyliczamy oceny wartości średnich cech podlok (wpływy z podatków lokalnych) i podrol (wpływy z podatku rolnego), które odpowiednio wynoszą 9856.581 i 90.07. Wartość wariancji $D_S^2(\bar{y}_{gS}, P_d)$ wynosi 14 121 453.451 i 665.967 odpowiednio w przypadku cech podlok i podrol. Zatem oceny błędów średnich szacunku estymatorów średnich populacji cech podlok i podrol wynoszą odpowiednio 3757.852 i 25.806. Uwzględniając wcześniejsze oszacowania wartości średnich w populacji wyliczamy oceny względnych błędów średnich szacunku, które odpowiednio wynoszą: 38.1% i 28.7%.

Zakładamy, że próba S jest automatycznie wyważona, co oznacza, że z każdej wybranej do próby S grupy U_p jest losowana próbka S_p o liczebności równej ustalonej frakcji f elementów tej grupy, czyli:

$$n_p = fN_p, \quad p = 1, \dots, G. \quad (3.65)$$

Wtedy określony wzorem (3.58) estymator wartości globalnej w populacji przekształca się do postaci:

$$\hat{y}_{gS} = \frac{G}{gf} \sum_{p \in S} \sum_{k \in S_p} y_k. \quad (3.66)$$

Oznaczając plan losowania próby dwustopniowej zrównoważonej przez P'_d na podstawie wzoru (3.60) wnioskujemy, że:

$$D^2(\hat{y}_{S_p}, P'_d) = \frac{G(G-g)}{g} v_*(z) + \frac{GN(1-f)}{gf} v_*^{(w)}, \quad (3.67)$$

gdzie wariancję $v_*(z)$ określa wzór (3.43), natomiast wariancję wewnątrzgrupową $v_*^{(w)}$ określa wyrażenie:

$$v_*^{(w)} = \sum_{p=1}^G w_p v_{*p}, \quad w_p = \frac{N_p}{N},$$

przy czym v_{*p} określa wzór (3.61).

W szczególności rozważmy schemat losowania próby dwustopniowej. Na pierwszym stopniu są losowane grupy elementów z prawdopodobieństwami proporcjonalnymi do liczebności elementów tworzących grupy zgodnie

ze schematem bezzwrotnego losowania próby Rao-Hartleya, który opisano w punkcie 3.1.2. Zatem zakładamy, że $\pi_{Ik} = gp_k \leq 1$, $k = 1, \dots, N$ i $p_k = N_k / \sum_{i=1}^N N_i$. Na drugim stopniu spośród już wylosowanych na pierwszym stopniu grup są losowane bezzwrotnie próby proste S_i o liczebności n_i , $i = 1, \dots, g$. Wówczas nieobciążonym estymatorem wartości globalnej w populacji jest statystyka (por. np. Bracha, 1996):

$$\tilde{y}_{\pi, S} = \sum_{i \in S} \frac{N_i}{\pi_i n_i} \sum_{e \in S_i} y_e,$$

której wariancję określa wzór:

$$D^2(\tilde{y}_{\pi, S}) = \sum_{k=1}^G \pi_k \left(1 - \frac{g-1}{g} \pi_k\right) \left(\frac{z_k}{\pi_k} - \frac{y_U}{g}\right)^2 + \\ + \sum_{k=1}^G N_k(N_k - n_k) \frac{v_{*k}}{\pi_k n_k} + O(N^{-1}),$$

przy czym v_{*k} określa wzór (3.61).

Estymator wariancji $D^2(\tilde{y}_{\pi, S})$ ma postać:

$$D_S^2(\tilde{y}_{\pi, S}) = \frac{1}{2(g-1)} \sum_{h, t \in S} \left(1 + g \sum_{k=1}^N p_k^2 - \pi_h - \pi_t\right) \left(\frac{\hat{z}_{S_h}}{\pi_h} - \frac{\hat{z}_{S_t}}{\pi_t}\right)^2 + \\ + \sum_{h \in S} \frac{N_h}{\pi_h n_h} (N_h - n_h) v_{*S_h},$$

przy czym $\hat{z}_{S_h}$ i v_{*S_h} określają odpowiednio wzory (3.59) i (3.63).

