Biblioteka WIADOMOŚCI STATYSTYCZNYCH

RYSZARD ZASĘPA

ZARYS METODY REPREZENTACYJNEJ

Tom 39

WARSZAWA 1991

ERRATA
DO „R. ZASĘPA — ZARYS METODY REPREZENTACYJNEJ"

Str.	Wiersz	Jest	Powinno być
14	2 od góry	jednej,	jednej części,
21	4 i 5 od dołu	$x-m$	$x-m$
22	3 od góry	$\left(1+\frac{t^2}{r}\right)\frac{(r+1)}{2}$	$\left(1+\frac{t^2}{r}\right)^{\frac{r+1}{2}}$
25	9 od góry	$P(X=x_k\|Y=y_e)$	$P(X=x_k, Y=y_e)$
27	16 od góry	$=0{,}129\cdot 0{,}5+$	$+0{,}129\cdot 0{,}5+$
28	9 od dołu	0.070	0.270
30	20 od dołu	operatu losowego	operatu losowania
40	12 od dołu	str. 2	str. 38
51	10 od dołu	Y i Y	Y i X
65	3 od dołu	$\frac{\sqrt{1-P}}{nP}$	$\frac{\sqrt{P(1-P)}}{\sqrt{n}}$
66	10 od góry	P_o	NP_0
68	9 od dołu	$N_k = L$	$N_h = N$
68	1 od dołu	Y^k	$\bar{Y}_h$
72	16 od dołu	$/K_h$	$/\sqrt{K_h}$
84	14 od góry	$N=k_n$	$N=kn$
86	9 od dołu	3. Korzystając	4. Korzystając
86	5 od dołu	4. Badając	3. Badając
94	8 od dołu	π_k (suma	π_k :(suma
101	12 od dołu	$-2\bar{x}\bar{y}$	$-2\sqrt{\bar{x}\bar{y}}$
102	5 od góry	przetransportować	przetransponować
109	17 od dołu	naturalna = α	naturalna, α
110	2 od dołu	$Y_i^2 =$	$y_j^2 =$
111	3 od góry	$D\bar{y}''$?	$D_{df}(\bar{y}'')$?
121	1 od góry	Y_i jednostki	Y_i bez jednostki
121	6 od dołu	y dla próby	y_i dla próby
124	5 od góry	uzyskanie kontaktu	niemożność uzyskania kontaktu
124	20 od góry	otrzymało	otrzymano
124	5 od dołu	$n_3=0$	$n_3=30$

GŁÓWNY URZĄD STATYSTYCZNY
POLSKIE TOWARZYSTWO STATYSTYCZNE

ZARYS METODY REPREZENTACYJNEJ

OUTLINE OF SAMPLING METHODS

Opracowanie redakcyjne
Stanisław Jońca

Redakcja techniczna
Zbigniew Karpiński

Okładkę projektował
mgr Krzysztof Dobrowolski

Zakład Wydawnictw Statystycznych 00-925 Warszawa, al. Niepodległości 208.
Nakład 700 egz, ark. druk. 8, Papier druk. kl. V g. 61 × 86
Przekazano do składu we wrześniu 1990 r. Druk ukończono w lipcu 1991
Cena zł. 25000,—
Druk ZWS, Warszawa, zam. 413/90
Informacje w sprawach zbytu publikacji — tel. 25-48-86.

PRZEDMOWA

Rada Główna Polskiego Towarzystwa Statystycznego przedstawia opracowanie prof. dra hab. Ryszarda Zasępy pt. „ZARYS METODY REPREZENTACYJNEJ".

Opracowanie prof. R. Zasępy stanowi kolejne wydawnictwo przygotowane merytorycznie i redakcyjnie przez Polskie Towarzystwo Statystyczne.

Rada Główna PTS wydaje nie tylko „BIULETYN INFORMACYJNY" i współredaguje „Wiadomości Statystyczne", ale także rozpoczęła wydawanie serii tłumaczeń na język polski ważniejszych publikacji zagranicznych z zakresu statystyki (zostało wydane opracowanie dra Richarda PLATKA nt. metodologiczne problemy braku odpowiedzi w ankietowych badaniach społecznych, a w przygotowaniu jest opracowanie Camillo Dagouma nt. analizy rozkładów płac). Będą także publikowane opracowania tematyczne ze statystyki polskiej w języku angielskim.

Rada Główna współuczestniczyła w przygotowaniu, a także finansowała niektóre wydawnictwa Rad Oddziałów PTS — m.in. „Sylwetki statystyków polskich" przygotowane przez Radę Oddziału Łódzkiego (wydane w języku: polskim — 1985; angielskim — 1989; rosyjskim — 1990) oraz „Ludność Galicji 1857—1910" dra Krzysztofa Zamarskiego z Krakowa (1989).

Obok tych wydawnictw w dalszym ciągu będą publikowane w serii „Biblioteka Wiadomości Statystycznych" materiały z ogólnopolskich konferencji naukowych organizowanych przez Polskie Towarzystwo Statystyczne. Dotychczas ukazały się materiały z konferencji w Jabłonnej (1988) nt. badań budżetu czasu, Zwartowie (1988) nt. badań metodą reprezentacyjną, Sulejowie (1989) nt. badania i analizy cen. W przyszłym roku ukażą się materiały z konferencji nt. wskaźników społecznych.

Przygotowanie, merytoryczne i redakcyjne, wszystkich publikacji jest finansowane ze środków PTS, natomiast wydawane są one — za zgodą i dzięki życzliwości Prezesa GUS — dra Franciszka Kubiczka — przez Zakład Wydawnictw Statystycznych GUS.

Opracowanie prof. Ryszarda Zasępy jest efektem kursu szkoleniowego dla wytypowanych pracowników GUS, zorganizowanego przez Biuro Badań i Analiz Statystycznych przy Radzie Głównej PTS, przy pewnej pomocy Wydziału Szkolenia GUS w okresie 24 maja—28 czerwca i 27 września—29 listopada 1989 r. W cyklu 16 wykładów zatytułowanych „ZARYS METODY REPREZENTACYJNEJ" — 14 wykładów przeprowadził

prof. dr R. Zasępa, a 2 wykłady — dr J. Bielecki z Uniwersytetu Gdańskiego.

Kurs zorganizowano przede wszystkim dla celów szkoleniowych, a wszystkie wykłady nagrano na taśmę wideo. Nagrania zrealizował i kopie przygotował red. Waldemar Zaręba z firmy EMI-VIDEO. Kopie nagrań przechowywane przez Centralną Bibliotekę Statystyczną mogą być wypożyczone nieodpłatnie wszystkim zainteresowanym tą problematyką.

W trakcie kursu przyjęto, że uczestnicy otrzymają konspekty wykładów. Te właśnie konspekty, uzupełnione o bibliografię przedmiotu oraz zadania i ćwiczenia, stały się podstawą do przygotowania niniejszej publikacji. Ze względów dydaktycznych i redakcyjnych publikację ograniczono do konspektów wykładów prof. R. Zasępy.

Niezależnie od pełnego cyklu wykładów, dotyczących zarysu metody reprezentacyjnej — BBiAS zorganizowało dwa obszerne wykłady syntetyczne 12 i 13 lipca 1989 r. nt. podstaw teoretycznych — prof. R. Zasępy (3 godziny) oraz nt. zastosowania metody reprezentacyjnej w praktyce — prof. J. Kordosa (3 godz. 25 min.). Wykład prof. J. Kordosa ilustrowany był materiałami pomocniczymi (schematami, definicjami, planszami itp.). Oba wykłady nagrano na taśmę wideo, której kopie mogą być udostępnione przez Centralną Bibliotekę Statystyczną.

Forma konspektu wykładu przygotowanego dla celów dydaktycznych tworzy specyficzny rodzaj materiału wydawniczego zamieszczonego w publikacji. Są to więc przede wszystkim s y n t e t y c z n e informacje dotyczące poszczególnych tematów, ściśle ustrukturyzowane, odciążone od akademickich dywagacji i uciążliwych dowodów poszczególnych twierdzeń. Zarówno wykłady, jak i konspekty przygotowano, przede wszystkim, z myślą o potrzebach statystyków-praktyków, pragnących stosować metodę reprezentacyjną w swojej codziennej pracy zawodowej.

Bezpośrednim organizatorem wykładów i nagrań oraz przygotowania publikacji był kierownik Biura Badań i Analiz Statystycznych przy Radzie Głównej PTS — mgr Wiesław Łagodziński.

Tom recenzował doc. dr hab. Andrzej BALICKI z Uniwersytetu Gdańskiego.

Rada Główna Polskiego Towarzystwa Statystycznego składa serdeczne podziękowania wszystkim osobom, które przyczyniły się do przygotowania i wydania niniejszej publikacji.

Prof. dr hab. Jan Kordos
PREZES RADY GŁÓWNEJ
POLSKIEGO TOWARZYSTWA
STATYSTYCZNEGO

Warszawa, sierpień 1990 r.

PREFACE

The Main Board of the Polish Statistical Association (PTS) presents the elaboration prepared by Prof. Ryszard Zasępa titled: ZARYS METODY REPREZENTACYJNEJ (OUTLINE OF SAMPLE METHOD). This publication is one of the successive issues of the PTS series. The Main Board of the PTS publishes not only the BIULETYN INFORMACYJNY (INFORMATION BULLETIN) it is also co-editor of the monthly „WIADOMOŚCI STATYSTYCZNE" (STATISTICAL NEWS); it has also began to issue translation of the most important foreign statistical publications into Polish. Also publications of Polish Statistics in English language will appear.

The Main Board of the PTS took part in preparing some publications of the PTS Regional Branches and financed their issue; for instance: SYLWETKI STATYSTYKÓW POLSKICH (BIOGRAPHIES OF POLISH STATISTICIANS) prepared by the PTS Regional Branch in Łódź (Polish version issued in 1985, English version — in 1989, Russian version — in 1990). LUDNOŚĆ GALICJI 1857—1910 (POPULATION OF GALICJA 1857—1910) by Krzysztof Zamarski, Kraków 1989.

Besides the above mentioned publications, the Main Board of the PTS prepared for publication papers presented during annual scientific conferences of the PTS in 1988, 1989 and 1990: PROBLEMY BADAŃ STATYSTYCZNYCH METODĄ REPREZENTACYJNĄ (PROBLEMS OF STATISTICAL SURVEYS BY SAMPLING METHOD), BADANIE BUDŻETU CZASU (TIME BUDGET SURVEYS) and BADANIE I ANALIZA CEN (SURVEY AND ANALYSIS OF PRICES). Papers on SOCIAL INDICATORS which were presented during scientific conference in 1990 will be published next year.

The elaboration prepared by Ryszard Zasępa (OUTLINE OF SAMPLE METHOD) came into being in the connection with the training course for the selected Central Statistical Office staff arranged by the Bureau of Statistical Research and Analyses of the PTS Main Board with some assistance of the CSO Training Section. During the training period (24 May—28 June and 27 September—29 November 1989) 16 lectures the series of which was entitled OUTLINE OF SAMPLE METHOD, Prof. R. Zasępa delivered 14 lectures and Dr. J. Bielecki — 2 lectures.

In principle, the course was arranged for training purposes; all the lectures were recorded on a video-tape. The recording and the tape copies were prepared by Mr. Waldemar Zaręba, editor in EMI-VIDEO Firm. The copies

are stored in the Central Statistical Library in Warsaw; they are to be borrowed without charge by everyone who is interested in sampling method problems.

During the course it was decided that all its participants would be provided with summaries of the lectures. Just these summaries supplemented with bibliography and exercises were the basis to prepare the presented publication. Independently from the complete series of lectures on the outline of sample survey method, the PTS Bureau of Statistical Research and Analyses arranged two extensive lectures (July 12 and 13 1989), namely: the lecture on theoretical principles of the sample survey method (3 hours) delivered by Prof. R. Zasępa and the lecture on practical applications of this method delivered by Prof. J. Kordos (3 hours, 25 minutes); this last was illustrated by auxiliary materials, such as schemes, definitions, large-scale illustrations etc. Both the lectures were also recorded on the video-tape; their copies are to be borrowed at the Central Statistical Library.

The form of lectures is a specific kind of editorial materials which was inserted in the presented publication. First of all, they contain synthetic information on the particular lecture subjects; they are rigorously structuralized free of academic consideration of the complicated proofs of theorems. Both the lectures and their summaries were prepared, first of all, for statisticians and other users who want to apply sampling method in their everyday professional work.

The direct arranger of the lectures, of the preparation of summaries and their recording and of the presented publication is the Chief of the Bureau of Statistical Research and Analyses of the PTS Main Board, M.Sc. Wiesław Łagodziński.

The review of the presented publication was prepared by Dr. A. Balicki, Gdańsk University; technical editor was Mr. S. Jońca.

The PTS Main Board would like to thank kindly everyone who took part in the preparation and edition of the presented publication.

Prof. Jan Kordos

PRESIDENT
POLISH STATISTICAL ASSOCIATION

Warsaw, October 1990

SPIS TREŚCI

CONTENTS

OD AUTORA

Pełne badania statystyczne obejmujące całe populacje, pociągają za sobą duże koszty, wymagają dużej liczby kwalifikowanych osób do zbierania posiadanych informacji i są bardzo uciążliwe dla osób czy instytucji udzielających tych informacji. Jest więc zrozumiała konieczność korzystania z odpowiednich technik badania częściowego. Najważniejszą z nich jest technika badania metodą reprezentacyjną. Pomimo olbrzymiego rozwoju teorii badania reprezentacyjnego jej stosowanie w warunkach krajowych napotyka na opory, wynikające głównie z niedostatecznej znajomości teorii.

Dla wypełnienia tej luki Biuro Badań i Analiz Statystycznych Polskiego Towarzystwa Statystycznego zorganizowało kurs szkoleniowy dla pracowników Głównego Urzędu Statystycznego, składający się z 16 wykładów, a w tym z 14 wykładów prowadzonych przez autora niniejszego opracowania.

Dla każdego z 14 podstawowych wykładów uczestnicy kursu szkoleniowego otrzymywali na bieżąco konspekt wykładu. Po zakończeniu kursu autor przejrzał i skorygował opracowane konspekty, uzupełnił je zadaniami oraz wykazem literatury uzupełniającej. Każdy wykład był poświęcony określonej grupie zagadnień. Jedynie wykład 7 odbywający się po przerwie wakacyjnej, został potraktowany odrębnie, jako przypomnienie ważniejszych dotąd wprowadzonych terminów, pojęć, problemów i ich rozwiązań.

Biorąc pod uwagę audytorium, autor w swych wykładach skoncentrował się na tych zagadnieniach, które są szczególnie ważne dla organizatorów badań reprezentacyjnych w GUS. Dlatego rozważania teoretyczne zostały maksymalnie ograniczone oraz pominięto wyprowadzenie wzorów i twierdzeń wskazując, gdzie można znaleźć ich dowody. Założono, że uczestnicy kursu będą korzystać z konsultacji ekspertów przy projektowaniu badania reprezentacyjnego, natomiast ich wiadomości powinny wystarczać na podjęcie decyzji, co do możliwości prowadzenia badania reprezentacyjnego określonej populacji, w świetle zasad teorii metody reprezentacyjnej, materiałów potrzebnych dla prawidłowego zaplanowania tego badania oraz jego przeprowadzenia, oceny wielkości próby w celu uzyskania odpowiedniej precyzji szacowanych parametrów populacji itp.

Dla ułatwienia korzystania z konspektu każdy wykład podzielony został na pewną liczbę ustępów wypunktowanych, jako para liczb składających się z numeru wykładu oraz numeru ustępu, przedzielonych kropką. Na przykład 11.4 oznacza ustęp czwarty wykładu 11. Powołując się na literaturę, której wykaz znajduje się na końcu, liczba w nawiasach kwadratowych pokazuje

pozycję danej publikacji w wykazie literatury. Np. [7] *oznacza, że chodzi o artykuł J. Kordosa wykazany w pozycji siódmej LITERATURY.*

Autor pragnie w tym miejscu serdecznie podziękować Panu Wiesławowi Łagodzińskiemu za wkład pracy organizacyjnej przy przygotowaniu i realizacji kursu szkoleniowego oraz konspektów do wykładów, jak również Panom Bronisławowi Lednickiemu i Florianowi Ruszkowskiemu za pomoc w czasie prowadzenia wykładów, a ponadto — Panu doc. dr. hab. Andrzejowi Balickiemu, recenzentowi, za zgłoszone uwagi krytyczne przyczyniające się do pojaśnienia tekstu oraz usunięcia zauważonych błędów i nieścisłości.

Prof. dr hab. Ryszard Zasępa

FROM THE AUTHOR

Complete statistical surveys which comprise whole populations involve high costs, numerous qualified staff to collect data and, on the other hand, they are burdensome for persons and institutions giving information. Thus, there is the necessity to use some techniques and methods of partial surveys. The most important of them is the sample survey method. In spite of the intense development of the theory of this method, its application in our country encounters many difficulties, mainly those resulting from unsatisfactory theory knowledge. To stop this gap, the Polish Statistical Association, Bureau of Statistical Research and Analyses arranged a training course for the Central Statistical Office (CSO) staff comprising 16 lectures of which 14 were delivered by the author of the presented publication.

Persons participating in the 14 basic lectures were provided before each running lecture with its summaries. After the course the author revised and corrected each summary and supplemented it with exercises and bibliographical references. Each lecture dealt with a set of selected problems; only the 7-th lecture, which was delivered after vacation, was treated separately to remind the erlier introducted terms, notions, problems and their solutions. Taking into account the needs of the audience the author focussed his attention on the problems which are especially important for the organizers of the CSO sample surveys. Hence, the theoretical deliberations were as much as possible reduced, also the deductions of formulas and theorems were omitted; in return, appropriate bibliography references were given.

It was assumed that the participants of the course would consult with the sampling experts whereas their knowledge in this field should be sufficient to take decisions concerning the applicability of sample survey of a given population in the light of the theory of sampling method, the materials necessary for proper survey planning, its performance, the estimation of sample size to enable the required precision of estimated parameters.

To make easier the use of lecture summary, each lecture was divided into several paragraphs marked with two numbers; the first one was lecture number and the second one denotes paragraph number. For instance: 11.4 denotes paragraph 4 of the 11-th lecture. Quotations of and references to the literature (bibliography is given at the end of the presented publication) are marked with the number in the square brackets which indicates the item of the quoted (or referred) publication. For instance, [7], indicates the publication of Prof. J. Kordos shown under the item 7 of the bibliography.

ACKNOWLEDGEMENTS

The author would like to thank cordially Mr. Wiesław Łagodziński for his contribution to the organizational work during the preparation and realization of the training course and of the summaries of lectures delivered during the course.

Also he expresses his cordial thanks to Mr. Bronisław Lednicki and to Mr. Florian Ruszkowski for their assistance during delivery of the lectures as well as to Assistant Professor Andrzej Balicki, for his critical comments which have enabled to make the texts clearer and to detect and correct errors or inaccuracies.

Prof. Ryszard Zasępa

Wykład 1. Badanie statystyczne i jego błędy

1.1. Przedmiotem badań statystycznych są tzw. zbiorowości statystyczne, to znaczy takie zbiorowości, których elementy (jednostki, obiekty) są ze sobą w pewien sposób powiązane, a równocześnie występują między nimi różnice, czyli są niejednorodne. Badane zbiorowości statystyczne będziemy nazywać populacjami generalnymi lub wprost **populacjami**.

1.2. Ograniczamy się do przypadków, gdy badane populacje są skończone. Oznaczając liczbę elementów populacji literą N duże, elementy populacji możemy w pewien sposób uporządkować w ciąg skończony. Numer k elementu będącego k-tym wyrazem tego ciągu uznajemy za **identyfikator** elementu. Formalnie, krótko możemy przedstawić populację N-elementową, jako zbiór:

$$U=\{1, 2, 3, ..., N-1, N\} \tag{1}$$

lub jeszcze krócej, jako:

$$U=\{k: 1\leqslant k\leqslant N\},\ k \text{ — liczba naturalna} \tag{1a}$$

1.3. Jednostki badanej populacji interesują nas ze względu na posiadanie określonego wariantu jakiejś cechy niemierzalnej (np. płeć męska osoby, wykształcenie wyższe, zawód rolnika) lub ze względu na wartość jakiejś cechy mierzalnej (np. wiek maszyny, liczba godzin przepracowanych przez osobę w ciągu określonego tygodnia). Badane cechy będziemy oznaczali dużymi, końcowymi literami alfabetu: X, Y, U, W, Z itd. Jeśli badamy jakąś cechę niemierzalną i wyróżniamy pewien wariant tej cechy, to oznaczając cechę literą X przyjmujemy symbol X_k ($k=1, 2, ..., N$) dla k-tego elementu, jako fakt posiadania ($X_k=1$) lub nieposiadania ($X_k=0$) wyróżnionego wariantu cechy X. Odpowiednio, badając cechę mierzalną Y symbolem Y_k oznaczamy wartość cechy Y dla k-tego elementu populacji.

1.4. Badaniem statystycznym obejmujemy często nie jedną, lecz kilka populacji. Np. spis ludności często obejmuje również spis budynków zamieszkanych, spis mieszkań oraz gospodarstw domowych i gospodarstw rolnych. Wówczas badamy pięć populacji. Badania statystyczne prowadzone w Głównym Urzędzie Statystycznym oraz urzędach wojewódzkich są z natury wielocelowe. Obiektem analizy statystycznej jest na ogół zarówno cała populacja, jak i pewne wyróżnione jej części, np. w populacji osób wyróżniamy z przyczyn analitycznych „podpopulację" kobiet w wieku prokreacji. Takie „podpopulacje" nazywamy **domenami** (lub dziedzinami studiów). Jeśli z innych

przyczyn niż analitycznych populację dzielimy rozłącznie i zupełnie, tzn. każda jednostka populacji należy do jednej i tylko jednej, wówczas nazywamy **warstwami**. Zauważmy, że domena nie będąca warstwą będzie się krzyżować z różnymi warstwami. **Przykład**: populację tworzą zakłady przemysłowe, które dzielimy według gałęzi (według klasyfikacji gospodarki narodowej); gałęzie stanowią warstwy. Domeną niech będą zakłady przemysłowe, zatrudniające co najmniej 50 osób. Takie zakłady mogą się znajdować w różnych gałęziach przemysłu (warstwach).

Uwaga 1.1. Zarówno warstwy, jak i domeny są pewnego rodzaju podpopulacjami zawartymi w badanej populacji generalnej.

1.5. Celem badania statystycznego jest przeważnie uzyskanie charakterystyk liczbowych, opisujących własności badanych jednostek. Te charakterystyki nazwiemy **parametrami populacji**. Pojęcie parametru populacji nie jest identyczne z pojęciem **parametru rozkładu** prawdopodobieństwa w statystyce matematycznej. Wyniki badania statystycznego na ogół przedstawia się w formie tablic statystycznych, których rubryki (kratki) określają wartości liczbowe odpowiednich parametrów populacji.

1.6. Najczęstsze parametry (populacje) to:

a) **wartość globalna** cechy mierzalnej (X)

$$X = X_1 + X_2 + \ldots + X_{N-1} + X_N = \sum_{k=1}^{N} X_k \tag{2}$$

b) **liczba N jednostek wyróżnionych;** ze względu na pewien wariant cechy niemierzalnej (X) określana również wzorem (2), z tym że X_k może przyjmować wartość 1 lub 0,

c) **wartość średnia** cechy mierzalnej (X)

$$\bar{X} = \frac{1}{N} \sum_{k=1}^{N} X_k = X/N \tag{3}$$

d) **frakcja P jednostek wyróżnionych;** ze względu na pewną cechę niemierzalną (X) określona wzorem (3), z tym że X_k może przyjmować wartość 1 lub 0; $P = N'/N$,

e) **iloraz wartości globalnych** (lub średnich) dwóch cech mierzalnych (X, Y)

$$R = \frac{Y}{X} = \frac{\bar{Y}}{\bar{X}} \tag{4}$$

f) **wariancja** cechy mierzalnej (X) określona jako:

$$\sigma_X^2 = \frac{1}{N} \sum_{k=1}^{N} (X_k - \bar{X})^2 \tag{5}$$

albo jako:

$$S_X^2 = \frac{1}{N-1} \sum_{k=1}^{N} (X_k - \bar{X})^2 = \frac{N}{N-1} \sigma_X^2 \tag{5a}$$

g) w przypadku wyróżnionego wariantu cechy niemierzalnej (X) odpowiednikami (5) i (5a) będą:

$$\sigma_X^2 = PQ, \quad Q = 1 - P \tag{6}$$

$$S_X^2 = \frac{N}{N-1} PQ \tag{6a}$$

h) **odchylenie standardowe** cechy (X) określone jako:

$$\sigma_X \text{ albo } S_X; \quad \sigma_X = \sqrt{\sigma_X^2}, \quad S_X = \sqrt{S_X^2} \tag{7}$$

i) **współczynnik korelacji** dwóch cech mierzalnych (X, Y)

$$\rho_{xy} = \frac{1}{N\sigma_x \sigma_y} \sum_{k=1}^{N} (X_k - \bar{X})(Y_k - \bar{Y}) \tag{8}$$

Przykłady 1.1. Niech badaną populację stanowią indywidualne gospodarstwa rolne, w pewnej gminie. Cechą badaną jest liczba trzody chlewnej. Liczbę trzody chlewnej w gminie obliczymy jako wartość globalną według wzoru (2). Dzieląc ją przez liczbę gospodarstw otrzymamy zgodnie z (3), średnią liczbę trzody chlewnej w gospodarstwie. Wariancję liczby trzody chlewnej w gospodarstwie otrzymamy stosując wzór (5) albo (5a). Z pierwiastka kwadratowego z wariancji wyniknie (7) odchylenie standardowe tej cechy.

1.2. W powyższej populacji badamy, czy gospodarstwo posiada ciągnik czy nie (w pierwszym przypadku $X_k = 1$, w drugim $X_k = 0$). Są to dwa warianty cechy niemierzalnej. Liczbę gospodarstw posiadających ciągniki określimy jako liczbę N' gospodarstw wyróżnionych (gdyż posiadają ciągnik), czyli według wzoru (2). Udział P gospodarstw, posiadających ciągnik w badanej populacji, obliczymy jako frakcję gospodarstw wyróżnionych według wzoru (3).

1.3. W opisanej wyżej w punkcie 1.1 populacji badamy ponadto dwie cechy gospodarstw: a) liczbę sztuk bydła (Y) oraz b) powierzchnię użytków rolnych w ha (X). W celu obliczenia ile sztuk bydła przypada na ha użytków rolnych w badanej gminie stosujemy wzór (4), przy czym Y oznacza liczbę sztuk bydła w gminie według wzoru (2), zastępując literę X literą Y, natomiast X — powierzchnię użytków rolnych w gminie według tego samego wzoru (2). Wynik R mnożymy zwykle przez 100, otrzymując wskaźnik „liczba sztuk bydła na 100 ha użytków rolnych". Może nas interesować także współczynnik korelacji liczby sztuk bydła w gospodarstwie i powierzchni użytków rolnych. Wówczas zastosujemy wzór (8).

Uwaga 1.2. Duża litera (X, Y, U, ...) ma podwójne znaczenie: a) jako symbol cechy oraz b) jako wartość globalna cechy. Z kontekstu wynikać będzie, o które znaczenie chodzi.

Uwaga 1.3. We wzorach (5)—(8) indeks „x" będzie opuszczany, o ile wiadomo o jaką cechę chodzi.

Uwaga 1.4. We wzorach (2)—(8) litery X, Y mogą być zastąpione dowolną inną literą alfabetu, np. przez Y, X bądź U, W itp.

1.7. Badanie statystyczne może być: **pełne** (wyczerpujące), kiedy badaniem obejmujemy całą populację generalną bądź **częściowe,** kiedy badaniem obejmujemy odpowiednio wybraną część populacji generalnej.

Stwierdzenie. Każde badanie statystyczne, dotyczące dużej populacji zawiera pewne błędy (zwane błędami statystycznymi).

Staranne przygotowanie i przeprowadzenie badania statystycznego pozwala znacznie zredukować zasięg błędów statystycznych. Na dokładność badania statystycznego wpływa przede wszystkim:

a) ścisłe określenie i ograniczenie badanej populacji; posiadanie dokumentacji adresowej jednostek badania bądź szerszych jednostek, za pośrednictwem których docieramy do jednostek populacji; dokumentacja ta powinna być pełna (bez opuszczeń!) i aktualna;
b) ścisłe sformułowanie pytań, definicji i klasyfikacji badanych cech;
c) technika zbierania danych;
d) organizacja badania, dobór personelu uczestniczącego w badaniu i jego szkolenie;
e) metody kontroli w trakcie różnych etapów badania;
f) metody opracowania zebranego materiału.

Badanie statystyczne powoduje koszty. Rozmiar tych kosztów zależy od zastosowanych technik postępowania, w różnych etapach badania. Pewne techniki mogą w poważnym stopniu zredukować powstawanie błędów w badaniu statystycznym, ale ich zastosowanie może być bardzo pracochłonne i powodować duże koszty. Planujący badanie statystyczne ma za zadanie takie rozłożenie kosztów badania na czynności występujące w różnych etapach jego realizacji, aby z jednej strony nie przekroczyć funduszu na badanie, a z drugiej — utrzymać ewentualne błędy badania w pożądanych granicach.

1.8. Każde badanie statystyczne odbywa się w konkretnych warunkach. Niektóre z nich są poza zasięgiem organizatorów badania, np. wydarzenia polityczne, ekonomiczne, klęski żywiołowe. Mogą one wpływać na dokładność oraz kompletność uzyskanych informacji statystycznych, inne mogą być regulowane, a co najmniej kontrolowane przez organizatorów badania statystycznego. Do nich należą warunki, wynikające z czynności wymienionych w punktach a)—f). Warunki te nazywane są **istotnymi warunkami badania statystycznego;** gdyby one były idealnie spełnione w czasie przeprowadzania **pełnego badania** dokładność wyników byłaby zapewniona, oczywiście, o ile badanie zostało prawidłowo zaprogramowane. W praktyce, jak to zostało wyżej podkreślone, zarówno przy zbieraniu danych, jak i ich opracowaniu powstają błędy. Np. dokumentacja, jaką się posługujemy w celu ujęcia jednostek populacji generalnej, może nie być aktualna z powodu niedokładnej rejestracji tych jednostek; w czasie spisu ludności rachmistrz spisowy, pomimo parokrotnych wizyt, nie zastaje nikogo w domu i w rezultacie może wystąpić pominięcie tych osób; niewłaściwa interpretacja pytania w kwestionariuszu (formularzu) statystycznym może powodować błąd w odpowiedzi; inny błąd odpowiedzi może wynikać z niedokładności pomiarów lub obliczeń, stanowiących podstawę informacji bądź z obawy, że

informacja prawdziwa zostanie wykorzystana przeciw interesom respondenta lub jednostki prawnej, którą reprezentuje; inne rodzaje błędów to: pomyłka w trakcie symbolizacji, pomyłka przy przenoszeniu danych badania na taśmę magnetyczną, błąd drukarski itp.

1.9. Powyższe przykłady błędów statystycznych odnoszą się do błędów występujących zarówno w badaniu pełnym, jak i w badaniu częściowym. Takie błędy nazywamy **błędami nielosowymi.** Błędy nielosowe, czyli błędy, które powstają w poszczególnych etapach badania statystycznego, dzielimy na: a) **błędy pokrycia,** tj. pominięcia, wielokrotne ujęcia lub błędne włączenia jednostki do badanej populacji, w tym błędy powstałe z braku odpowiedzi, czyli wskutek nieuzyskania odpowiedzi od jednostki, wchodzącej w skład badanej populacji; b) **błędy treści** (zawartości), obejmujące błędy odpowiedzi, błędy opracowywania danych oraz błędy analizy i prezentacji wyników.

1.10. Przypuśćmy, że szacujemy wartość globalną cechy mierzalnej w populacji N-elementowej. Niech X_k oznacza wartość tej cechy dla k-tego elementu populacji ($k=1, 2, ..., N$). Celem badania jest określenie wartości sumy:

$$X=\sum_{k=1}^{N} X_k \tag{9}$$

Uzyskana w czasie zbierania informacji statystycznej indywidualna wartość cechy dla k-tego elementu wynosi Y_k. Jeśli jest to **wartość prawdziwa,** to $Y_k=X_k$, ale jeżeli różni się ona od wartości prawdziwej, to $Y_k=X_k+d_k$. Ogólnie, d_k oznacza błąd obserwacji (odpowiedzi) (uwaga: $d_k\neq 0$ lub $d_k=0$), czyli mamy:

$$Y_k=X_k+d_k,\ k=1, 2, ..., N \tag{10}$$

Jeśli występują błędy pokrycia, powodujące pominięcia pewnych elementów populacji, to dla pewnych k brak jest informacji o wartości badanej cechy. Oznaczmy literą W zbiór tych k:

$$W=\{k_1, k_2, ..., k_l\} \tag{11}$$

Ostatecznie zamiast X otrzymujemy:

$$Y=\sum_{k\notin W}(X_k+d_k)=\sum_{k\notin W}X_k+\sum_{k\in W}d_k=X'+\Delta$$

$$Y=X+\left(\Delta-\sum_{i=1}^{l}X_{k_i}\right)=X+\Delta' \tag{12}$$

Wielkość Δ' jest błędem oceny wartości globalnej X; błąd ten może być dodatni lub ujemny; błąd ten nazywany jest również **obciążeniem** szacunku wartości globalnej. Przyjmuje się często przypuszczenie, że występujące indywidualne błędy odpowiedzi są różnych znaków i wzajemnie się znoszą. W praktyce zwykle te błędy są

jednokierunkowe, tendencyjne i dlatego nazywa się je także **błędami systematycznymi,** powodującymi obciążenie uzyskiwanych ocen wartości globalnych, przeciętnych i innych parametrów populacji generalnej. Zauważmy, że gdyby badanie zostało powtórzone w krótkim okresie, w tych samych istotnych warunkach, prawdopodobnie wynik oceny X byłby inny niż poprzednio, $Y = X + \Delta''$, $\Delta' \neq \Delta''$. **Wynik badania statystycznego jest więc szacunkiem określonego parametru populacji, często niedokładnym.**

1.11. Redukcja niektórych źródeł błędów nielosowych jest pracochłonna i kosztowna. Możliwa byłaby, gdyby ograniczyć badanie do odpowiednio wybranej części, czyli **próby,** a więc przez zastąpienie badania pełnego — badaniem częściowym. Jaką wybrać próbę, aby wyniki badania nie były obarczone dużymi błędami (innymi niż te, o których wyżej była mowa), spowodowanymi niereprezentatywnością wybranej części?

1.12. Wśród stosowanych metod wyboru próby do badań częściowych można wyróżnić: a) *wybór celowy,* b) *wybór losowy* oraz c) *wybór pseudolosowy.* Wybór celowy polega na subiektywnym wyborze do próby jednostek, które uważa się za „typowe" lub „przeciętne". Wybór **celowy,** jako wybór subiektywny, jest przeważnie wyborem tendencyjnym i nie daje próby o oczekiwanej reprezentatywności. Nie posiadamy miar, pozwalających ocenić czy zastosowany wybór celowy dostarcza próby, której struktura, według badanych cech, nie różni się istotnie od struktury populacji. Dlatego stosowanie wyboru celowego jest niebezpieczne. Przeciwnie, wybór **losowy** próby, w którym decyzja o wyborze danego elementu populacji generalnej do próby pochodzi od określonego mechanizmu losowego pozwala na ocenę dobroci szacunków, na podstawie danych próby. Opierając się na rozważaniach probabilistycznych teoria losowego wyboru próby została dobrze opracowana. Badania częściowe, w których próba została losowo wybrana, noszą nazwę **badań metodą reprezentacyjną** lub krótko, **badań reprezentacyjnych.** Do metod wyboru **pseudolosowego** należy zaliczyć tzw. **wybór kwotowy.** Ideą metody jest ustalenie kwot (udziałów) określonych podpopulacji, które powinny się znaleźć w próbie i wyznaczania ankieterom liczb jednostek populacji o odpowiednich cechach, które mają być zbadane. Wybór takich jednostek jest pozostawiony ankieterom. Wybór kwotowy ma podobne wady, jak wybór celowy.

1.13. Próba w badaniu reprezentacyjnym nie będzie miała z reguły identycznej struktury z populacją generalną. Dlatego oceny z próby (losowej) są obarczone pewnym błędem. Błąd ten nazywamy **błędem losowym.** W dalszych wykładach przekonamy się, jak oceniać wielkość błędu losowego w badaniu metodą reprezentacyjną. Błąd ten zależy od schematu losowego wyboru próby, od wielkości próby oraz od rozkładu badanej cechy i postaci szacowanego parametru. W badaniu reprezentacyjnym błędy statystyczne mogą być losowe oraz nielosowe (podobnie, jak w badaniu pełnym). Dodatkowe błędy nielosowe mogą powstać wskutek niedokładnego stosowania reguł wyboru losowego bądź stosowania niewłaściwych (obciążonych) estymatorów.

1.14. **LITERATURA UZUPEŁNIAJĄCA:** J. Kordos [5] str. 16—27, 39—54, [6] str. 13—50, R Zasępa [19] str. 23—36.

Wykład 2. Wybrane elementy rachunku prawdopodobieństwa

2.1. Doświadczenie, którego wynik, w obiektywnie identycznych warunkach może być różny i nie da się go przewidzieć, nazywamy **doświadczeniem losowym.** Każdemu doświadczeniu losowemu odpowiada określony zbiór możliwych, różnych wyników zwanych **zdarzeniami elementarnymi,** tj. zbiór zdarzeń elementarnych. Dowolny podzbiór zbioru zdarzeń elementarnych nazywa się **zdarzeniem losowym** albo (zdarzeniem); zdarzenia losowe oznaczamy dużymi literami A, B, C,... Przy pomocy operacji logicznych z danych zdarzeń możemy tworzyć nowe. **Sumą** zdarzeń A, B nazywamy zdarzenie[1]) „$A \cup B$", występujące wówczas, gdy zachodzi A bądź B, bądź oba równocześnie. **Iloczynem** zdarzeń A, B nazywamy zdarzenie „$A \cap B$" występujące wówczas, gdy równocześnie oba zachodzą. Zdarzeniem losowym **przeciwnym** zdarzenia A nazywamy zdarzenie „$\bar{A}$", występujące wtedy i tylko wtedy, gdy nie zachodzi A. Wyróżniamy ponadto zdarzenie **pewne,** czyli zawsze występujące w danym doświadczeniu oraz zdarzenie do niego przeciwne, czyli zdarzenie **niemożliwe,** które w wyniku doświadczenia nigdy nie może wystąpić. Zdarzenia A, B, które nie mogą zachodzić równocześnie nazywamy **rozłącznymi** (wyłączającymi się).

Przykład 2.1. Rzucamy kostką sześcienną. A — zdarzenie, polegające na wyrzuceniu jednego lub dwóch oczek, B — zdarzenie, polegające na wyrzuceniu parzystej liczby oczek; $A \cup B$ — wyrzucenie jednego oczka bądź parzystej liczby oczek, $A \cap B$ — wyrzucenie dwóch oczek, $\bar{A}$ — wyrzucenie więcej niż dwóch oczek. Zdarzenie, polegające na wyrzuceniu jednego oczka bądź większej liczby oczek jest zdarzeniem pewnym. C — zdarzenie, polegające na wyrzuceniu nieparzystej liczby oczek. Zdarzenia B, C są zdarzeniami rozłącznymi, ich iloczyn $B \cap C$ jest więc zdarzeniem niemożliwym.

2.2. **Prawdopodobieństwo** zdarzenia A jest pewną liczbą nieujemną $P(A)$, przypisaną temu zdarzeniu taką, że:

a) dla dowolnego ciągu zdarzeń A, B, C,... parami rozłącznych

$$P(A \cup B \cup C \cup \ldots) = P(A) + P(B) + P(C) + \ldots$$

b) jeśli E jest zdarzeniem pewnym, to $P(E) = 1$.

Teoria prawdopodobieństwa nie zajmuje się sprawą „poprawnego" określenia prawdopodobieństw zdarzeń byleby spełnione były warunki a) i b). Jeśli zbiór możliwych różnych wyników doświadczenia obejmuje n zdarzeń elementarnych „jednakowo możliwych" oraz m z nich odpowiada realizacji zdarzenia A, to **klasyczna** definicja prawdopodobieństwa określa $P(A) = m/n$.

Z definicji $P(A)$ wynika, że a) $0 \leqslant P(A) \leqslant 1$, b) $P(\bar{A}) = 1 - P(A)$, c) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

---

[1]) W tej definicji zakładamy, zgodnie z zasadą badań reprezentacyjnych, że zbiór zdarzeń elementarnych jest skończony albo co najmniej przeliczalny.

2.3. Obliczanie prawdopodobieństwa ułatwia często posługiwanie się pojęciem **prawdopodobieństwa warunkowego** zdarzenia A przy założeniu, że wystąpiło zdarzenie B; oznaczamy je $P(A|B)=P(A\cap B):P(B)$. Gdy $P(A|B)=P(A)$ zdarzenia A, B nazywamy **niezależnymi.** Wówczas $P(A\cap B)=P(A)\cdot P(B)$.

2.4. Jeśli w doświadczeniu losowym występuje zawsze jedno ze zdarzeń $B_1, B_2,..., B_n$ parami rozłącznych, to dla dowolnego zdarzenia A zachodzi wzór $P(A)=P(A|B_1)\cdot P(B_1)+P(A|B_2)\cdot P(B_2)+...+P(A|B_n)\cdot P(B_n)$. Wzór ten upraszcza obliczenie $P(A)$, gdy prawdopodobieństwa warunkowe $P(A|B_i)$ dla $i=1, 2,..., n$ są łatwe do obliczenia lub dane bezpośrednio.

2.5. Wynik doświadczenia losowego wyraża się często liczbą bądź układem liczb (wektorem). Wówczas mówimy, że mamy do czynienia ze **zmienną losową.** Zmienne losowe oznacza się dużymi końcowymi literami alfabetu: $X, Y, Z, U,...$ W przypadku zmiennej losowej X zakładamy, że jest określone prawdopodobieństwo, że przybierze ona wartość z przedziału [a, b), $P(a\leqslant X<b)$ dla $a<b$. Jeżeli zmienna losowa X przybiera tylko wartość i ze skończonego lub przeliczalnego zbioru $(x_1, x_2,..., x_n,...)$, nazywamy ją **zmienną losową dyskretną.** Jej rozkład opisuje **funkcja prawdopodobieństwa** $p_k=P(X=x_k)$ dla $k=1, 2, 3,...$ Jeżeli zmienna losowa X może przybierać dowolną wartość z pewnego przedziału liczbowego, nazywamy ją **zmienną losową ciągłą.**

2.6. Rozkład zmiennej losowej X opisujemy często przy pomocy funkcji zwanej **dystrybuantą** tej zmiennej, $F(x)=P(X<x)$ określającej prawdopodobieństwo tego, że zmienna losowa przybierze wartość mniejszą od x (danej liczby). Dystrybuanta $F(x)$ jest funkcją niemalejącą, określoną na zbiorze liczb rzeczywistych. W przypadku zmiennej losowej ciągłej X istnieje nieujemna funkcja $f(x)$ taka, że dla dowolnych $a<b$, $P(a\leqslant X<b)=\int_a^b f(x)\,dx$. Funkcję $f(x)$ nazywamy **funkcją gęstości** ciągłej zmiennej losowej X albo krótko **gęstością.** W tym przypadku dystrybuanta $F(x)=\int_{-\infty}^{x} f(x)\,dx$ oraz jej wykres jest linią ciągłą, przybierającą w przedziale liczb rzeczywistych wartości od 0 do 1. Funkcja gęstości $f(x)$ — równa pochodnej dystrybuanty w punkcie x, $f(x)=F'(x)$, jest ciągła.

2.7. Przypuśćmy, że populację tworzy N osób zarejestrowanych w kartotece indywidualnej. Wśród tych osób liczba kobiet wynosi M. Z kartoteki wybieramy raz za razem łącznie n kart w taki sposób, że przy kolejnym wyborze karty wszystkie pozostałe w kartotece karty mają taką samą szansę, że zostaną wybrane. Wybranych kart nie odkładamy z powrotem do kartoteki dopóki nie wybierzemy z niej łącznie n kart ($n<N$), notując liczbę m kobiet, figurujących na wybranych kartach.

Liczba m kobiet jest zmienną losową dyskretną, mogącą przybierać wartości: 0, 1, 2, ..., $min\,(M,n)$. Można udowodnić, że funkcja prawdopodobieństwa tej zmiennej losowej wynosi:

$$P(m=r|M,\ N,\ n)=\frac{\binom{M}{r}\binom{N-M}{n-r}}{\binom{N}{n}} \quad \text{dla } r=0,\ 1,\ 2,\ 3,...,\ min\,(M,n) \tag{13}$$

Symbol $\binom{M}{r}=\frac{M!}{r!(M-r)!}$, $M!=1\cdot 2\cdot 3,\ \ldots\cdot(M-1)\cdot M$, $0!=1$.

Rozkład określony w (13) nosi nazwę **rozkładu hipergeometrycznego.**

Przykład 2.2. $N=20$, $M=8$, $n=4$. Funkcję prawdopodobieństwa przedstawiamy w postawie tabelarycznej:

r	0	1	2	3	4
$p_r=P(m=r)$	0,102	0,363	0,382	0,139	0,014

Uwaga 2.1. Jeśli N, n rosną, to dystrybuanta rozkładu hipergeometrycznego coraz mniej różni się od dystrybuanty rozkładu normalnego. Mówimy zatem, że rozkładem granicznym zmiennej losowej o rozkładzie hipergeometrycznym jest rozkład normalny.