Pozostaje jeszcze m.in. problem optymalnego wyznaczania liczebności losowanych grup oraz liczebności losowanych na drugim stopniu prób. W tym celu formułuje się generalnie dwa rodzaje zadań. Pierwsze, częściej rozważane w praktyce, polega na minimalizacji odpowiednio definiowanej oczekiwanej funkcji kosztów losowania próby dwustopniowej i obserwacji w niej danych przy zadanej dokładności estymacji, czyli przy ustalonej postulowanej wartości wariancji estymatora. Drugie zadanie polega na minimalizacji wariancji estymatora przy ograniczonej wartości tych kosztów oczekiwanych. Te zadania były m.in. analizowane przez Konijn (1973) i Wywiał (1992, 1995) dla przypadku dwustopniowej próby prostej. Dodajmy, że Kokan (1963) oraz Kokan i Khan (1967) rozwiązywali ogólniejsze zadanie polegające na minimalizacji oczekiwanej funkcji kosztów przy ustalonych dopuszczalnych wartościach wariancji strategii $(\hat{y}_{gS}, P_d)$ estymacji wartości globalnych wielu zmiennych w populacji, lecz dla przypadku, gdy podczas drugiego stopnia losowania są wybierane próbki równoliczne z wcześniej dobranych grup.

3.3.3. Predykcja na podstawie grupowego modelu nadpopulacji

Rozważmy następujący model nadpopulacji:

$$\begin{cases} E_{\xi}(Y_i) = \mu_g, D_{\xi}^2(Y_i) = \sigma_g^2 & \text{dla } i \in U_g, g = 1, \dots, G, \\ \text{Cov } \xi(Y_i, Y_t) = 0 & \text{dla } i \neq t, t \in U, i \in U. \end{cases} \quad (3.68)$$

Z populacji jest dobierana próba dwustopniowa $s = [s_1 \dots s_g]$, gdzie s_i jest n_i -elementową próbką dobraną z grupy U_i , którą już wcześniej wyselekcjonowano spośród G grup. Wartość globalną $Y_U = \sum_{g=1}^G \sum_{i \in U_g} Y_i$ oceniamy na podstawie próby dwustopniowej za pomocą statystyki:

$$\tilde{Y}_s = \frac{G}{g} \sum_{s_i \in s} Z_{s_i}, \quad (3.69)$$

przy czym

$$Z_{s_i} = N_i \bar{Y}_{s_i}, \quad \bar{Y}_{s_i} = \frac{1}{n_i} \sum_{k \in s_i} Y_k, \quad (3.70)$$

a w szczególności, gdy $s_i = U_i$, $i = 1, \dots, G$, czyli całe grupy są losowane do próby, to

$$Z_{s_i} = Z_{U_i} = \sum_{k \in U_i} Y_k, \quad i = 1, \dots, G. \quad (3.71)$$

W przypadku arbitralnie dobranej próby statystyka $\tilde{Y}_s$ jest ξ -obciążonym predyktorem wartości globalnej. Można również pokazać, że ten predyktor wyznaczony z prostej próby dwustopniowej daje $p\xi$ -nieobciążone oceny wartości globalnej.

Można pokazać, że ξ -błąd średniokwadratowy predykcji ma postać:

$$MSE_{\xi}(\tilde{Y}_s) = E_{\xi}(\tilde{Y}_s - Y_U)^2 = \frac{G^2}{g^2} \sum_{s_i \in s} N_i^2 \frac{\sigma_i^2}{n_i} + \sum_{i=1}^G N_i \sigma_i^2 + \Delta_{\xi}^2, \quad (3.72)$$

gdzie $\Delta_{\xi} = E_{\xi}(\tilde{Y}_s) - E_{\xi}(Y_U)$ jest obciążeniem predyktora $\tilde{Y}_s$. W przypadku strategii predykcji $(\tilde{Y}_s, P_d(s))$, czyli predyktora $\tilde{Y}_s$ z prostej próby dwustopniowej wzory (3.60) i (3.61) pozwalają wyprowadzić następujące wyrażenie dla $p\xi$ błędu średniokwadratowego:

$$\begin{aligned} MSE_{P_{\xi}}(\tilde{Y}_s) &= E_P E_{\xi}(\tilde{Y}_s - Y_U)^2 = E_{\xi}(D_P^2(\bar{y}_{gs}, P_d)) = \\ &= \frac{(G-1)(G-g)}{gG(N-1)} \sum_{k=1}^G N_k \sigma_k^2 + \frac{G}{g} \sum_{p=1}^G \frac{N_p(N_p - n_p)}{n_p} \sigma_p^2 + v_{\#}, \end{aligned} \quad (3.73)$$

gdzie

$$v_{\#} = \frac{1}{N-1} \sum_{k=1}^G (N_k \mu_k - \bar{\mu})^2, \quad \bar{\mu} = \frac{1}{G} \sum_{k=1}^G N_k \mu_k.$$

Ograniczmy losowanie dwustopniowe do jednostopniowego, w którym do próby losuje się całe grupy. Wtedy $n_k = N_k$, $k = 1, \dots, G$, a wzór (3.73) redukuje się do następującego:

$$MSE_{P_{\xi}}(\tilde{Y}_s) = \frac{(G-1)(G-g)}{gG(N-1)} \sum_{k=1}^G N_k \sigma_k^2 + v_{\#} \approx \frac{G-g}{g} \bar{\sigma}^2 + v_{\#}, \quad (3.74)$$

gdzie $\bar{\sigma}^2 = \sum_{k=1}^G w_k \sigma_k^2$, $w_k = \frac{N_k}{N}$.

Weryfikację jednorodności rozkładu prawdopodobieństwa zmiennych losowych odpowiadających określonej grupie można przeprowadzić za pomocą np. testów rozważanych w punkcie 2.3.3.

Pytania i zadania

1. Określ plan bezzwrotnego losowania próby warstwowej.
2. Określ konstrukcję funkcji kosztów obserwacji cech.
3. Podaj proporcjonalną lokalizację prób w warstwach.
4. Przy jakich warunkach średnia z próby warstwowej jest lepszym estymatorem od średniej z próby prostej?
5. Od czego i jak zależy optymalna liczebność próby (losowanej z warstwy) otrzymanej z warunkowej minimalizacji wariancji średniej z próby warstwowej?
6. Od czego i jak zależy optymalna liczebność próby (losowanej z warstwy) otrzymanej z minimalizacji kosztów obserwacji zmiennej przy ustalonym dopuszczalnym poziomie wariancji średniej z próby warstwowej?
7. Podaj inne sposoby optymalnej lokalizacji prób w warstwach.
8. Zaproponuj konstrukcję przedziału ufności dla oceny wartości średniej w populacji na podstawie próby warstwowej.
9. Opisz systematyczny schemat losowania próby.
10. Wyjaśnij grupowy schemat losowania próby.
11. Zaproponuj konstrukcję przedziału ufności dla oceny wartości średniej w populacji na podstawie próby grupowej.
12. Wyprowadź wzór na wariancję średniej z prostej próby warstwowej.
13. Wyjaśnij, w jaki sposób wariancja średniej z próby grupowej zależy od wartości współczynnika korelacji wewnątrzgrupowej.
14. Wymień podstawowe własności współczynnika korelacji wewnątrzgrupowej.