2.8. Zmienna losowa ciągła X ma **rozkład normalny,** jeśli funkcja gęstości jest określona wzorem:

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}\ exp\left[-\frac{(x-m)^2}{2\sigma^2}\right] \text{ dla dowolnej liczby rzeczywistej } x \qquad (14)$$

przy czym $\sigma>0$ oraz m są pewnymi stałymi, tzw. **parametrami rozkładu** $N(m,\sigma^2)$. Wykres gęstości (wykres 1) rozkładu normalnego nosi nazwę krzywej Gaussa. Krzywa ta ma kształt dzwonu i jest symetryczna względem pionowej, przechodzącej przez punkt osi odciętych m. Rozkład $N(0,1)$, w którym $m=0$, $\sigma=1$ nosi nazwę **standardowego rozkładu normalnego.** Dystrybuantę i gęstość takiej zmiennej oznaczymy $\Phi(u)$ i $\varphi(u)$. Oznaczmy odpowiednio $F(x)$, $f(x)$ dystrybuantę i gęstość w punkcie x zmiennej losowej $N(m,\ \sigma^2)$. Mają miejsce związki:

a) $\varphi(-u)=\varphi(u)$,

b) $\Phi(-u)=1-\Phi(u)$,

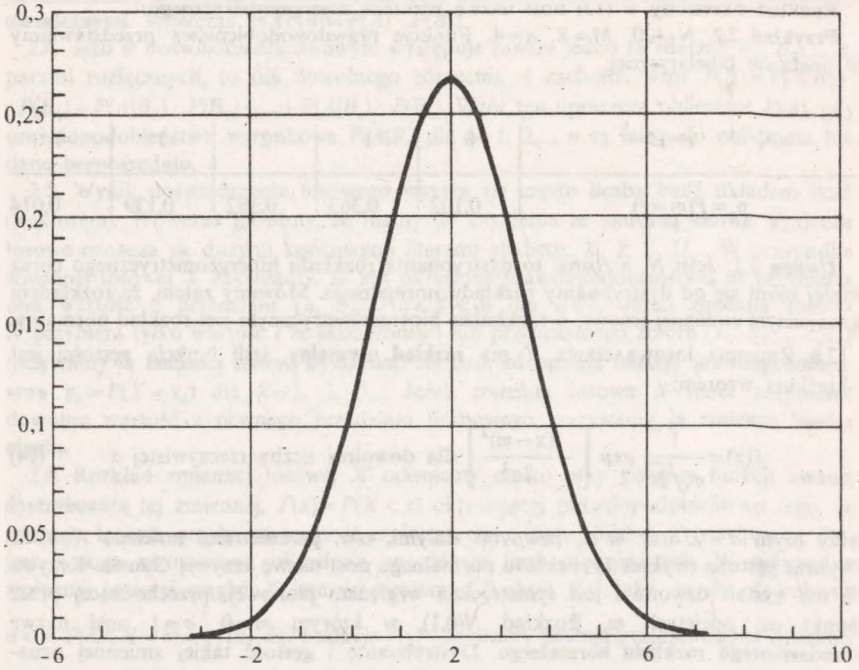
c) $f(x)=\frac{1}{\sigma}\varphi\left(\frac{x-m'}{\sigma}\right)$,

d) $F(x)=\Phi\left(\frac{x-m'}{\sigma}\right)$,

e) $P(a\leqslant X<b)=F(b)-F(a)$.

W tablicach statystycznych podaje się zwykle wartości $\varphi(u)$ oraz $\Phi(u)$. Pozwalają one uzyskać wartości $f(x)$ oraz $F(x)$, zgodnie z powyższymi zależnościami c) oraz d).

Wykres 1. FUNKCJA GĘSTOŚCI ROZKŁADU NORMALNEGO $N(21, 2, 25)$



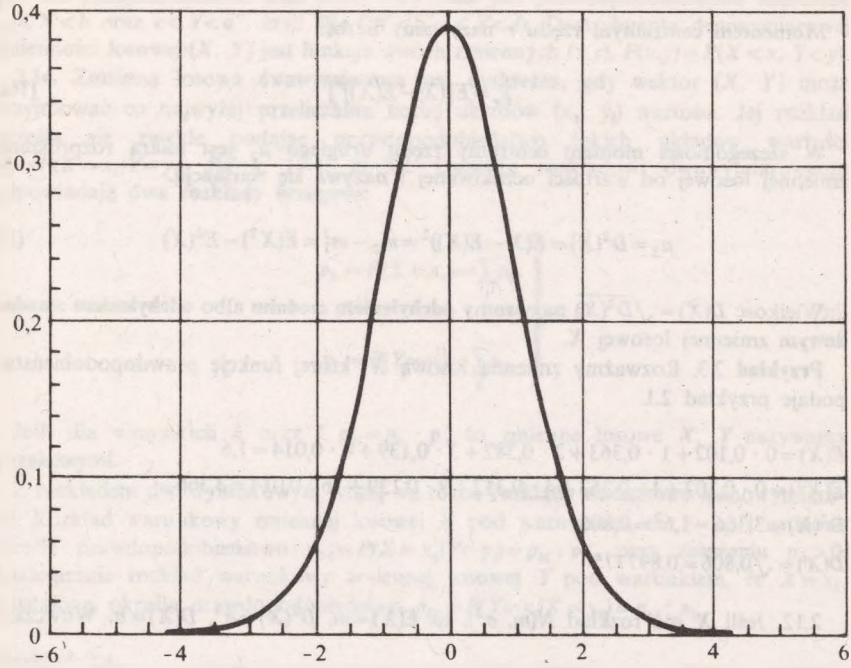
2.9. Zmienna losowa ciągła t ma **rozkład *t-Studenta* o r stopniach swobody** ($r>0$). Jeśli funkcja gęstości jest określona wzorem:

$$s_r(t)=\frac{2}{\sqrt{r}\,B\left(\frac{1}{2},\frac{r}{2}\right)\cdot\left(1+\frac{t^2}{r}\right)^{\frac{(r+1)}{2}}} \quad \text{dla } t \text{ rzeczywistego} \tag{15}$$

przy czym $B\left(\frac{1}{2},\frac{r}{2}\right)=\int_0^1 x^{-\frac{1}{2}}(1-x)^{\frac{r}{2}-1}dx$

Gęstość $s_r(t)$ jest symetryczna względem punktu $t=0$ (wykres 2). Jej wykres jest podobny do wykresu gęstości standardowego rozkładu normalnego, który jest rozkładem granicznym rozkładu *t-Studenta*. Tablice statystyczne podają wartości dystrybuanty dla $r\leqslant 120$. Przy $r>120$ dystrybuanta rozkładu *t-Studenta* nie różni się praktycznie od dystrybuanty rozkładu $N(0,1)$.

Wykres 2. FUNKCJA GĘSTOŚCI ROZKŁADU *t* — *STUDENTA* ($r=12$)



Uwaga 2.2. W praktyce spotykamy się ze zmiennymi losowymi o rozkładzie normalnym, w przypadku gdy na wartość zmiennej X ma wpływ duża ilość niezależnie działających czynników, z których każdy ma znikomy efekt i te efekty sumują się. W metodzie reprezentacyjnej znajduje zastosowanie również rozkład *t-Studenta*.

2.10. W zastosowaniach metody reprezentacyjnej często nie możemy podać rozkładu prawdopodobieństwa zmiennej losowej. Wówczas przybliżony opis rozkładu dają pewne wielkości, charakteryzujące rozkład, jak momenty rozkładu oraz ich funkcje. Zdefiniujemy je niżej.

2.11. **Wartością oczekiwaną** zmiennej losowej X nazywamy liczbę określoną, jako $E(X)=\sum_k x_k P(X=x_k)$ dla zmiennej losowej dyskretnej oraz jako $E(X)=\int_{-\infty}^{+\infty} xf(x)\,dx$ dla zmiennej losowej ciągłej.

Momentem zwykłym rzędu r zmiennej losowej X nazywamy w przypadku dyskretnym i ciągłym, odpowiednio liczbę $m_r=E(X^r)=\sum_k x_k^r P(X=x_k)$ bądź

$$m_r = E(X^r) = \int_{-\infty}^{+\infty} x^r f(x)\, dx \tag{16}$$

Momentem centralnym rzędu r nazywamy liczbę:

$$\mu_r = E[(X - E(X))^r] \tag{16a}$$

W szczególności moment centralny rzędu drugiego μ_2 jest miarą rozproszenia zmiennej losowej od wartości oczekiwanej i nazywa się **wariancją.**

$$\mu_2 = D^2(X) = E(X - E(X))^2 = m_2 - m_1^2 = E(X^2) - E^2(X) \tag{17}$$

Wielkość $D(X) = \sqrt{D^2(X)}$ nazywamy **odchyleniem średnim** albo **odchyleniem standardowym** zmiennej losowej X.

Przykład 2.3. Rozważmy zmienną losową X, której funkcję prawdopodobieństwa podaje przykład 2.1.

$E(X) = 0 \cdot 0{,}102 + 1 \cdot 0{,}363 + 2 \cdot 0{,}382 + 3 \cdot 0{,}139 + 4 \cdot 0{,}014 = 1{,}6$

$E(X^2) = 0 \cdot 0{,}102 + 1 \cdot 0{,}363 + 4 \cdot 0{,}382 + 9 \cdot 0{,}139 + 16 \cdot 0{,}014 = 3{,}366$

$D^2(X) = 3{,}366 - 1{,}6^2 = 0{,}806$

$D(X) = \sqrt{0{,}806} = 0{,}897775$

2.12. Jeśli X ma rozkład $N(m, \sigma^2)$, to $E(X) = m$, $D^2(X) = \sigma^2$, $D(X) = \sigma$. Wówczas:

$$P(m - k\sigma \leqslant X < m + k\sigma) = \begin{cases} 0{,}6827 & \text{dla } k = 1 \\ 0{,}9545 & \text{dla } k = 2 \\ 0{,}9973 & \text{dla } k = 3 \end{cases} \tag{18}$$

Im mniejszy jest parametr σ, tym węższy jest dla danego k ($k = 1, 2, 3$) przedział $[m - k\sigma, m + k\sigma)$, czyli tym bardziej są skupione wartości X dookoła wartości oczekiwanej (średniej) m.

Uwaga 2.3. *Iloraz odchylenia standardowego zmiennej losowej przez jej wartość oczekiwaną nosi nazwę* ***współczynnika zmienności:***

$$V(X) = \frac{D(X)}{E(X)}, \text{ jeśli } E(X) \neq 0 \tag{19}$$

Współczynnik zmienności mierzy dyspersję rozkładu w jednostkach wartości oczekiwanej.

2.13. Omawiane dotąd pojęcie zmiennej losowej dotyczyło zmiennej jednowymiarowej. Pojęcie to rozszerza się na przypadek wielowymiarowy, kiedy wynik dos-

wiadczenia losowego nie jest jedną liczbą, lecz układem liczb, czyli **wektorem.** W szczególności rozkład **dwuwymiarowej zmienności losowej** (X, Y) zostaje określony, gdy dla każdych czterech liczb $a<b$, $c<d$ dane jest prawdopodobieństwo zdarzenia „$a \leqslant X<b$ oraz $c \leqslant Y<d$", czyli $P(a \leqslant X<b,\ c \leqslant Y<d)$. Dystrybuanta dwuwymiarowej zmienności losowej (X, Y) jest funkcja dwóch zmiennych (x,y), $F(x,y)=P(X<x,\ Y<y)$.

2.14. Zmienna losowa **dwuwymiarowa** jest **dyskretna,** gdy wektor (X, Y) może przyjmować co najwyżej przeliczalną liczbę układów (x_k, y_l) wartości. Jej rozkład określa się zwykle podając prawdopodobieństwo takich układów wartości $p_{kl}=P(X=x_k|Y=y_l)$ dla $k,\ l=1,\ 2,\ 3,...$ Danemu rozkładowi dwuwymiarowemu odpowiadają dwa **rozkłady brzegowe:**

$$\left.\begin{aligned} p_{k\cdot} &= P(X=x_k)=\sum_l p_{kl} \\ p_{\cdot l} &= P(Y=y_l)=\sum_k p_{kl} \end{aligned}\right\} \tag{20}$$

Jeśli dla wszystkich k oraz l $p_{kl}=p_{k\cdot}\cdot p_{\cdot l}$ to zmienne losowe X, Y nazywamy **niezależnymi.**

Z rozkładem dwuwymiarowym wiążą się różne **rozkłady warunkowe** jednowymiarowe. Rozkład warunkowy zmiennej losowej X pod warunkiem, że $Y=y_l$, l ustalone określa prawdopodobieństwo $p_{k|l}=P(X=x_k|Y=y_l)=p_{kl}:p_{\cdot l}$, przy założeniu $p_{\cdot l}>0$. Analogicznie rozkład warunkowy zmiennej losowej Y pod warunkiem, że $X=x_k$, k ustalone, określa prawdopodobieństwo $p_{l|k}=P(Y=y_l|X=x_k)=p_{kl}:p_{k\cdot}$

Przykład 2.4.

k	p_{kl} dla $l=$				$p_{k\cdot}$
	1	2	3	4	
1	0,036	0,069	0,048	0,046	0,199
2	0,015	0,086	0,129	0,018	0,248
3	0,009	0,014	0,046	0,062	0,131
4	0,044	0,086	0,075	0,053	0,258
5	0,028	0,058	0,049	0,029	0,164
$p_{\cdot l}$ . . .	0,132	0,313	0,347	0,208	1,000

Zmienna losowa (X, Y) przyjmuje 20 różnych układów wartości (x_k, y_l), przy czym $p_{23}=0{,}129$, $p_{34}=0{,}062$, $p_{1\cdot}=0{,}199$, $p_{\cdot 3}=0{,}347$.

Funkcja prawdopodobieństwa rozkładów brzegowych i dwóch rozkładów warunkowych $p_{k|3}$ oraz $p_{l|2}$ przedstawiają się następująco:

k	1	2	3	4	5
$P(X=x_k)$	0,199	0,248	0,131	0,258	0,164

l	1	2	3	4
$P(Y=y_l)$	0,132	0,313	0,347	0,208

k	1	2	3	4	5
$P(X=x_{k\|3})$	0,138	0,372	0,133	0,216	0,141

l	1	2	3	4
$P(Y=y_{l\|2})$	0,060	0,347	0,520	0,073

Zauważmy, że zmienne losowe X, Y nie są niezależne, gdyż na przykład $p_{2\cdot}=0{,}248$, $p_{\cdot 3}=0{,}347$, a więc:

$$p_{23}=0{,}129\neq 0{,}086=0{,}248\cdot 0{,}347.$$

2.15. Jeśli dystrybuanta $F(x, y)$ dwuwymiarowej zmiennej losowej (X, Y) jest ciągła (obie jednowymiarowe zmienne losowe są ciągłe) i daje się przedstawić jako $F(x, y)=\int_{-\infty}^{x}\int_{-\infty}^{y} f(u, v)\,du\,dv$, zmienną (X, Y) nazywamy **dwuwymiarową zmienną losową ciągłą**. Funkcję $f(x, y)$ nazywamy **gęstością** tej zmiennej. Podobnie, jak w przypadku dyskretnym określamy jednowymiarowe rozkłady brzegowe. **Gęstość brzegową** zmiennej losowej X określamy jako $f_1(x)=\int_{-\infty}^{+\infty} f(x, y)\,dy$, a zmiennej losowej Y, jako $f_2(y)=\int_{-\infty}^{+\infty} f(x, y)\,dx$. Zmienne losowe X, Y nazywamy niezależnymi, jeżeli $f(x, y)=f_1(x)\,f_2(y)$. Przy ustalonych x_0, y_0 dla rozkładów warunkowych dwuwymiarowej zmiennej losowej określamy gęstość warunkową zmiennej losowej X pod warunkiem, że $Y=y_0$ jako $f(x|y_0)=f(x, y_0):f_2(y_0)$ oraz analogicznie gęstość warunkową zmiennej losowej Y pod warunkiem $X=x_0$, jako $f(y|x_0)=f(x_0, y):f_1(x_0)$.

2.16. **Momentem** m_{rs} rzędu $r+s$ zmiennej losowej dwuwymiarowej (X, Y) nazywamy wartość oczekiwaną zmiennej losowej $X^r\cdot Y^s$, czyli $m_{rs}=E(X^r\cdot Y^s)$. Momentem centralnym rzędu $r+s$ tej zmiennej nazywamy liczbę:

$$\mu_{rs}=E\{[X-E(X)]^r[Y-E(Y)]^s\} \tag{21}$$

przy czym $E(X)$, $E(Y)$ są wartościami oczekiwanymi w rozkładach brzegowych ($r, s=0, 1, 2, \ldots$) oraz $r+s>0$.

Zauważmy, że $m_{10}=E(X)$, $m_{01}=E(Y)$, $\mu_{10}=\mu_{01}=0$. Szczególnie ważną rolę odgrywa moment centralny μ_{11} zwany **kowariancją** i oznaczony $Cov(X, Y)$ lub C_{xy}

$$Cov(X, Y)=E\{[X-E(X)][Y-E(Y)]\}=E(XY)-E(X)\,E(Y) \tag{22}$$

Jeśli $Cov(X, Y)=0$ mówimy, że zmienne losowe są **nieskorelowane.** Jeśli X, Y są zmiennymi losowymi niezależnymi, to $Cov(X, Y)=0$, ale z faktu, że $Cov(X, Y)=0$ **nie wynika, że zmienne losowe są niezależne.**

2.17. Miarą stopnia zależności zmiennych losowych X i Y jest **współczynnik korelacji**

$$\rho=\rho(X, Y)=Cov(X, Y):\sqrt{D^2(X)\,D^2(Y)} \tag{23}$$

Wykazuje się, że $-1\leqslant\rho\leqslant 1$; $|\rho|=1$ wtedy i tylko wtedy, gdy między zmiennymi losowymi istnieje zależność liniowa $Y=aX+b$, przy czym gdy $a>0$ to $\rho=1$, a jeśli $a<0$, to $\rho=-1$. Jeżeli X, Y są niezależne bądź tylko nieskorelowane, to $\rho=0$.

Przykład 2.5. Przypuśćmy, że X przyjmuje następujące wartości: $x_1=0$, $x_2=0{,}5$, $x_3=1$, $x_4=1{,}5$, $x_5=2$ oraz że Y przyjmuje wartości: $y_1=-1$, $y_2=0$, $y_3=1$, $y_4=2$. Funkcja prawdopodobieństwa zmiennej losowej (X, Y) została podana w przykładzie 2.4.

$E(X)=m_{10}=0{,}199\cdot 0+0{,}248\cdot 0{,}5+0{,}131\cdot 1+0{,}258\cdot 1{,}5+0{,}164\cdot 2=0{,}97$
$E(Y)=m_{01}=0{,}132\cdot(-1)+0{,}313\cdot 0+0{,}347\cdot 1+0{,}208\cdot 2=0{,}631$
$E(XY)=-0{,}5\cdot 0{,}015=0{,}129\cdot 0{,}5+\ \ldots\ +0{,}029\cdot 4=0{,}5995$
$Cov(X, Y)=0{,}5995-0{,}97\cdot 0{,}631=-0{,}01257$
$E(X^2)=m_{20}=0{,}25\cdot 0{,}248+1\cdot 0{,}131+2{,}25\cdot 0{,}258+4\cdot 0{,}164=1{,}4295$
$D^2(X)=E(X^2)-E^2(X)=1{,}4295-0{,}97^2=0{,}4886$
$E(Y^2)=m_{02}=(-1)^2\cdot 0{,}132+1^2\cdot 0{,}347+2^2\cdot 0{,}208=1{,}311$
$D^2(Y)=1{,}311-0{,}631^2=0{,}9128$
$\rho(X, Y)=-0{,}01257:\sqrt{0{,}4886\cdot 0{,}9128}=-0{,}019$

2.18. Mając dwuwymiarowy rozkład zmiennej losowej (X, Y) możemy obliczać rozkłady funkcji zmiennych losowych, np. sumy $(X+Y)$. Wartość oczekiwana sumy (a, b — stałe):

$$E(aX+bY)=aE(X)+bE(Y) \tag{24}$$

Jeśli zmienne losowe X, Y są **niezależne,** to

$$E(XY)=E(X)\,E(Y) \tag{25}$$

W przypadku wariancji otrzymamy:

$$\left.\begin{aligned} &D^2(a+X)=D^2(X)\\ &D^2(aX)=a^2D^2(X)\\ &D^2(X+Y)=D^2(X)+D^2(Y)+2\,Cov(X, Y)\end{aligned}\right\} \tag{26}$$

Jeśli zmienne losowe X, Y są niezależne bądź nieskorelowane, to $D^2(X+Y)=D^2(X)+D^2(Y)$. Powyższe wzory można uogólnić na większą liczbę zmiennych losowych jednowymiarowych.

2.19. W praktyce istnieje czasem potrzeba wyznaczenia w płaszczyźnie (x, y) prostej, od której średnie odchylenie kwadratowe zmiennej losowej Y jest najmniejsze. Okazuje się, że taką własność ma prosta $y=\alpha x+\beta$, w której:

$$\alpha=\rho\frac{D(Y)}{D(X)}=\frac{Cov(X,Y)}{D^2(X)},\quad \beta=E(Y)-\alpha E(X)$$

Współczynnik α nosi nazwę **współczynnika regresji** zmiennej losowej Y względem X. Prosta nazywa się **prostą regresji drugiego rodzaju.**

2.20. **Zadania**

1) W urnie znajduje się 5 kul białych, 3 kule czarne oraz 2 kule niebieskie. Wybieramy losowo kulę i notujemy jej kolor. Określić zbiór zdarzeń elementarnych. Obliczyć prawdopodobieństwo wylosowania kuli białej.
2) Z powyższej urny wybieramy losowo 3 kule. Obliczyć prawdopodobieństwo zdarzenia A „wszystkie kule są tego samego koloru".
3) Doświadczenie jak wyżej. Określamy zmienną losową X jako liczbę wybranych kul koloru czarnego. Określić funkcję prawdopodobieństwa, dystrybuantę, wartość oczekiwaną oraz wariancję tej zmiennej losowej.
4) Zmienna losowa Y ma rozkład normalny $N(3; 0{,}25)$. Obliczyć prawdopodobieństwa: $P(2\leqslant Y<3)$, $P(2\leqslant Y<4)$, $P(Y>0)$.
5) Dwuwymiarowa zmienna losowa (X, Y) ma rozkład dyskretny o funkcji prawdopodobieństwa $p_{kl}=P(X=x_k, Y=y_l)$, jak niżej:

X	Y				Razem
	0	1	2	3	
0	0	0,2	0,15	0,15	0,5
1	0,125	0	0,038	0,038	0,201
2	0,088	0,035	0	0,026	0,149
3	0,089	0,035	0,026	0	0,15
Razem	0,302	0,070	0,214	0,214	1,000

Obliczyć współczynnik korelacji $\rho(X, Y)$.

2.21. **LITERATURA UZUPEŁNIAJĄCA:** J. Greń [2] str. 20—95, J. Jóźwiak, J. Podgórski [4] str. 11—100.

Wykład 3. Próbkowanie probabilistyczne, techniki losowania próby

3.1. Każdą skończoną N-elementową populację U możemy w pewien sposób uporządkować, nadając kolejne numery od 1 do N poszczególnym elementom populacji. Numer k w tym uporządkowaniu jest **identyfikatorem** odpowiedniego elementu populacji. Stąd populację możemy krótko przedstawić jako zbiór

$U=\{1, 2, 3, ..., N-1, N\}$ lub jeszcze krócej $U=\{k: 1 \leqslant k \leqslant N\}$. Jakikolwiek podzbiór populacji $s=\{k_1, k_2, ..., k_{n(s)}\}$, gdzie $1 \leqslant k_1 < k_2 < ... < k_{n(s)} \leqslant N$ nazywamy **próbą** o wielkości $n(s)$. Zbiór wszystkich prób oznaczymy dużą literą S, $S=\{s \in U\}$.

Przykład 3.1. Populacja U obejmuje $N=10$ elementów, np. pewnego rodzaju zakładów produkcyjnych w jakimś mieście. Zatem $U=\{k: 1 \leqslant k \leqslant 10\}$. Weźmy jako próbę $s=\{1, 4, 8\}$, czyli $k_1=1$, $k_2=4$, $k_3=8$; $n(s)=3$. Do próby s należą trzy zakłady o numerach: 1, 4, 8.

3.2. Weźmy pod uwagę zbiór S wszystkich możliwych prób. Na tym zbiorze określamy funkcję prawdopodobieństwa P, to znaczy dla każdej próby $s \in S$ określamy prawdopodobieństwo jej wyboru $p(s)$, $p(s) \geqslant 0$ tak, że $\sum_{s \in S} p(s)=1$. Rzecz jasna, próby, dla których $p(s)=0$ nigdy nie zostaną wybrane, natomiast jedna z prób, dla której $p(s)>0$ musi być wybrana. Taką funkcję prawdopodobieństwa nazywamy **wzorcem (planem) próbkowania losowego (probabilistycznego).**

Przykład 3.2. Populacja, jak w przykładzie 3.1. Wszystkich prób trójelementowych $(n/s/=3)$ jest $\binom{10}{3}=120$.
Niech:

$$s_1=\{1, 4, 7\};\ s_2=\{2, 5, 8\};\ s_3=\{3, 6, 9\};\ s_4=\{4, 7, 10\};\ s_5=\{1, 5, 8\};\ s_6=\{2, 6, 9\};$$
$$s_7=\{3, 7, 10\};\ s_8=\{1, 4, 8\};\ s_9=\{2, 5, 9\};\ s_{10}=\{3, 6, 10\};\ s_{11}, ..., s_{120}.$$

Określamy wzorzec próbkowania losowego następująco: $p(s_i)=0{,}1$ dla $i=1, 2, 3, ..., 10$ oraz $p(s_i)=0$ dla pozostałych prób.

3.3. Próby pobierane, zgodnie ze wzorcem próbkowania losowego, nazywamy **próbami losowymi.** Mechanizm losowy, realizujący określony wzorzec próbkowania losowego, nazywamy **schematem losowania próby** lub krótko, **schematem losowania.**

3.4. Badaniem metodą reprezentacyjną lub inaczej, **badaniem reprezentacyjnym** nazywamy badanie statystyczne, którego podstawą są informacje uzyskane z próby losowej.

3.5. Można skonstruować bardzo dużo różnych schematów losowania, jednak w praktyce stosuje się jedynie niektóre z nich i te będziemy omawiać. Losowanie próby może być postępowaniem jednoetapowym bądź wieloetapowym. W każdym etapie wybiera się losowo odpowiednio określone **jednostki losowania** (JL).

3.6. Jednostką losowania może być element populacji, czyli **jednostka badania bądź zespół jednostek badania.** W pierwszym przypadku losowanie nazywamy **indywidualnym,** w drugim — **losowaniem zespołowym.**

Przykład 3.3. Badamy liczbę uczniów szkół podstawowych na terenie pewnego miasta. Jednostką badania jest więc uczeń. Jeżeli jednostką losowania będzie także uczeń, losowanie jest indywidualne. Jeżeli jednostką losowania jest szkoła, losowanie jest zespołowe. Wówczas do próby należą wszyscy uczniowie wylosowanych szkół.

3.7. Losowanie konkretnej próby z badanej populacji, zgodnie z ustalonym schematem losowania, prowadzi się na pewnych materiałach (dokumentach), stanowiących **operat losowania.** Aby taki operat skonstruować należy najpierw określić zakres JL. Zasada: **każda jednostka badania należy do jednej i tylko do jednej jednostki losowania.** Żadna jednostka badania nie może być opuszczona, czyli nie może należeć do żadnej JL. Równocześnie, nie należy obejmować zakresem JL

jednostek nie będących elementami badanej populacji, niesłusznie traktowanych jako jej elementy.

Często *JL* są pewne jednostki obszarowe (terytorialne), np. rejony bądź obwody spisowe. Takie jednostki powinny mieć dokładnie określone granice, aby rachmistrz zbierający dane w terenie nie miał wątpliwości, czy dana jednostka badania należy do jego *JL*; może to być źródłem błędu pokrycia. *JL* powinny mieć dokładne identyfikatory (adresy). Zespół *JL* należy uporządkować, przydzielając każdej *JL* (identyfikatorowi) liczbę naturalną, będącą kolejnym numerem jednostki losowania w ustalonym przyporządkowaniu. Jeśli populację podzielono na *M JL*, to numery *JL* będą tworzyły ciąg 1, 2, ..., *M*. Zatem operatem losowania jest **wykaz** identyfikatorów z przyporządkowanymi im numerami oraz dokumentami, określającymi granice poszczególnych *JL*.

W praktyce operatem losowania mogą być odpowiednio przygotowane kartoteki, mapy, szkice itp. W przypadku losowania indywidualnego *JL* są jednostki badania, czyli operat losowania będzie wykazem kolejno uporządkowanych (ponumerowanych) jednostek badania z ich cechami adresowymi, jako identyfikatorami.

W pewnych schematach losowania stosujemy **warstwowanie,** tzn. podział populacji na wyłączające się podzbiory zwane **warstwami.** Wówczas operat losowania musi być osobny dla każdej warstwy. Np. dla badania krajowego warstwami mogą być województwa. Dla każdego województwa, jako warstwy populacji generalnej, konieczny będzie do losowania próby oddzielny **operat losowania,** obejmujący *JL* z danego województwa.

3.8. Sporządzenie operatu losowania powoduje powstawanie kosztów. Koszt operatu losowania indywidualnego jest często znacznie większy niż w przypadku operatu losowania zespołowego. Koszt operatu losowego powoduje, że niektóre „wymyślne" schematy losowania w danym badaniu statystycznym nie mogą być realizowane.

3.9. Koszt operatu losowania sprawia, że w praktyce często stosujemy **wielostopniowe losowanie próby.** W takim przypadku najpierw konstruujemy zespołowe **jednostki losowania pierwszego** stopnia (*JLPS*), a następnie losujemy pewną liczbę tych jednostek do próby. Jednostki **wylosowane** w losowaniu pierwszego stopnia dzielimy na mniejsze **jednostki losowania drugiego stopnia** *(JLDS)*. Z tych jednostek losujemy próbę; jest to losowanie drugiego stopnia. Jeśli te jednostki stanowią **ostateczną próbę,** czyli wchodzące w ich skład jednostki badania tworzą próbę, to takie losowanie jest **losowaniem dwustopniowym.** Jeżeli *JLDS* są jednostkami zespołowymi, to **wylosowane** do próby *JLDS* możemy podzielić na mniejsze jednostki losowania trzeciego stopnia (*JLTS*) i z nich losować ostateczną próbę. Wówczas losowanie próby nazywamy **losowaniem trójstopniowym.** W praktyce rzadko stosuje się losowanie o liczbie stopni większej niż trzy. W przypadku losowania dwustopniowego musimy skonstruować operat losowania pierwszego stopnia, a następnie operat losowania drugiego stopnia, ale ograniczony do części populacji, to jest do próby wylosowanej na pierwszym stopniu losowania. Koszt sporządzenia tych obu operatów jest zwykle wielokrotnie mniejszy, niż gdybyśmy losowali próbę jednostopniowo, a *JL* byłyby *JLDS*. Oczywiście w przypadku losowania trójstopniowego musimy skonstruować trzy

operaty losowania: a) pierwszego stopnia, b) drugiego stopnia, ale ograniczony do próby pierwszego stopnia oraz c) trzeciego stopnia, ale ograniczony do próby drugiego stopnia.

Przykład 3.4. Badamy stan zdrowotny ludności w wieku produkcyjnym pewnego województwa losując próbę trójstopniowo. *JLPS* są obwody spisowe; do próby pierwszego stopnia losujemy 20% obwodów. W wylosowanych obwodach dokonujemy obchodu mieszkań, które stanowią *JLDS*. Do próby drugiego stopnia losujemy 20% mieszkań; operatem losowania drugiego stopnia jest wykaz mieszkań z adresami w wylosowanych uprzednio obwodach spisowych. W wylosowanych mieszkaniach dokonujemy spisu ludności w wieku produkcyjnym; wykaz tych osób (*JLTS*) wraz z danymi adresowymi stanowi operat losowania trzeciego stopnia. Losujemy z nich próbę 10%. Wylosowane osoby stanowią ostateczną próbę. Łatwo zauważyć, że próba ta obejmuje około 0,4% ludności w wieku produkcyjnym.

3.10. Losowanie każdego stopnia może być **ze zwracaniem** bądź **bez zwracania.** Losowanie jest ze zwracaniem, gdy wylosowana *JL* bierze dalej udział w losowaniu, czyli losujemy kolejne *JL* do próby stale z całej populacji jednostek losowania. Jeśli natomiast *JL* raz wylosowana do próby nie bierze udziału w dalszym losowaniu, to takie losowanie nazywamy losowaniem bez zwracania. Gdy losujemy ze zwracaniem, to ten sam element populacji może wchodzić do próby wielokrotnie. Natomiast, gdy losujemy bez zwracania, to próba obejmuje jedynie różne jednostki badania.

Uwaga 3.1. Pojęcie próby i wzorca próbkowania losowego zdefiniowane wyżej w pkt. 3.1 i 3.2 nie obejmuje prób, jakie możemy uzyskać w schemacie losowania ze zwracaniem. Powyższe definicje łatwo rozszerzyć operując pojęciem **ciągu** zamiast pojęciem **podzbioru.** Dodamy więc możliwość określenia próby, jako ciągu $n(s)$ elementowego, $\underline{s}=(k_1, k_2, ..., k_{n(s)})$ takiego, w którym, jeśli M jest liczbą wszystkich *JL*, to $1 \leqslant k_i \leqslant M$ dla $i=1, 2, ..., n(s)$. Tak określoną próbę przyjęto nazywać **uporządkowaną,** a zbiór tego rodzaju prób będziemy oznaczać $\underline{S}$. Przeciwnie, próbę jako podzbiór populacji nazywa się próbą **nieuporządkowaną** (pkt 3.1). Definicja wzorca próbkowania losowego z pkt. 3.2 wymaga, formalnie, niewielkiego uzupełnienia przez uwzględnienie dodatkowe, że w przypadku prób uporządkowanych $\underline{s} \in \underline{S}$, $p(\underline{s}) \geqslant 0$, przy czym $\sum_{\underline{s} \in \underline{S}} p(\underline{s}) = 1$.

Przykład 3.5. Populacja, jak w przykładzie 3.1 oraz $n(\underline{s})=3$, czyli bierzemy pod uwagę próby — ciągi trójelementowe. Takich prób uporządkowanych jest $10^3=1000$. Na przykład $\underline{s}_1=(1, 1, 1)$, $\underline{s}_2=(1, 1, 2)$ itd. W pierwszym przypadku, za każdym razem wylosowano do próby pierwszą jednostkę badania, w drugim — za pierwszym i drugim razem pierwszą jednostkę, natomiast za trzecim razem — drugą jednostkę badania. Wśród prób uporządkowanych mieści się również próba z przykładu 3.1, $s=\{1, 4, 8\}$; będą jej odpowiadać próby uporządkowane: (1, 4, 8), (1, 8, 4), (4, 1, 8), (4, 8, 1), (8, 1, 4), (8, 4, 1), których jest $6=3!$ Jednym ze wzorców próbkowania losowego jest wzorzec „$p(\underline{s})=0{,}001$ dla $\underline{s} \in \underline{S}$", w którym każda trójelementowa próba uporządkowana ma takie samo prawdopodobieństwo, że ją wylosujemy, jako próbę w badaniu reprezentacyjnym.

3.11. Wybór *JL* może się odbywać z **jednakowymi** bądź z **różnymi** prawdopodobieństwami.

3.12. **Losowanie** może być **nieograniczone** bądź **ograniczone.** Losowanie jest nieograniczone, jeśli próbę losuje się z całej populacji oraz wylosowanie jakiejś *JL* do próby nie ogranicza wylosowania jakiejś innej *JL*. Jeżeli któryś z tych warunków nie jest spełniony, to losowanie jest ograniczone. Losowanie **warstwowe** oraz losowanie **systematyczne** należą do przypadków losowania ograniczonego.

3.13. **Losowanie warstwowe,** jak podkreślono wyżej w pkt. 3.7, polega na tym, że całą populację dzielimy na L rozłącznych warstw. Każda jednostka badania oraz *JL* należy do jednej i tylko do jednej warstwy. Niech N_h oznacza liczbę *JL* w warstwie o numerze h, $h=1, 2, ..., L$. Dla zrealizowania losowania warstwowego ustalamy wielkość próby z każdej warstwy n_h, tak żeby łączna próba z populacji N-elementowej, $N=\sum_h N_h$ wyniosła ustaloną wielkość $n=\sum_h n_h$. Próbę z każdej warstwy losujemy niezależnie, a więc musimy korzystać z odrębnego operatu losowania dla warstwy.

3.14. **Losowanie systematyczne** przeprowadza się następująco:

a) jeśli próba ma liczyć około n *JL*, obliczamy interwał losowania k, jako iloraz $N:n$ zaokrąglony w górę do liczby całkowitej;

b) losujemy liczbę naturalną r z przedziału od 1 do k, $1 \leqslant r \leqslant k$;

c) do próby wchodzą *JL* o numerach $r, r+k, r+2k, ..., r+(n-1)k$ (1)

Może się zdarzyć, że ostatni numer ciągu (1) jest większy od N. Wówczas go pomijamy, czyli próba obejmie $(n-1)$ *JL*.

Przykład 3.6. Populacja obejmuje $N=1020$ *JL*. Zamierzamy losować próbę $n=16$ elementową. Wobec $N:n=1020:16\cong 64$, interwał losowania wynosi 64. Przypuśćmy, że wylosowaliśmy $r=20$. Do próby wejdą *JL* o numerach 20, 84, 148, 212, ..., 980 próba obejmuje $n=16$ *JL*. Jeśli natomiast wylosowano $r=61$, to do próby wejdą *JL* o numerach 61, 125, 189, ..., 893, 957, a próba obejmie $n=15$ *JL* (następny numer wynosi 1021).

3.15. **Losowanie próby** danego stopnia odbywa się za pomocą **komputera,** posiadającego generator liczb losowych bądź za pomocą **tablicy liczb losowych.** W pierwszym przypadku odpowiednie liczby podaje wydruk komputera. Liczby te są numerami *JL* wylosowanymi do próby. Jeśli jakiś numer powtarza się r-krotnie oraz losowanie jest ze zwracaniem, to odnośna *JL* wchodzi do próby również r-krotnie.

W tablicach statystycznych publikowane są liczby (cyfry) losowe. Np. w zbiorze Ryszarda Zielińskiego Tablice Statystyczne (PWN, Warszawa 1972), na str. 315—324 podano tzw. „cyfry losowe" (Tabl. 62) w postaci pięciowierszowych bloków liczb pięciocyfrowych. Na każdej stronie bloki są ustawione w 10 kolumn oraz 10 bloków pięciowierszowych, czyli strona obejmuje 500 liczb pięciocyfrowych. W niniejszym konspekcie w tabl. 3.1 podano liczby figurujące na str. 316 powyższej publikacji.

3.16. Tablicę liczb losowych czytamy poczynając od dowolnego miejsca, liczby odczytujemy wierszami od lewej ku prawej, zgodnie z zasadą czytania polskiego tekstu. Jeśli potrzeba, można czytać inaczej, byleby stosować tę zasadę przez cały czas losowania danej próby, np. liczbę 77513 czytać jako 31577. Jeśli liczba *JL* w populacji (warstwie) jest liczbą trzycyfrową, można czytać trójki cyfr bądź opuszczać pierwsze dwie cyfry.

Omówione w tym punkcie i poprzednim losowanie jest losowaniem z jednakowymi prawdopodobieństwami wyboru ze zwracaniem. Jeśli chcemy losować próbę bez zwracania, to powtarzające się liczby skreślamy, pozostawiając pierwszy raz odczytaną liczbę. Liczby większe niż liczba JL w populacji pomijamy.

Przykład 3.7. Z populacji obejmującej 8650 JL wylosować próbę złożoną z 43 JL. Ponieważ $N=8650$ jest liczbą czterocyfrową odczytujemy liczby, pomijając pierwszą cyfrę. Rozpoczniemy czytanie od liczby figurującej w 16 wierszu oraz piątym bloku. Próbę obejmują JL o numerach:

1926, 4721, 8303, 3174, 3972, 5274, 6893, 1303, 2970, 4137, 3515, 0400, 1148, 3643, 4133, 4035, 2166, 3852, 0091, 1222, 0561, 2327, 8423, 6732, 6234, 7395, 6131, 0123, 1622, 5496, 7560, 1604, 5138, 6806, 7648, 5261, 4313, 5861, 5875, 1069, 5644, 7277, 8001.

Należy w operacie losowania odszukać jednostki losowania o powyższych numerach. Stanowią one konkretną próbę.

Przykład 3.8. Z populacji obejmującej $N=98$ JL wylosować próbę złożoną z $n=15$ JL. Ponieważ N jest liczbą dwucyfrową będziemy odczytywać pary cyfr poczynając od 21 wiersza, 2 bloku. Próbę obejmują JL o numerach: 02, 17, 68, 17, 19, 11, 71, 17, 16, 02, 92, 93, 77, 42, 19.

Próbę obejmują więc JL o numerach:

2 (2 razy), 17 (3 razy), 68, 19 (2 razy), 11, 71, 16, 92, 93, 77, 42. Jak widzimy, trzy jednostki losowania zostały wylosowane dwukrotnie albo trzykrotnie. Jeśli chcielibyśmy losować bez zwracania należy jeszcze dolosować 4 nowe JL. Czytając dalej tablicę mamy numery: 64, 04, 96, 55. Wówczas próbę stanowią JL o numerach: 2, 17, 68, 19, 11, 71, 16, 92, 93, 77, 42, 64, 4, 96, 55.

3.17. Powyższe przykłady dotyczyły losowania z jednakowymi prawdopodobieństwami wyboru. Losowanie ze zwracaniem z różnymi prawdopodobieństwami wyboru przebiega następująco:

Niech Π_k oznacza prawdopodobieństwo wylosowania za jednym razem k-tej $JL(1 \leqslant k \leqslant N)$, przy czym $\sum_{k=1}^{N} \Pi_k = 1$. Mnożymy prawdopodobieństwa Π_k przez jakąś „okrągłą" liczbę, np. $M=10000$. Zaokrąglamy liczby $M\Pi_k$ do liczb całkowitych tak, żeby suma zaokrąglonych $M\Pi_k$ dała liczbę M (w naszym przypadku 10 tysięcy). $M\Pi_k$ oznacza liczbę kolejnych numerów przyporządkowanych w czasie losowania danej (k-tej) JL. Numery przyporządkowujemy kolejno poszczególnym jednostkom losowania ze zbioru 1, 2, ..., M. Losowanie próby odbywa się analogicznie do losowania z jednakowymi prawdopodobieństwami ze zwracaniem.

Przykład 3.9. Z populacji obejmującej $N=30$ JL należy wylosować próbę $n=6$ JL z prawdopodobieństwami wyboru podanymi w tabl. 3.2. W rubr. (3) pokazano przyporządkowane numery. Ponieważ obecnie losujemy spośród $M=10000$ numerów sześć numerów, więc odczytujemy liczby czterocyfrowe, wówczas liczba czterocyfrowa 0000 oznacza M. Rozpoczynamy czytanie tablicy liczb losowych od wiersza 4, bloku 3. Pomijamy w czytaniu pierwszą cyfrę. Próbę wyznaczają: 0971, 7749, 0429, 2272, 5375, 5871.

Do próby wylosowane zostały JL o numerach: 6, 21, 4, 8, 14, 15.

Przykład 3.10. Z populacji obejmującej $N=1256$ JL należy wylosować próbę $n=157$ JL stosując losowanie systematyczne z jednakowymi prawdopodobieństwami wyboru. Obliczamy interwał losowania $k=1256:157=8$. Z tablicy liczb losowych odczytujemy cyfrę. Uzyskane miejsce w tablicy znajduje się w 34 wierszu i 9 bloku. Jest tam liczba 96299. Pierwsza cyfra $9>8$, więc ją pomijamy. Następna cyfra wynosi $6\leqslant 8$, czyli $r=6$. Do próby wchodzą JL o numerach: 6, 14, 22, 30, 38, ..., 1254.

Przykład 3.11. Z populacji $N=30$ JL, jak w przykładzie 3.9 należy wylosować próbę $n=6$ JL z prawdopodobieństwami wyboru podanymi w tabl. 3.1 **systematycznie.** W tym przypadku interwał losowania $k=10000:6=1666{,}6...$, $k=1667$. Odczytujemy liczbę czterocyfrową (pomijamy pierwszą cyfrę) w wierszu 34 w bloku 8. Wynosi ona $5676>1667$, następne liczby: 6299, $0836\leqslant 1667$, $r=836$, czyli do próby wejdą JL, którym przyporządkowano numery: 836, 2503, 4170, 5837, 7504, 9171, a więc wylosowano JL o numerach (identyfikatorach): 6, 8, 12, 15, 20, 27.

Uwaga 3.2. Jeśli prawdopodobieństwa wyboru $\Pi_k<\frac{1}{n}$ dla każdego $k=1, 2, ..., N$ (n — liczebność próby), to losowanie systematycznie dostarczy zawsze próby bez powtórzeń jakiejś JL.

3.18. **Zadania**

1) Z populacji obejmującej $N=856$ jednostek wylosować próbę $n=20$ elementową, stosując: a) losowanie proste ze zwracaniem, b) losowanie proste bez zwracania, c) losowanie systematyczne.

2) Populacja $N=50$ rodzin ma następujący rozkład wielkości mierzonej liczbą osób w rodzinie:

Liczba osób	1	2	3	4	5	6	7
Liczba rodzin	5	8	13	11	5	5	3

Wylosować z tej populacji próbę $n=10$ rodzin stosując: a) losowanie ze zwracaniem z prawdopodobieństwami proporcjonalnymi do liczby osób w rodzinie, b) losowanie systematyczne z takimi samymi prawdopodobieństwami wyboru jak wyżej.

Wskazówka. Uporządkować rodziny w ciąg niemalejący według liczby osób w rodzinie.

3.19. **LITERATURA UZUPEŁNIAJĄCA:** R. Zasępa [18] str. 25—43, [17] str. 48—66, 71—74.

TABL. 3.1. WYCINEK Z TABLICY LICZB LOSOWYCH R. ZIELIŃSKIEGO

09188	20097	32825	39527	04220	86304	83389	87374	64278	58044
90045	85497	51981	50654	94938	81997	91870	76150	68476	64659
73189	50207	47677	26269	62290	64464	27124	67018	41361	82760
75768	76490	20971	87749	90429	12272	95375	05871	93823	43178
54016	44056	66281	31003	00682	27398	20714	53295	07706	17813
08358	69910	78542	42785	13661	58873	04618	97553	31223	08420
28306	03264	81333	10591	40510	07893	32604	60475	94119	01840
53840	86233	81594	13628	51215	90290	28466	68795	77762	20791
91757	53741	61613	62269	50263	90212	55781	76514	83483	47055
89415	92694	00397	58391	12607	17646	48949	72306	94541	37408
77513	03820	86864	29901	68414	82774	51908	13980	72893	55507
19502	37174	69979	20288	55210	29773	74287	75251	65344	67415
21818	59313	93278	81757	05686	73156	07082	85046	31853	38452
51474	66499	68107	23621	94049	91345	42836	09191	08007	45449
99559	68331	62535	24170	69777	12830	74819	78142	43860	72834
33713	48007	93584	72869	51926	64721	58303	29822	93174	93972
85274	86893	11303	22970	28834	34137	73515	90400	71148	43643
84133	89640	44035	52166	73852	70091	61222	60561	62327	18423
56732	16234	17395	96131	10123	91622	85496	57560	81604	18880
65138	56806	87648	85261	34313	65861	45875	21069	85644	47277
38001	02176	81719	11711	71602	92937	74219	64049	65584	49698
37402	96397	01304	77586	56271	10086	47324	62605	40030	37438
97125	40348	87083	31417	21815	39250	75237	62047	15501	29578
21826	41134	47143	34072	64638	85902	49139	06441	03856	54552
73135	42742	95719	09035	85794	74296	08789	88156	64691	19202
07638	77929	03061	18072	96207	44156	23821	99538	04713	66994
60528	83441	07954	19814	59175	20695	05533	52139	61212	04455
83596	35655	06958	92983	05128	09719	77433	53783	92301	50498
10850	62746	99599	10507	13499	06319	53075	71839	06410	19362
39820	98952	43622	63147	64421	80814	43800	09351	31024	73167
59580	06478	75569	78800	88835	54486	23768	06156	04111	08408
35508	07341	23793	48763	90822	97022	17719	04207	95954	49953
30692	70668	94688	16127	56196	80091	82067	63400	05462	69200
65443	95659	18288	27437	49632	24041	08337	65676	96299	90836
27267	50264	13192	72294	07477	44606	17985	48911	97341	30358
91307	06991	19072	24210	36699	53728	28825	35793	28976	66252
68434	94688	84473	13622	62126	98408	12843	82590	09815	93146
48908	15877	54745	24591	35700	04754	83824	52692	54130	55160
06913	45197	42672	78601	11883	09528	63011	98901	14974	40344
10455	16019	14210	33712	91342	37821	88325	80851	43667	70883
12883	97343	65027	61184	04285	01392	17974	15077	90712	26769
21778	30976	38807	36961	31649	42096	63281	02023	08816	47449
19523	59515	65122	59659	86283	68258	69572	13798	16435	91529
67245	52670	35583	16563	79246	86686	76463	34222	26655	90802
60584	47377	07500	37992	45134	26529	26760	83637	41326	44344
53853	41377	36066	94850	58838	73859	49364	73331	96240	43642
24637	38736	74384	89342	52623	07992	12369	18601	03742	83873
83080	12451	38992	22815	07759	51777	97377	27585	51972	37867
16444	24334	36151	99073	27493	70939	85130	32552	54846	54759
60790	18157	57178	65762	11161	78576	45819	52979	65130	04860

TABL. 3.2. DANE POMOCNICZE DO PRZYKŁADU 3.9

Numer jednostki losowania k	Prawdopodobieństwo π_k	Liczba przyporządkowanych numerów $10000\ \pi_k$	Numery przyporządkowane	Numery wylosowane do próby
	(1)	(2)	(3)	(4)
1	0,0201	201	1— 201	
2	0,0102	102	202— 303	
3	0,0115	115	304— 418	
4	0,0150	150	419— 568	429
5	0,0240	240	569— 808	
6	0,0500	500	809— 1308	971
7	0,0760	760	1309— 2068	
8	0,0800	800	2069— 2868	2272
9	0,0600	600	2869— 3468	
10	0,0200	200	3469— 3668	
11	0,0430	430	3669— 4098	
12	0,0125	125	4099— 4223	
13	0,0540	540	4224— 4763	
14	0,0880	880	4764— 5643	5375
15	0,0630	630	5644— 6273	5871
16	0,0510	510	6274— 6783	
17	0,0280	280	6784— 7063	
18	0,0175	175	7064— 7238	
19	0,0195	195	7239— 7433	
20	0,0250	250	7434— 7683	
21	0,0130	130	7684— 7813	7749
22	0,0210	210	7814— 8023	
23	0,0205	205	8024— 8228	
24	0,0303	303	8229— 8531	
25	0,0132	132	8532— 8663	
26	0,0365	365	8664— 9028	
27	0,0212	212	9029— 9240	
28	0,0151	151	9241— 9391	
29	0,0202	202	9392— 9593	
30	0,0407	407	9594—10000	
Razem	1,0000	10000	×	×

Wykład 4. Problemy estymacji i precyzja oszacowań w badaniu reprezentacyjnym

4.1. Schemat losowania odpowiada określonemu wzorcowi próbkowania, czyli dla każdej możliwej do wyboru próby z populacji generalnej możemy obliczyć prawdopodobieństwo jej wylosowania. Dla wylosowanej próby uzyskujemy informacje, które należy przetworzyć, w celu oszacowania pewnych parametrów badanej populacji. Problem, jaką regułę (reguły) zastosować do szacowania tych parametrów nosi nazwę **problemu estymacji.**

4.2. Dowolną funkcję wartości cech mierzalnych (bądź wyróżnionych wariantów cech niemierzalnych) X, Y, Z, ... nazywamy **statystyką:**

$$t=f(X_k,\ Y_k,\ Z_k,\ ...;\quad k\in s) \tag{1}$$

Randomizacja (ulosowienie) wyboru próby przez zastosowanie danego schematu losowania próby powoduje, że statystyka jest zmienną losową i posiada pewien rozkład. Porównując rozkłady różnych statystyk można wybrać „najlepszą" statystykę dla szacowania wartości danego parametru populacji generalnej. Statystykę t stosowaną w danym badaniu do szacowania wartości parametru T nazywamy **estymatorem parametru** T, a jej wartość z próby — oceną parametru T. Stąd, estymator parametru T jest **zmienną losową.**

Przykład 4.1. Populację stanowią mieszkania w pewnym budynku. Liczba mieszkań $N=30$. Rzeczywiste liczby mieszkańców Y_k dla $k=1, 2, ..., 30$ wynoszą:

8, 7, 6, 6, 5, 5, 4, 4, 5, 4,
4, 3, 3, 6, 4, 3, 4, 3, 2, 5,
4, 3, 2, 3, 2, 9, 8, 8, 6, 6.

Losujemy próbę $n=4$ mieszkań z powyższej populacji stosując losowanie proste bez zwracania (*lpbz*). Liczba różnych możliwych do wylosowania prób wynosi:

$$\binom{N}{n}=\binom{30}{4}=\frac{30\cdot 29\cdot 28\cdot 27}{1\cdot 2\cdot 3\cdot 4}=27405$$

Weźmiemy pod uwagę statystykę:

$$\bar{y}=\frac{1}{4}\,(Y_{k_1}+Y_{k_2}+Y_{k_3}+Y_{k_4}) \text{ dla } s=\{k_1,\ k_2,\ k_3,\ k_4\}$$

$$1\leqslant k_1<k_2<k_3<k_4\leqslant 30 \tag{2}$$

Rozkład statystyki $\bar{y}$ przedstawia się następująco:

$\bar{y}$	Liczba prób	Prawdopodobieństwo wylosowania próby dającej odpowiednie $\bar{y}$ wartości statystyki
2,25	6	0,000
2,50 lub 2,75	242	0,009
3,00 lub 3,25	1343	0,049
3,50 lub 3,75	3274	0,119
4,00 lub 4,25	5088	0,186
4,50 lub 4,75	5811	0,212
5,00 lub 5,25	5127	0,187
5,50 lub 5,75	3528	0,129
6,00 lub 6,25	1899	0,069
6,50 lub 6,75	795	0,029
7,00 lub 7,25	240	0,009
7,50 lub 7,75	48	0,002
8,00 lub 8,25	4	0,000
Razem	27405	1,000

Rozkład statystyki $\bar{y}$ jest rozkładem dyskretnym. Prawdopodobieństwo, że $3{,}50 \leqslant \bar{y} \leqslant 5{,}75$ wynosi 0,833. Mamy więc dużą szansę, że konkretnie wylosowana próba da wartość statystyki $\bar{y}$ w tych granicach. Jednak, jak widać z zestawienia, $\bar{y}$ przyjmuje wartości od 2,25 do 8,25.

Uwaga 4.1. W teorii metody reprezentacyjnej przyjmuje się często konwencję, że dużymi literami alfabetu oznaczamy wielkości dotyczące populacji generalnej, natomiast — małymi literami alfabetu wielkości dotyczące próby. Dlatego parametr populacji T jest oznaczony dużą literą, a statystyka t małą literą. Przedstawiając wzór estymatora t według (1) operujemy wartościami populacji w tej jej części, jaką uzyskujemy z próby.

Innym sposobem zapisu statystyki $\bar{y}$ będzie formuła (2a), uwzględniająca wyniki kolejnego losowania do próby elementów badanej populacji; wartość cechy dla tych elementów z próby oznaczamy małymi literami, jako y_1, y_2, y_3, y_4 ... y_n. Jeśli pierwszym wylosowanym do próby jest k_1-szy element populacji, to $y_1 = Y_{k_1}$, drugim — k_2-gi element populacji, to $y_2 = Y_{k_2}$ itd.
Wówczas statystykę t możemy przedstawić, jako:

$$t = h(x_1, x_2, \ldots, x_n; y_1, y_2, \ldots, y_n; z_1, z_2, \ldots, z_n; \ldots) \tag{2a}$$

Podana w przykładzie 4.1 statystyka $\bar{y}$ wyraża się wzorem:

$$\bar{y} = \frac{1}{4}(y_1 + y_2 + y_3 + y_4)$$

Zauważmy, że y_1, y_2, y_3, y_4 są zmiennymi losowymi, ale dla konkretnej próby są to pewne liczby. Jeśli np. z populacji określonej w przykładzie 4.1 wylosowaliśmy do próby kolejno mieszkania o numerach: 8, 23, 15, 4, to $y_1 = Y_8 = 4$, $y_2 = Y_{23} = 2$, $y_3 = Y_{15} = 4$, $y_4 = Y_4 = 6$. Losując próbę drugi raz przypuśćmy, że wylosowaliśmy kolejno mieszkania o numerach: 12, 1, 26, 29, wówczas $y_1 = Y_{12} = 3$, $y_2 = Y_1 = 8$, $y_3 = Y_{26} = 9$, $y_4 = Y_{29} = 6$. Zatem wartości $y_i (i = 1 \ldots 4)$ mogą być inne dla różnych prób:

Numer próby	y_1	y_2	y_3	y_4	$\bar{y}$
1	4	2	4	6	4
2	3	8	9	6	6,5

4.3. Rzeczywisty rozkład estymatora jest rozkładem dyskretnym, gdyż próba obejmuje skończoną liczbę elementów populacji generalnej, a liczba możliwych do wylosowania prób jest także skończona. Okazuje się, że w przypadkach dużych prób, z jakimi mamy do czynienia w praktyce, rozkład estymatora można dobrze **aproksymować** za pomocą **rozkładu normalnego** o tych samych momentach pierwszego i drugiego rzędu. Jest to szczególnie użyteczne, gdy znamy wartość dwóch parametrów (rzeczywistego) rozkładu estymatora: wartości oczekiwanej oraz wariancji bądź odchylenia standardowego.

4.4. Poszukując „dobrego" estymatora jakiegoś parametru T musimy mieć kryterium porównawcze, którym będziemy się kierować. Przyjmujemy, że z dwóch statystyk lepszym estymatorem będzie ta statystyka, krórej wartości bardziej skupiają się dookoła wartości T szacowanego parametru. Rozrzut statystyki t wobec szacowanego parametru T mierzymy wartością **średniego błędu kwadratowego MSE:**

$$MSE(t)=E(t-T)^2=\sum_i (t_i-T)^2\ P(t=t_i)=D^2(t)+B^2(t) \tag{3}$$

przy czym:

$$D^2(t)=\sum_i [t_i-E(t)]^2\ P(t=t_i) \tag{4}$$

jest **wariancją** statystyki t, natomiast:

$$B(t)=E(t)-T \tag{5}$$

jest **obciążeniem** statystyki t, jako estymatora parametru T, przy czym $E(t)=\sum_i t_i\ P(t=t_i)$.

Estymator t parametru T jest **estymatorem nieobciążonym** tego parametru, gdy $E(t)=T$. Jeśli $E(t)\neq T$, czyli $B(t)=E(t)-T\neq 0$, estymator nazywamy **obciążonym.**

Przykład 4.2. Statystyka t ma rozkład, który można aproksymować rozkładem normalnym $N(m, \sigma^2)$. Gdybyśmy chcieli zastosować ją do estymacji wartości parametru T — to obciążenie tej statystyki wyniesie $m-T$, a wariancja σ^2, czyli: $MSE(t)=\sigma^2+(m-T)^2$.

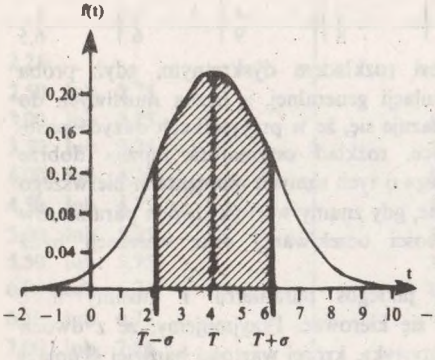
Niech $T=4$; $m=5{,}2$; $\sigma=1{,}8$; wówczas $MSE(t)=1{,}8^2+1{,}2^2=4{,}68$. Zgodnie z oznaczeniami i wzorami omówionymi w 2.8 otrzymamy:

$$P(T-\sigma\leqslant t<T+\sigma)=F(T+\sigma)-F(T-\sigma)=\Phi\left(\frac{T+\sigma-m}{\sigma}\right)-F\left(\frac{T-\sigma-m}{\sigma}\right)=$$

$$=\Phi(0{,}33)-\Phi(-1{,}67)=\Phi(0{,}33)-[1-\Phi(1{,}67)]=0{,}629300-1+0{,}952540=0{,}58184.$$

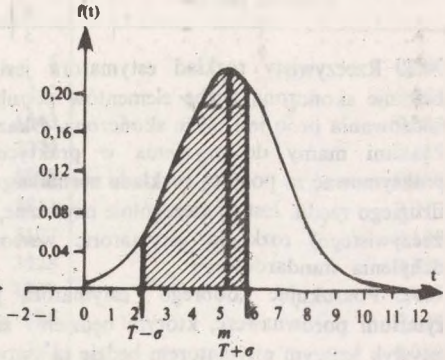
Gdyby $m=T=4$, czyli estymator byłby nieobciążony, wówczas $P(T-\sigma\leqslant t<T+\sigma)=$ $=0{,}6827$, a więc znacznie większe prawdopodobieństwo, że ocena będzie w granicach $(T-\sigma, T+\sigma)$.

Wykres 4.1 FUNKCJA GĘSTOŚCI NIEOBCIĄŻONEGO ESTYMATORA N ($m=T=4$; $\sigma=1{,}8$)



Powierzchnia zakreślonego pola pod krzywą jest równa $P(T-\sigma \leqslant t < T+\sigma)$

Wykres 4.2 FUNKCJA GĘSTOŚCI ESTYMATORA OBCIĄŻONEGO N ($m=5{,}2$; $\sigma=1{,}8$) PRZY $T=4 \neq m$



Powierzchnia zakreślonego pola pod krzywą jest równa $P(T-\sigma \leqslant t < T+\sigma)$

4.5. Porównując średnie błędy kwadratowe dwóch statystyk t_1 i t_2, jako ewentualnych estymatorów parametru t, przyjmujemy jako lepszą do estymacji tę statystykę, dla której średni błąd kwadratowy jest mniejszy bądź równy, ściślej, jeśli $MSE(t_1) \leqslant MSE(t_2)$ oraz dla pewnych rozkładów cechy $MSE(t_1) < MSE(t_2)$, to lepszym estymatorem jest statystyka t_1 niż statystyka t_2.

4.6. W przypadku nieobciążonego estymatora t parametru T, średni błąd kwadratowy $MSE(t)$ redukuje się do wariancji $D^2(t)$, czyli:

$$MSE(t) = D^2(t), \text{ gdy } E(t) = T \tag{6}$$

Przykład 4.3. Statystyka $\bar{y}$ z przykładu 4.1 jest nieobciążonym estymatorem średniej $\bar{Y}$ liczby osób w mieszkaniu, która wynosi $\bar{Y} = 4{,}733$. Korzystając z przedstawionej na str. 2 funkcji prawdopodobieństwa dla rozkładu statystyki $\bar{y}$ można wykazać, że $E(\bar{y}) = 4{,}733 = \bar{Y}$. Wariancja tej statystyki wynosi $D^2(\bar{y}) = 0{,}8064$.

4.7. Estymator obciążony szacuje przeciętnie częściej z błędami dodatnimi niż ujemnymi, gdy $B(t) = E(t) - T > 0$; przeciwnie, gdy $B(t) < 0$. Dawniej preferowano estymatory nieobciążone, obecnie ta tendencja zanika.

4.8. Inną ważną własnością jest tzw. **zgodność estymatora** parametru T. W statystyce matematycznej estymator t_n parametru T, z próby n-elementowej nazywamy **estymatorem zgodnym parametru** T, jeżeli dla dowolnie małej liczby dodatniej ε zachodzi związek:

$$\lim_{n \to \infty} P(|t_n - T| \leqslant \varepsilon) = 1 \tag{7}$$

Z definicji wynika, że próby mogą być dowolnie duże, podczas gdy w badaniu reprezentacyjnym zawsze $n < N$; N — liczebność badanej populacji. Dlatego w teorii

metody reprezentacyjnej, dla umożliwienia definicji (7), przeprowadzamy proces myślowy rozumiejąc pojęcie wzrastania n do nieskończoności ($n \to \infty$), jak następuje:

a) wyobrażamy sobie, że kiedy wzrasta nieograniczenie liczebność n próby, wzrasta także liczebność N populacji tak, że dla ustalonego ułamka r, $0 < r < 1$, $n : N < r$;

b) wzrost liczebności populacji nie powoduje zmian w jej strukturze według badanych cech, a więc parametr T pozostaje stały.

Uwaga 4.2. Proces wzrastania populacji możemy rozumieć jako wielokrotne reprodukowanie elementów populacji.

4.9. Z (7) wynika, że jeśli estymator jest zgodny, to przy dostatecznie dużej próbie mamy bliskie jedności prawdopodobieństwa, że otrzymana ocena będzie nieznacznie różnić się od rzeczywistej wartości szacowanego parametru. Dlatego w badaniach reprezentacyjnych z reguły stosujemy estymatory zgodne.

4.10. **Twierdzenie:** Jeśli a) $\lim |E(t_n) - T| = 0$ oraz b) $\lim D^2(t_n) = 0$, to t_n jest estymatorem zgodnym parametru T. Interpretacja tego twierdzenia może być następująca: jeżeli obciążenie estymatora maleje ze wzrostem próby, a jego wariancja również maleje i jest coraz bliższa zeru, to ten estymator jest estymatorem zgodnym danego parametru.

4.11. **Twierdzenie dalsze:** jeżeli $t_n^{(1)}, t_n^{(2)}, \ldots, t_n^{(k)}$ są estymatorami zgodnymi parametrów $T_1, T_2, \ldots, T_k$ odpowiednio, to dowolna funkcja ciągła $f(t_n^{(1)}, t_n^{(2)}, \ldots, t_n^{(k)})$ jest estymatorem zgodnym parametru $T = f(T_1, T_2, \ldots, T_k)$.

4.12. W praktyce badań statystycznych stosujemy estymatory nieobciążone lub obciążone, zgodne. W tym drugim przypadku próba powinna być dostatecznie duża tak, że obciążenie estymatora można uznać za nieistotne. Inaczej mówiąc, możemy traktować estymator jak nieobciążony. Wobec tego, że estymatory przy dostatecznie dużych próbach mają rozkład bardzo zbliżony do rozkładu normalnego, rozkład estymatora będziemy znali określając jego wariancję. Do tego celu konieczne jest co najmniej przybliżona znajomość pewnych parametrów populacji generalnej. Niezmiernie ważne jest obliczanie tych parametrów w przypadku każdego badania statystycznego, aby móc te dane wykorzystać przy planowaniu badania reprezentacyjnego.

4.13. Szacując parametr populacji generalnej opieramy się na uzyskanej z próby wartości stosowanego estymatora. Oczywiście chcielibyśmy wiedzieć czy ocena bardzo różni się od wartości parametru. Niestety, na takie pytanie nie możemy odpowiedzieć podając wielkość błędu, chyba że równocześnie przeprowadziliśmy pełne badanie statystyczne i możemy porównać ocenę z badania reprezentacyjnego z wynikiem badania pełnego. Możemy natomiast określić w pewien sposób precyzję szacunku metodą reprezentacyjną. Za **miarę precyzji szacunku** przyjmuje się:

a) odchylenie standardowe $D(t)$ estymatora t, zwane wówczas **błędem standardowym** lub **błędem średnim** szacunku bądź równoważnie;

b) współczynnik zmienności $V(t) = D(t) : E(t)$, $E(t) > 0$ zwany **względnym błędem standardowym (średnim) szacunku.** Zwykle mnożymy go przez 100, wyrażając w procentach. Otrzymujemy wówczas **procentowy błąd standardowy (średni) szacunku.**

Uwaga 4.3. Teoria metody reprezentacyjnej abstrahuje od błędów nielosowych

zakładając, że uzyskana z badania informacja jest pełna i bezbłędna. Dlatego mierzymy w niej wpływ błędów losowych na wynik szacunku, czyli część faktycznego błędu szacunku i mówimy o **precyzji** szacunku w odróżnieniu od pojęcia **dokładności** szacunku, na którą składają się błędy nielosowe wraz z błędem losowym z racji ograniczenia danych do wylosowanej próby.

4.14. Co nam daje znajomość błędu standardowego, czy procentowego błędu standardowego szacunku? Ze znajomości rozkładu normalnego wynika następujące: Niech b% oznacza wielkość procentowego błędu standardowego szacunku. Wówczas otrzymamy:

a) około 38 szans na 100, że **faktyczny** błąd szacunku nie przekroczy 0,5 b%,
b) około 68 szans na 100, że ten błąd nie przekroczy b%,
c) około 87 szans na 100, że ten błąd nie przekroczy 1,5 b%,
d) około 954 szans na 1000, że ten błąd nie przekroczy 2 b%,
e) około 997 szans na 1000, że błąd procentowy szacunku nie przekroczy 3 b%.

Z powyższego wynika, że jest ponad dwa razy większa szansa, że błąd szacunku będzie mniejszy od błędu **standardowego**, niż że będzie mu równy albo od niego większy. Jeśli błąd szacunku może mieć poważne konsekwencje praktyczne możemy założyć, że faktyczny błąd nie jest większy od dwóch błędów standardowych, czyli że „maksymalnie" możliwy błąd mieści się w granicach dwóch błędów standardowych. Możliwość, że faktyczny błąd oceny jest większy od dwóch błędów standardowych jest bardzo mała, gdyż można ją przyrównać do wyciągnięcia na „chybił-trafił" z urny kuli koloru czerwonego, gdy w urnie znajduje się 954 kul białych, a tylko 46 kul czerwonych. W wyjątkowych przypadkach przyjmuje się 3 błędy standardowe za „maksymalnie" możliwą wielkość błędu faktycznego oceny z próby.

Przykład 4.4. Z każdej z trzech populacji, w których szacowany parametr T wyniósł: 40 tys., 5 tys., 600, odpowiednio procentowy błąd standardowy wyniósł: $b=5\%$, $b=10\%$ oraz $b=20\%$. Z każdej z tych populacji wylosowano niezależnie pięć prób. Otrzymano wyniki:

Nr próby	Ocena t	Faktyczny błąd szacunku		Ocena t	Faktyczny błąd szacunku		Ocena t	Faktyczny błąd szacunku	
		liczba	%		liczba	%		liczba	%
	$T=40$ tys.		$b=5\%$	$T=5$ tys.		$b=10\%$	$T=600$		$b=20\%$
1 . .	37448	−2552	−6,4	5207	207	4,1	861	261	43,5
2 . .	38442	−1558	−3,9	5182	182	3,6	487	−113	−18,8
3 . .	40990	990	2,5	5380	380	7,6	922	322	53,7
4 . .	38728	−1272	−3,2	4684	−316	−6,3	686	86	14,3
5 . .	41694	1694	4,2	4096	−904	−18,1	738	138	23,0

Analiza tych liczb prowadzi do ciekawych wniosków (zob. [20] str. 8—10).

4.15. Wydaje się właściwe przyjęcie następującej konwencji: a) przy $b \leqslant 7{,}5\%$ uznajemy szacunki metodą reprezentacyjną za wystarczająco precyzyjne, b) jeśli $7{,}5\% < b \leqslant 15\%$ precyzja szacunków jest mała (słaba), c) $b > 15\%$ ocenę traktujemy jedynie, jako określenie rzędu wielkości parametru T.

4.16. Obok operowania pojęciem błędu standardowego szacunku stosuje się również **przedziały ufności** dla szacowanych parametrów. Przedziałem ufności $[\underline{U}(t), \overline{U}(t)]$ parametru T nazywamy przedział losowy, którego dolna $\underline{U}(t)$ i górna $\overline{U}(t)$ granice ufności są funkcjami estymatora t takimi, że prawdopodobieństwo $P\{\underline{U}(t) < T < \overline{U}(t)\} = 1 - \alpha$ jest stałe; liczbę $(1-\alpha)$ nazywamy **współczynnikiem ufności**; najczęściej przyjmuje się $\alpha = 0{,}05$ albo $\alpha = 0{,}01$. Po wylosowaniu próby, jeśli $t = t_0$, przedział wynosi $[\underline{U}(t_0), \overline{U}(t_0)]$. Przyjmując $\alpha = 0{,}05$, czyli $1 - \alpha = 0{,}95$, ten przedział nazywamy 95% przedziałem ufności. Uzyskany przedział może zawierać szacowaną wartość T parametru, ale może też tej wartości nie obejmować. Szansa, że 95% przedział ufności $[\underline{U}(t)\overline{U}(t)]$, obejmie T jest mniej więcej taka, jak wylosowanie kuli białej z urny, w której znajduje się 95 kul białych oraz 5 kul czerwonych. Stosując przedziały 95% w różnych badaniach reprezentacyjnych możemy się spodziewać, że w około 95% badań uzyskane przedziały zawierają szacowane wartości parametrów. W przypadku przedziału, przy współczynniku ufności $\alpha = 0{,}01$, czyli przedziału 99% powyższe stwierdzenia pozostają słuszne, jeśli „95" zastąpimy przez „99", a „5" przez „1".

4.17. **Zadania:**

1) Badamy populację N = 50 gospodarstw rolnych:

Numer (k) gospodarstwa	X_k	Y_k	Numer (k) gospodarstwa	X_k	Y_k
1—3	0,8	0	31	4	3
4	0,8	1	32—34	6	1
5—7	1,5	0	35—37	6	2
8—12	1,5	1	38	6	3
13	1,5	2	39—40	8,5	1
14	2,5	0	41—43	8,5	2
15—18	2,5	1	44—45	8,5	3
19—20	2,5	2	46	8,5	5
21	4	0	47—48	12,5	2
22—26	4	1	49	12,5	3
27—30	4	2	50	12,5	4

X — powierzchnia gospodarstwa w ha, Y — liczba krów w gospodarstwie.

Wylosować (*lpbz*) 25 niezależnych prób $n = 10$ gospodarstw rolnych. Dla każdej z nich obliczyć wartości średnie $\bar{x}$, $\bar{y}$. Wiedząc, że $\overline{X} = 4{,}764$ oraz $\overline{Y} = 1{,}48$ określić procentowe błędy szacunków $\delta_1 = 100\left|\frac{\bar{x} - \overline{X}}{\overline{X}}\right|$, $\delta_2 = 100\left|\frac{\bar{y} - \overline{Y}}{\overline{Y}}\right|$.

2) Zakładając powyżej, że procentowe błędy standardowe $b_1 = 100\, V(\bar{x}) = 20{,}1\%$, $b_2 = 100\, V(\bar{y}) = 20{,}5\%$ obliczyć (niezależnie dla każdej z badanych cech) frakcje prób, dla których $\delta_1 \leqslant k b_1$ oraz $\delta_2 \leqslant k b_2$ dla $k = 1, 2, 3$. Porównać te frakcje, odpowiednio, z prawdopodobieństwami $P(\delta_i \leqslant k b_i)$ dla $i = 1, 2$, równymi: 0,6827 dla $k = 1$; 0,9545 dla $k = 2$ oraz 0,9973 dla $k = 3$.

3) W zad. 1) przyjęliśmy średnią z próby $\bar{x}$ ($\bar{y}$) za estymator średniej $\overline{X}$ ($\overline{Y}$).

Przypuśćmy, że zamiast średniej z próby przyjmujemy medianę z próby, jako estymator średniej $\bar{X}$ ($\bar{Y}$). Wykorzystując dane wylosowanych 25 prób zbadać, czy procentowe błędy szacunków, opartych o medianę z próby, są przeciętnie większe niż szacunki oparte o średnią z próby?

4.18. **LITERATURA UZUPEŁNIAJĄCA:** R. Zasępa [18] str. 43—54, [17] str. 23—38.

Wykład 5. Estymatory proste i ich własności w losowaniu prostym

5.1. Indywidualne losowanie nieograniczone, z jednakowymi prawdopodobieństwami wyboru, przyjęto nazywać **losowaniem prostym.** Polega ono na tym, że: a) jednostką losowania jest jednostka badania, czyli element badanej populacji; b) wyboru dokonujemy raz za razem; c) dla każdego elementu populacji prawdopodobieństwo jego wybrania do próby jest takie samo; d) losujemy próbę z całej populacji.

5.2. W dalszym ciągu przyjmujemy, że populacja obejmuje N elementów, uporządkowanych w ciąg $(1, 2, ..., N)$ oraz wyboru kolejnych elementów do próby dokonujemy n razy. Wyróżniamy dwa schematy losowania prostego, odpowiadające dwom różnym wzorcom próbkowania losowego:

a) **losowanie proste ze zwracaniem** (*lpzz*), w którym wybieramy kolejno, niezależnie, jednostki do próby z całej populacji (a więc ta sama jednostka badania może być wybrana wielokrotnie). Wówczas dowolny ciąg $(k_1, k_2, ..., k_n)$, w którym $1 \leqslant k_i \leqslant N$ dla $i = 1, 2, ..., n$ ma tę samą szansę, że będzie wylosowaną próbą:

$$P[s=(k_1, k_2, ..., k_n)] = 1 : N^n \tag{1}$$

b) **losowanie proste bez zwracania** (*lpbz*), w którym element populacji wybrany do próby nie bierze udziału w dalszym wyborze (jednostka badania może być wybrana do próby tylko jeden raz). Dla każdego podzbioru n-elementowego z populacji badanej prawdopodobieństwo jego wylosowania jest takie samo;

$$P(s=\{k_1, k_2, ..., k_n\}) = 1 : \binom{N}{n}, \quad 1 \leqslant k_1 < k_2 < ... < k_n \leqslant N \tag{2}$$

5.3. Analizując oba wzorce próbkowania losowego stwierdzamy, że wzorzec *lpzz* zawiera wszystkie próby wzorca *lpbz*, a ponadto próby, w których niektóre elementy zostały wybrane wielokrotnie. Te ostatnie próby są informacyjnie gorsze od pozostałych, gdyż — jak to zostało udowodnione — krotność występowania jakiegoś elementu w próbie nie przynosi żadnej dodatkowej informacji dla poprawy szacunków metodą reprezentacyjną. To wskazuje, że schemat losowania prostego bez zwracania

(*lpbz*) jest **efektywniejszy** od schematu losowania prostego ze zwracaniem (*lpzz*). Zostało to udowodnione także formalnie przez hinduskiego statystyka D. Basu. Dalej będziemy się w zasadzie ograniczać do *lpbz*.

5.4. W celu realizacji losowania prostego operat losowania powinien zawierać informacje adresowe dla każdej jednostki badanej populacji tak, że pozwalają one uporządkować w ciąg poszczególne jednostki, czyli przyporządkować każdej jednostce populacji określony numer od 1 do N. Po wylosowaniu numerów, stanowiących próbę, wyznaczamy z operatu losowania jednostki tworzące próbę.

Przykład 5.1. W gminie chcemy wylosować próbę n indywidualnych gospodarstw rolnych, stosując losowanie proste bez zwracania. Jeśli gmina posiada kompletny i aktualny wykaz gospodarstw, które kolejno zostały ponumerowane, to ten wykaz może stanowić operat losowania próby. Numery (pozycje) wykazu stanowią identyfikatory poszczególnych gospodarstw. Posługując się tablicą liczb losowych odczytujemy z niej liczby nie większe od N, pomijając liczby pojawiające się po raz drugi. Odczytane ostatecznie liczby wyznaczają próbę. Wchodzą do niej te gospodarstwa rolne, które w operacie losowania mają numery wybrane do próby.

Przykład 5.2. Zadanie, jak w przykładzie 5.1, jednak nie istnieje wykaz gospodarstw rolnych, natomiast obszar gminy został podzielony na M obwodów spisowych, noszących kolejne numery i w nich istnieją wykazy indywidualnych gospodarstw ponumerowanych w ramach obwodu spisowego. Wówczas, znając liczbę gospodarstw w obwodzie, możemy je uporządkować w całej gminie następująco:

a) sporządzamy wykaz obwodów spisowych (*o.s.*),

b) podajemy w nim dla każdego *o.s.* liczbę indywidualnych gospodarstw rolnych,

c) kumulujemy te liczby.

Tak sporządzony dokument stanowi operat losowania próby. Numer przyporządkowany danemu gospodarstwu otrzymujemy dodając skumulowaną (łączną) liczbę gospodarstw w poprzednich *o.s.* do numeru gospodarstwa w *o.s.*, do którego należy. Np. jeśli gospodarstwo znajduje się w trzecim *o.s.* i posiada w nim numer 15, a łączna liczba gospodarstw w pierwszym i drugim *o.s.* wynosi 120, to temu gospodarstwu przypisujemy numer (identyfikator) wynoszący $120+15=135$. Po wylosowaniu próby identyfikujemy numery wylosowane do próby z odpowiadającymi im gospodarstwami rolnymi.

Przykład 5.3. Na terenie znajdują się cztery *o.s.*, w których liczby indywidualnych gospodarstw rolnych wynoszą odpowiednio: 23, 18, 31, 15. Konstruujemy operat losowania:

Numer *o.s.*	Liczba indywidualnych gospodarstw rolnych		Numery gospodarstw, wchodzących do próby z danego *o.s.*
	w *o.s.*	skumulowana	
1	23	23	3, 13, 21
2	18	41	15, 17
3	31	72	3, 14, 25
4	15	87	1, 12

Otrzymujemy łącznie $N=87$ gospodarstw, losujemy (*lpbz*) próbę $n=10$ gospodarstw. Z tablicy liczb losowych odczytujemy kolejne dwójki cyfr, zaczynając od 18 wiersza, pierwszego bloku, pomijając liczby większe od 87 (wraz z 00) oraz dwucyfrowe liczby już odczytane. Próbę wyznaczają numery: 84, 13, 38, 40, 44, 03, 55, 21, 66, 73. Ustawiamy je w ciąg rosnący:

$$\underset{o.s.\ 1}{3, 13, 21}, \quad \underset{o.s.\ 2}{38, 40}, \quad \underset{o.s.\ 3}{44, 55, 66}, \quad \underset{o.s.\ 4}{73, 84}$$

Numery gospodarstw wylosowanych do próby z poszczególnych *o.s.* podano w ostatniej kolumnie powyższego zestawienia.

Uwaga 5.1. Gdyby na danym terenie istniała kartoteka gospodarstw rolnych, to pomimo, że karty noszą jakąś numerację, przeważnie nie będzie to numeracja ciągła i dla konstrukcji operatu losowania konieczne będzie nowe ponumerowanie kolejne wszystkich kart.

5.5. Zajmiemy się zagadnieniem szacowania średniej $\bar{Y}$ cechy mierzalnej Y w populacji:

$$\bar{Y}=\frac{1}{N}\sum_{k=1}^{N} Y_k \tag{3}$$

Jeśli nie możemy wykorzystać żadnych dodatkowych informacji o badanej populacji najlepszym estymatorem jest średnia z próby $\bar{y}$:

$$\bar{y}=\frac{1}{n}\sum_{i=1}^{n} y_i=\frac{1}{n}\sum_{k\in s} Y_k \tag{4}$$

Średnia $\bar{y}$ jest **nieobciążonym** estymatorem średniej $\bar{Y}$, czyli $E(\bar{y})=\bar{Y}$, a jej **wariancja** wynosi:

$$D^2(\bar{y})=\left(1-\frac{n}{N}\right)\frac{S^2}{n} \tag{5}$$

przy czym:

$$S^2=\frac{1}{N-1}\sum_{k=1}^{N}(Y_k-\bar{Y})^2$$

jest **wariancją cechy** Y w populacji.

Błąd standardowy szacunku wynosi:

$$D(\bar{y})=\sqrt{1-\frac{n}{N}}\cdot\frac{S}{\sqrt{n}} \tag{6}$$

natomiast **procentowy błąd standardowy** wynosi:

$$100\,V(\bar{y})=100\sqrt{1-\frac{n}{N}}\,\frac{V(Y)}{\sqrt{n}} \tag{7}$$

gdzie $V(Y)=\frac{S}{\bar{Y}}$ jest współczynnikiem zmienności cechy Y w populacji. Zakładamy, że $\bar{Y}>0$.

5.6. Z wzoru (5) wynika, że wariancja $D^2(\bar{y})$ jest tym większa, im większa jest wariancja badanej cechy S^2, a ponadto, że ze wzrostem n próby wariancja estymatora maleje odwrotnie proporcjonalnie do wielkości próby, a nawet nieco więcej, z uwagi na czynnik $\left(1-\frac{n}{N}\right)$.

5.7. Planując badanie statystyczne metodą reprezentacyjną należy decydować o pożądanej wielkości próby, gwarantującej żądaną precyzję szacunku. Jeśli założymy, że maksymalny błąd szacunku średniej $\bar{Y}$ nie powinien przekroczyć wielkości d, przy ryzyku około 5%, czyli przyjmiemy, że próba powinna być tak duża, że w przybliżeniu $2D(\bar{y})=d$, otrzymujemy z (5), że:

$$n\approx\frac{N}{1+\frac{Nd^2}{4S^2}} \tag{8}$$

Aby się posłużyć tym wzorem należy określić S^2. Tej wariancji z reguły nie znamy.

5.8. Wariancję S^2 można oszacować na podstawie danych poprzedniego badania populacji generalnej o ile nie zaszły znaczne zmiany strukturalne w tej populacji, w okresie między poprzednim badaniem i badaniem planowanym. Innym rozwiązaniem jest pobranie wstępnej próby i z niej oszacowanie S^2. Okazuje się, że estymatorem nieobciążonym tej wariancji jest wariancja s^2 z próby. Wielkość wstępnej próby zależy od parametru rozkładu badanej cechy w populacji zwanego **współczynnikiem ekscesu** β_2:

$$\beta_2=\frac{1}{N\sigma^4}\sum_{k=1}^{N}(Y_k-\bar{Y})^4\approx\frac{1}{NS^4}\sum_{k=1}^{N}(Y_k-\bar{Y})^4$$

W przypadku rozkładu normalnego $\beta_2=3$. Wówczas wstępna próba powinna liczyć około 50 elementów, W praktyce β_2 może się okazać wyższe; wówczas wstępna próba powinna być odpowiednio większa. Np. dla $\beta_2=6$ powinna wynosić około 125 elementów. Jeśli nie jesteśmy w stanie przeprowadzić badania wstępnego i nie możemy użyć danych z przeszłości, istnieje możliwość szacunku S^2, gdy określimy w przybliżeniu Y_{max} — najwyższą wartość oraz Y_{min} — najniższą wartość cechy Y w populacji. Przyjmujemy:

$$S^2\approx\frac{(Y_{max}-Y_{min})^2}{16} \tag{9}$$

Dokładniejsze postępowanie wymaga znajomości typu rozkładu cechy Y.

Przykład 5.4. Populacja generalna obejmuje $N=12$ tys. elementów. Szacujemy średnią $\bar{Y}$, przyjmując za maksymalny dopuszczalny błąd szacunku $d=2$. Z wstępnej próby oszacowano wariancję jako $S^2=200$. Jak wielka powinna być próba (*lpbz*)?

Z wzoru (8) wynika, że:

$$\frac{Nd^2}{4S^2}=\frac{12000\cdot 4}{4\cdot 200}=60,\quad n\approx\frac{12000}{61}\approx 197$$

Próba powinna objąć około 197 elementów populacji generalnej.

Przypuśćmy, że nie mogliśmy przeprowadzić badania wstępnego dla szacunku S^2. Sądzimy, że $Y_{min}=0$, $Y_{max}=60$. Wówczas ze wzoru (9) otrzymamy $S^2\approx 225$. Łatwo obliczyć, że próba powinna objąć około 221 elementów.

5.9. Szacowanie wartości globalnej cechy mierzalnej Y odbywa się za pomocą estymatora $N\bar{y}$. Jest to estymator nieobciążony wartości globalnej Y, a jego wariancja $D^2(N\bar{y})=N^2D^2(\bar{y})$, natomiast błąd standardowy szacunku $D(N\bar{y})=ND(\bar{y})$. Procentowy błąd standardowy:

$$100\,V(N\bar{y})=100\,V(\bar{y})=100\sqrt{1-\frac{n}{N}}\,\frac{V(Y)}{\sqrt{n}} \tag{10}$$

Precyzja szacunku wartości globalnej mierzona procentowym błędem standardowym jest taka sama, jak szacunku średniej. Precyzja zależy od wielkości współczynnika zmienności badanej cechy $V(Y)$ oraz od wielkości n próby. Dla przykładu podajemy niżej wielkości $100\,V(N\bar{y})$ dla paru wartości $V(Y)$ oraz n przy $N=10000$.

$V(Y)$	Procentowy błąd standardowy szacunku przy $N=10000$, jeżeli liczebność próby wynosi $n=$				
	1000	500	200	100	50
0,2	0,6	0,87	1,40	1,99	2,82
0,5	1,5	2,18	3,50	4,97	7,05
0,8	2,4	3,49	5,60	7,96	11,29
1,0	3,0	4,36	7,00	9,95	14,11
1,5	4,5	6,54	10,50	14,92	21,16

Najczęściej $V(Y)<1$, wówczas próba wielkości $N=200$ dostarczy stosunkowo precyzyjnej oceny, wartości globalnej Y lub średniej $\bar{Y}$, gdyż względny błąd standardowy szacunku wyniesie poniżej 7%. Jeśli $V(Y)<0{,}5$, to nawet próba $n=100$ elementowa dostarczy precyzyjnej oceny tych parametrów.

Zauważmy, że o precyzji szacunku średniej bądź wartości globalnej nie decyduje istotnie to, jaka jest frakcja próby ($f=n:N$), lecz jaka jest liczebność n próby. Np. gdyby w powyższym przykładzie przyjąć $N=100000$, to przy danych $V(Y)$ oraz n precyzja praktycznie się nie zmieni. Dla próby $n=500$ oraz $V(Y)=0{,}5$ otrzymujemy $100\,V(N\bar{y})=100\,V(\bar{y})=2{,}179\cong 2{,}18$. Mówiąc mniej ściśle, jeśli frakcja próby nie jest zbyt wysoka np. jest mniejsza od 0,1, to taką samą precyzję dla dużej populacji, jak i dla wielokrotnie mniejszej populacji uzyskamy przy tej samej wielkości próby.

5.10. Rozważania z pkt. 5.5 i 5.9 łatwo przenieść na przypadek szacowania frakcji P elementów wyróżnionych (ze względu na posiadanie określonego wariantu cechy niemierzalnej) bądź liczby NP takich elementów, zamieniając $\bar{Y}$ na P, $\bar{y}$ na p, gdzie p oznacza frakcję elementów wyróżnionych w próbie. Zatem, frakcja p z próby jest nieobciążonym estymatorem frakcji P, czyli $E(p)=P$, a jej wariancja wynosi:

$$D^2(p)=\left(1-\frac{n}{N}\right)\frac{S^2}{n} \tag{11}$$

Obecnie $S^2=\frac{NP(1-P)}{N-1}=\frac{NPQ}{N-1}$, $Q=1-P$

Wzór (10) przekształci się na:

$$100V(Np)=100V(p)=100\sqrt{1-\frac{n}{N}}\sqrt{\frac{N(1-P)}{(N-1)nP}}\cong 100\sqrt{1-\frac{n}{N}}\cdot\sqrt{\frac{1-P}{P}}\cdot\frac{1}{\sqrt{n}} \tag{12}$$

Wyrażenie $\sqrt{\frac{1-P}{P}}$ jest odpowiednikiem $V(Y)$ we wzorze (10). Może ono przybierać duże wartości, gdyż wzrasta nieograniczenie, gdy P maleje. Obrazuje to następujące zestawienie:

P	0,5	0,1	0,05	0,01	0,005	0,001
$\sqrt{\frac{1-P}{P}}$	1	3	4,36	9,95	14,11	31,61

Stąd wniosek, że jeśli zażądamy dużej precyzji szacunku przy małej wartości szacowanej frakcji, próba musi być liczebnie wysoka. Np. jeśli w populacji $N=10000$ chcemy oszacować frakcję rzędu 0,01 z procentowym błędem standardowym nie przekraczającym 5%, to próba powinna objąć $n=7984$ elementy populacji. O ile rzeczywiście tak wysoka precyzja jest niezbędna dla marginesowych frakcji (liczb) wyróżnionych elementów populacji należy poszukiwać odpowiednio bardziej efektywnych schematów losowania próby niż *lpbz* dla szacowania P bądź NP. Może się okazać, że takich schematów nie znajdziemy i wówczas trzeba zrezygnować z badania reprezentacyjnego.