15. Opisz związki między współczynnikiem korelacji wewnątrzgrupowej a wariancjami wewnątrzgrupową i międzygrupową oraz wynikające stąd wnioski dotyczące interpretacji współczynnika korelacji wewnątrzgrupowej.
16. Porównaj precyzję średniej z próby grupowej i średniej z próby prostej.
17. Określ model nadpopulacji opisujący warstwową strukturę obserwacji.
18. Opisz strategię predykcji wartości globalnej z próby warstwowej.
19. Opisz model grupowy lub warstwowy nadpopulacji.
20. Wyjaśnij dwustopniowy schemat losowania próby.
21. Zaproponuj konstrukcję przedziału ufności dla oceny wartości średniej w populacji na podstawie próby dwustopniowej.
22. Od czego zależy wariancja średniej z próby dwustopniowej?
23. Wyjaśnij dwustopniowy schemat losowania próby zrównoważonej.
24. Wskaż możliwości wykorzystania estymatorów ilorazowych lub regresyjnych przy estymacji na podstawie próby dwustopniowej lub próby warstwowej.
25. Opisz strategię predykcji wartości globalnej z próby dwustopniowej.
26. Wykaż nieobciążoność średniej z prostej próby warstwowej oraz wyprowadź jej wzór na wariancję traktując ten estymator jako szczególny przypadek estymatora Horvitz-Thompsona.
27. Wartości cechy dodatkowej w populacji wynoszą: 1, 8, 4, 4, 22, 2, 0, 5, 3, 49. Wyznacz prawdopodobieństwa inkluzji rzędu pierwszego proporcjonalnie do tych liczb. Potem na ich podstawie wylosuj systematycznie próbę czteroelementową.

Indeks

- dane 29
- estymator 35
 - błąd średniokwadratowy 36
 - dopuszczalny 38
 - efektywny 38
 - Horvitz i Thompsona 57
 - ilorazowy 72
 - nieobciążony 36
 - regresyjny 77
 - strategia estymacji 40
 - zgodny 37
- liczebności prób 46, 47, 51, 52, 54
 - optymalne 106, 110
 - proporcjonalne 104
- model nadpopulacji 42
 - grupowy 122
 - prosty 55
 - regresyjny 75, 77, 81
 - uporządkowane 100
 - warstwowy 113
- plan losowania 17
 - Bernoulliego 22
 - bezzwrotnego próby prostej 21
 - Midzuna, Lahiriego, Sena 24
 - Poissona 23
 - proporcjonalny do statystyki pozycyjnej 27
 - próby dwustopniowej 119
 - próby grupowej 115
 - próby warstwowej 101
 - Singha, Srivastavy 26
 - systematycznego 93
 - zwrotnego próby prostej 21
- populacja 12
- prawdopodobieństwo inkluzji 18
- predyktor 43
 - błąd średniokwadratowy 44
 - ilorazowy 75
 - nieobciążony 44
 - regresyjny 82
 - strategia predykcji 44
- próba 15
 - losowa 18
 - losowana bezzwrotnie 15
 - losowana zwrotnie 15
 - przestrzeń 16
 - systematyczna 93, 97
- przedział ufności 39, 50
- schemat losowania próby 20
 - dwustopniowej 119
 - Hartleya i Rao 98
 - Midzuna, Lahiriego, Sena 24
 - prostej, bezzwrotny 21
 - prostej, zwrotny 22
 - prostej grupowej 115
 - prostej systematycznej 93
 - prostej warstwowej 101
 - Rao, Hartleya i Cochra 28
 - Singha i Srivastavy 26
 - wyliczający 20
- statystyka 31
- średnia
 - z populacji 13
 - z próby prostej 45
- test 69
 - Chowa 85
 - L_m 70
 - Tjetiena i Moore'a 69
- wariancja 13, 94
 - międzygrupowa 104, 117
 - wewnątrzgrupowa 94, 117
- wartość globalna 13
- współczynnik
 - jednorodności 94
 - korelacji wewnątrzgrupowej 117

Bibliografia

- Ardilly P., Tillé Y. (2006): *Sampling Methods. Exercises and Solutions*. Springer Verlag, New York.
- Basu D. (1971): *An Essay on the Logical Foundations of Survey Sampling*. In: *Foundations of Statistical Inference*. Edited by V.P. Godambe and D.A. Sprott. Holt, Rinehart and Winston of Canada Ltd., Toronto-Montreal.
- Bracha Cz. (1978): *Szacowanie parametrów liniowej funkcji regresji na podstawie próby losowanej bez zwracania z populacji skończonej*. „Przegląd Statystyczny”, Vol. 25.
- Bracha Cz. (1982): *Szacowanie parametrów liniowej funkcji regresji na podstawie próby losowanej dwustopniowo*. „Przegląd Statystyczny”, Vol. 29.
- Bracha Cz. (1983): *Regresja liniowa w badaniach reprezentacyjnych*. Prace i Materiały Instytutu Cybernetyki i Zarządzania. SGPiS, Warszawa.
- Bracha Cz. (1996): *Teoretyczne podstawy metody reprezentacyjnej*. Wydawnictwo Naukowe PWN, Warszawa.
- Bracha Cz. (1998): *Metoda reprezentacyjna w badaniu opinii publicznej i marketingu*. Efekt, Warszawa.
- Brewer K.R.W., Hanif M. (1983): *Sampling with Unequal Probabilities*. Springer Verlag, New York-Heidelberg-Berlin.
- Cassel C.M., Särndal C.E., Wretman J.H. (1977): *Foundation of Inference in Survey Sampling*. John Wiley & Sons, New York-London-Sydney-Toronto.
- Chaudhuri A., Vos J.W.E. (1988): *Unified Theory of Survey Sampling*. North Holland, Amsterdam-New York-Oxford-Tokyo.
- Chaudhuri A., Stenger H. (2005): *Survey Sampling. Theory and Methods*. Chapman & Hall/CRC Taylor & Francis Group, Boca Raton-London-New York-Singapore.
- Chaudhuri A., Dihidar K., Bose M. (2006): *On Feasibility of Basing Horvitz and Thompson's Estimator on a Sample by Rao, Hartley and Cochran's Scheme*. „Communications in Statistics – Theory and Methods”, Vol. 35, s. 2239–2244.
- Chow G.C. (1960): *Tests of Equality between Subsets of Coefficients in Two Linear Regression Models*. „Econometrica”, Vol. 28, s. 561–605.
- Chow G.C. (1995): *Ekonometria*. McGraw Book Company, New York.
- Chung J., DeLury D. (1950): *Confidence Limits for the Hypergeometric Distribution*. University of Toronto Press, Toronto.

- Cochran W. G. (1963): *Sampling Techniques*. John Wiley & Sons, New York.
- Czajkowski A., Domański Cz., Pruska K. (1985): *Zbiór zadań z metody reprezentacyjnej*. Wydawnictwo Uniwersytetu Łódzkiego, Łódź.
- Dalenius T. (1957): *Sampling in Sweden. Contribution to Methods and Theories of Sample Survey Practice*. Almqvist & Wiksells, Stockholm.
- Domański Cz., Pruska K. (2001): *Metody statystyki małych obszarów*. Wydawnictwo Uniwersytetu Łódzkiego, Łódź.
- Fisz M. (1967): *Wstęp do rachunku prawdopodobieństwa i statystyki matematycznej*. Wydawnictwo Naukowe PWN, Warszawa.
- Gabler S., Schweigkoffer R. (1990): *The Existence of Sampling Designs with Pre-assigned Inclusion Probabilities*. „Metrika”, Vol. 37, s. 87–96.
- Galdfeld S. M., Quandt R. E. (1965): *Some Tests for Homoscedasticity*. „Journal of the American Statistical Association”, Vol. 60, s. 537–547.
- Ghosh J. K. (1963): *A Game Theory Approach to the Problem of Optimum Allocation in Stratified Sampling with Multiple Characters*. „Calcutta Statistical Association Bulletin”, Vol. 12, s. 4–12.
- Godambe V. P., Joshi V. M. (1965): *Admissibility and Bayes Estimation in Sampling Finite Populations I*. „Annals of Mathematical Statistics”, Vol. 36, s. 1707–1722.
- Greń J. (1963): *Lokalizacja próby w wieloparametrowym losowaniu warstwowym*. „Przegląd Statystyczny”, Vol. 10, s. 291–302.
- Greń J. (1966): *O pewnym zastosowaniu programowania nieliniowego do metody reprezentacyjnej*. „Przegląd Statystyczny”, Vol. 13, s. 203–217.
- Greń J. (1969): *Wielowymiarowy estymator regresyjny średniej*. „Przegląd Statystyczny”, Vol. 16.
- Greń J. (1970): *Wielowymiarowy estymator regresyjny średniej dla skończonej populacji*. „Przegląd Statystyczny”, Vol. 17, s. 73–78.
- Grubbs F. (1950): *Sample Criteria for Testing Outlying Observations*. „Annals of Mathematical Statistics”, Vol. 21, s. 27–58.
- Hartley H. O. (1965): *Multiple Purpose Optimum Allocation in Stratified Sampling*. „Proceedings of the American Statistical Association, Social Statistics Section”, s. 258–261.
- Hartley H. O., Rao J. N. K. (1962): *Sampling with Unequal Probabilities and without Replacement*. „Annals of Mathematical Statistics”, Vol. 33, s. 350–374.
- Hellwig Z. (1995): *Elementy rachunku prawdopodobieństwa i statystyki matematycznej*. Wydawnictwo Naukowe PWN, Warszawa.
- Herzel A. (1986): *Sampling without Replacement with Unequal Probabilities: Sample Designs with Preassigned Joint Inclusion Probabilities of Any Order*. „Metron”, Vol. 44, No. 1–4, s. 49–68.