5.11. Dotąd rozpatrywaliśmy sytuację, gdy szacowany parametr dotyczy rozkładu cechy w całej populacji. W praktyce wyróżniamy często pewne podpopulacje, stanowiące **dziedziny studiów** (domeny). Wówczas szacowanie np. średniej wartości $\bar{Y}'$ może dotyczyć jedynie elementów, należących do domeny. Wobec tego procentowy błąd standardowy będzie zależał od **wielkości tej części próby, która wchodzi w skład domeny, a nie od liczebności całej próby.** Także współczynnik zmienności badanej cechy będzie ograniczony do domeny w populacji generalnej. Średnia $\bar{y}'$ z próby z domeny jest nieobciążonym estymatorem średniej $\bar{Y}'$ w domenie (podpopulacji), czyli

$E(\bar{y}')=\bar{Y}'$ oraz procentowy błąd standardowy średniej $\bar{y}'$ wynosi w przybliżeniu (wystarczającym w praktyce):

$$100V(\bar{y}')\approx 100\sqrt{1-\frac{n}{N}}\frac{V'(Y)}{\sqrt{nP_0}} \tag{13}$$

W tym wzorze n, N oznaczają liczebność próby, populacji, odpowiednio: P_0 jest frakcją domeny w populacji, a $V'(Y)$ współczynnikiem zmienności cechy Y w domenie. Stąd wynika, że liczebność domeny wynosi $N_0=NP_0$. Różnica pomiędzy prawą stroną (13) i (10) polega głównie na tym, że w mianowniku pod pierwiastkiem zamiast n występuje $nP_0<n$. Inaczej, jeśli domena obejmuje, na przykład 20% populacji, to próba domeny wyniesie około 20% próby, a więc jeśli $n=1000$, to dla omawianej domeny $nP_0=200$, próba obejmuje tylko około 200 elementów domeny.

Przykład 5.5. Badamy metodą reprezentacyjną dzietność kobiet w mieście. Populację badaną tworzą kobiety w wieku prokreacyjnym. Chcemy oszacować przeciętną liczbę urodzonych dzieci przez kobiety w wieku 20—24 lata. Domeną jest grupa kobiet w wieku 20—24 lata. Średnią liczbę $\bar{Y}'$ urodzonych przez nie dzieci otrzymamy z informacji o liczbie dzieci urodzonych przez poszczególne kobiety z domeny.

Załóżmy, że $N=30737$; $P_0=0{,}121$; $N_0=3720$; $\bar{Y}'=2{,}1$; $V(Y)=0{,}753$; $V'(Y)=0{,}773$. Losujemy (*lpbz*) próbę $n=1356$ kobiet w wieku 15—49 lat. Wówczas:

$$100V(\bar{y})=100\sqrt{1-\frac{1356}{30737}}\cdot\frac{0{,}753}{\sqrt{1356}}=2{,}00\%$$

natomiast:

$$100V(\bar{y}')=100\sqrt{1-\frac{1356}{30737}}\cdot\frac{0{,}773}{\sqrt{164{,}08}}=5{,}90\%$$

Szacując liczbę urodzonych dzieci za pomocą estymatora $N\bar{y}$:

$$100V(N\bar{y})=100V(\bar{y})=2\%$$

Szacując liczbę Y' urodzonych dzieci przez kobiety w wieku 20—24 lata za pomocą estymatora $Np_0\bar{y}'$, gdzie p_0 jest frakcją z próby kobiet domeny w wieku 20—24 lata, wzór (13) nie będzie dość dokładny. Wówczas:

$$\begin{aligned}100V(Np_0\bar{y}')&\approx 100\sqrt{1-\frac{n}{N}}\cdot\frac{\sqrt{V'^2(Y)+(1-P_0)}}{\sqrt{nP_0}}=\\ &=100\sqrt{1-\frac{1356}{30737}}\cdot\frac{\sqrt{0{,}597529+(1-0{,}121)}}{\sqrt{164{,}08}}=9{,}3\%\end{aligned} \tag{13a}$$

5.12. Z wzoru (13) łatwo otrzymać wzór na procentowy błąd standardowy frakcji elementów wyróżnionych w domenie:

$$P'_0=\frac{N'_0}{N_0}=\frac{N'_0}{NP_0}$$

Wzór ten ma postać:

$$100\,V(p'_0)\approx 100\sqrt{1-\frac{n}{N}}\,\frac{\sqrt{\frac{1-P'_0}{P'_0}}}{\sqrt{nP_0}} \tag{14}$$

Przez p'_0 oznaczamy frakcję $\frac{n'_0}{n_0}$, przy czym n_0 jest liczebnością domeny w próbie, natomiast n'_0 — liczbą jednostek wyróżnionych w próbie z domeny.

Przykład 5.6. N, n, P_0, jak w przykładzie 5.5. Niech frakcja elementów wyróżnionych w domenie, np. frakcja kobiet w wieku 20—24, które nie urodziły dziecka, wynosi $P'_0=0{,}27$. Procentowy błąd standardowy estymatora p'_0 tej frakcji — zgodnie z wzorem (14) — wynosi 1,6%. Szacowanie takiej samej frakcji dla całej populacji daje procentowy błąd standardowy 0,83%, czyli prawie dwa razy mniejszy niż poprzednio.

5.13. Dla oszacowania ilorazu R wartości globalnych (sum) bądź średnich dwóch cech mierzalnych Y i $\bar{Y}$;

$$R=\frac{Y}{X}=\frac{\bar{Y}}{\bar{X}} \tag{15}$$

posłużmy się statystyką:

$$r=\frac{\bar{y}}{\bar{x}} \tag{16}$$

Jest to estymator obciążony, ale zgodny, jeśli więc próba jest dostatecznie duża obciążenie nie ma praktycznego znaczenia. Wariancja estymatora r wynosi w przybliżeniu:

$$D^2(r)\approx R^2\left(1-\frac{n}{N}\right)\cdot\left[\frac{V^2(X)+V^2(Y)-2\rho_{xy}V(X)V(Y)}{n}\right] \tag{17}$$

$\rho_{xy}=\rho$ jest współczynnikiem korelacji cech X i Y;

$$\rho=\frac{1}{(N-1)S_xS_y}\sum_{k=1}^{N}(X_k-\bar{X})(Y_k-\bar{Y})$$

Uwaga 5.2. Nasuwa się pytanie, jak interpretować warunek „jeśli próba jest dostatecznie duża". W rozważanym zagadnieniu oznacza to, że próba powinna być tak duża, żeby procentowy błąd standardowy $100 \cdot V(\bar{x})$ był mniejszy niż 5%. Gdyby $\rho_{xy} \cong V(X) : V(Y)$, to wystarczy tak liczna próba, aby $100V(\bar{x}) < 15\%$.
Łatwo wykazać, że prawa strona (17) daje się przedstawić, jako:

$$D^2(r) \approx \left(1 - \frac{n}{N}\right) \cdot \frac{1}{n(N-1)\bar{X}^2} \cdot \sum_{k=1}^{N} (Y_k - RX_k)^2 \tag{17a}$$

Procentowy błąd standardowy szacunku ilorazu R wynosi w przybliżeniu:

$$100V(r) \approx 100 \sqrt{1 - \frac{n}{N}} \cdot \sqrt{\frac{V^2(X) + V^2(Y) - 2\rho_{xy} V(X) V(Y)}{n}} \tag{18}$$

Przykład 5.7. Na obszarze znajduje się $N = 7000$ gospodarstw domowych. W celu oszacowania miesięcznych wydatków żywnościowych na osobę planujemy przeprowadzenie badania reprezentacyjnego na próbie $n = 700$ gospodarstw domowych. Niech Y_k oznacza wydatki na żywność w k-tym gospodarstwie, natomiast X_k — liczbę osób w tym gospodarstwie — $k = 1, 2, ..., 7000$. Szacowany parametr:

$R = \frac{\bar{Y}}{\bar{X}} = \frac{\sum_k Y_k}{\sum_k X_k}$. Ile wyniesie procentowy błąd standardowy szacunku, jeśli sądzimy, że $V(X) = 0{,}64$, $V(Y) = 0{,}82$, $\rho_{xy} = 0{,}7$?

$$100V(r) \cong 100 \sqrt{1 - \frac{700}{7000}} \sqrt{\frac{0{,}64^2 + 0{,}82^2 - 2 \cdot 0{,}7 \cdot 0{,}64 \cdot 0{,}82}{700}} = 2{,}1\%,$$

5.14. **Zadania.** 1) Szacujemy średnią $\bar{Y}$ w populacji, obejmującej $N = 20$ tys. jednostek. Jaka powinna być minimalna liczebność próby, aby błąd standardowy szacunku nie przekroczył $d = 0{,}25$, jeśli $n' = 60$ elementowa próba wstępna wynosi:

$$\sum_{i=1}^{60} y_i = 729 \quad \sum_{i=1}^{60} y_i^2 = 12\,500?$$

2) Szacujemy frakcję F mieszkań o powierzchni co najmniej 60 m^2. Jak duża powinna być próba z populacji $N = 15000$ mieszkań, jeżeli przypuszczamy, że $0{,}2 \leqslant P \leqslant 0{,}4$ oraz żądamy, aby procentowy błąd standardowy szacunku nie przekroczył 7%?

5.15. **LITERATURA UZUPEŁNIAJĄCA:** R. Zasępa [18] str. 64—75, 98—106, J. Steczkowski [14] str. 135—148, 158—167.

Wykład 6. Estymatory ilorazowe i regresyjne w losowaniu prostym bez zwracania

6.1. W poprzednim wykładzie zostały omówione własności średniej z próby $\bar{y}$ w przypadku szacowania (*lpbz*) średniej $\bar{Y}$. Przypuśćmy, że dla badanej populacji posiadamy informację o wartości średniej $\bar{X}$ cechy mierzalnej X, która, jak przypuszczamy, jest skorelowana z badaną cechą Y. Wówczas można tę informację wykorzystać konstruując **estymator ilorazowy średniej** $\bar{Y}$:

$$\bar{y}_q = \bar{y} \cdot \frac{\bar{X}}{\bar{x}} = r\bar{X}, \quad r = \frac{\bar{y}}{\bar{x}} \tag{1}$$

Przykład 6.1. Przeprowadzamy badanie reprezentacyjne pogłowia zwierząt gospodarskich w gminie (lpbz). Dla oszacowania liczby krów (cecha Y) chcemy wykorzystać informację o liczbie krów w indywidualnych gospodarstwach rolnych ze spisu ubiegłorocznego (cecha X). Wówczas dla każdego gospodarstwa z próby notujemy dodatkowo liczbę krów X_k w ubiegłym roku i estymatorem liczby Y krów w roku bieżącym będzie estymator ilorazowy:

$$N\bar{y}_q = rN\bar{X} = rX$$

Uwaga 6.1. W punkcie 6.1. zakłada się, że wartość statystyki $\bar{x}$ uzyskuje się z badania reprezentacyjnego. W pewnych przypadkach może powstać obawa, że dane o wartości cechy dodatkowej uzyskiwane w czasie tego badania mogą nie pokrywać się z danymi badania pełnego, będącego podstawą obliczenia średniej $\bar{X}$. Wówczas — tak jak w przykładzie 6.1 — posiadanie informacji o wartościach cechy X, dla poszczególnych elementów populacji generalnej, można wykorzystać, o ile identyfikatory jednostek z próby pozwalają na takie postępowanie.

6.2. Estymator ilorazowy $\bar{y}_q$ jest obciążonym, ale zgodnym estymatorem średniej $\bar{Y}$. Jeżeli próba jest dostatecznie duża, to obciążenie estymatora można pominąć. Próba jest dostatecznie duża, gdy względny błąd standardowy $\bar{X}$, $V(\bar{x}) \leqslant 0{,}05$.
Wówczas MSE $(\bar{y}_q) \cong D^2(\bar{y}_q)$:

$$D^2(\bar{y}_q) = \left(1 - \frac{n}{N}\right)\left[\frac{R^2 S_x^2 + S_y^2 - 2\rho_{xy} R S_x S_y}{n}\right] = \tag{2}$$

$$= \left(1 - \frac{n}{N}\right) \cdot \frac{1}{n} \cdot \frac{1}{N-1} \sum_{k=1}^{N} \left(Y_k - RX_k\right)^2 \tag{2a}$$

Przypominamy oznaczenia: $R=\frac{\bar{Y}}{\bar{X}}$, $S_x^2=\frac{1}{N-1}\sum_{k=1}^{N}\left(X_k-\bar{X}\right)^2$, analogicznie obliczamy S_y^2; n, N — liczebność próby, populacji, odpowiednio:

$$\rho_{xy}=\frac{1}{(N-1)S_xS_y}\sum_{k=1}^{N}(X_k-\bar{X})\ (Y_k-\bar{Y})$$

jest współczynnikiem korelacji cech X i Y.

6.3. Wobec (2a) estymatorem obciążonym, ale zgodnym wariancji $D^2(\bar{y}_q)$ jest:

$$\check{D}^2(\bar{y}_q)=\left(1-\frac{n}{N}\right)\cdot\frac{1}{n}\cdot\frac{1}{n'-1}\left[\sum_{i=1}^{n'}y_i^2+r^2\sum_{i=1}^{n}x_i^2-2r'\sum_{i=1}^{n}x_iy_i\right] \tag{3}$$

w przypadku próby n'-elementowej, przy czym r' obliczamy według (1) z tej próby (np. wstępnej próby, gdy chcemy zorientować się, jak duża przy danym n może być wariancja $D^2(\bar{y}_q)$ estymatora ilorazowego $\bar{y}_q$). Praktyka wskazuje, że jeśli próba (wstępna próba) jest tak duża, że $V(\bar{x})\leqslant 0{,}1$ oraz $V(\bar{y})\leqslant 0{,}1$, to ocena wariancji $D^2(\bar{y}_q)$ na podstawie (3) jest wystarczająco precyzyjna dla potrzeb ustalenia „minimalnej" próby gwarantującej precyzję oceny $\bar{Y}$.

Przykład 6.2. Zamierzamy oszacować metodą reprezentacyjną średnią liczbę bydła $\bar{Y}$ na indywidualne gospodarstwo rolne w populacji obejmującej $N=6200$ gospodarstw, o której wiemy, że jest tam $N\bar{X}=19840$ sztuk trzody chlewnej. Tę informację zamierzamy wykorzystać stosując estymator ilorazowy $\bar{y}_q=r\bar{X}$. W celu zorientowania się, jak dużą należałoby wylosować próbę (*lpbz*), wylosowano wstępną próbę $n=50$ gospodarstw. Z tej próby otrzymano (w [18] na str. 79—81 figurują dane oraz szczegółowe obliczenia) stosując wzór (3):

$$\check{D}^2(\bar{y}_q)=\left(1-\frac{n}{N}\right)\cdot\frac{1{,}99}{n}$$

Ponadto okazało się, że $V(\bar{x})\cong 0{,}1$ oraz $V(\bar{y})\cong 0{,}1$. Wyniki z wstępnej próby są dostatecznie precyzyjne dla ustalenia na ich podstawie liczebności n-planowanego badania reprezentacyjnego.

Przypuśćmy, że przy ryzyku 0,05 zgadzamy się na błąd oceny $d=0{,}1$, tzn. że nie chcemy pomylić się przy szacunku więcej niż o $d=0{,}1$, przy prawdopodobieństwie większej pomyłki wynoszącym 0,05. Odpowiada to warunkowi, że $2\check{D}(\bar{y}_q)=0{,}1$. Pozostaje do rozwiązania równanie:

$$\left(1-\frac{n}{N}\right)\cdot\frac{1{,}99}{n}=0{,}0025,\qquad N=6200$$

$$\frac{1}{n}-\frac{1}{N}=\frac{0{,}0025}{1{,}99},\qquad \frac{1}{n}=0{,}00016129+0{,}00125628\qquad n=706$$

Gdybyśmy postawili warunek, że taki sam błąd oceny chcemy zapewnić przy ostrzejszym kryterium, mianowicie ryzyku 0,0027, co odpowiada temu, że $3\hat{D}(\bar{y}_q)=0{,}1$, czyli $D(\bar{y}_q)=0{,}03$, wówczas otrzymamy $n=1391$. Zatem, w pierwszym przypadku próba powinna liczyć około 706 gospodarstw, w drugim przypadku — około 1391 gospodarstw.

6.4. Na pytanie, czy zawsze estymator ilorazowy $\bar{y}_q$ jest efektywniejszy od średniej $\bar{y}$ z próby odpowiada **twierdzenie:**

Przy szacowaniu średniej $\bar{Y}$ estymator ilorazowy $\bar{y}_q$ jest efektywniejszy od średniej $\bar{y}$ z próby, czyli $D^2(\bar{y}_q)<D^2(\bar{y})$, jeżeli:

$$\rho_{xy}>\frac{1}{2}\,\frac{V(X)}{V(Y)} \tag{4}$$

Oba estymatory są tak samo efektywne, gdy w (4) znak większości ($>$) zastąpimy przez znak równości ($=$).

Jeśli:

$$\rho_{xy}<\frac{1}{2}\,\frac{V(X)}{V(Y)} \tag{5}$$

to średnia z próby $\bar{y}$ jest efektywniejsza od $\bar{y}_q$.

Uwaga 6.2. W powyższym twierdzeniu zakłada się, że $\bar{X}>0$ oraz $\bar{Y}>0$. Te warunki są zwykle w praktyce spełnione.

6.5. Z twierdzenia 6.4. wynika, że informacja o średniej $\bar{X}$ w populacji nie zawsze jest przydatna do konstrukcji estymatora ilorazowego $\bar{y}_q$. Chodzi o to, aby cecha X była możliwie silnie skorelowana z cechą badaną Y, co zapewnia spełnienie nierówności (4). Spełnienie tej nierówności jest tym łatwiejsze, jeżeli $V(X)$ jest mniejszy od $V(Y)$. Z dwóch informacji dodatkowych, w których cechy są podobnie skorelowane z cechą Y należy wybrać tę, dla której $V(X)$ jest mniejsze.

Jeśli ma miejsce (4) to zysk na efektywności w procentach wynosi:

$$100\cdot\frac{D^2(\bar{y})-D^2(\bar{y}_q)}{D^2(\bar{y}_q)}\approx 100\cdot\frac{V(X)}{V(Y)}\left[2\rho_{xy}-\frac{V(X)}{V(Y)}\right] \tag{6}$$

I tak np., gdy $\dfrac{V(X)}{V(Y)}=0{,}8$; $\rho=0{,}7$; to zysk ten wynosi 47%.

6.6. Wychodząc od estymatora (1) łatwo uzyskujemy **estymator ilorazowy** $\hat{y}_q$ **wartości globalnej** Y cechy badanej. Mianowicie:

$$\hat{y}_q=N_{\bar{y}}\frac{\bar{X}}{\bar{x}}=r\,X, \qquad r=\frac{\bar{y}}{\bar{x}} \tag{7}$$

Własności tego estymatora są analogiczne do własności estymatora $\bar{y}_q$, a jego wariancję otrzymujemy z (2) lub (2a) mnożąc odnośne wzory przez N^2.

6.7. Inną metodą wykorzystania dodatkowej informacji do budowy efektywniejszego estymatora jest wykorzystanie regresji liniowej. **Estymatorem liniowym regresyjnym** $\bar{y}_{lr}$ średniej $\bar{Y}$ jest:

$$\bar{y}_{lr} = \bar{y} + b(\bar{X} - \bar{x}), \quad b = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} \tag{8}$$

Współczynnik b w (8) jest zwykłym **współczynnikiem regresji z próby cechy** Y względem cechy X.

Uwaga 6.3. W wykładzie 2. *Wybrane elementy rachunku prawdopodobieństwa* była mowa o współczynniku regresji drugiego rodzaju, który w przypadku populacji skończonej ma postać:

$$B = \rho_{xy}\frac{S_y}{S_x} = \frac{\sum_k (X_k - \bar{X})(Y_k - \bar{Y})}{\sum_k (X_k - \bar{X})^2} \tag{9}$$

6.8. **Twierdzenie.** Estymator $\bar{y}_{lr}$ jest obciążonym, ale zgodnym estymatorem średniej $\bar{Y}$. Jeżeli próba jest dostatecznie duża, to obciążenie estymatora można pominąć. Wariancja estymatora wynosi w przybliżeniu:

$$D^2(\bar{y}_{lr}) \approx \left(1 - \frac{n}{N}\right) \cdot \frac{S_y^2(1 - \rho_{xy}^2)}{n} \tag{10}$$

6.9. Wariancję (10) możemy oszacować z wstępnej próby n'-elementowej za pomocą zgodnego estymatora:

$$\check{D}^2(\bar{y}_{lr}) = \left(1 - \frac{n}{N}\right) \cdot \frac{s_u'^2}{n} \tag{10a}$$

gdzie $s_u'^2$ jest wariancją z wstępnej próby zmiennej losowej $u = y - b'x$; po łatwych przekształceniach otrzymujemy:

$$s_u'^2 = \frac{1}{n'-1}\left[(n'-1)s_y'^2 - \frac{\left(\sum_i x_i y_i - n'\bar{x}\bar{y}\right)^2}{(n'-1)S_x'^2}\right] \quad n' \geqslant 50$$

Przykład 6.3. Dla danych z przykładu 6.2 otrzymujemy $s_u'^2 = 1{,}6$, czyli:

$$\check{D}^2(\bar{y}_{lr}) = \left(1 - \frac{n}{N}\right)\frac{1{,}6}{n}$$

Ta wariancja jest mniejsza niż:

$$\check{D}^2(\bar{y}) = \left(1 - \frac{n}{N}\right) \cdot \frac{1{,}99}{n}$$

6.10. Estymatory ilorazowe i regresyjne stosujemy, gdy próba jest „duża" i wówczas $MSE(\cdot) \cong D^2(\cdot)$. Okazuje się, że estymator $\bar{y}_{lr}$ jest efektywniejszy od estymatora $\bar{y}_q$, jeśli:

$$\rho_{xy} \neq \frac{V(X)}{V(Y)} \tag{11}$$

Gdy nie zachodzi (11), czyli $\rho_{xy} = \frac{V(X)}{V(Y)}$, to oba estymatory są tak samo efektywne. Wówczas linia regresji drugiego rodzaju cechy Y względem cechy X przechodzi przez początek układu.

6.11. **Twierdzenie,** Estymator regresyjny $\bar{y}_{lr}$ jest efektywniejszy od średniej próby, przy szacowaniu średniej $\bar{Y}$, jeżeli $\rho_{xy} \neq 0$. W przypadku gdy $\rho_{xy} = 0$, oba estymatory są tak samo efektywne (mają takie same wariancje).

6.12. Procentowy zysk na efektywności przy użyciu $\bar{y}_{lr}$ zamiast $\bar{y}$ do szacowania średniej $\bar{Y}$ wynosi:

$$100 \cdot \frac{D^2(\bar{y}) - D^2(\bar{y}_{lr})}{D^2(\bar{y})} \approx . \; 100\rho_{xy}^2 \; \% \tag{12}$$

Tak np. jeżeli $\rho_{xy} = 0{,}5$, to wariancja $D^2(\bar{y}_{lr})$ jest o 25% mniejsza od wariancji średniej $\bar{y}$ z próby. Przy $\rho_{xy} = 0{,}7$ wariancja $D^2(\bar{y}_{lr})$ zmniejsza się o 49% w stosunku do wariancji $D^2(\bar{y})$.

6.13. Korzystanie z estymatora ilorazowego bądź regresyjnego podraża koszt badania statystycznego: 1) badamy dwie cechy zamiast jednej — co zwiększa koszt obserwacji i opracowania zebranych informacji; 2) w przypadku estymatora regresyjnego dochodzi koszt obliczenia współczynnika regresji b z próby. Porównując efektywność tych estymatorów z efektywnością średniej z próby zakładaliśmy tę samą liczebność próby. Planując badanie reprezentacyjne należy porównywać efektywność estymatorów, przy tych samych kosztach badania statystycznego. Dopiero wówczas możemy decydować, który estymator będzie w konkretnym badaniu najefektywniejszy.

6.14. W praktyce stosowanie estymatora regresyjnego $\bar{y}_{lr}$ jest rzadkie. Często stosuje się natomiast estymator ilorazowy $\bar{y}_q$.

6.15. **Zadania.** 1) W gminie znajduje się $N=2500$ indywidualnych gospodarstw rolnych o przeciętnej powierzchni gospodarstwa $\bar{X}$ równej 3,2 ha. Jak wielka powinna być próba, aby szacując przeciętną liczbę $\bar{Y}$ bydła na gospodarstwo, przy użyciu estymatora ilorazowego $\bar{y}_q$, procentowy błąd standardowy szacunku nie przekraczał 5%, jeżeli wstępna próba $n=50$ gospodarstw dała wyniki:

$$\sum_{i=1}^{50} x_i = 150; \quad \sum_{i=1}^{50} y_i = 110; \quad \sum_{i=1}^{50} x_i^2 = 631; \quad \sum_{i=1}^{50} y_i^2 = 422; \quad \sum_{i=1}^{50} x_i y_i = 460$$

2) Warunki jak wyżej. O ile zwiększy się procentowy błąd standardowy, gdy użyjemy estymatora prostego $\bar{y}$ (pomijamy informacje o cesze X)?

3) Dane jak w zadaniu 1). O ile zmniejszy się procentowy błąd standardowy przy tej samej wielkości próby, gdy zastosujemy do szacowania $\bar{Y}$ estymator regresyjny $\bar{y}_r$?

6.16. **LITERATURA UZUPEŁNIAJĄCA:** R. Zasępa [18] str. 76—98, 106—111, J. Steczkowski [14] str. 148—158, Cz. Bracha [1]. str. 30—39.

Wykład 7. Przypomnienie ważniejszych pojęć, problemów i technik omówionych w poprzednich wykładach

7.1. W wykładach analizujemy pewną technikę częściowego badania statystycznego zwaną badaniem metodą reprezentacyjną albo krótko — **badaniem reprezentacyjnym.** W badaniu reprezentacyjnym wybór próby jest wyborem losowym, generowanym przez statystyka. Posługując się aparatem rachunku prawdopodobieństwa oraz statystyki matematycznej możemy analizować różne wzorce próbkowania pod względem ich efektywności, a więc wnioskować, który schemat losowania próby jest w danych warunkach najkorzystniejszy. Tej zalety nie posiada żadna inna metoda badania częściowego ani tzw. wybór celowy, ani wybór kwotowy bądź inna metoda wyboru próby niż wybór losowy.

7.2. Zbiorowość statystyczną, będącą przedmiotem badania nazywamy populacją generalną lub wprost: **populacją.** W naszych rozważaniach zakładamy, że badana populacja jest **skończona,** a jej elementy, czyli **jednostki badania** można uporządkować w skończony ciąg; w większości przypadków w praktyce takie uporządkowanie jest bardzo pracochłonne i nie zawsze konieczne, ale ważne jest, że jest teoretycznie możliwe. Jeśli populacja generalna obejmuje N elementów, to w przyjętym uporządkowaniu każdy element populacji ma określony numer k; $k=1, 2, ..., N$. Dwa różne elementy populacji mają różne numery k_1, k_2, $k_1 \neq k_2$. Zatem numer k jest **identyfikatorem** jednostki badania.

7.3. Jednostki badanej populacji różnią się pomiędzy sobą wartościami pewnych cech mierzalnych bądź posiadaniem określonego wariantu cechy niemierzalnej. Celem badania statystycznego jest określenie wartości **parametrów populacji,** będących liczbowym opisem rozkładów badanych cech oraz związków między nimi. Badane parametry najczęściej są publikowane w postaci **tablic statystycznych** zwanych też **podstawowymi tablicami** badania statystycznego.

7.4. Cechy oznaczamy dużymi końcowymi literami alfabetu: X, Y, Z, U, W, ... Wartość cechy X do k-tego elementu populacji generalnej oznaczamy symbolem X_k przyjmując, że X_k jest wartością cechy mierzalnej; natomiast $X_k=1$, gdy chodzi o cechę niemierzalną i k-ty element populacji posiada wyróżniony wariant tej cechy, zaś $X_k=0$. gdy posiada inny wariant cechy. Jednostkę populacji, posiadającą wyróżniony wariant cechy niemierzalnej, nazywamy też **jednostką wyróżnioną.**

7.5. Przykłady parametrów populacji:

a) wartość globalna X cechy mierzalnej $X=X_1+X_2+...+X_N$,

b) liczba $N'=NP$ jednostek wyróżnionych,

c) wartość średnia cechy mierzalnej $\bar{X}=1/N(X_1+X_2+...+X_N)$,

d) frakcja P jednostek wyróżnionych,

e) iloraz R wartości globalnych cech X i Y, $R=Y:X$.

Ponadto, analizując rozkłady cech oraz precyzję wyników badania reprezentacyjnego bardzo ważną rolę odgrywają parametry:

f) wariancja cechy (σ^2 lub S^2),

g) odchylenie standardowe (σ lub S),

h) współczynnik zmienności $\left(V=\frac{\sigma}{\bar{X}} \text{ bądź } V=\frac{S}{\bar{X}}\right)$,

i) współczynnik korelacji ρ_{xy} cech X i Y,

$$\rho_{xy}=\frac{1}{N\sigma_x\sigma_y}\sum_{k=1}^{N}(X_k-\bar{X})(Y_k-\bar{Y})$$

7.6. Rozważając możliwości losowego wyboru próby, tj. części badanej populacji (pewne elementy populacji mogą być w próbie powtórzone wielokrotnie!) dochodzimy do pojęcia **zbioru S prób** oraz pojęcia **wzorca (planu) próbkowania losowego.** Dla każdej próby $s\in S$ wzorzec określa prawdopodobieństwo $p(s)\geqslant 0$ jej wyboru. Jeśli $p(s)=0$, to próby s nigdy nie wylosujemy, przy danym wzorcu próbkowania. Jeśli $p(s)>0$, to przy bardzo dużej liczbie niezależnych realizacji wzorca próba s będzie wylosowana w około 100 p(s)% przypadków. Formalnie, wzorzec próbkowania jest **funkcją prawdopodobieństwa** określoną na zbiorze S prób.

7.7. Przy danym wzorcu próbkowania mechanizm losowy, polegający na kolejnym wyborze do próby jednostek populacji bądź jej zespołów nazywamy **schematem losowania.** Wybierane kolejno jednostki (bądź zespoły jednostek) do próby nazywamy **jednostkami losowania** (JL). Każda jednostka badania (element populacji) powinna należeć dokładnie do jednej JL.

7.8. Losowanie próby prowadzi się na dokumentach, określających JL wraz z ich identyfikatorami (cechami adresowymi). Dokumenty te zwane **operatem losowania**

mogą mieć formę wykazów, kartotek, map z granicami *JL* itp. W operacie losowania *JL* muszą być uporządkowane w ciąg, czyli każda jednostka losowania musi posiadać określony numer w tym ciągu. Operat losowania powinien być **kompletny** i **aktualny**, to jest nie powinno w nim być pominięć ani wielokrotnego ujęcia tej samej jednostki badania lub obejmowania jednostek nie wchodzących w skład badanej populacji. Niekompletność bądź nieaktualność operatu losowania jest przyczyną **nielosowych błędów pokrycia.**

7.9. Losowanie próby może być **jednostopniowe** albo **wielostopniowe.** W przypadku losowania **dwustopniowego** najpierw konstruujemy zespoły jednostek populacji, stanowiące jednostki losowania pierwszego stopnia (*JLPS*). Korzystając z odpowiedniego operatu losowania losujemy próbę I stopnia. Wylosowane do próby *JLPS* dzielimy na mniejsze jednostki losowania drugiego stopnia (*JLDS*). Konstruujemy więc nowy operat losowania **ograniczony** do próby I stopnia. Próba wylosowanych *JLDS* stanowi próbę ostateczną. Wchodzą do niej te jednostki populacji, które należą do wylosowanej próby II stopnia. Schemat losowania nazywamy wówczas dwustopniowym. Postępowanie analogiczne do podziału *JLPS*, ale odnoszące się do *JLDS* pozwala na losowanie **trójstopniowe** itd.

7.10. Próbę można losować z całej populacji lub z poszczególnych warstw. **Warstwy** stanowią rozłączne części populacji, ale w sumie ją pokrywające, np. dla całego kraju warstwami mogą być województwa, ewentualnie osobne warstwy mogą stanowić obszary miast oraz obszary wsi województwa. Warstwy stanowią często tzw. **dziedziny studiów,** czyli **domeny.** Z reguły domenami są miasta (wsie) województw. W przypadku schematu losowania warstwowego (jednostopniowego) musimy posługiwać się **warstwowymi operatami losowania.**

7.11. Jednostki losowania mogą być wybierane do próby z jednakowymi bądź z różnymi prawdopodobieństwami wyboru, ze zwracaniem bądź też bez zwracania. Technikę losowania omówiono szczegółowo w wykładzie 3, wraz z przykładami praktycznymi.

7.12. Uzyskane z próby informacje należy odpowiednio przetworzyć, aby uzyskać **ocenę** t danego **parametru** T populacji. Dowolną funkcję wyników z próby nazywamy **statystyką.** Statystykę stosowaną do szacowania wartości parametru T nazywamy **estymatorem** tego parametru. Z natury losowego wyboru próby wynika, że estymator danego parametru jest **zmienną losową.** W przypadku dużych prób, z jakimi przeważnie mamy do czynienia w praktyce, rozkład estymatora można dobrze aproksymować za pomocą **rozkładu normalnego** o tych samych momentach pierwszego i drugiego rzędu.

7.13. Dyspersję estymatora t możemy mierzyć wartością **średniego błędu kwadratowego** $MSE(t)=B^2(t)+D^2(t)$, przy czym $B(t)=E(t)-T$ jest **obciążeniem** estymatora parametru T, a $D^2(t)$ **wariancją** estymatora. Jeśli estymator t jest **nieobciążony,** czyli $B(t)=0$, wówczas $MSE(t)=D^2(t)$.

7.14. W praktyce posługujemy się **estymatorami zgodnymi** nieobciążonymi bądź takimi, że przy dużej liczebności próby obciążenie estymatora jest względnie małe, a ich wariancja $D^2(t)$ zmierza do zera, gdy liczebność próby wzrasta nieograniczenie. Wobec tego, przy dużych próbach, stosowane estymatory zgodne obciążone możemy

uważać za praktycznie nieobciążone, czyli $MSE(t) \approx D^2(t)$, natomiast $\sqrt{MSE(t)} = D(t)$ — odchyleniowi standardowemu estymatora t parametru T.

7.15. **Ocena** parametru T jest **uzyskaną dla próby wartością estymatora** t. Nie możemy określić wielkości błędu popełnianego, przyjmując ocenę za wartość parametru. Gdybyśmy wylosowali inną próbę ocena wypadłaby najprawdopodobniej inna, jako druga realizacja zmiennej losowej t. Możemy natomiast w określony sposób scharakteryzować dobroć oceny, szacując wariancję $D^2(t)$ zgodnego estymatora t parametru T. Precyzję szacunku określamy wartością odchylenia standardowego $D(t)$ zwanego **błędem standardowym** szacunku lub **błędem średnim** albo — wyrażając ten błąd w jednostkach wartości oczekiwanej estymatora $E(t)$ — wartością współczynnika zmienności $V(t) = D(t) : E(t)$, zwanego także **względnym błędem standardowym** szacunku. Zwykle, mnożąc tę wartość przez 100 otrzymujemy **procentowy błąd standardowy** szacunku.

Uwaga 7.1. Precyzją szacunku mierzymy wpływ czynników losowych na ewentualny błąd oceny, natomiast nie jest brany pod uwagę wpływ błędów nielosowych na wynik oceny. Na **dokładność** szacunku składają się błędy nielosowe wraz z błędem losowym.

7.16. Gdy próba jest odpowiednio **duża** także rozkład estymatora t parametru T można aproksymować rozkładem normalnym, przy czym t jest estymatorem **nieobciążonym** parametru T, zachowanie się ocen szacujących ten parametr jest następujące:

a) w około 38 próbach na 100 bezwzględny błąd oceny nie przekroczy 0,5 $D(t)$,
b) w około 68 próbach na 100 bezwzględny błąd oceny nie przekroczy $D(t)$,
c) w około 87 próbach na 100 bezwzględny błąd oceny nie przekroczy 1,5 $D(t)$,
d) w około 954 próbach na 1000 bezwzględny błąd oceny nie przekroczy 2 $D(t)$,
e) w około 997 próbach na 1000 bezwzględny błąd oceny nie przekroczy 3 $D(t)$.

Uwaga 7.2. Nie ma naukowo jednoznacznej reguły, kiedy próba jest wystarczająco duża, aby można było rozkład estymatora traktować jako w przybliżeniu normalny, gdyż zależny jest od rodzaju parametru i rozkładu cech w populacji generalnej. Zwykle próbę poniżej 25 jednostek losowania (JL) uznaje się za „małą próbę", natomiast powyżej 100 JL — jako „dużą próbę". W przypadku losowania wielostopniowego chodzi o $JLPS$, a dla domeny — o $JLPS$ z nią związane.

7.17. W praktyce na ogół przyjmuje się, że 2 $D(t)$ jest „maksymalnym" błędem oceny z próby. Jest to konwencja dopuszczająca ryzyko 46 szans na 1000, że nastąpi jeszcze większy błąd oceny. Liczbę $\alpha = 46 : 1000 = 0{,}046$ nazywamy **współczynnikiem ufności,** a przedział $[t - 2\,D(t),\ t + 2\,D(t)] - 95{,}4\%$ **przedziałem ufności** dla T. Zauważmy, że w praktyce nie znamy dokładnej wartości odchylenia standardowego $D(t)$, które w określony sposób szacujemy z próby, obliczając długość półprzedziału ufności „2 $D(t)$". Dla konkretnej próby faktyczny błąd oceny t będzie przeważnie znacznie mniejszy niż 2 $D(t)$, co ilustruje przykład 4.4, podany w konspekcie wykładu 4 (w którym w miejsce $D(t)$ operuje się procentowym błędem standardowym szacunku $V(t)$.

7.18. W wykładach 5 i 6 przeanalizowany został schemat losowania prostego, przy czym z uwagi na większą efektywność ograniczono się do losowania prostego bez zwracania (*lpbz*). Dla realizacji losowania prostego niezbędny jest operat losowania, w którym wszystkie jednostki populacji generalnej zostały uporządkowane i dla

każdej z nich podane zostały cechy adresowe, umożliwiające identyfikację poszczególnych jednostek badania.

7.19. W *lpbz* każda jednostka badania (będąca *JL*) ma taką samą szansę wylosowania do próby, przy czym może być do niej wylosowana tylko jeden raz. Zakładamy, że liczebność populacji wynosi N jednostek, a do próby losujemy n jednostek. Jednostka o numerze k, $1 \leqslant k \leqslant N$, ma prawdopodobieństwo jej wylosowania do próby $P(k) \approx n:N$.

7.20. W celu oszacowania średniej wartości $\bar{Y}$ cechy Y, o ile nie dysponujemy żadnymi informacjami dodatkowymi, za estymator przyjmujemy średnią $\bar{y}$ z próby:

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \tag{1}$$

Przez y_i, $1 \leqslant i \leqslant n$, rozumiemy wartość cechy Y dla elementu populacji wylosowanego za i-tym razem.

Błąd standardowy estymatora $\bar{y}$, $D(\bar{y})$ wynosi ($\bar{y}$ jest nieobciążonym estymatorem średniej $\bar{Y}$):

$$D(\bar{y}) = \sqrt{1 - \frac{n}{N}} \frac{S}{\sqrt{n}} \tag{2}$$

gdzie:

$$S = \sqrt{S^2} \text{ oraz } S^2 = \frac{1}{N-1} \sum_{k=1}^{N} (Y_k - \bar{Y})^2 \tag{3}$$

oznaczają odpowiednio odchylenie standardowe oraz wariancję cechy Y w nr populacji.

7.21. Ze wzoru (2) wynika, że błąd standardowy estymatora średniej $\bar{Y}$ jest wprost proporcjonalny do odchylenia standardowego S cechy Y. Przy danych N i n wraz ze wzrostem S wzrasta proporcjonalnie $D(\bar{y})$. Występujący we wzorze (2) czynnik $\sqrt{1 - \frac{n}{N}}$ jest bliski 1, **o ile tylko próba nie obejmuje więcej niż 5% populacji generalnej.** Błąd standardowy przy danym S jest wówczas odwrotnie proporcjonalny do $\sqrt{n}$. Jeśli np. próbę zmniejszymy czterokrotnie, to błąd standardowy $D(\bar{y})$ wzrośnie dwukrotnie. Zauważmy, że w omawianym przypadku ($n \leqslant 0{,}05N$) wielkość populacji (N) nie wpływa na błąd standardowy $D(\bar{y})$, ściślej: wpływa tak nieznacznie, że ten wpływ możemy uznać za nie istniejący. Natomiast, przy próbach obejmujących **znaczną część populacji**, np. co najmniej 20%, czynnik $\sqrt{1 - \frac{n}{N}}$ maleje w sposób znaczący, w porównaniu z jednością. Dla przykładu, jeśli $n = 0{,}2\ N$, to $\sqrt{1 - \frac{n}{N}} = 0{,}89$, a dla próby 50% równa się 0,71; błąd standardowy $D(\bar{y})$ zmniejsza się jeszcze bardziej niż odwrotnie proporcjonalnie do wzrostu $\sqrt{n}$.

7.22. Planując badanie reprezentacyjne chcemy je ograniczyć do jak najmniejszej próby, ale gwarantującej — z nieznacznym ryzykiem — nieprzekroczenie założonej maksymalnej wielkości d błędu przy ocenie $\bar{Y}$. Przyjmując ryzyko 5%, odpowiadające przyrównaniu d do $2\,D(\bar{y})$, próba powinna objąć co najmniej n elementów:

$$n \approx \frac{N}{1+\frac{Nd^2}{4S^2}} \tag{4}$$

Do zastosowania wzoru (4) niezbędna jest wielkość S^2. W praktyce sprowadza się to do oceny wariancji S^2 badanej cechy Y. Możemy tego dokonać, opierając się na danych **wstępnej próby,** zawierającej około 50 jednostek badania $n' = 50$. Gdy badanie wstępne odpowiednio szybko nie może być przeprowadzone należy ocenę S^2 oprzeć na danych z przeszłości, a gdy takie nie istnieją bądź obawiamy się większych zmian w dyspersji cechy pozostaje mniej dokładna ocena oparta na wzorze:

$$S^2 \approx \frac{(Y_{max} - Y_{min})^2}{16} \tag{5}$$

gdzie: Y_{max}, Y_{min} oznaczają odpowiednio przypuszczalną największą i najmniejszą wartość badanej cechy. W przypadku gdy rozkład cechy Y jest podobny do rozkładu normalnego, w mianowniku (5) zamiast „16" należałoby przyjąć „36"; gdy rozkład cechy Y jest podobny do symetrycznego rozkładu trójkątnego mianownik wzoru (5) można zastąpić liczbą „24"; gdy przypuszczamy, że rozkład cechy Y jest zbliżony do rozkładu jednostajnego należałoby przyjąć w mianowniku (5) „12" zamiast „16".

7.23. Jeśli dysponujemy informacją dodatkową o wartości średniej $\bar{X}$ dodatkowej cechy mierzalnej X możemy tę informację wykorzystać, stosując estymator ilorazowy $\bar{y}_q$ bądź estymator regresyjny $\bar{y}_{lr}$. Estymator ilorazowy średniej $\bar{Y}$ definiujemy, jako:

$$\bar{y}_q = \bar{y}\,\frac{\bar{X}}{\bar{x}} = r\bar{X} \qquad r = \frac{\bar{y}}{\bar{x}} \tag{6}$$

Jest to estymator obciążony, zgodny. Jeśli próba jest tak **duża,** że współczynnik zmienności $V(\bar{x}) \leqslant 0{,}05$, czyli względny błąd standardowy szacunku średniej $\bar{X}$ z próby $V(\bar{x}) = D(\bar{x}) : \bar{X} \leqslant 0{,}05$, wówczas obciążenie estymatora $\bar{y}_q$ jest małe i można je pominąć oraz:

$$D^2(\bar{y}_q) \approx \left(1-\frac{n}{N}\right)\left[\frac{R^2S_x^2 + S_y^2 - 2\rho_{xy}RS_xS_y}{n}\right] \tag{7}$$

gdzie: $R = \frac{\bar{Y}}{\bar{X}}$, $\rho_{xy} = \frac{1}{(N-1)S_xS_y}\sum_{k=1}^{N}(X_k - \bar{X})(Y_k - \bar{Y})$

ρ_{xy} jest współczynnikiem korelacji cech X i Y w populacji generalnej.

Z porównania (7) i (2) wynika **twierdzenie:** jeśli $\rho_{xy} > \frac{1}{2}\frac{V(X)}{V(Y)}$ $V(X)$ jest współczynnikiem zmienności cechy X, $V(Y)$ — cechy Y w populacji generalnej, $V(X)=S_x:\bar{X}$, $V(Y)=S_y:\bar{Y}$), to $D^2(\bar{y}_q) < D^2(\bar{y})$.