- Horvitz D. G., Thompson D. J. (1952): *A Generalization of Sampling without Replacement from Finite Universe*. „Journal of the American Statistical Association”, Vol. 47, s. 663–685.
- Huddleston H. F., Claypool P. L., Hocking R. R. (1970): *Optimal Sample Allocation to Strata Using Convex Programming*. „Applied Statistics”, Vol. 19, s. 273–278.
- Hughes E., Rao J. N. K. (1979): *Some Problems of Optimal Allocation in Sample Surveys Involving Inequality Constraints*. „Communication in Statistics”, Vol. A8, s. 1551–1571.
- Jaganathan R. (1965): *The Programming Approach in Multiple Characters Studies*. „Econometrica”, Vol. 33, s. 236–237.
- Jaganathan R. (1965a): *A Method for Solving a Nonlinear Programming Problem in Sample Surveys*. „Econometrica”, Vol. 33, s. 841–846.
- Joshi P. C. (1969): *Bounds and Approximations for the Moments of Order Statistics*. „Journal of the American Statistical Association”, 64, s. 1617–24.
- Joshi P. C. (1986): *Tests for Two Outliers in a Linear Model*. „Biometrika”, Vol. 73, s. 236–39.
- Kish L. (1961): *Efficient Allocation of Multipurpose Sample*. „Econometrica”, Vol. 29, s. 363–385.
- Kish L. (1965): *Survey Sampling*. John Wiley & Sons, Inc. New York-London-Sydney.
- Kokan A. R. (1963): *Optimum Allocation in Multivariate Surveys*. „Journal of the Royal Statistical Society”, Vol. A 126, s. 557–565.
- Kokan A. R., Khan S. (1967): *Optimum Allocation in Multivariate Surveys: An Analytical Solution*. „Journal of the Royal Statistical Society”, Vol. B 29, s. 115–125.
- Konijn H. S. (1962): *Regression Analysis in Sample Surveys*. „Journal of the American Statistical Association”, Vol. 57.
- Konijn H. S. (1973): *Statistical Theory of Sample Survey and Analysis*. North-Holland Publishing Company, Inc., Amsterdam-London.
- Kordos J. (1988): *Jakość danych statystycznych*. PWE, Warszawa.
- Krysicki W., Bartos J., Dyczka W., Królikowska K., Wasilewski M. (1989): *Rachunek prawdopodobieństwa i statystyka matematyczna. Cz. II. Statystyka matematyczna*. Wydawnictwo Naukowe PWN, Warszawa.
- Lahiri G. W. (1951): *A Method for Sample Selection Providing Unbiased Ratio Estimator*. „Bulletin of the International Statistical Institute”, Vol. 33, s. 133–140.
- Lohr S. L. (1999): *Sampling: Design and Analysis*. Duxbury Press, International Thompson Publishing Company.
- Maddala G. S. (2006): *Ekonometria*. John Wiley, New York.
- Midzuno H. (1950): *An Outline of the Theory of Sampling Systems*. „Annals of the Institute of Mathematical Statistics”, Vol. 1, s. 149–156.