Zatem, przy wysokiej korelacji cechy dodatkowej z cechą badaną estymator ilorazowy jest efektywniejszy od średniej z próby. Jeżeli jednak korelacja tych cech jest tak mała, że $\rho_{xy} < \frac{1}{2}\frac{V(X)}{V(Y)}$, średnia z próby $\bar{y}$ jest efektywniejsza od estymatora ilorazowego $\bar{y}_q$.

7.24. W wykładzie 6 omówiony został estymator regresyjny $\bar{y}_{lr}$:

$$\bar{y}_{lr}=\bar{y}+b(\bar{X}-\bar{x}), \qquad b=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \tag{8}$$

który jest obciążony, zgodny. Przy dużej próbie obciążenie można pominąć oraz:

$$D^2(\bar{y}_{lr}) \approx \left(1-\frac{n}{N}\right)\cdot\frac{(1-\rho_{xy}^2)S_y^2}{n} \tag{9}$$

Z porównania (9) i (2) wynika, że jeśli $\rho_{xy} \neq 0$, to $D^2(\bar{y}_{lr}) < D^2(\bar{y})$, a ponadto z porównania (7) i (9) wynika, że jeśli $\rho_{xy} \neq V(X):V(Y)$, to $D^2(\bar{y}_{lr}) < D^2(\bar{y}_q)$. Gdy $\rho_{xy}=0$, $D^2(\bar{y}_{lr})=D^2(\bar{y})$, a gdy $\rho_{xy}=V(X):V(Y)$ $D^2(\bar{y}_{lr})=D^2(\bar{y}_q)$. Pomimo tych pozytywnych własności estymatora $\bar{y}_{lr}$ w praktyce jest on rzadko stosowany z uwagi na stosunkowo skomplikowaną postać współczynnika b regresji z próby, który występuje w estymatorze $\bar{y}_{lr}$ zgodnie ze wzorem (8).

7.25. Jednym z celów badania reprezentacyjnego jest ocena parametrów w odniesieniu do określonych części populacji, będących dziedzinami studiów, czyli **domenami.** Dla przykładu, badając indywidualne gospodarstwa rolne, domeną mogą być gospodarstwa o najmniejszej powierzchni np. do 2 ha. Szacując średnią liczbę $\bar{Y}$ trzody chlewnej w domenie, gdy próba z całej populacji obejmuje n gospodarstw, natomiast frakcja gospodarstw domeny wynosi P_0, dla domeny próba jest około nP_0-elementowa, a więc znacznie mniejsza niż dla całej badanej populacji. Przykładowo próba obejmuje $n=2100$ gospodarstw oraz $P_0=0{,}3$, **dla domeny** próba wynosi $n'=630$ gospodarstw; pomijając w (2) czynnik $\sqrt{1-\frac{n}{N}}$ oraz oznaczając przez S_1 odchylenie standardowe cechy Y w domenie mamy dla niej:

$$D(\bar{y}') \approx \frac{S_1}{25.1}, \text{ zaś dla całej populacji } D(\bar{y})=\frac{S}{45{,}8} \tag{10}$$

7.26. Szacowanie wartości globalnej cechy mierzalnej Y odbywa się za pomocą estymatora $N\bar{y}$, $N\bar{y}_e$ bądź $N\bar{y}_{tr}$. Posługując się pojęciem **procentowego błędu standardowego** estymatora otrzymujemy $100V(N\bar{y}) = 100V(\bar{y})$. Procentowy błąd standardowy estymatora $N\bar{y}$ wartości globalnej Y jest taki sam, jak w przypadku estymatora średniej $\bar{Y}$.

7.27. Estymatorem wartości globalnej cechy Y dla domeny jest wyrażenie $N_{P_0}\bar{y}'$, gdzie p_0 jest frakcją domeny z próby, natomiast $\bar{y}'$ jest średnią wartością cechy Y z próby ograniczonej do domeny. Procentowy błąd standardowy estymatora wynosi w przybliżeniu:

$$100V(N_{P_0}\bar{y}') \approx 100\sqrt{1-\frac{n}{N}} \cdot \frac{\sqrt{V'(Y)+(1-P_0)}}{\sqrt{nP_0}} \tag{11}$$

przy czym $V'(Y)$ oznacza wartość współczynnika zmienności cechy Y w domenie, zaś p_0 — frakcję jednostek, tworzących domenę w populacji. Zwróćmy uwagę, że w przypadku estymatora $N\bar{y}$ wartości globalnej X dla całej populacji:

$$100V(N\bar{y}) = 100 \cdot \sqrt{1-\frac{n}{N}} \cdot \frac{V(Y)}{\sqrt{n}} \tag{11a}$$

$V(Y) = S_y : \bar{Y}$ [przy tym $100V(N\bar{y}) = 100V(\bar{y})$].

7.28. Szacując iloraz $R = \frac{\bar{Y}}{\bar{X}} = \frac{Y}{X}$ przyjmujemy za estymator statystykę $r\left(r = \frac{\bar{y}}{\bar{x}}\right)$. Jest to estymator obciążony, zgodny. Jeśli próba jest tak duża, że $V(\bar{X}) \leqslant 0{,}05$, obciążenie estymatora można pominąć, jako nieistotne oraz:

$$D^2(r) \approx R^2\left(1-\frac{n}{N}\right) \cdot \left[\frac{V^2(X)+V^2(Y)-2\rho_{xy}V(X)\,V(Y)}{n}\right] \tag{12}$$

7.29. Nieobciążonym estymatorem frakcji P jednostek wyróżnionych (ze względu na posiadanie określonego wariantu cechy Y niemierzalnej) jest frakcja p jednostek wyróżnionych w próbie. Błąd standardowy frakcji p wynosi w przybliżeniu:

$$D(p) \approx \sqrt{1-\frac{n}{N}} \cdot \frac{\sqrt{1-P}}{nP} \tag{13}$$

natomiast procentowy błąd standardowy:

$$100V(p) = 100V(Np) \approx 100\sqrt{1-\frac{n}{N}} \cdot \sqrt{\frac{1-P}{P}}\,\frac{1}{\sqrt{n}} \tag{14}$$

Wykazano (zob. pkt 5.10), że wyrażenie $\sqrt{\frac{1-P}{P}}$ rośnie, gdy P maleje. Dla $P=0{,}1$ jest ono równe 3, natomiast dla $P=0{,}005$ wzrasta do 14,11. Szacowanie małych frakcji wymaga bardzo dużych prób. Jeśli np. w populacji $N=100000$ chcemy oszacować frakcję rzędu 0,005 z procentowym błędem standardowym $100V(p)\approx 10\%$, próba powinna objąć około $n=16604$ elementy populacji.

7.30. Szacowanie liczby N' jednostek wyróżnionych ($N'=NP$) odbywa się za pomocą statystyki Np, której błąd standardowy jest N razy większy od prawej strony wzoru (13), a procentowy błąd standardowy jest taki sam, jak w przypadku szacowania frakcji P, określony w przybliżeniu wzorem (14).

7.31. W przypadku szacowania frakcji P'_0 jednostek wyróżnionych w domenie obejmującej P_0 jednostek badania posługujemy się, jako estymatorem, frakcją p'_0 z próby w domenie. Procentowy błąd standardowy estymatora wynosi w przybliżeniu:

$$100V(p'_0)\approx 100\sqrt{1-\frac{n}{N}}\cdot\frac{\sqrt{\frac{1-P'_0}{P'_0}}}{\sqrt{nP_0}} \qquad (15)$$

7.32. Wzory na wariancję bądź błąd standardowy (procentowy błąd standardowy) estymatora pozwalają zorientować się, jak wielka powinna być próba, aby przy szacowaniu danego parametru nie pomylić się więcej niż o z góry zadaną wielkość, przy tym bierze się pod uwagę jedynie tzw. **błąd losowy**, natomiast abstrahuje od możliwych błędów nielosowych. Dotąd przeanalizowany został jedynie schemat losowania prostego bez zwracania (*lpbz*). W następnych wykładach przeanalizujemy efektywność innych schematów losowania próby w porównaniu z *lpbz*.

7.33. **Zadania.** 1. W mieście zamieszkuje 5426 kobiet. Należy oszacować liczbę Y dzieci urodzonych przez te kobiety z próby w *lpbz*. Jak wielka powinna być próba, aby procentowy błąd standardowy szacunku wyniósł około 3%, o ile wstępna próba $n'=50$ kobiet dała wynik:

Liczba dzieci urodzonych przez kobietę	0	1	2	3	4	5	6
Liczba kobiet	5	11	15	10	5	3	1

2. Z populacji, obejmującej $N=20000$ jednostek wylosowano (*lpbz*) wstępną próbę $n'=60$ jednostek, dla których wartości cech X i Y wyniosły, jak niżej:

i	x_i	y_i	i	x_i	y_i	i	x_i	y_i	i	x_i	y_i
1	8	7	16	6	5	31	28	34	46	25	34
2	14	17	17	22	25	32	7	12	47	15	13
3	21	27	18	4	5	33	11	8	48	19	8
4	11	4	19	13	13	34	12	9	49	10	13
5	9	12	20	15	11	35	6	3	50	13	10
6	15	16	21	18	19	36	23	21	51	15	15
7	18	10	22	9	3	37	14	13	52	16	29
8	11	17	23	18	25	38	13	18	53	14	11
9	24	31	24	16	20	39	8	13	54	20	35
10	3	3	25	5	7	40	17	18	55	11	12
11	16	17	26	10	8	41	12	16	56	10	10
12	12	12	27	10	12	42	14	22	57	19	24
13	13	17	28	12	13	43	21	17	58	10	22
14	20	23	29	14	7	44	6	12	59	17	12
15	12	13	30	17	24	45	11	11	60	16	10

(A) Oszacować — na podstawie tej próby — minimalną liczebność próby, tak aby: a) przy szacowaniu średniej $\bar{X}$ nie pomylić się więcej niż o $d=0{,}5$; b) przy szacowaniu średniej $\bar{Y}$ nie pomylić się więcej niż o $d=0{,}7$; za estymator średniej przyjmujemy średnią z próby.

(B) Oszacować wariancję estymatora ilorazowego $\bar{y}_q$ przy założeniu, że próba obejmuje $n=1500$ jednostek oraz $\bar{X}=14{,}5$.

(C) Oszacować wariancję estymatora regresyjnego $\bar{y}_{lr}$ przy takim samym założeniu.

(D) Jaka powinna być minimalna liczebność próby, aby stosując estymator regresyjny $\bar{y}_{lr}$ nie pomylić się przy szacowaniu średniej $\bar{Y}$ więcej niż o $d=0{,}7$, jeżeli wiadomo, że $\bar{X}=14{,}5$?

Wskazówki — do zad. 1) — skorzystać ze stwierdzenia 7.26 oraz ocenić $\bar{Y}$ i S^2 ze wstępnej próby;
— do zad. 2) — przyjąć $d=2D(.)$ oraz ocenić S_x^2, S_y^2, ρ_{xy} z danych wstępnej próby.

7.34. **LITERATURA UZUPEŁNIAJĄCA:** R. Zasępa [18] str. 11—113, [9] str. 7—28, 32—37, 42—44, [20] str. 6—19, J. Steczkowski [14] str. 33—48, 50—54, 56—57, 124—173.

Wykład 8. Indywidualne losowanie warstwowe

8.1. Przy omawianu schematu losowania prostego bez zwracania (*lpbz*) wykazana została możliwość polepszenia precyzji szacunku średniej $\bar{Y}$ (wartości globalnej Y) przez wykorzystanie posiadanej dla populacji generalnej informacji o wartości średniej $\bar{X}$ (wartości globalnej X) cechy dodatkowej X, skorelowanej z cechą badaną Y. Poprawę precyzji szacunku uzyskujemy konstruując estymator ilorazowy bądź

regresyjny szacowanego parametru, pod warunkiem że próba jest dostatecznie duża. Obecnie przeanalizujemy jeden z możliwych sposobów podwyższenia precyzji szacunku przez zastosowanie bardziej efektywnego schematu losowania próby.

8.2. W przypadku *lpbz* każdy *n*-elementowy podzbiór jednostek badania ma takie samo prawdopodobieństwo wylosowania jako próby. Rzecz jasna, pewne *n*-elementowe części populacji będą niereprezentatywne jako próby. Pożądane jest eliminowanie z możliwości wyboru jak największej liczby takich prób. W tym kierunku prowadzi stosowanie schematu **losowania warstwowego.**

8.3. Warstwowaniem nazywamy podział badanej populacji na **rozłączne części** zwane **warstwami.** W przypadku losowania warstwowego próbę losujemy niezależnie z każdej warstwy. Operat losowania musi się więc składać z cząstkowych (warstwowych) operatów losowania. Podkreślamy, że losowanie z warstw nie musi być prowadzone według tego samego schematu losowania. Np. w części warstw losowanie może być indywidualne, natomiast w innej części może być losowaniem zespołowym albo losowaniem wielostopniowym. **W niniejszym wykładzie zakładamy, że losowanie próby z każdej warstwy jest** *lpbz*. Realizacja losowania próby wymaga wówczas, aby w operacie losowania dowolnej warstwy były wyszczególnione wszystkie jednostki badania wchodzące do tej warstwy i aby były one uporządkowane w ciąg, czyli każdej z tych jednostek był przyporządkowany numer k zgodnie z tym uporządkowaniem oraz cechy adresowe, identyfikujące odnośną jednostkę badania.

8.4. Przyjmujemy oznaczenia: a) populacja generalna obejmuje N jednostek badania, które są jednostkami losowania (JL), b) populację podzielono na L warstw, które numerujemy kolejno $h=1, 2, ..., L$, c) w h-tej warstwie znajduje się N_h jednostek badania, czyli $\sum_{h=1}^{L} N_h = L$, d) każda jednostka badania (JL) nosi pewien numer k, $1 \leqslant k \leqslant N_h$, czyli para (h, k) jest formalnie identyfikatorem danej jednostki badania w populacji, gdzie $1 \leqslant h \leqslant L$, e) wartość cechy Y dla elementu (h, k) oznaczymy Y_{hk}.

8.5. Średnią wartość cechy Y w h-tej **warstwie** oznaczymy symbolem $\bar{Y}_h$:

$$\bar{Y}_h = \frac{1}{N_h} \sum_{k=1}^{N_h} Y_{hk} \tag{1}$$

natomiast wariancję S_{hy}^2 lub S_h^2 wyznaczymy z wzoru:

$$S_h^2 = \frac{1}{N_h - 1} \sum_{k=1}^{N_h} (Y_{hk} - \bar{Y}_h)^2 \text{ bądź } \sigma_h^2 = \frac{N_h - 1}{N_h} S_h^2 \tag{2}$$

Średnia $\bar{Y}$:

$$\bar{Y} = \frac{1}{N} \sum_{h=1}^{L} \sum_{k=1}^{N_h} Y_{hk} = \sum_{h=1}^{L} W_h \bar{Y}^h, \; W_h = \frac{N_h}{N} \tag{3}$$

W_h jest **frakcją elementów** h-tej **warstwy.**

8.6. Niech $\bar{y}_h$ będzie średnią z próby pochodzącej z h-tej warstwy:

$$\bar{y}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} y_{hi} \tag{4}$$

gdzie n_h oznacza liczebność próby z h-tej warstwy. Można wykazać, że w klasie liniowych nieobciążonych estymatorów średniej $\bar{Y}$ estymatorem o najmniejszej wariancji jest $\bar{y}_{(w)}$:

$$\bar{y}_{(w)} = \sum_{h=1}^{L} W_h \bar{y}_h \tag{5}$$

Estymator ten jest nieobciążony, a jego wariancja wynosi:

$$D^2(\bar{y}_{(w)}) = \sum_{h=1}^{L} W_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{S_h^2}{n_h} \tag{6}$$

8.7. Rozdział próby pomiędzy warstwy jest zależny od celów badania reprezentacyjnego. Jeśli koszty jednostkowe zbierania informacji są różne w różnych warstwach i na całkowity koszt badania składają się: a) koszt K_o planowania i organizacji badania wraz z przygotowaniem operatu losowania oraz wylosowaniem próby, b) koszty jednostkowe K_h zebrania informacji oraz jej opracowania — różne w różnych warstwach, to całkowity koszt badania reprezentacyjnego wyniesie:

$$K = K_o + \sum_{h=1}^{L} K_h n_h \tag{7}$$

Rozdział próby między warstwy $(n_1, n_2, ..., n_L)$ taki, aby przy całkowitym koszcie badania K wariancja $D^2(\bar{y}_{(w)})$ była jak najmniejsza, nazwiemy **optymalnym.** Optymalny rozdział próby otrzymujemy, gdy:

$$n_h = \frac{K - K_o}{\sum_{h=1}^{L} \sqrt{K_h} W_h S_h} \cdot \frac{W_h S_h}{\sqrt{K_h}} \qquad h = 1, 2, ..., L \tag{8}$$

Schemat losowania warstwowego, w którym rozdział próby pomiędzy warstwy odbywa się według (8), nazywamy **optymalnym schematem** losowania warstwowego. Wówczas:

$$D^2_{opt.}(\bar{y}_{(w)}) = \frac{1}{K - K_o} \left(\sum_{h=1}^{L} \sqrt{K_h} W_h S_h \right)^2 - \frac{1}{N} \sum_{h=1}^{L} W_h S_h^2. \tag{9}$$

8.8. Rozdział **próby *n*-elementowej** pomiędzy warstwy, nie biorący pod uwagę kosztów badania, taki, żeby wariancja $D^2(\bar{y}_{(w)})$ była jak najmniejsza uzyskujemy, przyjmując:

$$n_h = n\frac{N_h S_h}{\sum_{h=1}^{L} N_h S_h} = n\frac{W_h S_h}{\sum_{h=1}^{L} W_h S_h} \quad \text{dla } h=1, 2, ..., L \tag{10}$$

Wówczas:

$$D^2_{Neym.}(\bar{y}_{(w)}) = \frac{1}{N^2}\left[\frac{1}{n}\left(\sum_{h=1}^{L} N_h S_h\right)^2 - \sum_{h=1}^{L} N_h S_h^2\right] \tag{11}$$

Rozdzielając *n*-elementową próbę między warstwy zgodnie z (10) schemat losowania warstwowego nazywamy **optymalnym schematem Neymana.**

8.9. Zgodnie z (5) estymator $\bar{y}_{(w)}$ wymaga, w przypadku wyżej omówionych sposobów rozdziału próby, osobnego opracowania zebranych informacji w każdej warstwie, obliczenia średnich $\bar{y}_h$ oraz przemnożenia ich przez pewne wagi W_h i zsumowania. Ponadto, dla zastosowania wzoru (8) albo (10) niezbędna jest znajomość przynajmniej przybliżona odchyleń standardowych S_h bądź równoważnie, wariancji S_h^2 cechy Y w poszczególnych warstwach. Jak szacować wariancję wiemy z wykładu 5 (zob. pkt 5.8). Należy jeszcze wziąć pod uwagę, że w praktyce szacujemy bardzo dużą liczbę parametrów. Jeśli badane cechy są ze sobą skorelowane wystarczy wziąć pod uwagę kilka cech najważniejszych i ustalając dla każdej z nich wielkość próby z każdej warstwy te wielkości uśrednić bądź uwzględnić tylko rozdział próby według jednej najważniejszej cechy, z którą najbardziej są skorelowane pozostałe badane cechy. J. Greń postulował, aby ustalać liczby nieujemne $n_1, n_2, ..., n_L$, które minimalizują sumę $\sum_{h=1}^{L} n_h$, przy warunkach $D^2(\bar{y}_{(w)}^{(r)}) \leqslant d_r^2$, dla $r=1, 2, ..., p$ (bierzemy pod uwagę p cech). Rozwiązanie tego zagadnienia wymaga stosowania tzw. metody rzutowanego gradientu. Zastosowanie tej metody jest kłopotliwe przy większej liczbie szacowanych średnich. W Ośrodku Elektronicznym GUS został opracowany w języku Algol program na EMC opisany w [16] na str. 93—101.

8.10. Gdy przypuszczamy, że odchylenia standardowe S_h w warstwach dla cechy Y nie różnią się znacznie między sobą, czyli możemy przyjąć $S_h = const.$ dla $h=1, 2, ..., L$ wzór (10) upraszcza się. Rozdział próby pomiędzy warstwy możemy realizować zgodnie ze wzorem:

$$n_h = nW_h, \quad \frac{n_h}{N_h} = \frac{n}{N} = const. \quad \text{dla } h=1, 2, ..., L \tag{12}$$

czyli, próba z każdej warstwy powinna obejmować taką samą frakcję jednostek

badania. Np. jeśli łączna próba ma objąć 5% populacji, to także w każdej warstwie próba jest pięcioprocentowa. Stosując rozdział próby pomiędzy warstwy zgodnie z (12) losowanie próby nazywamy **schematem proporcjonalnego losowania warstwowego.** W tym schemacie każda jednostka badania ma takie samo prawdopodobieństwo wylosowania do próby. Wówczas mówimy, że **próba jest automatycznie wyważona.** Zwróćmy uwagę, że w przypadku *lpbz* również próba jest automatycznie wyważona.

Łatwo zauważyć, że stosując schemat proporcjonalnego losowania warstwowego estymator $\bar{y}_{(w)}$ uprości się, jako:

$$\bar{y}_{(w)} = \frac{1}{n} \sum_{h=1}^{L} \sum_{i=1}^{n_h} y_{hi} = \bar{y} \tag{13}$$

oraz:

$$D^2_{prop.}(\bar{y}_{(w)}) = D^2_{prop.}(\bar{y}) = \left(1 - \frac{n}{N}\right) \frac{\sum_{h=1}^{L} W_h S_h^2}{n} = \left(1 - \frac{n}{N}\right) \frac{S_w^2}{n} \tag{14}$$

przy czym S_w^2:

$$S_w^2 = \sum_{h=1}^{L} W_h S_h^2 \tag{15}$$

jest **przeciętną wariancją wewnątrzwarstwową.**

8.11. Jeśli współczynniki zmienności $V(Y_h)$ cechy Y różnią się w warstwach nieznacznie, wówczas optymalny schemat Neymana przybiera postać:

$$n_h = \frac{n\, Y_h}{Y} \quad \text{dla } h = 1, 2, ..., L \tag{16}$$

przy czym Y_h, Y oznaczają globalną wartość cechy Y w h-tej warstwie oraz odpowiednio w całej populacji. Ponieważ wartości Y_h, Y nie są znane w czasie planowania badania, zamiast cechy Y przyjmujemy pewną cechę dodatkową, skorelowaną z cechą Y, dla której posiadamy informacje z przeszłości.

Przykład 8.1. Szacujemy średnią $\bar{Y}$ stosując indywidualne losowanie warstwowe z populacji obejmującej $N = 12$ tys. jednostek badania. Przeznaczamy na badanie $K = 2$ miliony zł, przy czym koszty stałe przypuszczalnie wyniosą $K_o = 200$ tys. zł. Populacja została podzielona na $L = 5$ warstw. Liczebności N_h, oszacowane odchylenia standardowe cechy Y, S_h oraz koszty jednostkowe K_h ($h = 1, 2, ..., 5$) podajemy dalej. Ustalić rozdział próby pomiędzy warstwy: a) optymalny, b) Neymana, c) proporcjonalny.

h	1	2	3	4	5	Razem
N_h	1000	3000	5000	2000	1000	12000
W_h	0,0833	0,25	0,4167	0,1667	0,0833	1,0000
S_h	1	1,5	2,5	3	2	×
K_h złotych	200	400	900	1225	1600	×
$W_h S_h = A_h$	0,0833	0,3750	1,0418	0,5001	0,1666	2,1668
$\sqrt{K_h} W_h S_h$	1,1780	7,5000	31,2540	17,5035	6,6640	64,0995
$W_h S_h / \sqrt{K_h}$	0,005890	0,018750	0,034727	0,014289	0,004165	×
$n_h - opt.$	165	527	975	401	117	2185
$A_h K_h$	16,66	150,00	937,58	612,63	266,56	1983,43
n'_h — Neymana	76	340	945	454	151	1966
$W_h K_h$	16,66	100,00	375,03	204,21	133,28	829,18
n''_h — prop.	181	543	904	362	181	2171

Obliczenia dla optymalnego rozdziału próby prowadzimy według wzoru (8). Iloraz:

$$\frac{K - K_0}{\sum_{h=1}^{L} \sqrt{K_h} W_h S_h} = \frac{1800000}{64{,}0995} = 28081$$

Mnożąc przez tę liczbę dane z wiersza $W_h S_h / K_h$ otrzymujemy, że z pierwszej warstwy należy wylosować 165 jednostek, z drugiej 527 jednostek itd. Łączna próba wyniesie 2185 jednostek. Łatwo sprawdzić, że $\sum_{h=1}^{L} K_h n_h = 1799725$, a więc mieści się w założonych wydatkach na badanie.

W przypadku rozdziału próby według zasady Neymana należy określić wpierw wielkość tej próby. Określamy ją pamiętając, że liczebność próby z h-tej warstwy jest wprost proporcjonalna do $W_h S_h = A_h$. Gdyby przyjąć, że współczynnik proporcjonalności $n \Big/ \sum_{h=1}^{L} W_h S_h = 1$ koszt zebrania informacji wyniósłby $\sum_{h=1}^{h} A_h K_h = 1983{,}43$. W rzeczywistości powinien on wynieść 1,8 mln zł, a więc $1983{,}43 \cdot \lambda = 1{,}8$ mln. Stąd $\lambda = 907{,}5$[1]), czyli $n_h = 907{,}5\, A_h$. Wówczas z pierwszej warstwy należy wylosować 76 jednostek badania, z drugiej 340 jednostek itd. Łączna próba wyniesie 1966 jednostek, a $\sum_{h=1}^{L} K_h n'_h = 1799450$ zł.

Stosując proporcjonalny rozdział próby według wzoru (12) należy również ustalić n'', zgodnie z warunkami kosztów. Gdyby przyjąć wielkość próby z h-tej warstwy równą W_h koszt zebrania materiałów wyniósłby $\sum_{h=1}^{L} W_h K_h = 829{,}18$. Aby otrzymać

[1]) $\lambda \sum_{h=1}^{L} A_h = 907{,}5 \cdot 21668 = 1966$ jest wielkością próby n'.

koszt około 1,8 mln zł, liczebność próby należy odpowiednio powiększyć, zgodnie z formułą 829,18 $\lambda' = 1800000$. Otrzymujemy $\lambda' = 2170{,}8$, czyli wielkość próby łącznej zaokrąglona do liczby całkowitej wynosi $n'' = 2171$ jednostek badania. Liczebność prób z poszczególnych warstw otrzymujemy mnożąc W_h przez n''. Z pierwszej warstwy należy wówczas wylosować 181 jednostek badania, z drugiej 543 jednostki itd. $\sum_{h=1}^{L} K_h n_h'' = 1800050$ zł.

Obliczmy, jak będą się różnić wariancje estymatora $\bar{y}_{(w)}$ średniej w każdym z trzech wymienionych schematów rozdziału próby między warstwy. Stosując rozdział optymalny, uwzględniający koszty, otrzymujemy zgodnie ze wzorem (9):

$$D^2_{opt.}(\bar{y}_{(w)}) = \frac{1}{1800000}(64{,}0995)^2 - \frac{1}{12000} \cdot 5{,}0837 = 0{,}001859$$

$$D_{opt.}(\bar{y}_{(w)}) = 0{,}0431$$

Stosując optymalny rozdział próby między warstwy Neymana, zgodnie z (11), po włączeniu czynnika stałego $1/N$ do wyrazów w nawiasie kwadratowym, otrzymujemy:

$$D^2_{Neym.}(\bar{y}_{(w)}) = \frac{1}{n'}\left(\sum_{h=1}^{5} W_h S_h\right)^2 - \frac{1}{N}\sum_{h=1}^{5} W_h S_h^2 = \frac{1}{1966}(2{,}1668)^2 - \frac{1}{12000} \cdot 5{,}0837 = 0{,}001964$$

$$D_{Neym.}(\bar{y}_{(w)}) = 0{,}0443$$

Natomiast proporcjonalny rozdział próby między warstwy daje zgodnie z (14):

$$D^2_{prop.}(\bar{y}_{(w)}) = \left(1 - \frac{2171}{12000}\right)\frac{5{,}0837}{2171} = 0{,}001918, \quad D_{prop.}(\bar{y}_{(w)}) = 0{,}0438$$

Z porównania błędów standardowych wynika, że różnice w wielkości błędu standardowego wypadły w niniejszym przykładzie stosunkowo małe. Nie jest to jednak regułą.

Uwaga 8.1. Jak należało się spodziewać, najmniejszy błąd standardowy szacunku uzyskujemy, stosując optymalny rozdział próby między warstwy, uwzględniający różnice w kosztach jednostkowych K_h zebrania materiału. Nie powinien budzić zdziwienia fakt, że $D^2_{Neym.}(\bar{y}_{(w)}) > D^2_{prop.}(\bar{y}_{(w)})$. Zależność $D^2_{Neym.}(\bar{y}_{(w)}) < D^2_{prop.}(\bar{y}_{(w)})$ występuje, gdy liczebność próby jest **taka sama.** Przyjmując $n'' = 1966$ otrzymamy $D^2_{prop.}(\bar{y}_{(w)}) = 0{,}002162$ oraz $D_{prop.}(\bar{y}_{(w)}) = 0{,}04650 > D_{Neym.}(\bar{y}_{(w)}) = 0{,}0443$. Proporcjonalny rozdział próby między warstwy byłby następujący: $n_1 = 164$, $n_2 = 491$, $n_3 = 819$, $n_4 = 328$ oraz $n_5 = 164$.

8.12. Jak dzielić populację na warstwy? Zarówno intuicja, jak i analiza wzoru (6), podającego wielkość wariancji $D^2(\bar{y}_{(w)})$ wskazują, że należy przy tym podziale dążyć, aby warstwy były **wewnętrznie** jak najbardziej **jednorodne,** natomiast aby **różniły się** między sobą pod względem **wartości średnich** $\bar{Y}_h$ cechy Y. Do warstwowania populacji należy wykorzystać wszelkie posiadane informacje, np. dane z poprzednich badań

populacji. Nie wymaga się żadnej jednolitości informacji. Jeśli brak jest pewnych informacji dla jakiejś frakcji populacji możemy te jednostki wydzielić w osobną warstwę. O ile to możliwe — ze względu na koszty operatu losowania — należy wydzielać w osobne warstwy części populacji, stanowiące **domeny** (dziedziny studiów). Jeżeli dla celów warstwowania można wykorzystać dane dotyczące różnych cech, należy preferować cechy między sobą mało skorelowane, natomiast silnie skorelowane z cechami badanymi. Jeśli możemy wykorzystać informacje o dwóch cechach: jednej mierzalnej, a drugiej niemierzalnej, należy za cechę warstwującą przyjąć cechę niemierzalną. Informacje o cesze mierzalnej należy wykorzystać bądź do głębszego powarstwowania, bądź przy konstrukcji efektywniejszego estymatora, np. stosując estymator ilorazowy zamiast estymatora prostego.

Dzieląc populację na warstwy należy mieć na uwadze, że losowanie próby z warstwy wymaga odpowiedniego operatu losowania. Koszt przygotowania takiego operatu wzrośnie z liczbą warstw, a więc zbyt duża liczba warstw może być w praktyce niepożądana. Ponadto, warstwy powinny być dostatecznie duże, żeby próba z każdej z nich obejmowała co najmniej 2 jednostki badania. Stosując losowanie proporcjonalne należy dążyć, aby w próbie warstwy miały liczebności n_h dokładnie równe $n W_h$, szczególnie dla stosunkowo małych warstw, gdyż w przypadku zaokrągleń próba nie jest dobrze wyważona.

Przykład 8.2. Planując reprezentacyjny spis rolny przeprowadzony w czerwcu 1989 r. wzięto pod uwagę postulat, że domenami będą gminy, dla których jest potrzebna informacja o zasiewach i zwierzętach w indywidualnych gospodarstwach rolnych. Zatem gminy stanowiły warstwy. Wobec posiadania informacji o powierzchni gospodarstw w 1987 r. **dla każdego** gospodarstwa rolnego **w rejestrze komputerowym**, postanowiono wykorzystać dane o powierzchni ogólnej gospodarstwa dla ustawienia indywidualnych gospodarstw gminy w ciąg o malejącej powierzchni (od gospodarstw największych do najmniejszych) i utworzenia z kolejnych 40 gospodarstw warstw. Z każdej warstwy losowano do próby 2 gospodarstwa, czyli próba warstwowa losowana proporcjonalnie objęła 5% gospodarstw. Informacje adresowe wylosowanych gospodarstw zostały przekazane rachmistrzom spisowym.

Ponieważ gospodarstwa o podobnej powierzchni mniej się między sobą różnią niż gospodarstwa o odmiennej powierzchni ogólnej, można się spodziewać, że dokonana stratyfikacja jest właściwa dla oszacowania powierzchni zasiewów, natomiast może nie być taką przy szacowaniu pogłowia zwierząt gospodarskich. Dlatego stosowano złożone estymatory (estymatory ilorazowe), wykorzystujące informacje ze spisu 1988 r.

8.13. Podobnie, jak w przypadku *lpbz* także w losowaniu warstwowym wykorzystanie pomocniczej informacji może powodować wzrost efektywności estymatora. Pomijając estymator regresyjny opiszemy sposób zastosowania estymacji ilorazowej.

Jeżeli posiadamy informacje o wartości średniej cechy dodatkowej X dla każdej z L warstw, na które została podzielona badana populacja, czyli znamy $\bar{X}_h$ dla $h = 1, 2, ..., L$ oraz cecha X, jak się spodziewamy, jest skorelowana z badaną cechą Y, możemy skonstruować dwa rodzaje estymatorów średniej $\bar{Y}$:

1) konstruując **rozdzielony**[1]) **estymator ilorazowy** $\bar{y}_{q(w)}$

$$\bar{y}_{q(w)} = \sum_{h=1}^{L} W_h \frac{\bar{y}_h}{\bar{x}_h} \bar{X}_h \tag{17}$$

2) konstruując **łączny**[2]) **estymator ilorazowy** $\bar{y}'_{q(w)}$

$$\bar{y}'_{q(w)} = \frac{\sum_{h=1}^{L} W_h \bar{y}_h}{\sum_{h=1}^{L} W_h \bar{x}_h} \cdot \bar{X} \tag{18}$$

Obydwa estymatory są estymatorami obciążonymi, zgodnymi parametru $\bar{Y}$. Jeśli próba obejmuje stosunkowo dużo jednostek badania z każdej warstwy, powiedzmy 50 lub więcej, obciążenie estymatora $\bar{y}_{q(w)}$ jest praktycznie bez znaczenia i można je pominąć. Wariancja estymatora $D^2(\bar{y}_{q(w)})$ wynosi w przybliżeniu:

$$D^2(\bar{y}_{q(w)}) \approx \sum_{h=1}^{L} \left(1 - \frac{n_h}{N_h}\right) W_h^2 \left[\frac{R_h^2 S_{hx}^2 + S_{hy}^2 - 2\rho_{hxy} R_h S_{hx} S_{hy}}{n_h}\right] \tag{19}$$

gdzie: $R_h = \frac{\bar{Y}_h}{\bar{X}_h}$, $\quad \rho_{hxy} = \frac{1}{(N_h - 1) S_{hx} S_{hy}} \sum_{k=1}^{N_h} (X_{hk} - \bar{X}_h)(Y_{hk} - \bar{Y}_h)$ (20)

Jeśli natomiast z poszczególnych warstw losujemy do próby po kilkanaście elementów, obciążenie estymatora $\bar{y}_{q(w)}$ może być istotne i wówczas należy raczej skorzystać z estymatora $\bar{y}'_{q(w)}$.

Jeżeli próba n jest dostatecznie duża, obciążenie łącznego estymatora ilorazowego możemy uznać za praktycznie równe zeru.
Wówczas:

$$D^2(\bar{y}'_{q(w)}) \approx \sum_{h=1}^{L} \left(1 - \frac{n_h}{N_h}\right) W_h^2 \left[\frac{R^2 S_{hx}^2 + S_{hy}^2 - 2\rho_{hxy} R S_{hx} S_{hy}}{n_h}\right] \tag{21}$$

gdzie: $R = \frac{\bar{Y}}{\bar{X}}$

8.14. Doświadczenie wskazuje, że prawe strony (19) i (21) na ogół mało się między sobą różnią. Dlatego najczęściej stosuje się łączny estymator ilorazowy $\bar{y}'_{q(w)}$. Jeśli R_h

[1]) Ten termin uważam za lepiej oddający ideę estymatora niż dotąd użyty w moich opracowaniach termin „estymator ilorazowy prosty".
[2]) Dawniej nazywałem go terminem estymator „ilorazowy złożony".

różnią się znacznie między sobą oraz próba z każdej warstwy jest stosunkowo duża, należy stosować rozdzielony estymator ilorazowy $\bar{y}_{q(w)}$.

8.15. Kiedy teoretycznie możliwy i efektywny podział populacji na warstwy, o znanych liczebnościach, nie może być zrealizowany z powodu zbyt wysokich kosztów sporządzenia operatu losowania i w rezultacie stosujemy nieograniczony schemat *lpbz*, podział populacji na warstwy może być wykonany na wylosowanej próbie. Takie postępowanie, tj. warstwowanie po wylosowaniu próby nazywamy **poststratyfikacją.**

Przypuśćmy, że szacujemy średnią $\bar{Y}$ i elementy **próby** podzieliliśmy na G warstw, dla których znamy liczebności w (populacji generalnej) N_h, $h=1, 2, ..., G$, a więc także frakcje $W_h=N_h/N$. Estymatorem średniej $\bar{Y}$ niech będzie:

$$\bar{y}'=\sum_{h=1}^{G} W_h \bar{y}_h \tag{22}$$

Formalnie $\bar{y}'$ jest identyczny z $\bar{y}_{(w)}$, lecz próba jest obecnie losowana, zgodnie z *lpbz*, z całej populacji. Okazuje się, że gdy przeciętna liczebność $\bar{n}=n/G$ próby z warstwy będzie dostatecznie duża, np. większa od 20, to wariancja $D^2(\bar{y}')$ będzie nieznacznie (mniej niż o 5%) przewyższać wariancję $D^2_{prop.}(\bar{y}_{(w)})$, w przypadku proporcjonalnego losowania warstwowego (a więc warstwowania przed losowaniem próby).

Poststratyfikację należałoby stosować szczególnie w przypadku badań reprezentacyjnych, towarzyszących badaniom spisowym, jak np. przy szacowaniu parametrów badania kobiet w NSP 1988 r. Wylosowana próba około 100 tys. kobiet, kiedy warstwami były województwa w podziale na miasta i wieś, może być dodatkowo powarstwowana i o ile zabezpieczone będą potrzebne informacje ze spisu pełnego, dotyczące liczebności podwarstw mierzonej liczbą kobiet objętych badaniem dzietności, można będzie uzyskać bardziej precyzyjne szacunki niż bez uwzględnienia tej dodatkowej poststratyfikacji.

W praktyce międzynarodowej dość często stosuje się odpowiednią poststratyfikację w badaniach metodą reprezentacyjną.

8.16. Planując badania reprezentacyjne chcielibyśmy wiedzieć, jak wielką próbę ($n=?$) należy przyjąć, aby z pewnym ryzykiem błąd oceny nie przekraczał wielkości d. Przyjmując ryzyko bliskie 0,046 należy wówczas przyrównać $2D\bar{y}_{(w)}$ do d. Otrzymujemy:

$$n\approx\frac{\sum_{h=1}^{L}\frac{W_h^2 S_h^2}{w_h}}{\frac{d^2}{4}+\frac{1}{N}\sum_{h=1}^{k} W_h S_h^2}, \qquad w_h=\frac{n_h}{n} \tag{23}$$

Zależnie od tego, jaki przyjmujemy schemat rozdziału populacji między próby (wzory: 8, 10, 12, 16), podstawiamy w (23) odpowiednie wartości, jako w_h. Dla przykładu w przypadku losowania proporcjonalnego (12) mamy $w_h=W_h$ oraz:

$$n_{prop.} \approx \frac{\sum_{h=1}^{L} W_h S_h^2}{\frac{d^2}{4} + \frac{1}{N} \sum_{h=1}^{L} W_h S_h^2} \tag{24}$$

Stosując optymalny rozdział próby między warstwy, uwzględniający różne koszty zbierania informacji dla warstw (wzór 8), ilorazy W_h nie zmieniają się ze zmianami n. Można więc dla dowolnie ustalonego łącznego kosztu K badania obliczyć n_h według (8) oraz stąd otrzymać w_h, które należy podstawić do prawej strony (23). Uzyskaną wartość n należy przemnożyć przez kolejne $w_h (h = 1, 2, ..., h)$ i otrzymać n_h. Stąd łatwo obliczyć koszt badania $K - K_o = \sum_{h=1}^{L} K_h n_h$. Może się okazać, że łączny koszt badania przekroczy koszty przeznaczone na to badanie. Wówczas istnieją dwie możliwości: a) zmniejszyć wymagania co do precyzji szacunków, aby zmieścić się w dostępnych funduszach na badanie bądź b) żądać podwyższenia tych funduszów. Jeśli żadna z tych możliwości nie może być spełniona, należy zrezygnować z badania reprezentacyjnego. Przed tym jednak trzeba dobrze przeanalizować żądanie co do „maksymalnego błędu szacunku d".

8.17. W niniejszym wykładzie omówione zostało zagadnienie szacowania średniej $\bar{Y}$ cechy mierzalnej Y w przypadku losowania warstwowego. Ponieważ wartość globalna $Y = N\bar{Y}$, wnioski łatwo transponować na przypadek szacowania tej wartości. Np. prostym estymatorem dla tego parametru jest $y_{(w)} = N\bar{y}_{(w)}$, a błąd standardowy estymatora $y_{(w)}$, $D(y_{(w)}) = ND\bar{y}_{(w)}$.

Przypadek szacowania frakcji P elementów wyróżnionych sprowadza się do przypadku szacowania średniej $\bar{Y}$ przy założeniu, że cecha Y może przyjąć wartość 1, gdy dany element jest wyróżniony, natomiast 0, gdy jest przeciwnie. Zatem, analogicznie do (5) $p_{(w)} = \sum_{h=1}^{L} W_h p_h$, gdzie p_h jest frakcją elementów wyróżnionych w próbie z h-tej warstwy, $p_{(w)}$ jest nieobciążonym estymatorem frakcji P. W dalszych wzorach należy pamiętać, że — o ile $N_h/(N_h - 1) \approx 1$, S_h^2 należy zastąpić iloczynem $P_h Q_h$. Np. stosując losowanie proporcjonalne z (14) otrzymujemy:

$$D^2_{prop.}(p_{(w)}) \approx \left(1 - \frac{n}{N}\right) \frac{\sum_{h=1}^{L} W_h P_h Q_h}{n}, \qquad Q_h = 1 - P_h \tag{25}$$

Jeżeli naszym zadaniem jest szacowanie liczby NP jednostek wyróżnionych w populacji, za estymator możemy przyjąć $Np_{(w)}$.

Uwaga 8.2. We wzorach (23)—(25) figurują wartości parametrów S_h^2, P_h, Q_h. W praktyce będą one najczęściej nie znane. Wówczas przyjmujemy w ich miejsce oceny ze wstępnej próby bądź z badań prowadzonych poprzednio.

8.18. **Zadania.** 1. Planujemy badanie reprezentacyjne, stosując losowanie warstwowe próby. Przeznaczamy na badanie $K = 8$ mln zł. Koszty stałe oceniamy

$K_o = 500$ tys. zł. Szacujemy średnią wartość $\bar{Y}$. Określić wielkość próby i jej optymalny podział między warstwy, jeśli koszty jednostkowe K_h, liczebność N_h oraz oceny wariancji S_h^2 w poszczególnych warstwach wynoszą:

h	1	2	3	4	5	6
K_h (tys. zł)	0,9	1,6	2,5	3,6	4,9	6,4
N_h	1000	3000	5000	4000	3500	1500
S_h^2	1	2,25	6,25	9,00	16,00	25,00

Jak wielką próbę można by wylosować z każdej warstwy w przypadku zastosowania:
a) optymalnego schematu Neymana,
b) losowania proporcjonalnego?
Ile wynosi błąd standardowy średniej $\bar{y}_{(w)}$ w każdym z tych schematów lokalizacji próby?

2. W celu oceny liczby pewnych wyrobów sztukowych, znajdujących się w magazynach 25 tys. sklepów, podzielono sklepy na cztery warstwy w zależności od liczby tych wyrobów w magazynach w poprzednim badaniu. Wylosowano próbę, stosując losowanie warstwowe i uzyskano następujące wyniki:

h	1	2	3	4
N_h	20000	3250	1020	730
n_h	3050	600	350	250
$\bar{y}_h$	4,1	13,0	25,0	38,2
S_h^2	34,8	92,2	174,2	320,4

Oszacować liczbę Y wyrobów w magazynach oraz błąd standardowy szacunku. Ocenić zysk na odchyleniu standardowym estymatora w stosunku do tak samo dużej próby, ale wylosowanej według proporcjonalnej lokalizacji próby z warstw.

3. Stosując schemat proporcjonalnego losowania warstwowego uzyskano następujące dane z 10% próby:

h	1	2	3	4	5	6
n_h	150	250	300	500	400	400
p_h	0,5	0,8	0,9	0,3	0,1	0,05

Oszacować frakcję P elementów wyróżnionych oraz błąd standardowy szacunku. O ile zwiększy się ten błąd, gdyby takiej samej wielkości próbę losować nieograniczenie z całej populacji?

8.19. **LITERATURA UZUPEŁNIAJĄCA:** R. Zasępa [18] str. 120—157, J. Steczkowski [14] str. 174—205.

Wykład 9. Zespołowe losowanie nieograniczone z jednakowymi prawdopodobieństwami wyboru oraz losowanie systematyczne

9.1. Dotąd omawialiśmy schematy losowania indywidualnego, których realizacja wymagała posiadania operatu losowania, wyszczególniającego pojedyncze jednostki badania. Jeśli taki operat nie istnieje, jego konstrukcja może być bardzo kosztowna. Ponadto, nawet posiadając indywidualny operat losowania, losowana z niego próba — w przypadku badań terenowych — będzie rozproszona w terenie i zebranie informacji od jednostek wylosowanych do próby może być bardzo pracochłonne oraz wymagać zbyt dużej liczby personelu, jaki należałoby zatrudnić, w celu zebrania informacji statystycznej w żądanym okresie. Dużym ułatwieniem może być **losowanie zespołowe,** gdy jednostki losowania tworzą zespoły jednostek badania.

Przykład 9.1. W 1982 r. przeprowadzone *Reprezentacyjne badanie ludności rolniczej* opierało się na próbie wylosowanej zespołowo. Zespoły stanowiące *JL* były obwodami spisowymi. W wylosowanym do próby obwodzie spisowym spisywano ludność rolniczą w nim zamieszkałą. Próba zawierała 8150 obwodów spisowych.

9.2. Grupowanie jednostek populacji generalnej w zespoły może być dokonywane w różny sposób. Zwykle jest ono w naturalny sposób związane z jednostkami badania. Np., gdy populację generalną tworzy ludność, zespołami stanowiącymi zespołowe jednostki losowania (*JL*) mogą być mieszkania, gospodarstwa domowe, obwody spisowe, miasta i gminy wiejskie; gdy populację generalną tworzą indywidualne gospodarstwa rolne zespołowymi *JL* mogą być obwody spisowe, miejscowości. Gdy badamy uczniów zespołowymi *JL* mogą być szkoły, klasy, jak również miejscowości. Badając zarobki ludności, pracującej zarobkowo, zespołowymi *JL* mogą być zakłady pracy itd.

9.3. W niniejszym wykładzie omówione zostanie zespołowe losowanie nieograniczone bez zwracania z jednakowymi prawdopodobieństwami wyboru, czyli losowanie proste zespołów bez zwracania (*lpzbz*). Wówczas operat losowania powinien dla każdego zespołu, stanowiącego *JL*, podawać właściwe mu cechy adresowe, a ponadto numer tego zespołu, zgodnie z przyjętym uporządkowaniem *JL* w operacie losowania.

9.4. W dalszych rozważaniach przyjmujemy, że elementy populacji zostały podzielone na N zespołów. Każdy zespół posiada określony numer k, $k=1, 2, ..., N$. W k-tej zespołowej *JL* znajduje się M_k jednostek badania w określony sposób uporządkowanych, czyli identyfikatorem jednostki badania jest para numerów (k, l) $l=1, 2, ..., M_k$, dla $k=1, 2, ..., N$. Wartość cechy Y, dla (k, l)-tej jednostki badania, oznaczymy jako Y_{kl}. Wartość globalna cechy Y w k-tej *JL*, $Y_k=\sum_{l=1}^{M_k} Y_{kl}$ (przyjmujemy dotychczasową konwencję, że w przypadku cechy niemierzalnej $Y_{kl}=1$, gdy (k, l)-ta jednostka badania jest wyróżniona, natomiast $Y_{kl}=0$ w przypadku przeciwnym;

wówczas Y_k jest liczbą jednostek wyróżnionych w k-tej JL), **wartość globalna cechy** Y **w populacji** $Y=\sum_{k=1}^{N} Y_k$. Średnią wartość cechy Y wyznaczymy, jako:

$$\bar{Y}=\frac{\sum_{k=1}^{N} Y_k}{\sum_{k=1}^{N} M_k}=\frac{1}{M}\sum_{k=1}^{N} Y_k=\frac{Y}{M} \tag{1}$$

Jest to **średnia na jednostkę badania** (M jest liczebnością populacji $M=\sum_{k=1}^{N} M_k$). **Średnią wartością cechy** Y **na jednostkę losowania** $\hat{\bar{Y}}$ jest:

$$\hat{\bar{Y}}=\frac{Y}{N}=\frac{1}{N}\sum_{k=1}^{N} Y_k \tag{2}$$

Wariancja cechy Y, S_y^2 wynosi:

$$S_y^2=S^2=\frac{1}{M-1}\sum_{k=1}^{N}\sum_{l=1}^{M_k}(Y_{kl}-\bar{Y})^2 \tag{3}$$

Wariancja zespołowa cechy Y pomiędzy jednostkami losowania (zespołami):

$$S^{S'}=\frac{1}{N-1}\sum_{k=1}^{N}(Y_k-\hat{\bar{Y}})^2 \tag{4}$$

Symbolem $\bar{M}$ oznaczymy średnią liczbę jednostek badania w jednostce losowania:

$$\bar{M}=\frac{M}{N}=\frac{1}{N}\sum_{k=1}^{N} M_k \tag{5}$$

Dla próby oznaczamy: n — liczba JL w próbie, y_i — wartość globalna cechy w i-tej z kolei wylosowanej JL ($i=1, 2, ..., n$), m_i — liczebność tej JL (jeśli za i-tym razem wylosowano k-tą JL to $m_i=M_k$).

9.5. Szacujemy średnią $\bar{Y}$. Za estymator przyjmujemy średnią $\bar{y}^{(z)}$ z próby:

$$\bar{y}^{(z)}=\frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} m_i} \tag{6}$$

Twierdzenie 9.1.

(a) Jeżeli jednostki losowania **są jednakowo liczne,** tj. $M_k = \bar{M}$ estymator (6) jest nieobciążonym estymatorem średniej $\bar{Y}$, a jego wariancja wynosi:

$$D^2(\bar{y}^{(z)}) = \left(1 - \frac{n}{N}\right) \frac{z^{S^2}}{n\bar{M}^2} \tag{7}$$

wówczas w (6) $m_i = \bar{M}$, czyli $\sum_{i=1}^{n} m_i = n\bar{M}$

(b) Jeżeli jednostki losowania **nie są jednakowo liczne,** to estymator (6) jest obciążonym, zgodnym estymatorem średniej $\bar{Y}$. Jeżeli n jest dostatecznie duże (tak, że $V(\bar{m}) \leqslant 0{,}05$), obciążenie estymatora można pominąć oraz:

$$D^2(\bar{y}^{(z)}) \approx \left(1 - \frac{n}{N}\right) \cdot \frac{1}{n\bar{M}^2} \cdot \frac{1}{N-1} \sum_{k=1}^{N} (Y_k - M_k \bar{Y})^2 \tag{8}$$

9.6. W celu porównania wariancji (8) z wariancją w przypadku losowania próby $n\bar{M}$ w losowaniu prostym bez zwracania okazuje się, że:

$$D^2(\bar{y}^{(z)}) \approx \left(1 - \frac{n}{N}\right) \frac{S^2}{n\bar{M}} [1 + (\bar{M} - 1)\kappa] \tag{8a}$$

Z (8a) stwierdzamy, że:

$$D^2(\bar{y}^{(z)}) \approx [1 + (\bar{M} - 1)\kappa] D^2(\bar{y}) \tag{9}$$

Występujący we wzorach (8a) i (9) **współczynnik** κ jest pewną **miarą jednorodności** JL:

$$\kappa = \frac{\sum_{k=1}^{N} \sum_{l_1 \neq l_2}^{M_k} (Y_{kl_1} - \bar{Y})(Y_{kl_2} - \bar{Y})}{(M-1)(\bar{M}-1)S^2} = \frac{\sum_{k=1}^{N} (Y_k - M_k \bar{Y})^2}{(M-1)(\bar{M}-1)S^2} - \frac{1}{\bar{M}-1} \tag{10}$$

Przykład 9.2. Populacja generalna obejmuje $M = 12$ jednostek badania, które pogrupowano na dwa różne sposoby (Y_{kl}, dla $k = 1,2,3,4$ oraz $l = 1, 2, ..., M_k$) jest wartością cechy Y dla kolejnych jednostek badania w zespołach:

a) (1,3); (0,2,2,1); (1,2,1); (0,2,3);

b) (0,0); (1,1,1,1); (2,2,2); (3,3,2).

W przypadku a) zespoły są wewnętrznie bardziej niejednorodne niż w przypadku b). Łatwo obliczyć, że w populacji $\bar{Y} = 1{,}5$; $S^2 = 1$; $\bar{M} = 3$ oraz $M_1 = 2$; $M_2 = 4$; $M_3 = 3$; $M_4 = 3$.

Dla grupowania a) mamy $\sum_{k=1}^{N} (Y_k - M_k \bar{Y})^2 = (4-3)^2 + (5-6)^2 + (4-4{,}5)^2 + (5-4{,}5)^2 = 2{,}5$.

Podstawiając dane do prawej strony (10) otrzymujemy:
$\kappa = \frac{2,5}{11,2,1} - \frac{1}{2} = -0,39$. Dla grupowania b) uzyskujemy $\kappa = 0,75$.

Przyjmując $n=2$ zespoły do próby wylosujemy przeciętnie $n\bar{M}=6$ jednostek badania. Próba sześcioelementowa losowana indywidualnie w *lpbz* daje wariancję średniej z próby $D^2_{ind}(\bar{y}) = \left(1 - \frac{6}{12}\right) \cdot \frac{1}{6} = \frac{1}{12}$

W przypadku losowania zespołowego (a):

$$D^2(\bar{y}^{(z)}) \approx \left(1 - \frac{2}{4}\right) \cdot \frac{1}{2 \cdot 3} [1 + (3-1)(-0,39)] = 0,018\,(3) < \frac{1}{12} = 0,08\,(3)$$

W przypadku losowania zespołowego (*b*):

$$D^2(\bar{y}^{(2)}) \approx \left(1 - \frac{2}{4}\right) \cdot \frac{1}{2 \cdot 3} [1 + (3-1) \cdot 0,75] = 0,208\,(3) > 0,08\,(3)$$

Z powyższych wyliczeń wynika, że w przypadku (a) losowanie zespołowe jest znacznie efektywniejsze niż losowanie indywidualne ($D^2(\bar{y}^{(z)})$ jest ponad 4,5 raza mniejsza niż $D^2_{ind}(\bar{y})$. Jest przeciwnie w przypadku (b), kiedy $D^2(\bar{y}^{(2)})$ jest około 2,5 raza większa niż $D^2_{ind}(\bar{y})$.

9.7. Z analizy wzoru (10) — jego prawej strony — wynika, że najmniejszą wartością κ jest $-\frac{1}{\bar{M}-1}$. Ta wartość wystąpi wówczas, gdy dla **każdego** zespołu stanowiącego *JL* globalna wartość cechy *Y* w *k*-tym zespole ($k=1, 2, ..., N$), $Y_k = M_k \bar{Y}$, czyli $\bar{Y}_k = Y_k : M_k = \bar{Y}$. To oznacza, że najkorzystniejszym podziałem na zespoły, będące *JL* jest podział, w którym średnia wartość cechy *Y* jest dla każdego zespołu taka sama, równa $\bar{Y}$. Wówczas nawet jednoelementowa próba ($n=1$) daje ocenę dokładną, $D(\bar{y}^{(z)})=0$.

Jeśli *JL* są **jednakowo liczne**, $M_k = \bar{M}$, to największą wartością κ jest 1, czyli:

$$-\frac{1}{\bar{M}-1} \leqslant \kappa \leqslant 1 \tag{11}$$

Jeśli *JL* **nie są** jednakowo liczne, to κ może przyjąć znacznie większe niż 1 wartości.

Przykład 9.3. Współczynnik κ, jako miara jednorodności zespołów osiąga dla danej populacji największą wartość, gdy w każdym zespole jednostki badania mają tę samą wartość cechy *Y*, mianowicie $\bar{Y}_k$ dla *k*-go zespołu stanowiącego *JL*. Wówczas łatwo wykazać, że:

$$\kappa = \frac{\sum_{k=1}^{N} (Y_k - M_k \bar{Y})^2}{\sum_{k=1}^{N} \frac{\bar{M}-1}{M_k} (Y_k - M_k \bar{Y})^2} - \frac{1}{\bar{M}-1} \tag{12}$$

Zakładamy, że populacja generalna obejmuje 120 jednostek badania pogrupowanych w $N=6$ zespołowych JL w pełni jednorodnych, ze względu na wartości cechy Y.

M_k	$\bar{Y}_k$	Y_k	M_kY	$(Y_k-M_k\bar{Y})^2$	$\frac{\bar{M}-1}{M_k}(Y_k-M_k\bar{Y})^2$
4	1	4	12,23	67,7329	321,73
8	0,5	4	24,47	419,0209	995,17
12	3,5	42	36,70	28,0900	44,48
50	6,0	300	152,92	21632,5264	8220,36
40	0,2	8	122,33	13071,3489	6208,89
6	1,5	9	18,35	87,4225	276,84
$M=120$	$\bar{Y}=3,0583$	367	×	35306,1416	16067,47
$\bar{M}=20$					

$$\kappa=\frac{35306,1416}{16067,47}-\frac{1}{19}=2,14$$

9.8. Wzór (9) prowadzi do następującego wniosku:
Przy danej wielkości próby losowanie zespołowe jest efektywniejsze od losowania indywidualnego, gdy $\kappa<0$. tak samo efektywne, gdy $\kappa=0$; mniej efektywne niż losowanie indywidualne, gdy $\kappa>0$. Można udowodnić, że jeżeli grupowanie w zespoły, stanowiące jednostki losowania jest losowe, to $\kappa\approx 0$.

9.9. W praktyce zespołowe JL są na ogół związane z dostępnością materiałów administracyjnych bądź innych, które mogą być użyteczne, jako operat losowania. W praktyce GUS bardzo często zespołowymi JL są **obwody spisowe** bądź **regiony spisowe**, dla których figurują pewne informacje dodatkowe i które są kolejno ponumerowane w województwach (w podziale na obwody miejskie oraz obwody wiejskie); odnośne dane figurują w rejestrze komputerowym (w Ośrodku Elektronicznym GUS). Zespoły składają się najczęściej z jednostek badania bardziej do siebie podobnych pod względem wartości badanych cech niż to ma miejsce w populacji. Inaczej mówiąc, **przeważnie współczynnik** κ, będący miarą jednorodności JL jest dodatni, $\kappa>0$.

Okazuje się, że przy $\kappa>0$ oraz wzroście liczebności (M_k) JL współczynnik κ maleje, ale iloczyn $(\bar{M}-1)\kappa$ rośnie. Z amerykańskich badań materiałów spisu ludności 1940 r. wynika, że np. dla cechy „liczba osób w gospodarstwie domowym" w przypadku zespołów o 3 gospodarstwach domowych $\kappa=0,230$, o 9 gospodarstwach domowych $\kappa=0,186$, o 27 gospodarstwach domowych $\kappa=0,142$, a o 62 gospodarstwach domowych $\kappa=0,066$; natomiast, gdy rozpatrujemy iloczyn $(\bar{M}-1)\kappa$ dla $\bar{M}=3, 9, 27, 62$ otrzymujemy $(\bar{M}-1)\kappa=0,460$; 1,49; 3,69; 4,03. Jeśli więc zespoły byłyby równoliczne oraz $\bar{M}=62$, to precyzja próby danej wielkości, mierzona wariancją średniej z próby, byłaby około 5 razy gorsza niż gdybyśmy zastosowali *lpbz*.

9.10. Z powyższych rozważań wynika, że pożądane jest, aby zespołowe JL nie różniły się co do ich wielkości (M_k). Jeśli te różnice są znaczne i posiadamy informacje o wielkości zespołów można poprawić precyzję szacunków, ujmując zespoły podobnej

wielkości w warstwy. Jedną z technik postępowania jest uporządkowanie zespołowych JL w ciąg od największej do najmniejszej i utworzenie $\frac{n}{2}$ warstw, obejmujących kolejne $\frac{2N}{n}$ JL oraz wylosowanie z każdej warstwy po dwie JL. Łącząc niektóre małe zespoły w większe można zawsze doprowadzić do tego, żeby $\frac{2N}{n}$ było liczbą całkowitą, a ponadto nie ma problemu, aby przyjąć, że liczebność próby n jest liczbą parzystą. Tego rodzaju postępowanie zostało przyjęte np. przy losowaniu próby obwodów spisowych do reprezentacyjnego badania dzietności kobiet przy okazji *NSP* 1970 r.

9.11. **LOSOWANIE SYSTEMATYCZNE.** Jak wiemy, losowanie systematyczne polega na: a) ustaleniu **interwału losowania** k, b) wylosowaniu liczby naturalnej r z przedziału $[1, k]$ i c) wybraniu do próby tych jednostek badania z listy, będącej operatem losowania, które są oddalone od r-tej jednostki, o wielokrotność liczby k. Do próby wchodzą więc jednostki badania o numerach:

$$r, r+k, r+2k, ..., r+(n-1)k \tag{13}$$

Załóżmy, że $N=k_n$, przy czym n jest wielkością próby.

9.12. Z określonego uporządkowania jednostek badania wynika, że schemat systematycznego losowania próby n-elementowej jest równoważny schematowi losowania zespołowego, w którym do próby losujemy **jedną** zespołową JL z populacji podzielonej na k **równolicznych** (n-elementowych) **zespołów.**

9.13. Szacujemy średnią $\bar{Y}=\frac{1}{N}\sum_{l=1}^{N} Y_r$, przy pomocy średniej z próby $\bar{y}$:

$$\bar{y}=\frac{1}{n}\sum_{i=1}^{n} y_i;\ \text{jeśli}\ y_1=Y_r,\ \text{to:}$$
$$y_i=Y_{r+(i-1)k}\ \text{dla}\ i=1, 2, ..., n \tag{14}$$

Jest to estymator określony wzorem (6), przy odpowiednio zmienionych występujących w nim oznaczeniach (mianowicie: $n\to 1$, $y_1\to\sum_{i=1}^{n} y_i$ oraz $m_1\to n$). Wariancja średniej (14) wynosi — zgodnie z (8a):

$$D^2_{syst}(\bar{y})\approx\frac{S^2}{n}[1+(n-1)\kappa]\left(1-\frac{1}{k}\right) \tag{15}$$

Przy obecnych oznaczeniach (**współczynnik** κ **jednorodności wewnątrzzespołowej zwany obecnie współczynnikiem korelacji wewnątrzklasowej,** określony według wzoru (10), ma postać:

$$\kappa=\frac{\sum_{r=1}^{k}\sum_{l_1\neq l_2}^{n}(Y_{r+(l_1-1)k}-\bar{Y})(Y_{r+(l_2-1)k}-\bar{Y})}{(n-1)(kn-1)S^2} \tag{16}$$

9.14. Z wzoru (15) wynika, że losowanie systematyczne jest mniej więcej tak samo efektywne jak *lpbz*, gdy uporządkowanie jednostek badania w operacie losowania jest losowe. Gdy jednostki badania bliższe sobie są bardziej do siebie podobne, pod względem wartości badanej cechy (Y) niż jednostki oddalone, można się spodziewać, że $\kappa < 0$, czyli losowanie systematyczne jest efektywniejsze od *lpbz*.
W pozostałych przypadkach losowanie systematyczne jest mniej efektywne ($\kappa > 0$) od *lpbz*. Szczególnie niebezpieczna jest sytuacja, gdy uporządkowanie jednostek badania jest cykliczne, czyli po określonej liczbie jednostek badania wartości cechy Y powtarzają się i interwał losowania jest wielokrotnością okresu cyklu. Wówczas n-elementowa próba losowania systematycznie jest **równoważna jednoelementowej próbie** losowanej zgodnie ze schematem *lpbz*. Jeśli więc obawiamy się występowania jakiejś cykliczności w wartościach cechy Y, w uporządkowaniu przyjętym w operacie losowania, **nie należy stosować losowania systematycznego.**

9.15. Planując badanie statystyczne chcielibyśmy tak określić liczebność próby, aby — z pewnym, małym ryzykiem błąd oceny nie przekraczał dopuszczalnej wielkości. Niestety tego postulatu nie można spełnić w przypadku losowania systemytycznego próby, gdyż nie wiemy czy zastane uporządkowanie jednostek badania jest korzystne czy nie — dla losowania systematycznego o danym interwale. Aby przynajmniej określić precyzję szacunku **ex post**, tzn. po zrealizowaniu badania reprezentacyjnego należy zamiast jednej próby n-elementowej o interwale k wylosować g podprób o interwale g razy większym, czyli gk. Oznaczmy średnią z i-tej podpróby $i = 1, 2, ..., g$ przez $\bar{y}_i$. Nieobciążonym estymatorem średniej $\bar{Y}$ jest:

$$\bar{y} = \frac{1}{g} \sum_{i=1}^{g} \bar{y}_i \tag{17}$$

a nieobciążonym estymatorem wariancji $D^2_{syst}(\bar{y})$ jest statystyka:

$$v = \frac{1 - \frac{1}{k}}{g(g-1)} \sum_{i=1}^{g} (\bar{y}_i - \bar{y})^2 \tag{18}$$

Statystyka $\sqrt{v}$ jest obciążonym estymatorem błędu standardowego $D_{syst}(\bar{y})$.
Oznaczmy symbolem w rozstęp średnich z podprób

$$w = \max_i \bar{y}_i - \min_i \bar{y}_i \tag{19}$$

tj. różnicę pomiędzy największą ocena $\max_i \bar{y}_i$ oraz najmniejszą oceną $\min_j \bar{y}_1$. Nieobciążonym estymatorem błędu standardowego jest wówczas statystyka u:

$$u = \frac{w}{\lambda} \sqrt{\frac{1}{g}\left(1 - \frac{1}{k}\right)} \tag{20}$$

gdzie: λ jest pewną stałą, której wielkość zależy od g, czyli od liczby podprób. Wartość λ, dla $g = 5$ wynosi $\lambda = 2{,}326$, dla $g = 10$ $\lambda = 3{,}078$.

Przykład 9.4. W celu oszacowania średniej $\bar{Y}$ wylosowano z populacji systematycznie 5 podprób, przy interwale losowania $k_g = 100$. Łączna próba objęła 5% populacji. Losowanie odbyło się następująco: z pierwszych 100 jednostek badania wylosowano jednostki o numerach: 14, 95, 32, 91, 03; do pierwszej podpróby weszły więc jednostki o numerach: 14, 114, 214, 314, ..., 1914 (populacja liczyła $N = 2000$ jednostek badania), do drugiej podpróby weszły jednostki o numerach: 95, 195, 295, ..., 1995 itd. Średnie z podprób wyniosły:

$$\bar{y}_1 = 4{,}5;\ \bar{y}_2 = 3{,}4;\ \bar{y}_3 = 6{,}2;\ \bar{y}_4 = 7{,}0;\ \bar{y}_5 = 5{,}4$$

$$\text{Stąd: } \bar{y} = \frac{1}{5}(4{,}5 + 3{,}4 + 6{,}2 + 7{,}0 + 5{,}4) = 5{,}3$$

$$w = 7{,}0 - 3{,}4 = 3{,}6;\ u = \frac{3{,}6}{2{,}326}\sqrt{\frac{1}{5}\left(1 - \frac{1}{20}\right)} = 0{,}67$$

Około 95% przedział ufności dla $\bar{Y}$ wynosi: [3,96; 6,64].

9.16. **Zadania.** 1. Oszacować wariancję $D^2(\bar{y}^{(z)})$ estymatora przeciętnej $\bar{Y}$(ha) powierzchni gospodarstwa rolnego na terenie województwa, obejmującego $N = 5789$ obwodów spisowych, jeżeli próba ma objąć $n = 580$ obwodów, a wstępna próba $n' = 60$ obwodów dała wyniki:

$$\sum_{i=1}^{60} m_i = 2378;\ \sum_{i=1}^{60} m_i^2 = 98483;\ \sum_{i=1}^{60} y_i = 13591$$

$$\sum_{i=1}^{60} m_i y_i = 648319;\ \sum_{i=1}^{60} y_i^2 = 4424798$$

2. Dane, jak w poprzednim zadaniu. Oszacować współczynnik κ jeżeli przypuszczamy, że $S_y \approx 3$ ha. O ile mniej efektywne jest losowanie obwodów spisowych od losowania (*lphz*) gospodarstw rolnych?

3. Korzystając z danych poprzedniego zadania, oszacować:

a) liczbę ludności Y w badanym mieście,

b) 95% przedział ufności dla parametru Y,

c) błąd standardowy oceny, posługując się rozstępem ocen z podprób.

4. Badając mieszkania w pewnym mieście wylosowano systematycznie pięć jednoprocentowych podprób. Badane cechy: liczba izb (X) oraz liczba osób zamieszkałych (Y). Wyniki badania zestawiono na następnej stronie (w jednym wierszu i dla danej podpróby figurują wartości cech X i Y w wylosowanym mieszkaniu):

Numer podpróby									
1		2		3		4		5	
x	y	x	y	x	y	x	y	x	y
1	1	1	7	2	6	2	2	2	7
3	5	2	4	2	1	2	6	2	3
1	5	2	4	1	3	1	1	1	1
2	2	2	4	2	3	3	3	2	3
2	5	3	5	2	5	2	3	3	2
2	3	3	5	2	2	4	5	3	5
2	2	1	6	1	2	3	6	2	3
3	4	2	2	2	2	1	3	1	3
2	3	2	3	2	3	2	6	3	3
3	3	1	2	2	6	1	4	1	2
1	2	2	4	1	3	2	4	1	1
1	2	2	4	2	6	1	5	3	3
2	2	4	3	2	6	1	2	3	4
1	4	2	3	3	5	2	8	2	4
2	8	2	6	2	4	3	4	3	4
2	2	2	6	1	1	2	3	2	8
2	3	3	4	1	5	2	6	1	2
2	4	3	3	3	6	4	4	2	5
2	4	2	5	2	5	2	5	2	4
3	3	3	4	3	4	3	3	4	4
2	3	2	2	2	3	1	1	1	2
3	8	2	4	1	1	2	5	4	5
3	4	2	5	3	2	2	7	2	3
1	1	3	3	3	7	2	3	1	3
2	4	2	8	2	4	1	1	2	9
Razem									
50	87	55	106	49	95	51	100	53	93

Należy oszacować:

a) przeciętną $\bar{X}$ liczbę izb na mieszkanie,

b) przeciętną $\bar{Y}$ liczbę osób na mieszkanie oraz wariancje estymatorów tych średnich.

9.17. **LITERATURA UZUPEŁNIAJĄCA:** R. Zasępa [18], str. 165–186, 188–195, 202–221, [17], str. 228–254, 260–282, J. Steczkowski [14], str. 206–244.

Wykład 10. Losowanie zespołowe z różnymi prawdopodobieństwami wyboru

10.1. Przypomnijmy, że losowanie zespołowe polega na: a) pogrupowaniu jednostek badania w zespoły takie, że każda jednostka badania należy do jednego i tylko jednego zespołu, b) zespoły stanowią jednostki losowania (*JL*), c) próbę stanowią te jednostki badania, które należą do wylosowanych *JL*. Poprzednio rozważaliśmy sytuację, gdy schemat losowania był **losowaniem prostym** *JL* **bez zwracania** (*lzpbz*), czyli każda *JL* miała **takie samo prawdopodobieństwo** dostania się do próby oraz *JL* wylosowana do próby nie brała udziału w dalszym losowaniu próby.

10.2. Rozpatrując szacowanie średniej $\bar{Y}$ wartości cechy Y za estymator przyjęliśmy średnią z próby $\bar{y}^{(z)}$. Badając — **przy danej wielkości próby** — zależność pomiędzy wariancją estymatora w przypadku *lzpbz* oraz wariancją estymatora w przypadku *lpbz* stwierdziliśmy, że ($\bar{M}$ oznacza średnią liczebność *JL*):

$$\frac{D^2(\bar{y}^{(z)})}{D^2_{ind}(\bar{y})} \approx 1 + (\bar{M} - 1)\kappa \tag{1}$$

gdzie κ oznacza współczynnik (miarę) jednorodności *JL*. Gdy *JL* **są jednakowo liczne** współczynnik ten zwany wówczas **współczynnikiem korelacji wewnątrzklasowej** przybiera wartości z przedziału $\left[-\frac{1}{\bar{M}-1}, 1\right]$, czyli $\frac{1}{\bar{M}-1} \leqslant \kappa \leqslant 1$. Gdy przeciwnie, *JL* **nie są jednakowo** liczne, współczynnik κ może mieć wartość większą od 1. Stąd wniosek, że schemat losowania zespołowego powinien w jakiś sposób kontrolować wielkości *JL* wybieranych do próby.

10.3. Jeśli chcemy uzależniać szanse wyboru do próby od wielkości *JL* możemy wykonać to, ustalając **różne** prawdopodobieństwa wyboru. Powstaje wówczas zagadnienie, jakie te prawdopodobieństwa powinny być.

10.4. W dotychczasowych wykładach rozważaliśmy głównie zagadnienie szacowania średniej wartości cechy Y zwracając uwagę, że wnioski łatwo jest przenieść na szacowanie wartości globalnej tej cechy. Ze względów rachunkowych, w przypadku losowania zespołowego, wygodniej jest skoncentrować się na szacowaniu wartości globalnej Y. Szacowanie średniej $\bar{Y}$ będzie odpowiadało szacowaniu ilorazu wartości globalnych dwóch cech.

10.5. **Rozpatrzymy najpierw przypadek losowania zespołowego ze zwracaniem**, jako prostszego w realizacji, a następnie przejdziemy do przypadku losowania bez zwracania.

10.6. Przyjmujemy oznaczenia wprowadzone w poprzednim wykładzie: N — liczba zespołów stanowiących *JL*, w k-tym zespole znajduje się M_k jednostek badania, Y_k, $\bar{Y}_k$ są wartościami cechy Y w k-tej *JL* ($k = 1, 2, \ldots, N$); wartością globalną i wartością

średnią. Liczbę jednostek badania w populacji oznaczamy jako M, $M=\sum_{k=1}^{N} M_k$. Niech π_k oznacza prawdopodobieństwo wylosowania do próby **za jednym razem** k-tej JL. Oczywiście, $\sum_{k=1}^{N} \pi_k = 1$. Przez $\pi_{(i)}$, oznaczymy prawdopodobieństwo wylosowania za jednym razem tej $J\alpha$, która została do próby wylosowana za i-tym razem, $i=1, 2, ..., n$; n oznacza liczbę kolejnych losowań do próby.

Przykład 10.1. Populacja obejmuje $N=10$ jednostek losowania (JL). Do próby losowano JL $n=3$ razy. Przypuśćmy, że wylosowano kolejno JL o numerach: 8, 2, 3. Wówczas $\pi_{(1)}=\pi_8$; $\pi_{(2)}=\pi_2$, $\pi_{(3)}=\pi_3$. Za drugim razem wynik losowania próby był następujący: 5, 3, 5. Zatem piąta JL została do próby wylosowana dwa razy. Wówczas $\pi_{(1)}=\pi_5$, $\pi_{(2)}=\pi_3$, $\pi_{(3)}=\pi_5$.

10.7. Dla próby oznaczamy przez y_i wartość globalną cechy Y dla wylosowanej za i-tym razem JL.

Przykład 10.2. W pierwszym losowaniu z przykładu 10.1 otrzymaliśmy $y_{(1)}=Y_8$; $y_{(2)}=Y_2$; $y_{(3)}=Y_3$, a w drugim — $y_{(1)}=Y_5$; $y_{(2)}=Y_3$, $y_{(3)}=Y_5$.

Z obu powyższych przykładów wynika, że $\pi_{(i)}$ dla $i=1, 2, ..., n$ są zmiennymi losowymi. Również y_i jest zmienną losową dla ustalonego i.

10.8. Szacujemy Y — wartość globalną cechy badanej:

$$Y=\sum_{k=1}^{N} Y_k \tag{1}$$

Dowodzi się, że $y^{(z)}$:

$$y^{(z)}=\frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{\pi_{(i)}} \tag{2}$$

jest nieobciążonym estymatorem wartości globalnej Y, a jego wariancja wynosi:

$$D^2(y^{(z)})=\frac{1}{n}\sum_{k=1}^{N}\pi_k\left(\frac{Y_k}{\pi_k}-Y\right)^2=\frac{1}{n}\left(\sum_{k=1}^{N}\frac{Y_k^2}{\pi_k}-Y^2\right) \tag{3}$$

Wariancja $D^2(y^{(z)})$ zależy od przyjętego układu prawdopodobieństw $\pi_1, \pi_2, ..., \pi_N$. Jeśli przyjmiemy $\pi_k=\frac{Y_k}{Y}$, to prawa strona wzoru (3) jest równa zeru, a więc precyzja estymatora (2) jest doskonała. Jednak, projektując badanie reprezentacyjne nie znamy ani Y_k, ani Y, a więc powyższemu postulatowi nie potrafimy zadość uczynić.

Jeśli dla każdego zespołu znamy wartości pewnej cechy Z, o której możemy sądzić, że jest silnie skorelowana z cechą badaną Y, wówczas możemy przyjąć:

$$\pi_k=\frac{Z_k}{Z}, \text{ dla } k=1, 2, ..., N, \; Z=\sum_{k=1}^{N} Z_k \tag{4}$$

a to powinno poprawić precyzję estymatora (2), który przyjmie postać:

$$y_{HH}^{(z)}=\frac{Z}{n}\sum_{i=1}^{n}\frac{y_i}{z_i} \tag{5}$$

$z_i=Z_k$, gdy za i-tym razem wylosowano k-tą JL.

Ten estymator został najpierw wprowadzony przez amerykańskich statystyków M.H. Hansena i W.N. Hurwitza i dlatego dodaliśmy indeks „HH". Łatwo wyprowadzić z (3), że wariancja:

$$D^2(y_{HH}^{(z)})=\frac{1}{n}\left(Z\sum_{k=1}^{N}\frac{Y_k^2}{Z_k}-Y^2\right) \tag{6}$$

W praktyce często korzystamy z przeszłej informacji o wielkości zespołów $M_k(k=1, 2, ..., N)$, przyjmując:

$$\pi_k=\frac{M_k}{M} \quad \text{dla } k=1, 2, ..., N \tag{7}$$

Wówczas:

$$y_{HH}^{(z)}=\frac{M}{n}\sum_{i=1}^{n}\frac{y_i}{m_i}=\frac{M}{n}\sum_{i=1}^{n}\bar{y}_i \tag{8}$$

Wariancja estymatora (8) wynosi:

$$D^2(y_{HH}^{(z)})=\frac{1}{n}\left(M\sum_{k=1}^{N}\frac{Y_k^2}{M_k}-Y^2\right) \tag{9}$$

Nieobciążonym estymatorem tej wariancji, w przypadku losowania wstępnej próby n'-elementowej z prawdopodobieństwami wyboru (7) proporcjonalnymi do wielkości zespołu (JL), jest statystyka:

$$u=\frac{1}{n(n'-1)}\left[M^2\sum_{i=1}^{n'}\bar{y}_i^2-n'(y'^{(z)}_{HH})^2\right] \tag{10}$$

przy czym $y'^{(z)}_{HH}$ jest estymatorem wartości globalnej cechy Y określonym wzorem (8) dla próby wstępnej n'-elementowej, natomiast n jest wielkością próby w projektowanym badaniu reprezentacyjnym.

Rzecz jasna, statystyka u przy $n'=n$ może być wykorzystana do oceny wariancji estymatora $y_{HH}^{(z)}$ po przeprowadzeniu badania reprezentacyjnego, jak również do oceny błędu standardowego szacunku.

10.9. Losowanie zespołowe ze zwracaniem, w którym prawdopodobieństwa wyboru są określone według (7), nosi nazwę **losowania z prawdopodobieństwami proporcjonalnymi do wielkości zespołów** i oznaczane jest symbolem $PPWZ$ (w literaturze angielskiej używa się symbolu PPS).

10.10. Wiadomo, że losowanie ze zwracaniem jest mniej efektywne niż losowanie bez zwracania. Naturalnym jest więc poszukiwanie schematu **losowania zespołowego z różnymi prawdopodobieństwami wyboru bez zwracania.** Okazuje się, że znalezienie takiego schematu losowania przy założonych prawdopodobieństwach wyboru jest trudne. Dotąd przedstawiono około 50 takich procedur losowania, jednak albo są one zbyt skomplikowane w realizacji, albo nie spełniają warunku założonych prawdopodobieństw wyboru, albo sprawiają duże trudności przy obliczaniu wariancji estymatora wartości globalnej bądź przy szacowaniu tej wariancji. W praktyce GUS stosuje się 2 wybrane schematy losowania zespołowego z różnymi prawdopodobieństwami wyboru. Pierwszy schemat pochodzi od W.G. Madowa (1949), przy czym pewna modyfikacja pozwalająca uzyskać asymptotyczne wzory na wariancję estymatora wartości globalnej i estymatory wariancji została zaproponowana przez H.O. Hartleya i J.N.K. Rao [3]. Drugi schemat losowania został wysunięty przez trzech autorów J.N.K. Rao, H.O. Hartleya i W.G. Cochrana [13], którzy wykazali własności estymatora wartości globalnej, wariancji tego estymatora oraz estymatora wariancji.

10.11. **Zmodyfikowana metoda Madowa.** Mamy N-elementową populację zespołowych JL, tzn. każda JL ma określony numer k, $1 \leqslant k \leqslant N$ w operacie losowania. To uporządkowanie zmieniamy losowo, tzn. ze zbioru liczb naturalnych od 1 do N przyporządkowujemy losowo jedną liczbę „i" każdej JL ($i=1, 2, ..., N$). Oznaczmy symbolem X miarę wielkości zespołów (JL); mamy więc dane X_i dla $i=1, 2, ..., N$. Oznaczmy przez π_i iloraz:

$$\pi_i = \frac{X_i}{\sum_{i=1}^{N} X_i} = \frac{X_i}{X} \tag{11}$$

a jeśli losujemy próbę n-elementową, to niech:

$$p_i = n\pi_i \tag{12}$$

p_i jest prawdopodobieństwem wylosowania do próby i-tej JL (***uwaga:*** obecnie identyfikatorem JL jest przyporządkowany jej losowo numer i). Przyjmujemy założenie $p_i < 1$. Losowanie próby jest losowaniem systematycznym. W tym celu tworzymy skumulowany ciąg (α_i) prawdopodobieństw p_i, przedstawiony w postaci tabeli:

α_1	α_2	α_3	.....	α_N
p_1	p_1+p_2	$p_1+p_2+p_3$	.....	$\sum_{i=1}^{N} p_i = n$

n — liczebność próby.

Z przedziału [0, 1) losujemy liczbę r, $0 \leqslant r < 1$. Do próby wchodzą te $JL(i)$, dla których są spełnione nierówności:

$$\alpha_{i-1} \leqslant r + k < \alpha_i, \quad k = 0, 1, 2, \ldots, n-1 \quad (\alpha_0 = 0) \tag{13}$$

Przykład 10.3. Populacja obejmuje $N = 8$ zespołowych JL. Za miarę X wielkości JL przyjmujemy liczbę jednostek badania. Dla ich uporządkowania losowego odczytujemy z tablicy liczb losowych, załączonej do wykładu 3 cyfry, poczynając od 22 wiersza, 7 bloku, pomijając cyfry raz odczytane oraz 0,9. Otrzymujemy kolejne cyfry:

4, 7, 3, 2, 6, 5, 8, 1

Zatem dla $k = 1$; $i = 8$, dla $k = 2$; $i = 4$, dla $k = 3$; $i = 3$ itd.

i	k	X_i	π_i	p_i	α_i	Do próby wchodzą
1 . . .	4	81	0,27	0,81	0,81	+
2 . . .	7	55	0,19	0,57	1.38	+
3 . . .	3	42	0,14	0,42	1,80	
4 . . .	2	15	0,05	0,15	1,95	
5 . . .	6	26	0,09	0,27	2,22	
6 . . .	5	45	0,15	0,45	2,67	+
7 . . .	8	16	0,05	0,15	2,82	
8 . . .	1	20	0,06	0,18	3.00	
Razem		300	1,00	3,00	×	

Z powyższej populacji losujemy próbę $n = 3$ JL. Odczytujemy z tablicy liczb losowych liczbę z 36 wiersza i 8 bloku. Jest to 35793. Interpretujemy ją jako 0,35793 i zaokrąglamy do dwóch cyfr po przecinku, otrzymując: $r = 0{,}36$; $0 \leqslant 0{,}36 < 0{,}81 = \alpha_1$, $r + 1 = 1{,}36$; $\alpha_1 \leqslant 1{,}36 < \alpha_2 = 1{,}38$, $r + 2 = 2{,}36$; $\alpha_5 \leqslant 2{,}36 < 2{,}67 = \alpha_6$.

Do próby weszły JL o numerach losowo przyporządkowanych $i = 1{,}2$ oraz 6, czyli JL o identyfikatorach w operacie losowania $k = 4{,}5$ oraz 7.

10.12. W zmodyfikowanej metodzie Madowa estymatorem nieobciążonym wartości globalnej cechy Y jest estymator (2), a jego wariancja wynosi w przybliżeniu:

$$D^2(y_M^{(z)}) \approx \frac{1}{n}\left(\sum_{k=1}^{N} \frac{Y_k^2}{\pi_k} - Y^2\right) - \frac{n-1}{n} \sum_{k=1}^{N} \pi_k^2 \left(\frac{Y_k}{\pi_k} - Y\right)^2 \tag{14}$$

Wzór (14) określa $D^2(y_M^{(z)})$ **jedynie dla** dużych liczebności N populacji generalnej oraz dla liczebności n prób stosunkowo małych, w porównaniu z N.

10.13. Omawiany wyżej schemat losowania próby jest stosowany w badaniach metodą reprezentacyjną w Głównym Urzędzie Statystycznym w przypadku losowania próby I stopnia. Zaletą jego jest łatwość realizacji schematu oraz ścisłe zachowanie założonych prawdopodobieństw wyboru. Wadą jest trudność uzyskania ścisłego wzoru na wariancję estymatora wartości globalnej, gdyż prawa strona (14) jest obciążonym estymatorem wariancji. Z porównania prawych stron (3) i (14) wynika, że

w przypadku schematu Madowa zysk na wariancji wynosi około:

$$\frac{n-1}{n}\sum_{k=1}^{N}\pi_k^2\left(\frac{Y_k}{\pi_k}-Y\right)^2$$

10.14. Trzej autorzy wymienieni w pkt. 10.10 Rao, Hartley i Cochran zaproponowali inną niż wyżej opisana, procedurę losowania próby, którą przyjęto określać **schematem losowania RHC.** Schemat ten przedstawia się następująco:

1. Jeśli losujemy próbę n JL, a populacja obejmuje N zespołowych JL, pierwszym krokiem jest losowy rozdział JL na n rozłącznych możliwie podobnej wielkości grup o liczebnościach N_1, N_2, ... N_n takich, że $N_1+N_2+...+N_n=N$.
2. Z każdej z tak utworzonych grup losujemy niezależnie po jednej JL **z prawdopodobieństwami proporcjonalnymi** do π_k ($k=1, 2, ... N$). Jeśli k-ta JL znalazła się w i-tej grupie ($i=1, 2, ..., n$), do której wpadły jednostki o numerach $k_1, k_2, ..., k_{N_i}$, to prawdopodobieństwo, że zostanie ona wylosowana do próby wyniesie $p_{(i)}=\pi_k:(\pi_{k_1}+\pi_{k_2}+...+\pi_{k_{N_i}})$.

Przykład 10.4. Dane, jak w poprzednim przykładzie ($N=8$, $n=3$). Tworzymy 3 grupy: dwie pierwsze po 3 JL oraz trzecia, zawierająca 2 JL. Ponieważ chodzi o losowe przyporządkowanie JL do poszczególnych grup, możemy to wykonać następująco: 1) podobnie, jak w metodzie Madowa, zmieniamy losowo uporządkowanie JL, 2) w tym ciągu do pierwszej grupy zaliczamy N_1 pierwszych numerów JL, do drugiej grupy N_2 następnych numerów JL itd. .... do ostatniej grupy pozostałe numery JL. Zatem postępujemy, podobnie jak w przykładzie 10.3, odczytując kolejne cyfry, od miejsca następującego po ostatniej uprzednio odczytanej cyfrze w tablicy liczb losowych. Zakończyliśmy poprzednie losowanie czytając cyfrę 1 w liczbie 23 wiersza, pierwszego bloku (była to liczba 97125). Zatem otrzymujemy następujący ciąg numerów od 1 do 8:

$$\underbrace{2, 5, 4,}_{i=1}\ \underbrace{3, 8, 7,}_{i=2}\ \underbrace{1, 6}_{i=3}$$

Z każdej z tych grup będziemy losowali po jednej JL.

Grupa 1. $\pi_2=0{,}05$; $\pi_5=0{,}15$; $\pi_4=0{,}27$ oraz $\pi_2+\pi_5+\pi_4=0{,}47$, dla $k=2$, $p_{(1)}=\frac{0{,}05}{0{,}47}=0{,}11$; dla $k=5$, $p_{(1)}=\frac{0{,}15}{0{,}47}=0{,}32$; dla $k=4$, $p_{(1)}=\frac{0{,}27}{0{,}47}=0{,}57$.

Z tablicy liczb losowych odczytujemy liczbę dwucyfrową. Jeśli odczytamy liczbę od 00 do 10 — wylosowano do próby JL o numerze $k=2$, jeśli odczytamy liczbę od 11 do 42 — wylosowano do próby JL o numerze $k=5$, jeśli odczytamy liczbę od 43 do 99 — wylosowano do próby JL o numerze $k=4$.

Grupa 2. $\pi_3=0{,}14$; $\pi_8=0{,}05$; $\pi_7=0{,}19$ oraz $\pi_3+\pi_8+\pi_7=0{,}38$, dla $k=3$, $p_{(2)}=\frac{0{,}14}{0{,}38}=0{,}37$; dla $k=8$, $p_{(2)}=\frac{0{,}05}{0{,}38}=0{,}13$; dla $k=7$, $p_{(2)}=\frac{0{,}19}{0{,}38}=0{,}50$.

Odczytujemy, jak wyżej, następną liczbę dwucyfrową. Jeśli będzie to liczba od 00 do 36 — wylosowano do próby $k=3$, jeśli od 37 do 49 — wylosowano do próby $k=8$, jeśli od 50 do 99 — wylosowano do próby $k=7$.