- Murthy M. N. (1977): *Sampling Theory and Practice*. Statistical Publishing Society, Calcutta.
- Miszczak W. (2004): *Projektowanie próby*. Wydawnictwo Akademii Ekonomicznej, Wrocław.
- Miszczak W., Ostasiewicz W., Wawrzynek J. (2008): *Projektowanie próby z elementami planowania eksperymentów*. Wydawnictwo Akademii Ekonomicznej, Wrocław.
- Neyman J. (1933): *Zarys teorii i praktyki badania struktury ludności metodą reprezentacyjną*. Instytut Spraw Społecznych Warszawa.
- Neyman J. (1937): *Outline of Theory of Statistical Estimation Based on the Classical Theory of Probability*. „Philosophical Transaction of the Royal Society”, Vol. A 236, s. 333–380.
- Pawłowski Z. (1972): *Wstęp do statystycznej metody reprezentacyjnej*. Wydawnictwo Naukowe PWN, Warszawa.
- Račkauskas A. (2006): *Hölderian Properties of Partial Sums of Regression Residuals*. „Metrika”, Vol. 63, s. 191–205.
- Račkauskas A., Suquet C. (2006): *Testing Epidemic Changes of Infinite Dimensional Parameters*. „Statistic Inference for Stochastic Processes”, Vol 9, s. 11–134.
- Rao C. R. (1982): *Modele liniowe statystyki matematycznej*. Wydawnictwo Naukowe PWN, Warszawa.
- Rao J. N. K., Hartley H. O., Cochran W. G. (1962): *On a Simple Procedure of Unequal Probability Sampling without Replacement*. „Journal of the Royal Statistical Association”, Vol. B 24, No. 2, s. 482–491.
- Rao J. N. K. (2003): *Small Area Estimation*. John Wiley & Sons Inc. New York.
- Rao T. V. H. (1962): *An Existence Theorem in Sampling Theory*. „Sankhya”, Vol. A 24, s. 327–330.
- Royall R. M. (1970): *On Finite Population Sampling Theory under Certain Linear Regression Models*. „Biometrika”, Vol. 57, s. 377–378.
- Royall R. M. (1971): *Linear Regression Models Finite Population Sampling Theory*. In: *Foundation of Statistical Inference*. V. P. Godambe, D. A. Sprott. Holt, Rinehart, Winston, Toronto.
- Särndal C. E., Swensson B., Wretman J. (1992): *Model Assisted Survey Sampling*. Springer Verlag, New York-Berlin-Heidelberg.
- Sen A. R. (1953): *On the Estimation of Variance in Sampling with Varying Probabilities*. „Journal of the Indian Society of Agricultural Statistics”, Vol. 5, No. 2, s. 119–127.
- Sinha B. K. (1973): *On Sampling Schemes to Realize Preassigned Sets of Inclusion Probabilities of Selection*. „Biometrika”, Vol. 54, s. 499–513.
- Singh P., Srivastava A. K. (1980): *Sampling Schemes Providing Unbiased Regression Estimators*. „Biometrika”, Vol. 67, s. 205–209.
- Smaliak S. A., Titarienko B. P. (1980): *Odporne metody estymacji* (po rosyjsku). Statistika, Moskwa.