Grupa 3. $\pi_1=0{,}06$; $\pi_6=0{,}09$; $\pi_1+\pi_6=0{,}15$, dla $k=1$, $p_{(3)}=\dfrac{0{,}06}{0{,}15}=0{,}40$; dla $k=6$, $p_{(3)}=\dfrac{0{,}09}{0{,}15}=0{,}60$.

Odczytywanie liczby dwucyfrowej, jak wyżej. Jeśli będzie to liczba od 00 do 39 — wylosowano do próby $k=1$, jeśli inna liczba — $k=6$. Odczytane kolejno trzy liczby dwucyfrowe są następujące:

20, 47, 15

Do próby wylosowane zostały JL o numerach: $k_1=5$, $k_2=8$, $k_3=1$.

10.15. Statystyka[1]):

$$y_{RHC}^{(z)}=\sum_{i=1}^{n}\frac{y_{(i)}}{p_{(i)}} \tag{15}$$

jest nieobciążonym estymatorem wartości globalnej Y. Wariancja estymatora (15) jest najmniejsza, gdy losowe grupy są jednakowo liczne, czyli $N_1=N_2=\ldots=N_n=\dfrac{N}{n}$. Wówczas:

$$D^2(y_{RHC}^{(z)})=\left(1-\frac{n-1}{N-1}\right)\cdot\frac{1}{n}\left(\sum_{k=1}^{N}\frac{Y_k^2}{\pi_k}-Y^2\right) \tag{16}$$

Nieobciążonym estymatorem tej wariancji jest statystyka:

$$v=\frac{1}{n-1}\left(1-\frac{n}{N}\right)\cdot\sum_{i=1}^{n}\frac{\pi_{(i)}}{p_{(i)}}\left(\frac{y_{(i)}}{\pi_{(i)}}-y_{RHC}^{(z)}\right)^2 \tag{17}$$

Uwaga 10.1. Symbol $p_{(i)}$ występujący w (15) i (17) oznacza — gdy z i-tej grupy wylosowano do próby k-tą jednostkę losowania — iloraz π_k (suma prawdopodobieństw π_k dla $k\in i$-tej grupy), czyli gdy do i-tej grupy wylosowano JL o numerach $k_1, k_2, \ldots, k_{N_i}$, to $p_{(i)}=\pi_k:(\pi_{k_1}+\ldots+\pi_{k_{N_i}})$. Wówczas $\dfrac{\pi_{(i)}}{p_{(i)}}=\pi_{k_1}+\pi_{k_2}+\ldots+\pi_{k_{N_i}}$ oraz $\pi_{(i)}=\pi_k$.

10.16. Z porównania prawej strony (14) i (16) wynika, że estymator $y_{RHC}^{(z)}$ jest mniej efektywny od estymatora $y_M^{(z)}$, gdy:

$$\sum_{k=1}^{N}\pi_k^2\left(\frac{Y_k}{\pi_k}-Y\right)^2>\frac{1}{N-1}\sum_{k=1}^{N}\pi_k\left(\frac{Y_k}{\pi_k}-Y\right)^2$$

[1]) $y_{(i)}$ oznacza wartość globalną cechy Y dla zespołu wylosowanego do próby z i-tej grupy.

Można wykazać, że w pewnych przypadkach ta nierówność występuje, ale w innych przypadkach nie. Biorąc pod uwagę, że wzór (16) daje dokładną wariancję estymatora wartości globalnej Y, w przeciwieństwie do (14), który jest słuszny jedynie dla bardzo dużych N wydaje się, że należy raczej stosować schemat *RHC*, a nie schemat Madowa.

10.17. **Zadania.** 1. Populacja obejmuje $N=40$ obwodów spisowych. Dla każdego z nich $k=1, 2, ..., 40$ podano niżej: liczbę mieszkań M_k, liczbę ludności Y_k oraz liczbę X_k ludności czynnej zawodowo.

k	M_k	Y_k	X_k	k	M_k	Y_k	X_k
1	164	634	327	21	64	235	126
2	77	224	119	22	47	222	141
3	43	159	94	23	45	187	99
4	32	109	59	24	58	210	145
5	117	352	183	25	23	92	57
6	23	80	45	26	33	111	64
7	41	109	81	27	53	163	92
8	24	60	33	28	42	161	90
9	63	222	113	29	67	221	134
10	63	261	159	30	54	273	149
11	26	91	56	31	93	329	220
12	90	322	195	32	18	65	44
13	32	118	61	33	6	289	8
14	49	161	85	34	84	298	174
15	56	203	118	35	64	215	140
16	62	222	130	36	41	166	99
17	51	174	87	37	118	468	253
18	113	325	161	38	63	185	131
19	88	356	197	39	103	320	184
20	53	192	103	40	85	294	159

Załóżmy, że w celu oceny liczby ludności Y oraz liczby X ludności czynnej zawodowo losujemy z powyższej populacji próbę $n=5$ obwodów spisowych. Obliczyć wariancje oraz procentowe błędy standardowe estymatorów tych parametrów, jeśli stosujemy losowanie obwodów z prawdopodobieństwami π_k proporcjonalnymi do liczby M_k mieszkań w obwodzie: a) ze zwracaniem, b) zmodyfikowaną metodą Madowa, c) metodą *RHC*.

2. Z populacji opisanej w zadaniu 1) wylosować trzy próby $n=5$ obwodów spisowych, stosując schematy losowania jak wyżej w pkt. a)—c). Na podstawie danych każdej z tych prób ocenić parametry Y i X. Porównać te oceny z rzeczywistymi wartościami i obliczyć procentowe błędy ocen. W jakim stosunku są te błędy do obliczonych poprzednio procentowych błędów standardowych?

10.18. **LITERATURA UZUPEŁNIAJĄCA:** K. Kondrat [9] str. 128—143, Hartley H.O., Rao J.N.K. [3]. Rao J.N.K., Hartley H.O., Cochran W.G. [13].

Wykład 11. Losowanie dwustopniowe

11.1. Omówione, w poprzednich wykładach, schematy losowania próby dotyczyły schematów losowania jednostopniowego. Okazało się, że przy danej wielkości próby mierzonej **liczbą jednostek badania** losowanie indywidualne jest na ogół efektywniejsze od losowania zespołowego. Jeśli jednak w praktyce stosujemy losowanie zespołowe, a nie losowanie indywidualne, to jest to powodowane: 1) dużymi kosztami sporządzenia indywidualnego operatu losowania, 2) zwiększonymi kosztami zbierania informacji, gdy jednostki badania wylosowane do próby są rozrzucone w terenie, 3) trudniejszą organizacją zbierania danych oraz usuwania spostrzeżonych błędów nielosowych (poprawianie błędów odpowiedzi w drodze kontaktu z respondentem, zmniejszanie liczby odmów przez wielokrotne próby uzyskania odpowiedzi itp.).

Przykład 11.1. W ramach Narodowego Spisu Powszechnego 1970 po raz pierwszy zostało w Polsce szeroko ujęte badanie dzietności kobiet metodą reprezentacyjną. Losowanie indywidualne kobiet, podlegających badaniom, mogło być przeprowadzone na materiałach dokonywanego spisu. Byłoby to jednak dość skomplikowane zadanie dla ówczesnego personelu spisującego. Zdaniem kierownictwa przyjęcie takiego schematu losowania[1]) „stwarzałoby poważne trudności organizacyjne i mogło prowadzić do powstawania błędów nielosowych (wynikających z pomyłek rachmistrzów), które w rezultacie mogły w poważny sposób wpłynąć na zmniejszenie precyzji wyników badania". Zastosowano więc schemat losowania zespołowego, warstwowego, w warstwach *lpbz*. Jednostką losowania był obwód spisowy. Próba objęła około 5% obwodów, w których spisano wszystkie kobiety podlegające badaniu. W wylosowanych do próby obwodach spisowych jako rachmistrzów spisowych zatrudniono kobiety, gdzie było to możliwe pracownice służby zdrowia bądź inne osoby o wyższych kwalifikacjach do spisania informacji o dzietności.

11.2. W przypadku losowania dużych *JL* (zespołowych) wchodzące do nich jednostki badania są bardziej do siebie podobne niż to ma miejsce dla całej populacji. Wobec tego lepiej losować więcej zespołowych *JL*, ale nie brać do próby wszystkich należących do nich jednostek badania, lecz tylko ich część. To prowadzi do **dwustopniowego losowania próby.** Procedura postępowania jest następująca:

a) losujemy do próby pewną liczbę zespołowych jednostek losowania; nazwiemy to postępowanie **losowaniem pierwszego stopnia,** a losowane *JL* jednostkami losowania **pierwszego stopnia** (*JLPS*);
b) wylosowane do próby *JLPS* dzielimy na mniejsze jednostki losowania zwane **jednostkami losowania drugiego stopnia** (*JLDS*); *JLDS* mogą być jednostkami zespołowymi bądź jednostkami badania;
c) przeprowadzamy **losowanie drugiego stopnia.** Wylosowane do próby *JLDS* tworzą ostateczną próbę; wchodzą do niej te jednostki badania, które należą do wylosowanych na drugim stopniu *JLDS*.

[1]) *Narodowy Spis Powszechny 8 XII 1970. Wyniki ostateczne. Ludność*, zeszyt Nr 24. *Dzietność kobiet*. Polska cz. I. Warszawa, grudzień 1971, str. V.

11.3. Z powyższego wynika, że w przypadku losowania dwustopniowego musimy dysponować dwoma operatami losowania. Operat losowania pierwszego stopnia powinien składać się z materiałów pozwalających na odpowiednie uporządkowanie *JLPS* oraz na pełną identyfikację tych jednostek (granice wytyczające jednostki obszarowe, kiedy są one *JLPS*; cechy adresowe). W zależności od schematu losowania (nieograniczony czy warstwowy) *JLPS* powinny być w operacie losowania właściwie ponumerowane. Z tego operatu losowania wybieramy próbę pierwszego stopnia. Po wylosowaniu próby pierwszego stopnia sporządzamy operat losowania drugiego stopnia. Ten operat jest ograniczony tylko do części *JLPS*, mianowicie do tych *JLPS*, które na pierwszym stopniu losowania zostały wybrane do próby. Koszt sporządzenia operatu losowania jest więc tylko częścią kosztu, jakiego wymagałoby sporządzenie operatu losowania dla całej populacji *JL*, będących obecnie *JLDS*. Losowanie drugiego stopnia, w którym próba pierwszego stopnia odgrywa rolę „populacji", może być losowaniem zespołowym bądź indywidualnym, warstwowym bądź nieograniczonym, z jednakowymi bądź z różnymi prawdopodobieństwami wyboru.

W celu uproszczenia dalszych rozważań ograniczymy się do dwóch schematów losowania dwustopniowego:

a) losowanie pierwszego stopnia jest losowaniem zespołowym według schematu *lpbz*,
b) losowanie pierwszego stopnia jest losowaniem zespołowym z różnymi prawdopodobieństwami wyboru bez zwracania według procedury Rao-Hartleya-Cochrana.

W obu schematach losowanie drugiego stopnia jest losowaniem warstwowym (warstwy stanowią wylosowanie do próby *JLPS*), w warstwach losowaniem *lpbz*. *JLDS* są jednostkami badania.

11.4. **Oznaczenia.** Populacja obejmuje N *JLPS*, w k-tej *JLPS* znajduje się M_k jednostek badania (*JLDS*). Jednostkę badania identyfikujemy przy pomocy pary liczb (k, l), gdzie k oznacza numer *JLPS*, natomiast l numer jednostki badania (*JLDS*), w tej *JLPS*, czyli $k=1, 2, ..., N$ oraz dla danego k, $l=1, 2, ..., M_k$. Wartość cechy Y dla (k, l)-tej jednostki badania oznaczamy, jako Y_{kl}. Globalną wartość cechy Y w k-tej *JLPS* oznaczymy przez Y_k, a w całej populacji — symbolem Y:

$$Y=\sum_{k=1}^{N} Y_k=\sum_{k=1}^{N}\sum_{l=1}^{M_k} Y_{kl} \tag{1}$$

Ze względu na dwa stopnie losowania próby wprowadzamy osobne oznaczenia dla dwóch średnich:

$$\bar{Y}=\frac{Y}{N} \quad \text{oraz} \quad \bar{\bar{Y}}=\frac{Y}{M}, \quad M=\sum_{k=1}^{N} M_k \tag{2}$$

Odpowiednie oznaczenie dla średniej wartości cechy Y w k-tej *JLPS* będzie:

$$\bar{Y}_k=\frac{Y_k}{M_k} \tag{3}$$

Średnia liczba jednostek badania w *JLPS* wynosi:

$$\bar{M}=\frac{M}{N} \tag{4}$$

Wśród wariancji cechy Y rozróżniamy:
a) wariancję S_{1y}^2 pomiędzy *JLPS*

$$S_{1y}^2=S_1^2=\frac{1}{N-1}\sum_{k=1}^{N}(Y_k-\bar{Y})^2 \tag{5}$$

b) wariancję S_{2k}^2 wewnętrzną k-tej *JLPS*

$$S_{2ky}^2=S_{2k}^2=\frac{1}{M_k-1}\sum_{l=1}^{M_k}\left(Y_{kl}-\bar{Y}_k\right)^2 \quad \text{dla } k=1, 2, \ldots, N \tag{6}$$

c) średnią wariancję wewnętrzną S_{2y}^2 *JLPS*

$$S_{2y}^2=S_2^2=\frac{1}{M}\sum_{k=1}^{N}M_kS_{2k}^2 \tag{7}$$

Dla próby przyjmujemy oznaczenia małymi literami. Liczbę *JLPS* wylosowanych do próby pierwszego stopnia oznaczymy przez n. Jeżeli k-ta *JLPS* została do tej próby wylosowana, to liczbę jednostek badania (*JLDS*) wybieranych z niej do ostatecznej próby oznaczymy jako m_k. Przyjmując uporządkowanie próby pierwszego stopnia w kolejności wyboru *JLPS* zastąpimy wskaźnik k wskaźnikiem i, analogicznie dla próby drugiego stopnia wskaźnik l, zastąpimy wskaźnikiem j. Zatem $i=1, 2, \ldots, n$ oraz dla danego i, $j=1, 2, \ldots, m_{(i)}$. Subskrypt i podajemy w nawiasach dla odróżnienia symbolu m_k od symbolu $m_{(i)}$. Odpowiednio symbol $M_{(i)}$ będzie oznaczać liczbę jednostek badania w *JLPS* wybranej do próby za i-tym razem. Mamy $m_{(i)}\leqslant M_{(i)}$ oraz

$$\sum_{i=1}^{n}m_i=m \tag{8}$$

jest **łączną liczbą** jednostek badania w próbie. Para (i, j) określa jednostkę badania, pochodzącą z wybranej za i-tym razem *JLPS* oraz wybranej z niej za j-tym razem *JLDS*. Wartość cechy Y dla tej jednostki badania oznaczymy jako y_{ij} oraz średnią wartość cechy Y w *JLPS* wybranej do próby pierwszego stopnia za i-tym razem oznaczymy, jako:

$$\bar{y}_i=\frac{1}{m_{(i)}}\sum_{j=1}^{m_{(i)}}y_{ij} \tag{9}$$

Oszacowaną dla tej jednostki wartość globalną cechy Y oznaczymy symbolem $\hat{y}_i$

$$\hat{y}_i = M_{(i)} \bar{y}_i \tag{10}$$

A. SCHEMAT LOSOWANIA DWUSTOPNIOWEGO, W KTÓRYM LOSOWANIE PIERWSZEGO STOPNIA JEST *lzpbz*

11.5. Omówimy szacowanie wartości globalnej cechy Y w populacji generalnej.

Twierdzenie 11.1. Statystyka:

$$\hat{\bar{y}} = \frac{N}{n} \sum_{i=1}^{n} \frac{M_{(i)}}{m_{(i)}} \sum_{j=1}^{m_{(i)}} y_{ij} \tag{11}$$

jest nieobciążonym estymatorem wartości globalnej cechy Y

$$E(\hat{\bar{y}}) = Y$$

Wariancja tego estymatora wynosi:

$$D^2(\hat{\bar{y}}) = N^2 \left[\left(1 - \frac{n}{N} \right) \frac{S_1^2}{n} + \frac{1}{N} \sum_{k=1}^{N} M_k (M_k - m_k) \frac{S_{2k}^2}{n m_k} \right] \tag{12}$$

11.6. W praktyce przeważnie organizatorzy badania życzą sobie, aby próba była automatycznie wyważona, gdyż to znacznie upraszcza opracowanie wyników. Taką próbę otrzymamy, gdy przyjmiemy:

$$m_k = f M_k, \quad 0 < f < 1 \text{ dla ustalonego } f \tag{13}$$

to znaczy, że z każdej wylosowanej do próby *JLPS* do próby drugiego stopnia losujemy taki sam procent, $100f\%$ *JLDS*. Wówczas:

$$\hat{\bar{y}} = \frac{N}{f n} \sum_{i=1}^{n} \sum_{j=1}^{m_{(i)}} y_{ij} \tag{14}$$

oraz:

$$D^2(\hat{\bar{y}}) = N^2 \left[\left(1 - \frac{n}{N} \right) \frac{S_1^2}{n} + \bar{M}^2 (1 - f) \frac{S_2^2}{n \bar{m}} \right] \tag{15}$$

Uwaga 11.1. W praktyce, ścisłe zastosowanie (13) na ogół nie będzie możliwe. Wówczas niektóre liczby m_k, wynikające z prawej strony (13), nie będą całkowite i będą wymagały zaokrągleń, a więc próba nie będzie w pełni wyważona. Można się spodziewać, że różnica pomiędzy (11) i (14) będzie rzędu ułamka procentu, gdy n jest stosunkowo duże, np. rzędu setek.

11.7. Wzory (12) i (15) wskazują, że wariancja zależy od liczby n *JLPS* losowanych do próby oraz od układu liczb $m_1, m_2, \ldots, m_N$ bądź w przypadku próby automatycznie wyważonej od przeciętnej liczby $\bar{m}$ losowanych do próby *JLDS* z *JLPS*. Planując

badanie reprezentacyjne powinniśmy ustalić — jak wielka powinna być próba dla zagwarantowania wymaganej precyzji szacunku. Jest to możliwe, gdy wylosujemy najpierw próbę wstępną n'-elementową zgodnie z przyjętym układem losowania drugiego stopnia: $m_1, m_2, ..., m_N$. Nie podajemy wzorów statystyk, które mogą być estymatorami składników wariancji. Wzory te znaleźć można np. w [18] na str. 231.

11.8. Innym zagadnieniem jest ustalenie optymalnego układu liczb $n, m_1, m_2, ..., m_N$, minimalizującego (12). W związku z tym należy przyjąć pewną **funkcję kosztów badania reprezentacyjnego** w przypadku dwustopniowego losowania próby. Oznaczmy przez K_1 koszt jednostkowy sporządzenia operatów losowania pierwszego i drugiego stopnia wraz z kosztem wylosowania próby pierwszego stopnia, mierzony na *JLPS*. Niech K_2 będzie kosztem jednostkowym losowania próby drugiego stopnia, zebrania danych z próby oraz ich opracowania, mierzony na *JLDS*. W koszcie całkowitym K badania reprezentacyjnego pomijamy koszty stałe niezależne od wielkości próby. Wówczas przeciętny (oczekiwany) koszt badania K wyniesie:

$$K = K_1 n + K_2 \cdot \frac{1}{N} \sum_{k=1}^{N} n m_k \tag{16}$$

Zminimalizujemy wariancję (12), gdy (ograniczając się do próby automatycznie wyważonej — chociaż można podać wzory ogólne) przyjmiemy:

$$m_k = \sqrt{\frac{K_1}{K_2}} \cdot \frac{M_k}{\sqrt{\delta}} \quad \text{dla} \quad k = 1, 2, ..., N$$

$$n = \frac{K}{K_1 + K_2 \bar{M} f} \tag{17}$$

przy czym $\delta = \frac{S_1^2}{S_2^2} - \bar{M} > 0$

Łatwo zauważyć, że $f = \sqrt{\frac{K_1}{K_2 \delta}}$

11.9. W praktyce badań reprezentacyjnych okazało się, że planujący nie potrafią dobrze określić kosztów K_1 i K_2, dlatego nie korzysta się z możliwości ustalenia liczb losowań zgodnie ze wzorami (7). Ponieważ K_1 jest zwykle znacznie mniejszy niż K_2 oraz $\delta > 1$, f powinno być możliwie małym ułamkiem, natomiast powiększenie próby powinno być wykonywane przez powiększanie n, czyli liczby *JLPS*, jakie losujemy do próby.

11.10. W niniejszym wykładzie omawialiśmy szacowanie wartości globalnej Y cechy Y. Dla szacowania większości różnych parametrów populacji okazuje się, że one mają postać ilorazów wartości globalnej cechy Y przez wartość globalną cechy X.

Przykład 11.2. Populację generalną stanowią gospodarstwa domowe. Losujemy próbę dwustopniowo. *JLPS* są obwody spisowe, *JLDS* — mieszkania. Liczba gospodarstw domowych nie jest wiadoma. Dla określenia przeciętnej wielkości gospodarstwa domowego — oznaczmy ją symbolicznie jako R — posłużymy się wzorem:

$$R=\frac{Y}{X}$$

gdzie: Y oznacza liczbę osób w gospodarstwach domowych, natomiast X — liczbę gospodarstw domowych. Estymatorem jest:

$$r=\frac{\hat{y}}{\hat{x}}$$

przy czym $\hat{y}$ jest oszacowaną liczbą osób w gospodarstwach domowych, natomiast $\hat{x}$ jest oszacowaną liczbą gospodarstw domowych. Obydwa oszacowania otrzymujemy zgodnie z (11).

11.11. Jeśli próba jest dostatecznie duża, ściślej — jeśli liczba n *JLPS* jest dostatecznie duża (rzędu co najmniej kilkudziesięciu), mamy przybliżony wzór na tzw. **wariancję względną** estymatora r (kwadrat względnego błędu standardowego szacunku) $r=\frac{\hat{y}}{\hat{x}}$ parametru $R=\frac{Y}{X}$:

$$V_r^2 \approx V_{\hat{x}}^2 + V_{\hat{y}}^2 - 2_{\hat{x}\hat{y}} \tag{18}$$

gdzie: $V_{\hat{x}}^2=\frac{D^2(\hat{x})}{X^2}$, $V_{\hat{y}}^2=\frac{D^2(\hat{y})}{Y^2}$

oraz: $V_{\hat{x}\hat{y}}=\frac{Cov(\hat{x},\ \hat{y})}{XY}=\rho_{\hat{x}\hat{y}}\ \frac{D(\hat{x})\ D(\hat{y})}{XY}$

$\rho_{\hat{x}\hat{y}}$ jest współczynnikiem korelacji.

Jeśli istnieje bardzo mała korelacja pomiędzy $\hat{x}$ oraz r, a tego rodzaju związek występuje w praktyce często, można przyjąć przybliżenie:

$$V_r^2 \approx V_{\hat{y}}^2 - V_{\hat{x}}^2 \tag{19}$$

Wariancja względna ilorazu, a więc także względny błąd standardowy ilorazu jest przeważnie znacznie mniejszy niż w przypadku szacowania wartości globalnej Y.

11.12. W celu uproszczenia naszych rozważań zakładaliśmy, że *JLDS* są jednostki badania. W praktyce, jednostkami losowania drugiego stopnia są często zespoły jednostek badania.

Przykład 11.3. W czasie NSP 1988 badano metodą reprezentacyjną dzietność kobiet. Próbę losowano dwustopniowo; *JLPS* były obwody spisowe, a *JLDS* — mieszkania.

Wprowadzone wyżej wzory, dotyczące schematu losowania dwustopniowego, w którym *JLDS* są jednostki badania łatwo przetransportować na przypadek, w którym *JLDS* są zespołami jednostek badania. W tym celu Y_{kl} powinno oznaczać wartość globalną cechy Y dla (k, l)-tej *JLDS*. Wówczas przeciętna wartość cechy Y, $\bar{Y}$ będzie ilorazem Y/X wartości globalnej tej cechy przez wartość globalną cechy „liczba jednostek badania w (k, l)-tej *JLDS*".

B. SCHEMAT LOSOWANIA DWUSTOPNIOWEGO, W KTÓRYM LOSOWANIE PIERWSZEGO STOPNIA JEST LOSOWANIEM Z RÓŻNYMI PRAWDOPODOBIEŃSTWAMI WYBORU BEZ ZWRACANIA WEDŁUG SCHEMATU RAO-HARTLEYA-COCHRANA

11.13. Omówiony zostanie schemat, w którym losowanie pierwszego stopnia jest losowaniem z prawdopodobieństwami proporcjonalnymi do wielkości *JLPS*. Oznaczmy więc przez π_k:

$$\pi_k = \frac{M_k}{M} \quad \text{dla } k = 1, 2, \ldots, N \tag{20}$$

W celu dokonania losowania pierwszego stopnia najpierw przyporządkowujemy *JLPS* n grupom G_i $(i = 1, 2, \ldots, n)$ w sposób **losowy.** Zakładamy, że liczba *JLPS* w populacji, wynosząca N jest podzielna przez n. Do takiej sytuacji łatwo doprowadzić przez łączenie ze sobą pewnych zespołów jednostek badania, tworząc nowe *JLPS*. Każda z n grup składa się z tej samej liczby W *JLPS*, $W = N : n$.

Niech $p_{(i)}$ oznacza prawdopodobieństwo wylosowania tej *JLPS*, która została wylosowana do próby z i-tej grupy G_i. Jeżeli jest to k-ta *JLPS*, to $p_{(i)} = \pi_{(i)} : \sum_{l \in G_i} \pi_{(l)}$, $\pi_{(i)} = \pi_k$. Z tej *JLPS* losujemy $m_{(i)}$ *JLDS*; kolejność ich wylosowania do próby symbolizujemy numerem t, $t = 1, 2, \ldots, m_{(i)}$, a łącznie parą (i, t). Wartość cechy Y dla tej *JLDS* oznaczymy jako y_{it}. Nieobciążonym estymatorem wartości globalnej Y jest:

$$\hat{Y}_{RHC} = \sum_{i=1}^{n} \frac{M_{(i)}}{p_{(i)} m_{(i)}} \sum_{t=1}^{m_{(i)}} y_{it} \tag{21}$$

Wariancja estymatora (21) wynosi:

$$D^2(\hat{Y}_{RHC}) = \left(1 - \frac{n-1}{N-1}\right)\left(\sum_{k=1}^{N} \frac{Y_k^2}{n\pi_k} - \frac{Y^2}{n}\right) + \frac{1}{n} \sum_{k=1}^{N} \frac{M_k^2}{\pi_k}\left(\frac{1}{m_k} - \frac{1}{M_k}\right) S_{2k}^2 -$$

$$- \frac{n-1}{n(N-1)} \sum_{k=1}^{N} M_k^2 \left(\frac{1}{m_k} - \frac{1}{M_k}\right) S_{2k}^2 \left(\frac{1}{\pi_k} - N\right) \tag{22}$$

Można wykazać, że na ogół wariancja (22) jest — przy tej samej wielkości próby — mniejsza niż gdyby losowanie pierwszego stopnia było z jednakowymi prawdopodobieństwami wyboru: $\pi_k = 1:N$.

11.14. Nieobciążonym estymatorem wariancji $D^2(\hat{y}_{RHC})$ jest statystyka:

$$v = \frac{1-\frac{n}{N}}{n-1} \sum_{i=1}^{n} \frac{\pi_{(i)}}{p_{(i)}} \left(\frac{M_{(i)}\bar{y}_i}{\pi_{(i)}} - \hat{y}_{RHC} \right)^2 + \sum_{i=1}^{n} \frac{M_{(i)}^2}{p_{(i)}} \left(\frac{1}{m_{(i)}} - \frac{1}{M_{(i)}} \right) s_{2i}^2 \qquad (23)$$

s_{2i}^2 jest wariancją z próby cechy Y w i-tej *JLPS*, wylosowanej do próby z i-tej grupy G_i.

11.15. **Zadania.** 1) Populacja obejmuje $M = 5300$ indywidualnych gospodarstw rolnych, znajdujących się na terenie $N = 100$ obwodów spisowych. Na podstawie próby losowanej dwustopniowo oszacować łączną powierzchnię ogólną Y(ha) gospodarstw, jeżeli w próbie było $n = 10$ obwodów, a z każdego z nich losowano po $m_{(i)} = 5$ gospodarstw. Na I stopniu losowania zastosowano schemat *lzpbz*, a na II stopniu — z każdego wylosowanego do próby obwodu — *lpbz*. Dane próby:

i	$M_{(i)}$	y_{ij} dla $j=$				
		1	2	3	4	5
1	56	7,5	6,5	3,8	8,0	8,2
2	32	4,3	5,0	6,2	12,1	10,5
3	45	6,4	12,0	9,3	6,3	6,6
4	46	12,3	8,1	6,3	3,4	7.5
5	30	4,4	2,6	7,7	7,5	4,1
6	85	5,0	4,1	2,5	4,6	5,2
7	56	2,1	5,0	3,8	4,7	10,7
8	49	5,7	2,5	5,6	5,1	5,8
9	78	5,6	7,4	5,5	7,3	9,3
10	45	9,9	6,5	6,7	9,4	7.9

2) Zakładając dla próby dane, jak w zadaniu poprzednim, oszacować łączną powierzchnię ogólną Y(ha) gospodarstw, jeżeli losowanie I stopnia przeprowadzono według schematu *RHC*. W czasie losowego podziału obwodów na $n = 10$ zespołów p_i wyniosło:

i	1	2	3	4	5	6	7	8	9	10
$p_{(i)}$	0,085	0,057	0,116	0,122	0,138	0,136	0,063	0,120	0,047	0,116

Ocenić wariancję $D^2(\hat{y}_{RHC})$.

11.16. **LITERATURA UZUPEŁNIAJĄCA:** R. Zasępa [18] str. 224—245, [17] str. 284—365, J. Steczkowski [14] str. 253—276.

Wykład 12. Losowanie dwufazowe oraz badania powtarzalne

12.1. Analizowane dotąd schematy losowania próby dotyczyły jednofazowych badań, kiedy po wylosowaniu próby zbierano dla niej potrzebne informacje, opracowywano dane i na ich podstawie szacowano różne parametry populacji generalnej. W przypadku badań technicznie trudnych do przeprowadzenia, kosztownych, wymagających ograniczenia badania do stosunkowo małej próby, możemy precyzję ocen znacznie poprawić, stosując metodę badania dwufazowego.

12.2. W badaniu dwufazowym dzielimy badane cechy na te, które zostaną zbadane w pierwszej fazie oraz na te, które zostaną zbadane w drugiej fazie. Program pierwszej fazy badania powinien składać się z cech stosunkowo łatwych do obserwacji, dla których koszty jednostkowe badania są niskie. W drugiej fazie badania znajdą się cechy trudniejsze do obserwacji. **Losowanie dwufazowe** polega na: a) wylosowaniu „większej" próby do zebrania informacji dla pierwszej fazy badania oraz b) wylosowaniu z niej „mniejszej" próby do uzyskania informacji dla drugiej fazy badania. Próba pierwszej fazy stanowi, jak gdyby populację dla drugiej próby, a więc jej dane mogą być wykorzystane dla uzyskania jak najkorzystniejszego schematu losowania oraz dla konstrukcji najefektywniejszych estymatorów szacowanych w drugiej fazie parametrów populacji generalnej.

Uwaga 12.1. Reprezentacyjne badanie dwufazowe jest strukturalnie różne od losowania dwustopniowego; w losowaniu dwufazowym jednostki losowania są zwykle takie same, jak jednostki losowania ostatecznego stopnia do próby pierwszej fazy badania, natomiast w przypadku losowania dwustopniowego *JLDS* są podzbiorami *JLPS*.

Przykład 12.1. Typowym badaniem dwufazowym jest program Zintegrowanego Systemu Badań Gospodarstw Domowych (zob. [12] str. 269—279). Zgodnie z jego założeniami — w pierwszej fazie badań losuje się dwustopniowo tzw. „próbę matkę". *JLPS* jest rejon statystyczny, liczący minimum 250 mieszkań. W całym kraju jest ponad 24 tys. rejonów, pogrupowanych w warstwy, którymi są województwa w podziale na miasta i wieś (łącznie 98 warstw). Na taśmie magnetycznej emc zostały zapisane, dla każdego rejonu: a) cechy adresowe, b) szacunkowa liczba mieszkań. Ten rejestr stanowił operat losowania I stopnia. Do „próby matki" wylosowano cztery podpróby po 450 rejonów statystycznych (ściślej: terenowych punktów badań), czyli razem 1800 *JLPS*. Losowanie próby było warstwowe, a w warstwach z prawdopodobieństwami proporcjonalnymi do liczby mieszkań według zmodyfikowanego schematu Madowa (omówionego w wykładzie 10). Losowanie drugiego stopnia było losowaniem dwufazowym. W pierwszej fazie wylosowano podpróby mieszkań (*JLDS*): a) do badań demograficzno-społecznych objętych Zintegrowanym Systemem BGD, b) do przygotowania operatu losowania próby do badań budżetów rodzinnych (druga faza). W związku z tym w *tpb* wylosowanych na pierwszym stopniu wylosowano: 1) pięć podprób, po 5 mieszkań do badań demograficznych i społecznych, 2) 150

mieszkań do drugiej fazy badania, tj. losowania próby do badania budżetów rodzinnych. W tych ostatnich mieszkaniach zebrano informacje (pierwsza faza badania) o: a) grupie społeczno-zawodowej gospodarstw domowych, b) liczbie osób w gospodarstwie, c) dochodzie na osobę, d) powierzchni gospodarstwa rolnego, w przypadku gospodarstw domowych rolników indywidualnych. Te dane zostały wykorzystane do podziału gospodarstw domowych na domeny studiów (gospodarstwa pracownicze, chłopów-robotników, chłopskie, emerytów i rencistów); wewnątrz każdej z nich porządkowano gospodarstwa według liczby osób, a w przypadku takiej samej liczby osób — według dochodu na osobę. Gospodarstwa chłopskie porządkowano według ogólnej powierzchni gospodarstwa rolnego.

Próbę do drugiej fazy losowano systematycznie (48 gospodarstw z każdych 150 gospodarstw domowych).

Próby, o których mowa w pkt. 1) zostają wykorzystane do różnych badań dodatkowych, a więc badań drugiej fazy.

12.3. Reprezentacyjne badanie dwufazowe wymaga dwukrotnego losowania próby, ponieważ w sumie posługujemy się informacjami z dwóch prób: większej i mniejszej, będącej podpróbą większej próby. W celu ułatwienia dalszych rozważań założymy, że **losowanie próby do pierwszej fazy** badania jest *lpbz*, natomiast losowanie z niej podpróby do drugiej **fazy** badania odbywa się **według takiego samego schematu losowania bądź** jest indywidualnym **losowaniem warstwowym (z warstw** *lpbz*).

12.4. Szacujemy średnią $\bar{Y}$. Pobieramy n_1-elementową próbę (*lpbz*) dla I fazy badania. Uzyskujemy z niej ocenę $\bar{x}_1$ średniej cechy dodatkowej $\bar{X}$. Z wylosowanej próby n_1-elementowej losujemy próbę n-elementową (*lpbz*). Z mniejszej próby szacujemy $\bar{Y}$, wykorzystując oprócz średniej $\bar{x}_1$ dodatkowo średnią z próby ostatecznej $\bar{x}$ oraz $\bar{y}$, a ponadto ocenę współczynnika regresji cechy Y względem cechy X:

$$b=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \tag{1}$$

Za estymator średniej $\bar{Y}$ przyjmujemy:

$$\bar{y}''=\bar{y}+b(\bar{x}_1-\bar{x}) \tag{2}$$

Jest to estymator obciążony, zgodny. Jeżeli n jest dostatecznie duże (co najmniej 200) obciążenie estymatora jest mało istotne oraz wariancja:

$$D^2_{df}(\bar{y}'')\approx\frac{S_y^2}{n}\left[1-\rho_{xy}^2\left(1-\frac{n}{n_1}\right)\right] \tag{3}$$

gdzie: ρ_{xy} jest współczynnikiem korelacji cech X i Y.

Zauważmy, że dla $\rho_{xy}=1$ prawa strona (3) redukuje się do $\frac{S_y^2}{n_1}$. Wówczas losowanie

dwufazowe próby n_1-elementowej byłoby mniej więcej tak samo efektywne, jak losowanie jednofazowe „dużej" próby n_1-elementowej. To wskazuje, że przy dużej korelacji cech X i Y losowanie dwufazowe daje duży zysk na wariancji, która przy próbie $n<n_1$-elementowej byłaby rzędu S_y^2/n.

Przykład 12.2. Niech $N=1000000$, $n_1=10000$, $n=1000$, $\rho_{xy}=0{,}7$ oraz $S_y^2=2$. Wówczas:

$$D_{df}^2(\bar{y}'')\approx\frac{2}{1000}\left[1-0{,}49\left(1-\frac{1000}{10000}\right)\right]=0{,}001118$$

Natomiast w przypadku jednofazowym:

$$D^2(\bar{y})=\left(1-\frac{n}{N}\right)\frac{S_y^2}{n}=\left(1-\frac{1000}{1000000}\right)\cdot\frac{2}{1000}=0{,}001998$$

Zysk na wariancji:

$$\frac{[D^2(\bar{y})-D_{df}^2(\bar{y}'')]\cdot 100}{D^2(\bar{y})}=44\%$$

12.5. Wariancja $D_{df}^2(\bar{y}'')$ zależy od układu liczb (n_1, n). Niech K_1 oznacza koszt jednostkowy obserwacji statystycznej pierwszej fazy badania wraz z kosztem wylosowania próby oraz opracowania danych tej fazy, natomiast K_2 odpowiedni koszt jednostkowy drugiej fazy badania. Pomijając koszty stałe, łączny koszt K badania dwufazowego wyniesie:

$$K=K_1 n_1+K_2 n \tag{4}$$

Przy koszcie badania K określonym w (4) wariancja $D_{df}^2(\bar{y}'')$ jest najmniejsza, gdy:

$$n_1=\frac{K}{K_1+K_2\sqrt{\frac{1-\rho_{xy}^2}{\rho_{xy}^2}}\sqrt{\frac{K_1}{K_2}}} \tag{5}$$

oraz:

$$n=n_1\sqrt{\frac{1-\rho_{xy}^2}{\rho_{xy}^2}}\sqrt{\frac{K_1}{K_2}} \tag{5a}$$

Badanie dwufazowe jest efektywniejsze od badania jednofazowego, gdy spełniona jest nierówność:

$$\rho_{xy}^2>\frac{4K_1K_2}{(K_1+K_2)^2} \tag{6}$$

Przykład 12.3. Niech $K_1 = 50$ zł, $K_2 = 600$ zł. Wówczas prawa strona (6) wynosi 0,284, czyli jeśli $\rho_{xy}^2 > 0{,}284$, tj. $\rho_{xy} > 0{,}533$ to badanie dwufazowe jest efektywniejsze niż badanie jednofazowe. Gdyby $K_1 = 100$ zł $K_2 = 600$ zł, warunek (6) oznacza, że powinna być spełniona nierówność $\rho_{xy} > 0{,}7$.

12.6. Załóżmy obecnie, że losowanie próby pierwszej fazy jest, jak poprzednio, *lpbz*. Informacje o cesze dodatkowej X wykorzystujemy, dzieląc zakres zmienności tej cechy na L przedziałów i przydzielając do h-tej warstwy te jednostki z próby, dla których wartości cechy X należą do h-tego przedziału ($h = 1, 2, ..., L$). Próbę **drugiej fazy** badania losujemy stosując schemat **proporcjonalnego losowania warstwowego.** Niech w'_h jest frakcją elementów z próby pierwszej fazy należących do h-tej warstwy, natomiast $\bar{y}_h$ — średnią z próby z h-tej warstwy, uzyskaną z badania w drugiej fazie. Wówczas statystyka:

$$\bar{y}'_{(w)} = \sum_{h=1}^{L} w'_h \bar{y}_h \tag{7}$$

jest nieobciążonym estymatorem średniej $\bar{Y}$ oraz:

$$D^2_{df}(\bar{y}'_{(w)}) \approx \left(1 - \frac{n}{N}\right)\frac{S_w^2}{n} + \left(1 - \frac{n_1}{N}\right)\frac{S_b^2}{n_1} \tag{8}$$

gdzie:

$$S_w^2 = \sum_{h=1}^{L} W_h S_h^2, \quad S_b^2 = \sum_{h=1}^{L} W_h (\bar{Y}_h - \bar{Y})^2 \tag{9}$$

S_w^2 jest przeciętną wariancją wewnątrzwarstwową cechy Y, natomiast S_b^2 jest wariancją międzywarstwową tej cechy.

Ustalając optymalne n_1 i n (wzory pomijamy; można je znaleźć np. w [18] na str. 254) okazuje się, że przy danych kosztach badania reprezentacyjnego (4) losowanie dwufazowe jest efektywniejsze od losowania jednofazowego, gdy:

$$\frac{S_b}{S_w} > \frac{2\sqrt{\frac{K_1}{K_2}}}{1 - \frac{K_1}{K_2}} \tag{10}$$

Można się spodziewać spełnienia tej nierówności, gdy podział próby I fazy na warstwy według danych o cesze X będzie korzystny ze względu na badaną cechę Y, czyli gdy te cechy będą silnie skorelowane.

12.7. Dotychczas omówione zostały zagadnienia **jednorazowych** badań reprezentacyjnych. W praktyce statystycznej badania są często **powtarzalne**, co pewien okres, ze względu na zachodzące z biegiem czasu zmiany szacowanych parametrów.

W przypadku powtarzania badania reprezentacyjnego wymagają wyjaśnienia sprawy:
1) czy kolejne badanie reprezentacyjne powinno objąć tę samą próbę co badanie poprzednie?
2) jeżeli nie, to czy nie należałoby przynajmniej pewną część próby z pierwszego badania pozostawić w następnym badaniu?
3) czy i w jaki sposób można wykorzystać dane poprzednich badań w celu zwiększenia precyzji w przyszłym badaniu?

Udzielenie odpowiedzi na te pytania nie jest łatwe i wymaga poruszenia wielu zagadnień teoretycznych, między innymi dotyczących dynamiki zmian. Pominiemy więc rozważania teoretyczne, jedynie podkreślając wybrane wnioski z tych rozważań. Wypada zaznaczyć, że w badaniach teoretycznych abstrahuje się zwykle od zmian (w czasie) w składzie populacji generalnej, co jest nieadekwatne procesom zachodzącym w praktyce.

12.8. Z rozważań teoretycznych wynika, że szacując średnią wartość $\bar{Y}$ badanej cechy **w drugim momencie** (okresie), przy wykorzystaniu wyników poprzedniego badania reprezentacyjnego, **nie należy powtarzać zbyt dużej części próby**; istotny zysk na precyzji szacunku osiągamy, gdy powtarzamy co najwyżej połowę próby oraz korelacja pomiędzy wartościami cech w dwóch badanych momentach (okresach) jest co najmniej rzędu 0,7.

Natomiast inaczej przedstawia się sytuacja, gdy jesteśmy zainteresowani, w ocenie **różnicy** pomiędzy średnimi $\bar{Y}$ w dwóch różnych momentach (okresach). Wówczas najefektywniejsze jest **dwukrotne badanie** tej samej próby.

Jeżeli celem badania przeprowadzanego po raz drugi jest zarówno oszacowanie poziomu $\bar{Y}$, jak i różnicy w poziomie pomiędzy pierwszym i drugim momentem (okresem) badania powstaje dylemat, jaką część próby należy powtórzyć. Jeżeli obie oceny są jednakowo ważne, najrozsądniejszym wydaje się powtórzenie połowy próby. W dwóch innych przypadkach istniejących preferencji powtórzona część próby powinna być odpowiednio mniejsza albo odpowiednio większa.

W pewnych przypadkach zależy nam na ocenie średniej ze średnich w dwóch różnych momentach (okresach) badania. Np. szacujemy przeciętne wydatki rodziny na żywność w ciągu miesiąca i pierwszym okresem badania reprezentacyjnego był pierwszy kwartał roku, a drugim — drugi kwartał. Średnia z tych przeciętnych wydatków będzie oceną przeciętnych miesięcznych wydatków rodziny na żywność w pierwszym półroczu. Jeżeli korelacja pomiędzy szacowanymi wydatkami jest dodatnia (co z reguły zachodzi w praktyce) próba w drugim okresie badania reprezentacyjnego **nie powinna obejmować** żadnego elementu badanego w pierwszym okresie.

Wnioski, wypowiedziane wyżej, dotyczą badań powtarzalnych i można je rozszerzyć na przypadek więcej niż dwóch badań powtarzalnych.

12.9. Przedmiotem badań statystycznych mogą być zjawiska, które mogą wystąpić w dowolnym momencie, w określonej o mniej więcej stałym składzie populacji. Na przykład w populacji szpitali badanym zjawiskiem może być przyjęcie chorego do szpitala; w przypadku populacji zakładów produkujących obuwie badanym zjawiskiem może być wyprodukowanie pary obuwia itd. Celem badania statystycznego jest

wówczas oszacowanie częstości występowania danego zjawiska w pewnej jednostce czasu bądź oszacowanie natężenia zjawiska. W rozważanych przypadkach badanie statystyczne może wymagać ciągłej obserwacji statystycznej w ustalonym okresie, czyli być **badaniem ciągłym.**

12.10. Przypuśćmy, że populacja generalna obejmuje N jednostek badania, zaś badany okres (np. rok kalendarzowy) podzielony jest na M mniejszych okresów rozłącznych (np. okresów tygodniowych), które nazwiemy $t_1, t_2, ..., t_M$. Badamy cechę Y, która w okresie $l (l = 1, 2, ..., M)$ przyjmuje wartość zero, gdy w tym okresie **nie wystąpiło** badane zjawisko, natomiast wartość W, gdy pewna miara badanego zjawiska, które wystąpiło w badanym okresie co najmniej raz wynosi W. Gdy chodzi o k-tą jednostkę badania oraz l-ty okres, wartość cechy Y oznaczymy, jako Y_{kl}.