- Steczkowski J. (1995): *Metoda reprezentacyjna w badaniach zjawisk ekonomiczno-społecznych*. Wydawnictwo Naukowe PWN, Warszawa.
- Szreder M. (2004): *Metody i techniki sondażowych badań opinii*. PWE, Warszawa.
- Theil H. (1971): *Principles of Econometrics*. John Wiley and Sons, New York-London-Sydney-Toronto.
- Tietjen G., Moore H. (1972): *Some Grubb's Type Statistics for Detection of Several Outliers*. „Technometrics”, Vol. 14, s. 583–597.
- Tillé Y. (2006): *Sampling Algorithms*. Springer Verlag, New York.
- Tripathi T. P. (1973): *Double Sampling for Induction Probabilities and Regression Method of Estimation*. „Journal of the Indian Statistical Association”, Vol. 10, s. 33–46.
- Tschuprov A. A. (1923): *On the Mathematical Expectation of the Moments of Frequency Distribution in the Case of Correlated Observations*. „Metron”, Vol. 2, s. 461–493 i s. 646–683.
- Valliant R., Dorfman A. H., Royall R. M. (2000): *Finite Population Sampling and Inference, a Prediction Approach*. John Wiley and Sons, New York.
- Volter K. M. (1985): *Introduction to Variance Estimation*. Springer Verlag, New York.
- Vos J. W. E. (1980): *Mixing of Direct, Ratio, and Product Method Estimators*. „Statistica Neerlandica”, Vol. 34, s. 209–219.
- Wilson A. L. (1978): *When Is the Chow Test UMP*. „The American Statistician”, Vol. 32, s. 66–68.
- Wywiał J. L. (1988): *Test dla hipotezy o wartości oczekiwanej wielowymiarowej zmiennej o sferycznym rozkładzie normalnym*. „Przegląd Statystyczny” Vol. 35, s. 341–355.
- Wywiał J. L. (1989): *Weryfikacja hipotezy o liniowości trendu na podstawie reszt*. „Wiadomości Statystyczne”, nr 10, s. 29–31.
- Wywiał J. L. (1991): *O planie losowania próby proporcjonalnym do średniej z próby cechy pomocniczej*. „Wiadomości Statystyczne”, nr 6, s. 21–22.
- Wywiał J. L. (1992): *Statystyczna metoda reprezentacyjna w badaniach ekonomicznych (Optymalizacja badań próbkowych)*. Wydawnictwo Akademii Ekonomicznej, Katowice.
- Wywiał J. L. (1993): *Predykcja średniej w populacji pogrupowanej według cech dodatkowych*. „Przegląd Statystyczny”, Vol. XL, zeszyt 3–4, s. 303–308.
- Wywiał J. L. (1995): *Wielowymiarowe aspekty metody reprezentacyjnej*. Ossolineum, Wrocław-Warszawa-Kraków.
- Wywiał J. L. (2000): *On Precision of Horvitz-Thompson Strategies*. „Statistics in Transition”, Vol. 4, No. 5, s. 779–798.
- Wywiał J. L. (2003): *Some Contributions to Multivariate Methods in Survey Sampling*. University of Economics, Katowice.

- Wywiał J. L. (2004): *Quantile Regression Sampling Strategy*. W: *Metoda Reprezentacyjna w Badaniach Ekonomiczno-Społecznych*. Red. J. Wywiał. Wydawnictwo Akademii Ekonomicznej, Katowice, s. 32–42.
- Wywiał J. L. (2007): *Simulation Analysis of Accuracy Estimation of Population Mean on the Basis of Strategy Dependent on Sampling Design Proportionate to the Order Statistic of an Auxiliary Variable*. „Statistics in Transition – new series”, Vol. 8, No. 1, s. 125–137.
- Wywiał J. L. (2008): *Sampling Design Proportional to Order Statistic of Auxiliary Variable*. „Statistical Papers”, April, Vol. 49, No. 2, s. 277–289.
- Yates F., Grundy P. M. (1953): *Selection without Replacement from within Strata with Probability Proportional to Size*. „Journal of the Royal Statistical Society”, Vol. B 15, s. 235–261.
- Zasępa R. (1962): *Badania statystyczne metodą reprezentacyjną*. PWE, Warszawa.
- Zasępa R. (1972): *Metoda reprezentacyjna*. PWE, Warszawa.
- Żądło T. (2008): *Elementy statystyki małych obszarów z programem R*. Wydawnictwo Akademii Ekonomicznej, Katowice.



ISBN 978-83-7246-524-5