Przykład 12.4. Badamy populację rodzin, a badaną cechą są wydatki na meble. Dla k-tej rodziny $(k = 1, 2, ..., N)$ w l-tym miesiącu; Y_{kl} oznacza wydatki w danym miesiącu na meble $(l = 1, 2, ..., 12)$. Jeśli w l-tym miesiącu k-ta rodzina zakupiła mebel za 50 tys. zł, to $Y_{kl} = 50000$; gdy żadnego kupna mebli nie było w tej rodzinie, to $Y_{kl} = 0$.

12.11. Przyjmijmy, że celem badania statystycznego jest oszacowanie pewnych parametrów będących funkcjami MN zmiennych, np. średniej wartości $\bar{Y}$ cechy Y **w badanym okresie:**

$$\bar{Y} = \frac{1}{MN} \sum_{k=1}^{N} \sum_{l=1}^{M} Y_{kl}$$

bądź średniej wartości cechy Y **w pewnym podokresie badania:**

$$\bar{Y}_{\alpha} = \frac{1}{M_1 N} \sum_{k=1}^{N} \sum_{l=\alpha}^{\alpha + M_1 - 1} Y_{kl}$$

gdzie: l przebiega zbiór pewnych numerów podokresów (przeważnie kolejnych) badania, np. $l = \alpha, \alpha + 1, ..., \alpha + M_1 - 1$; $1 \leqslant \alpha$ — liczba naturalna $= \alpha + M_1 - 1 \leqslant M$ bądź różnicy $\bar{Y}_{\alpha_1} - \bar{Y}_{\alpha_2}$ średnich w dwóch różnych podokresach badania, itd.

Jeśli oszacowanie mamy wykonać metodą reprezentacyjną, powstaje zagadnienie, jak losować próbę dla każdego z M podokresów badania? Z uwagi na różnorodność szacowanych parametrów zwykle przyjmuje się, że tak samo liczną próbę bada się w każdym z kolejnych podokresów, na przykład próbę n-elementową. Takie ujęcie badania reprezentacyjnego jest korzystne z przyczyn organizacyjnych, gdyż rozkłada równomiernie w czasie pracę związaną z obserwacją statystyczną.

12.12. Badanie ciągłe metodą reprezentacyjną może być zorganizowane w różny sposób:

a) przez cały okres badania próba obejmuje te same jednostki, czyli brak jest jakiejkolwiek rotacji; badanie tak zorganizowane nazywamy **badaniem stabilnym;**
b) w każdym podokresie, z wyjątkiem początkowego, pewna frakcja P próby jest badana ponownie, a pozostała część próby obejmuje nowo wylosowane jednostki badania; takie badanie nazywamy **badaniem o częściowej rotacji;**
c) w każdym podokresie badamy inną próbę, a więc nie obejmującą jednostek raz badanych; to badanie nazywamy **badaniem o całkowitej rotacji.**

Przykład 12.5. Prowadzone w GUS badanie reprezentacyjne budżetów rodzinnych w latach 1986—1989 jest badaniem rotacyjnym o częściowej rotacji. W każdym terenowym punkcie badań wylosowano próbę 48 gospodarstw domowych. Ta próba została podzielona na **część stałą** złożoną z 4 gospodarstw badanych w tym samym kwartale (łącznie $4 \times 4 = 16$ gospodarstw domowych) oraz **część zmienną**, obejmującą po dwa gospodarstwa na kwartał ($2 \times 4 \times 4 = 32$ gospodarstwa domowe). Gospodarstwa części stałej są badane co roku w tym samym **kwartale**, natomiast gospodarstwa części zmiennej są badane **jednorazowo w określonym kwartale**. Stąd wynika, że w każdym kwartale bada się w terenowym punkcie badania 6 gospodarstw domowych, z których 4 gospodarstwa są co roku w próbie powtarzane, a 2 gospodarstwa są nowe. Próba wylosowanych 900 terenowych punktów badania obejmuje w każdym kwartale 5400 gospodarstw domowych, czyli rocznie 21600 gospodarstw domowych. W 1987 r. nastąpiło zwiększenie próby na terenach wiejskich o 50%; badanie obejmuje 1068 terenowych punktów badań.

Omawiane badanie budżetów rodzinnych jest badaniem **o rotacji kwartalnej.** Dla danych **rocznych** jest to badanie **o całkowitej rotacji.** Natomiast **dla okresu 4 lat** jest ono badaniem **o częściowej rotacji.**

12.13. Z rozważań (pkt. 12.8.) można sądzić, że gdy celem badania jest poznanie **dynamiki,** czyli ustalenie zmian w wartościach średniej, najkorzystniejsze jest badanie **stabilne,** w którym przez cały okres badania (np. roczny) obserwacja statystyczna dotyczy jednej i tej samej próby. Jeżeli natomiast celem badania jest **oszacowanie poziomów** (np. średnich wartości cech) w pewnych podokresach bądź w całym okresie badania, badanie stabilne jest mało efektywne i powinno być zastąpione badaniem rotacyjnym.

12.14. Dłuższe badanie ankietowe tych samych jednostek badania powoduje ich niechęć do uczestniczenia w badaniu od samego początku (odmowy) oraz może powodować wzrost systematycznych błędów odpowiedzi. W sumie, badanie populacji generalnej z danych wylosowanej próby ogranicza się do badania jakiejś wyodrębnionej z populacji generalnej jej części. Inaczej mówiąc, im więcej odmów czy na początku badania czy też w jego trakcie, tym bardziej obciążone pozostają wyniki badania reprezentacyjnego i jego oceny szacowanych parametrów. Dlatego badania stabilne powinny być zastępowane badaniami rotacyjnymi o możliwie krótkim okresie rotacji. Rotacja może być całkowita bądź częściowa, ale o niedużej frakcji części powtarzalnej próby.

12.15. **Zadania.** 1. Z populacji $N = 1200000$ indywidualnych gospodarstw rolnych wylosowano $n_1 = 75$ tys. gospodarstw (*lpbz*), otrzymując ocenę przeciętnej $\bar{X}$ wielkości gospodarstwa $\bar{x}_1 = 6{,}25$ ha; w drugiej fazie badania, w celu oszacowania przeciętnej $\bar{Y}$ liczby trzody chlewnej na gospodarstwo wylosowano z tej próby $n = 2055$ gospodarstw uzyskując dane:

$$\sum_{i=1}^{n} y_i = 5150 \qquad \sum_{i=1}^{n} Y_i^2 = 23864 \qquad \sum_{i=1}^{n} x_i = 12597{,}8$$

$$\sum_{i=1}^{n} x_i^2 = 114700{,}56 \qquad \sum_{i=1}^{n} x_i y_i = 45855{,}64$$

a) oszacować $\bar{Y}$ oraz $D_{af}(\bar{y}'')$; b) zakładając, że korzystamy jedynie z próby wylosowanej do drugiej fazy (co jest równoważne jednofazowemu *lpbz*) oszacować $\bar{Y}$ oraz $D(\bar{y})$. Jak bardzo różni się ocena $\hat{D}(\bar{y})$ od $\hat{D}\,\bar{y}''$? Ocenić współczynnik ρ korelacji cech X i Y.

2. Szacujemy średnią $\bar{Y}$ cechy Y w populacji $N=100$ tys. jednostek, przeznaczając na badanie $K=15$ mil zł. Jednostkowe koszty badania dwufazowego wynoszą: $K_1=1000$ zł, $K_2=20000$ zł. Losujemy próbę prostą (*lpbz*) $n_1=4635$ elementową i na podstawie danych z próby, dotyczących cechy dodatkowej X dzielimy ją na cztery warstwy, stanowiące następujące frakcje próby: 0,1; 0,3; 0,4 i 0,2. Stosując losowanie proporcjonalne warstwowe wylosowano $n=518$ elementową próbę, z której otrzymano: $\bar{y}_1=2$; $\bar{y}_2=3{,}5$; $\bar{y}_3=5$; $\bar{y}_4=7$. Oszacować średnią $\bar{Y}$ oraz procentowy błąd standardowy oceny, jeśli oszacowane wstępnie wariancje $s_b^2=64$, $s_w^2=16$. Sprawdzić spełnienie nierówności (10).

12.16. **LITERATURA UZUPEŁNIAJĄCA:** R. Zasępa [18] str. 247–270, J. Kordos [7].

Wykład 13. Planowanie i realizacja badania reprezentacyjnego

13.1. Każde badanie statystyczne ma określony cel, który powinien być możliwie jak najdokładniej sprecyzowany. Cel badania byłoby najlepiej określić, formułując główne hipotezy robocze, których weryfikację zamierzamy oprzeć na wynikach przygotowywanego badania statystycznego. W przypadku badań prowadzonych w Głównym Urzędzie Statystycznym ich wyniki są wykorzystywane przez różne instytucje administracyjne i badawcze. Często badania statystyczne są wielocelowe. Dlatego cel badania przedstawia się zwykle podając zakres informacji, które zamierzamy uzyskać, w postaci zestawu makiet tablic podstawowych, pokazujących jakie dane statystyczne planuje się uzyskać, w jakich przekrojach i klasyfikacjach. Rzecz jasna, dane te są powiązane z treścią formularzy (kwestionariuszy) badania statystycznego.

13.2. Jak już było podkreślone w pierwszym wykładzie, każde masowe badanie statystyczne zawiera pewne błędy. Planując badanie statystyczne powinniśmy zwrócić szczególną uwagę na te czynności, które przyczyniają się do redukcji błędów statystycznych. W badaniach reprezentacyjnych błędy te można ująć w dwie obszerne klasy: a) błędów losowych oraz b) błędów nielosowych. Błędy nielosowe występują zarówno w badaniach pełnych, jak i w badaniach reprezentacyjnych, czy innych badaniach częściowych.

13.3. W teorii metody reprezentacyjnej abstrahuje się od zagadnienia błędów nielosowych. Przyjmuje się w niej założenie, że próba jest wybrana ściśle według założonego schematu losowania, a jej zawartość będzie przedmiotem badania, bez

opuszczeń i bez błędów odpowiedzi. Takie podejście pozwala na uproszczenie rozważań i skoncentrowanie się na zagadnieniach precyzji szacunku metodą reprezentacyjną. Jednak wiadomo, że w praktyce badań reprezentacyjnych założenie przyjęte w teorii nie będzie nigdy spełnione i ten fakt należy odpowiednio uwzględniać przy planowaniu badania reprezentacyjnego.

13.4. Każde badanie statystyczne pociąga za sobą określone koszty. Jedynie dobra ocena elementarnych kosztów różnych czynności badania statystycznego pozwala na taką jego organizację, aby przy łącznych kosztach badania uzyskać jak najlepszą dokładność wyników. Na tę dokładność, w przypadku badania reprezentacyjnego, wywierają wpływ nie tylko błędy losowe, ale także błędy nielosowe. W pewnych przypadkach większą rolę będą odgrywały błędy losowe, w innych — błędy nielosowe. Na ogół sądzi się, że w przypadku dużych prób, obejmujących setki *JLPS* ważniejszy wpływ na dokładność badania reprezentacyjnego mają błędy nielosowe niż błędy losowe; gdy próby są stosunkowo nieduże jest przeciwnie.

13.5. Po sprecyzowaniu celu badania reprezentacyjnego plan badania reprezentacyjnego powinien m.in. ustalić: a) metodę zbierania informacji statystycznej, b) schemat losowania próby i jej wielkość, c) reguły estymacji, d) zasady opracowania danych, e) algorytm oceny precyzji wybranych ocen badania, f) plan czynności, dotyczących ocen dokładności wyników badania ze względu na błędy nielosowe.

13.6. Planując badanie reprezentacyjne należy ściśle zdefiniować badaną populację (badane populacje) oraz jednostki badania. Jest to niezmiernie ważne dla kontroli błędów nielosowych. Jednostkami badania mogą być np. zakłady pracy określonej branży przemysłu, a także pracownicy tych zakładów. Niektóre grupy jednostek badania są, z różnych przyczyn, wyłączone z badania. Np. w badaniach gospodarstw domowych nie biorą udziału tzw. gospodarstwa zbiorowe, personel sił zbrojnych itp. Jeśli takich wyłączeń nie uwzględnimy przy opracowaniu planu badania powstają przy realizacji badania błędy pokrycia oraz wzrastają błędy standardowe szacunków.

13.7. Głównymi metodami zbierania informacji statystycznej są: a) metoda obserwacji oraz pomiaru, b) metoda bezpośredniego wywiadu oraz c) metoda samospisywania. W krajach o wysokim stopniu telefonizacji stosuje się też metodę wywiadu telefonicznego. Pierwsza metoda jest metodą najdroższą; jej zaletą jest minimalizacja błędów odpowiedzi. Tańsza jest metoda bezpośredniego wywiadu często stosowana w badaniach społecznych i demograficznych; jej zaletą jest ograniczenie błędów odpowiedzi, ale nie tak duże, jak w przypadku metody pomiaru. Najtańsza jest metoda samospisywania. Jej zaletą jest możliwość ograniczenia czynności przesyłania i odbioru pocztą kwestionariusza badania. Takie podejście, z uwagi na zmniejszenie kosztów badania statystycznego, jest coraz częściej stosowane w praktyce międzynarodowej w badaniach masowych, np. w spisach powszechnych. Przekazanie kwestionariuszy do wypełnienia i ich odbiór po wypełnieniu może być wykonywany przez osobny personel badania.

W zależności od zastosowanej metody zbierania informacji statystycznej różne schematy losowania próby mogą się okazać efektywniejsze.

13.8. Planując badanie reprezentacyjne należy wybrać najefektywniejszy schemat losowania próby oraz najefektywniejszy estymator szacowanego parametru, czyli

najefektywniejszą **strategię próbkowania,** spośród możliwych strategii do zastosowania w danej sytuacji. Przypuśćmy, że szacujemy parametr T i rozważamy dwie strategie: H_1, w której estymator oznaczymy t_1 oraz schemat losowania A, symbolicznie $H_1(t_1, A)$ oraz H_2, w której estymator oznaczymy t_2 oraz schemat losowania B, symbolicznie $H_2(t_2, B)$. Mówimy, że strategia $H_1(t_1, A)$ jest efektywniejsza od strategii $H_2(t_2, B)$, jeżeli średni błąd kwadratowy estymatora t_1 jest mniejszy niż średni błąd kwadratowy estymatora t_2, czyli $MSE(t_1|A) < MSE(t_2|B)$. Jeżeli $MSE(t_1|A) = MSE(t_2|A)$ mówimy, że obie strategie są tak samo efektywne. W przypadku estymatorów nieobciążonych, czyli **strategii nieobciążonych** zamiast średnich błędów kwadratowych porównujemy wariancje $D^2(t_1|A)$ oraz $D^2(t_2|B)$. Podobnie postępujemy, gdy estymatory t_1, t_2 są obciążone, zgodne, a próba jest dostatecznie duża, aby obciążenia estymatorów uznać za nieistotne.

Jeśli estymatory t_1, t_2 wykorzystują tę samą informację dodatkową, zamiast mówić o efektywności strategii będziemy mówili o efektywności schematów losowania.

13.9. Porównując efektywność schematów losowania **przy danej wielkości próby** mamy następujące prawidłowości:

a) losowanie indywidualne jest efektywniejsze niż losowanie zespołowe, a różnica w efektywności wzrasta ze wzrostem wielkości zespołowych JL;

b) losowanie warstwowe jest efektywniejsze niż losowanie nieograniczone; różnica w efektywności jest tym większa im bardziej są jednorodne warstwy pod względem wartości badanych cech;

c) losowanie zespołowe z odpowiednio dobranymi różnymi prawdopodobieństwami wyboru jest efektywniejsze niż losowanie z jednakowymi prawdopodobieństwami wyboru;

d) losowanie dwustopniowe jest efektywniejsze niż losowanie jednostopniowe zespołów stanowiących $JLPS$;

e) losowanie bez zwracania jest efektywniejsze niż losowanie ze zwracaniem; różnica w efektywności jest tym większa im większą część populacji obejmuje próba.

13.10. Wyżej wymienione prawidłowości występują, gdy porównujemy efektywność różnych schematów losowania, przy tej samej wielkości próby. Pomijając zagadnienie błędów nielosowych, racjonalne planowanie badania reprezentacyjnego polega na wyborze najefektywniejszego schematu losowania, przy ustalonym **całkowitym koszcie badania** (po ewentualnym wyodrębnieniu kosztów czynności związanych z redukcją błędów nielosowych)[1]). Wówczas wielkość próby wypadnie różna w zależności od rozważanego schematu losowania. Np. koszt jednostkowy (na jednostkę badania) będzie przeważnie wielokrotnie wyższy w przypadku losowania indywidualnego niż zespołowego bądź z powodu wysokiego kosztu skonstruowania operatu losowania, bądź kosztu zebrania informacji statystycznej. W rezultacie próba losowana indywidualnie musiałaby być znacznie mniejsza niż próba losowana zespołowo. W pewnych konkretnych sytuacjach efektywność schematu losowania zespołowego może więc wypaść większa niż losowania indywidualnego. To wskazuje na wielką

[1]) W rzeczywistości sytuacja jest bardziej skomplikowana, gdyż przy pewnych schematach losowania próby jednostkowe koszty redukcji błędów nielosowych (np. błędów pokrycia), będą znacznie wyższe (koszty niezbędnych powtórnych wizyt jednostek rozrzuconych w terenie) niż przy innych schematach losowania próby.

wagę analizy kosztów badania, przy planowaniu badania reprezentacyjnego.

13.11. W celu wylosowania próby niezbędne są materiały, mogące stanowić, ewentualnie po pewnej przeróbce, operat losowania. W zależności od rozważanego schematu losowania operat losowania powinien mieć odpowiednią formę. Na ogół wyróżnia się **operaty listujące** (w formie wykazów jednostek badania) oraz **operaty zespołowe**, przeważnie obszarowe. Jednostki losowania powinny być w operacie uporządkowane, czyli kolejno ponumerowane oraz zawierać informacje, pozwalające na dokładne zidentyfikowanie (cechy adresowe, granice w przypadku operatu obszarowego itp.). W celu dobrego przeprowadzenia badania reprezentacyjnego operat losowania powinien być wolny od takich defektów, jak: niedokładność, niekompletność oraz podwójne figurowanie niektórych jednostek badania. Niezmiernie pożądane jest, aby operat losowania zawierał informacje o wartościach pewnych cech **dla poszczególnych** *JL*, które można wykorzystać jako wartości cech dodatkowych, dla uzyskania efektywniejszego schematu losowania próby oraz efektywniejszych reguł estymacyjnych.

13.12. Rzadko operat losowania konstruuje się niezależnie dla każdego kolejnego badania statystycznego. Przeważnie zmiany dokonuje się w związku z przeprowadzonym spisem i różnymi sposobami stara się go aktualizować w późniejszych okresach. Doświadczenie wskazuje, że jeśli operat może być aktualizowany ze względu na „nabytki" (nowe *JL* bądź jednostki badania) oraz „ubytki" (*JL* bądź jednostki badania znikające), błędy z usterek operatu nie są zwykle poważne. Aktualizacja operatu losowania jest trudniejsza w odniesieniu do operatów listowych niż w odniesieniu do operatów zespołowych.

Przykład 13.1. Sytuacja jest dość skomplikowana w powtarzalnych badaniach ekonomicznych, szczególnie dotyczących badania przedsiębiorstw i zakładów. Jak podkreślają statystycy z Biura Spisów Stanów Zjednoczonych[2]) zakłady (prywatne) wypadają albo wchodzą do działalności stosunkowo często; pozostając w działalności (handel, produkcja, usługi) mogą się rozszerzać bądź kurczyć. Spółki stale zmieniają skład przez łączenie się, nabytki i tworzenie nowych zakładów.

W miesięcznych badaniach reprezentacyjnych przedsiębiorstw handlowych wykorzystywany jest wieloraki operat losowania oparty na operacie listującym oraz operacie powierzchniowym. W USA istnieje Standardowy Wykaz Statystyczny przedsiębiorstw aktualizowany danymi z wewnętrznych urzędów podatkowych i z administracji ubezpieczeniami społecznymi, głównie w celu aktualizacji identyfikacji pracodawców (każdy zakład posiadający płatnych pracowników musi posiadać rachunek Ubezpieczenia Społecznego (Social Security) oraz odpowiadający Identyfikator Pracodawcy) do włączenia nowych firm zatrudniających płatny personel. Operat obszarowy wykorzystuje się dla pokrycia zbiorowości nie ujętej przez operat listowy. Badanie Przedsiębiorstw Hurtu spoczywa w całości na operacie listującym.

Najważniejsze są notowania „nabytków" i „ubytków". Przy obrocie zakładów około 20% rocznie, pominięcie nabytków lub ubytków poważnie wpływa na błąd oceny zarówno poziomu, jak i dynamiki.

[2]) Zob. [15].

13.13. W celu właściwego podjęcia decyzji, co do schematu losowania próby w przygotowywanym badaniu reprezentacyjnym oraz wielkości próby, konieczna jest znajomość pewnych parametrów rozkładów ważniejszych cech, takich jak: wariancje tych cech pomiędzy *JL* czy pomiędzy jednostkami badania wewnątrz *JL*, korelacja cech itp. Oceny tych parametrów mogą być dokonane na podstawie danych z poprzedniego badania populacji, o ile takie badanie było niedawno przeprowadzone. Badanie tych materiałów, nie koniecznie pełne, dostarczy wystarczająco precyzyjnych ocen dla zadecydowania, jaki schemat losowania próby należy zaakceptować. Jeśli dla szacunków wymienionych wyżej parametrów rozkładów nie istnieją materiały statystyczne niezbędne jest przeprowadzenie **badania wstępnego** na stosunkowo niezbyt dużej próbie.

13.14. Dotychczasowe doświadczenia pokazują, że problem ustalenia minimalnej wielkości próby gwarantującej żądaną precyzję szacunków, powoduje praktyczne trudności. Przede wszystkim należy: a) wybrać od kilku do kilkunastu ważniejszych, szacowanych w planowanym badaniu, parametrów populacji, b) określić maksymalnie dopuszczalne błędy przy szacowaniu parametrów, c) ustalić poziom ryzyka przekroczenia tych „maksimów"; dopiero posiadanie takich informacji oraz danych omawianych w poprzednim punkcie pozwala określić, jak wielka próba spełnia postawione warunki.

Jak wiadomo, celem przygotowywanego badania reprezentacyjnego jest z reguły szacowanie olbrzymiej liczby parametrów populacji, czyli liczby pozycji makiet tablic wynikowych. Samo przedstawienie tych tablic planującemu badanie reprezentacyjne nic nie daje. Tymczasem, tylko organizatorzy badania znając warunki mogą wybrać określoną liczbę najważniejszych parametrów.

Przypuśćmy, że taki wybór został dokonany. Konieczne jest wówczas określenie, dla każdego z wybranych parametrów, „maksymalnego" dopuszczalnego błędu, a więc wyrażenie go w procentach wartości szacowanego parametru z podaniem „orientacyjnej" wielkości tego parametru. Równocześnie należy ustalić wielkość ryzyka błędu szacunku większego niż „maksymalny". Zwykle przyjmuje się to ryzyko na poziomie $\alpha = 0{,}05$, co oznacza, że szansa na tzw. maksymalny błąd szacunku wynosi 5%; inaczej można to wyrazić: gdybyśmy przy $\alpha = 0{,}05$ szacowali 100 różnych parametrów, to dla około pięciu parametrów błąd będzie większy niż „maksymalny" (założony), a dla około 95 parametrów błąd będzie się mieścił w wyznaczonych granicach.

Uwaga 13.1. Przy określaniu żądanej precyzji estymacji należy unikać stawiania przesadnych warunków do wielkości błędu losowego estymacji. Np. gdy spodziewamy się błędu rzędu 5% wskutek błędów odpowiedzi nie należy żądać, aby „maksymalny" błąd losowy szacunku był równy 5% wartości szacowanego parametru; wystarczy ustalić granice błędu losowego na 10%. Przesadne wymaganie precyzji estymacji może przekreślić możliwość badania reprezentacyjnego bądź nie potrzebnie podrożyć jego koszty.

13.15. Stosunkowo proste jest określenie minimalnej wielkości próby, gdy stosujemy schemat indywidualnego *lpbz*, jak to pokazuje następujący przykład.

Przykład 13.2. Jak wielką próbę ($n = ?$) należy wylosować, stosując indywidualne *lpbz* z populacji $N = 100000$, aby przy szacowaniu wartości globalnej Y nie

pomylić się więcej niż o 10%, z ryzykiem $\alpha = 0{,}05$? Z wstępnej próby otrzymaliśmy $S^2 = 2$, $\bar{y} = 1{,}5$.

Postawiony warunek jest równoważny temu, aby $2D(Y) = 0{,}1Y$, czyli $4D^2(\hat{y}) = 0{,}01Y^2$.

$$D^2(\hat{y}) = N^2\left(1 - \frac{n}{N}\right)\frac{S^2}{n} \rightarrow 4D^2(\hat{y}) = 4N^2\left(\frac{1}{n} - \frac{1}{N}\right)S^2$$

Otrzymujemy w przybliżeniu:

$$4 \cdot 10^{10}\left(\frac{1}{n} - \frac{1}{10^5}\right) \cdot 2 = 0{,}01 \cdot 10^{10} \cdot 1{,}5^2$$

$$\frac{1}{n} - \frac{1}{10^5} = 0{,}0028125 \rightarrow n = 354$$

Należy wylosować próbę wielkości $n = 354$ elementy.

Przypuśćmy, że zależy nam również na oszacowaniu frakcji P jednostek wyróżnionych z błędem nie przekraczającym 5% jej wartości. Jaka powinna być minimalna wielkość próby, jeśli przypuszczamy, że P wynosi około 0,1?

$$2D(p) = 0{,}05P \quad 4D^2(p) = 0{,}0025P^2$$

Ale:
$$D^2(p) \approx \left(1 - \frac{n}{N}\right) \cdot \frac{P(1-P)}{n}$$

Otrzymujemy warunek:

$$4\left(\frac{1}{n} - \frac{1}{N}\right) \cdot 0{,}1 \cdot 0{,}9 = 0{,}0025 \cdot 0{,}1^2 \rightarrow n = 12587$$

Próba powinna obejmować $n = 12587$ elementów populacji generalnej.

13.15. Dla porównania efektywności różnych strategii amerykański statystyk L. Kish wprowadził pojęcie **efektu strategii,** miary oznaczanej symbolicznie *„Deff (·)"* od wyrażenia angielskiego „design effect". Efekt strategii oblicza się jako iloraz wariancji danej strategii $D^2(t|H)$ do wariancji strategii opartej na indywidualnym *lpbz*, $D^2(t|lpbz)$. Dla przykładu, jeśli szacujemy wartość globalną cechy Y, to dla populacji N-elementowej i próby n-elementowej otrzymamy:

$$D^2(\hat{y}|lpbz) = N^2\left(1 - \frac{n}{N}\right)\frac{S^2}{n}$$

Chcąc tę strategię porównać ze strategią losowania zespołowego prostego bez zwracania, otrzymamy:

$$Deff(\hat{y}) \approx 1 + (\bar{M} - 1)\kappa$$

gdzie: $\bar{M}$ jest przeciętną wielkością zespołu stanowiącego *JL*, natomiast κ jest miarą jednorodności wewnątrzzespołowej określonej według wzoru (10) z wykładu 9. Jeśli wielkości zespołów stanowiących *JL* nie różnią się znacznie, jak wiemy, $-\frac{1}{\bar{M}-1} \leqslant \kappa \leqslant 1$. Znając minimalną wielkość n próby, w przypadku losowania indywidualnego, możemy określić minimalną wielkość próby w przypadku losowania zespołowego, o ile potrafimy ocenić $\bar{M}$ oraz κ.

Przykład 13.3. Dane, jak w przykładzie 13.2. Jak wielka powinna być próba przy szacowaniu wartości globalnej Y?

Przyjmujemy, że oceniamy $\bar{M}=50$ oraz $\kappa=0{,}2$. Z (1) otrzymujemy:

$$Deff(\hat{y}) \approx 1+(50-1)\cdot 0{,}2=10{,}8$$

Próba powinna obejmować $354 \cdot 10{,}8$ jednostek badania, czyli 3823 jednostki badania, inaczej $m=76$ zespołowych *JL*.

Uwaga 13.2. Szacowanie wartości κ dla różnych cech badanych w czasie NSP 1970 r. przeprowadził B. Lednicki [8]. W przypadku cech demograficznych κ przyjmowało wartości od 0,06 do 0,716.

Przykład 13.4. W dniu 30 III 1974 r. przeprowadzono spis ludności i mieszkań metodą reprezentacyjną. *JL* był obwód spisowy. Losowanie było warstwowe (populacja liczyła 39 warstw), a próba z danej warstwy wynosiła 5% albo 10% obwodów. Tablice wynikowe podawały dane o liczbach ludności wyróżnionych kategorii. Ponadto, oszacowano procentowe błędy standardowe szacunków w zależności od wielkości pozycji. W oparciu o dane dla województwa koszalińskiego oraz dla m. st. Warszawy i województwa warszawskiego, w porównaniu z wielkościami błędu standardowego w przypadku losowania indywidualnego (*lpbz*), obliczono przeciętne oceny błędów standardowych[3]). Wartości *Deff* zostają podane w poniższej tablicy.

Z danych tabl. 1 wynika, że efekt *Deff* jest coraz mniejszy, gdy pozycja tablicy wynikowej maleje. Różnice pomiędzy wartościami dla próby 10% i dla próby 5% wynikają prawdopodobnie z różnic pomiędzy wartościami miary κ jednorodności wewnątrzobwodowej. Obie próby nie pochodzą z tej samej, lecz z różnych populacji. Próba 10% została wylosowana z województwa koszalińskiego, natomiast próba 5% z m. st. Warszawy i województwa warszawskiego.

Tabl. 1. Wartości *Deff* dla strategii losowania zastosowanej w Mikrospisie 1974, przy szacowaniu liczb ludności wyróżnionej kategorii

Przeciętna liczba ludności wyróżnionej kategorii	Wartości *Deff*	
	próba 10%	próba 5%
100000	40,6	8,5
50000	18,1	6,7
20000	7,6	1,7
10000	3,5	1,8
5000	2,3	1,8
2000	2,0	1,1
1000	1,4	1,2
500	1,2	1,0

13.16. W wykładzie nie omawiam zagadnienia, jak wysoką część kosztów badania reprezentacyjnego należy przeznaczyć na działalność związaną z redukcją błędów nielosowych. Jest to zagadnienie samo w sobie, poza problemami metody reprezentacyjnej.

13.18 **LITERATURA UZUPEŁNIAJĄCA:** Cz. Bracha [1] str. 22—24, 30—48, 82—84, J. Kordos [11] str. 98—109, A. Kubiczek [11] str. 160—169, J. Bielecki [11] str. 50—64.

---

[3]) Zob. R. Zasępa [10] str. 12—28.

Wykład 14. Planowanie i realizacja badania reprezentacyjnego (dok.)

14.1. W poprzednim wykładzie omówione zostały ważniejsze zagadnienia związane z planowaniem badania reprezentacyjnego, ściślej z ustaleniem schematu losowania próby i jej wielkości. Jak wiadomo, prawidłowe wykonanie tego zadania wymaga posiadania odpowiednich informacji, bez których zaprojektowany plan badania reprezentacyjnego może być daleki od optymalnego. Do nich należą:

1) określenie elementarnych kosztów jednostkowych różnych czynności badania statystycznego;

2) informacja o materiałach poprzedniego badania pełnego tej samej populacji, jaką projektujemy objąć badaniem reprezentacyjnym oraz o innych materiałach dostępnych, które mogą być wykorzystane, ewentualnie po pewnym udoskonaleniu (np. po ich aktualizacji, uporządkowaniu, wykreśleniu jednostek figurujących w nich wielokrotnie itp.), jako operat losowania próby;

3) informacja o zakresie możliwych do wykorzystania w planowanym badaniu reprezentacyjnym danych, dotyczących obecnie badanych cech bądź cech z nimi wysoko skorelowanych, dla jednostek losowania próby; te dane mogą być wykorzystane do:
   a) podziału populacji na warstwy,
   b) ustalenia prawdopodobieństw wyboru JI_i,
   c) konstrukcji estymatorów złożonych, np. estymatorów ilorazowych;

4) informacja o istniejących danych, dotyczących **parametrów rozkładu** badanych cech w projektowanym badaniu reprezentacyjnym, takich jak: wartość globalna (średnia), wariancja pomiędzy oraz wewnątrz JI_i, korelacja pomiędzy badanymi cechami itp.; brak tego rodzaju informacji jest niekorzystny i może wymagać przeprowadzenia badania wstępnego;

5) określenie najważniejszych, szacowanych parametrów oraz wielkości „maksymalnych" dopuszczalnych błędów ocen z ustalonym ryzykiem przekroczenia,

np. 100 $\alpha = 5\%$; jest to istotne dla ustalenia ostatecznej wielkości próby oraz łącznych kosztów badania reprezentacyjnego.

14.2. W poprzednim wykładzie pominięto bardzo ważne praktycznie zagadnienie — jaką część kosztów przeznaczonych na projektowane badanie reprezentacyjne należałoby przeznaczyć na czynności związane z redukowaniem błędów nielosowych? Wobec tego, że odpowiednia decyzja **powinna być podjęta** w czasie planowania badania reprezentacyjnego, nie można ustalać reguł decyzyjnych, gdyż warunki badań statystycznych są różne. W jednych badaniach — ze względu na bardzo niekorzystny wpływ błędów nielosowych na dokładność wyników — należy znaczną część środków finansowych przeznaczyć na czynności kontrolne, m.in. dodatkowe kontakty z respondentami, aby zredukować spodziewane błędy nielosowe, nawet kosztem znacznego zmniejszenia wielkości próby. W innych badaniach sytuacja może być odwrotna.

Na ogół uważa się, że w przypadku mniejszych populacji należy położyć większy nacisk na zwiększenie liczebności próby, natomiast w przypadku większych populacji błąd nielosowy przeważnie zdominuje błąd losowy, a więc należy położyć większy nacisk na redukcję błędów nielosowych.

14.3. Rozważmy, jak różne rodzaje źródeł błędów nielosowych wpływają na oceny szacowanych parametrów. Dla uproszczenia przyjmujemy, że stosujemy *lpbz* w celu oceny średniej wartości badanej cechy *Y*. Zastanowimy się najpierw, jakie nielosowe błędy badania reprezentacyjnego mogą być wywołane przez **wady operatu losowania** (przyjmujemy, że jest to operat listujący jednostek badania):

a) pominięcia pewnych jednostek należących do badanej populacji, b) dwukrotne (a nawet więcej razy) zawieranie niektórych jednostek oraz c) obejmowanie jednostek faktycznie już nie istniejących (wskutek niedokładnej aktualizacji materiałów operatu losowania).

W przypadku a) żadna z pominiętych jednostek w operacie losowania nie może być wylosowana do próby, czyli zamiast populacji będącej celem badania, badamy jakąś jej część. Jeśli rozkład cechy *Y* jest inny w pominiętej części niż w całej badanej populacji szacunki parametrów będą obciążone, a stosowane estymatory będą miały inne rozkłady niż w przypadku braku jakichkolwiek pominięć w operacie losowania. Zilustrujemy to na przykładzie.

Przykład 14.1. Populacja obejmuje $N = 11$ jednostek badania, natomiast w operacie listującym pominięto **ostatnią** jednostkę (nr 11). Losujemy próbę $n = 2$-elementową stosując estymator:

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i = \frac{1}{2}(y_1 + y_2) = \frac{1}{2}(Y_k + Y_l) \tag{1}$$

Poniżej zestawiono wartości cechy *Y* dla populacji oraz wartości estymatora (1).

Nr jedn. $k(l)$	1	2	3	4	5	6	7	8	9	10	11
$Y_k(Y_l)$	3	4	1	2	2	3	3	4	1	4	5

Wartości $\bar{y}$ dla poszczególnych prób
(bez jednostki nr 11)

l \ k	1	2	3	4	5	6	7	8	9
2	3,5	×	×	×	×	×	×	×	×
3	2,0	2,5	×	×	×	×	×	×	×
4	2,5	3,0	1,5	×	×	×	×	×	×
5	2,5	3,0	1,5	2,0	×	×	×	×	×
6	3,0	3,5	2,0	2,5	2,5	×	×	×	×
7	3,0	3,5	2,0	2,5	2,5	3,0	×	×	×
8	3,5	4,0	2,5	3,0	3,0	3,5	3,5	×	×
9	2,0	2,5	1,0	1,5	1,5	2,0	2,0	2,5	×
10	3,5	4,0	2,5	3,0	3,0	3,5	3,5	4,0	2,5

Razem $25{,}5+26{,}0+13{,}0+14{,}5+12{,}5+12{,}0+9{,}0+6{,}5+2{,}5 = 121{,}5$

$E(\bar{y}) = 121{,}5 : 45 = \mathbf{2{,}7} \neq \bar{Y} = \mathbf{2{,}91}$

Obciążenie $B = 2{,}7 - 2{,}91 = -0{,}21$, co stanowi 7,2% rzeczywistej wartości średniej $\bar{Y}$. Łatwo obliczyć, że:

$$D^2(\bar{y}) = \left(1 - \frac{2}{10}\right)\frac{S_1^2}{2} = 0{,}538 \qquad \text{(ponieważ } S_1^2 = 1{,}3444\text{)}.$$

Estymator (1) traktowany w teorii metody reprezentacyjnej jako nieobciążony, wskutek wadliwego operatu losowania staje się estymatorem obciążonym oraz $MSE(\bar{y}) = 0{,}538 + 0{,}21^2 = \mathbf{0{,}582}$.

14.4. Jeśli w operacie losowania pewne jednostki badania figurują dwukrotnie (bądź więcej razy), to te jednostki mają dwa razy (bądź więcej razy) większą szansę dostania się do próby niż pozostałe jednostki, a więc struktura prób zostanie w pewien sposób zdeformowana i estymator (1) stanie się estymatorem obciążonym, ilustruje to następny przykład.

Przykład 14.2. Korzystamy z danych poprzedniego przykładu zakładając, że jednostka o numerze 11 nie istnieje ani w populacji badanej, ani w operacie losowania. Założymy ponadto, że jednostka figurująca pod numerem 3 jest powtórzona w tym operacie pod numerem 9. Zatem populacja obejmuje faktycznie jednostki o numerach od 1 do 8 oraz 10. Szacowana średnia $\bar{Y}$ wynosi zatem $\bar{Y} = \frac{26}{9} = 2{,}89$. Wartość oczekiwana średniej z próby jest taka sama, jak poprzednio: $E(\bar{y}) = 2{,}7$. Obciążenie wynosi więc $B = 2{,}7 - 2{,}89 = -0{,}19$, co stanowi 6,6% szacowanej średniej $\bar{Y}$.

14.5. Wymieniony w punkcie 14.3 przypadek c) istnienia w operacie losowania jednostek faktycznie już nie istniejących („wybyłych") również powoduje zniekształcenie procesu estymacyjnego, jak tego dowodzi następujący przykład.

Przykład 14.3. Populacja wykazana w operacie losowania $N = 10$-elementowa, jak w przykładzie 14.1 (po usunięciu jednostki $k = 11$). Próba $n = 2$-elementowa. Szacujemy

średnią $\bar{Y}$ jednostki $k=9$, która nie wchodzi w skład badanej próby populacji. Zatem populacja badana jest $N=9$-elementowa oraz faktycznie $\bar{Y}=\frac{26}{9}=2{,}89$. Wartości $\bar{y}$ dla wszystkich możliwych do wylosowania prób 2-elementowych podane są niżej:

l \ k	1	2	3	4	5	6	7	8	9a	9b	9c
2	3,5	×	×	×	×	×	×	×	×	×	×
3	2,0	2,5	×	×	×	×	×	×	×	×	×
4	2,5	3,0	1,5	×	×	×	×	×	×	×	×
5	2,5	3,0	1,5	2,0	×	×	×	×	×	×	×
6	3,0	3,5	2,0	2,5	2,5	×	×	×	×	×	×
7	3,0	3,5	2,0	2,5	2,5	3,0	×	×	×	×	×
8	3,5	4,0	2,5	3,0	3,0	3,5	3,5	×	×	×	×
9a	?	?	?	?	?	?	?	?	×	×	×
9b	3,0	4.0	1.0	2,0	2,0	3,0	3,0	4,0	×	×	×
9c	2,0	2,5	1,0	1,5	1,5	2,0	2,0	2,5	×	×	×
10	3,5	4,0	2,5	3,0	3,0	3,5	3,5	4,0	?	4,0	2,5

Zauważamy, że 9 prób na 45 obejmuje jednostkę badania o numerze $k=9$ w operacie losowania. Wówczas powstaje pytanie — co robić, ponieważ nie można obliczyć oceny zgodnie z (1)? Dlatego dla tych prób w powyższej tablicy postawiliśmy w wierszu 9a oraz w kolumnie 9a znak zapytania. Często w takich przypadkach proponuje się dwa rozwiązania:

b) pominięcie nie istniejącej wylosowanej jednostki,

c) zastąpienie nie istniejącej wylosowanej jednostki jednostką sąsiednią.

W przypadku *b*) średnia z próby n-elementowej:

$$\bar{y}=\frac{1}{n}\sum_{i=1}^{n} y_i \text{ zostaje zastąpiona przez średnią } \bar{y}' \text{ z}$$

uzyskanych obserwacji, czyli z $n'=n-1$ obserwacji

$$\bar{y}'=\frac{1}{n'}\sum_{i=1}^{n'} y' \text{ dla próby obejmującej nie istniejącą jednostkę.} \qquad (2)$$

W naszym przypadku:

$$y'=y_{k+1} \text{ gdy } l=9 \text{ oraz } k=1,\ 2,\ \ldots\ 8 \qquad (3)$$

Wówczas zamiast znaków zapytania otrzymamy wartości ocen z próby, jak podano w wierszu i kolumnie 9b. Poniżej podajemy obliczenia wartości oczekiwanej oraz wariancji średniej $\bar{y}$.

$\bar{y}'$	1	1,5	2	2,5	3	3,5	4	Razem
Liczba prób t	1	2	6	9	12	9	6	45
$t\bar{y}'$	1	3	12	22,5	36	31,5	24	130
$t\bar{y}'^2$	1	4,5	24	56,25	108,0	110,25	96	400

$$E(\bar{y}') = \frac{1}{45}\sum t\bar{y}' = 130 : 45 = 2{,}89 = \bar{Y}$$

$$D^2(\bar{y}') = \frac{1}{45} \cdot 400 - 2{,}89^2 = 0{,}5368\ ;\ D(\bar{y}') = 0{,}733$$

Gdyby nie wchodzącą do populacji jednostkę usunąć z operatu losowania, jako nieaktualną, populacja objęłaby 9 jednostek badania po usunięciu z operatu jednostki $k=9$. Wówczas rozkład średniej $\bar{y}$ z próby charakteryzowałby się innymi parametrami (w sensie wartości) niż wyżej. Dla obliczenia $E(\bar{y})$ oraz $D^2(\bar{y})$ należy z tablicy prób usunąć wiersz, dotyczący $l=9$ oraz kolumnę dotyczącą $k=9$. Odpowiednie obliczenia podajemy niżej:

$\bar{y}$	1,5	2	2,5	3	3,5	4	Razem
Liczba prób t . . .	2	4	9	9	9	3	36
$t\bar{y}$	3	8	22,5	27	31,5	12	104
$t\bar{y}^2$ . . .	4,5	16	56,25	81	110,25	48	316

$E(\bar{y}) = 104 : 36 = 2{,}89 = \bar{Y}$

$D^2(\bar{y}) = 316 : 36 - 2{,}89^2 = 0{,}4321$

$D(\bar{y}) = 0{,}657$

Obydwa estymatory $\bar{y}'$ i $\bar{y}$ są nieobciążonymi estymatorami średniej $\bar{Y}$, jednak ich wariancje różnią się: $D^2(\bar{y}') > D^2(\bar{y})$.
Wariancja estymatora $\bar{y}'$ jest o 24% wyższa od wariancji $D^2(\bar{y})$, a odchylenie standardowe (błąd standardowy) $D(\bar{y}')$ jest o 11,6% wyższy od błędu standardowego $D^2(\bar{y})$.

Przypuśćmy, że zamiast postępowania według b) postąpimy według c), zastępując nie istniejącą jednostkę $k=9$ jednostką sąsiednią, dla której $Y=1$ (tego rodzaju sugestię często obserwujemy w praktyce!). Wówczas w tablicy prób zastąpimy dane wiersza (oraz kolumny) 9b przez liczby wpisane do wiersza (i kolumny) 9c. Estymator oznaczamy symbolem $\bar{y}''$.

$\bar{y}''$	1	1,5	2	2,5	3	3,5	4	Razem
t	1	4	7	12	9	9	3	45
$t\bar{y}''$	1	6	14	30	27	31,5	12	121,5
$t\bar{y}''^2$	1	9	28	75	81	110,25	48	352,25

$E(\bar{y}'') = 121{,}5 : 45 = 2{,}7 \neq \bar{Y} = 2{,}89$; $B = E(\bar{y}'') - \bar{Y} = -0{,}19$

$D^2(\bar{y}'') = 352{,}25 : 45 - 2{,}7^2 = 0{,}538$

$MSE(\bar{y}'') = 0{,}538 + 0{,}19^2 = 0{,}574$

$\sqrt{MSE(\bar{y}'')} = 0{,}758$

Zastępstwo nie istniejącej, ale wylosowanej do próby jednostki badania inną (np. sąsiednią) jednostką badania powoduje obciążenie estymatora $\bar{y}$ średniej $\bar{Y}$. W naszym przypadku, średni błąd kwadratowy $MSE(\bar{y})$ jest o 33% wyższy od wariancji $D^2(\bar{y})$, która odpowiada sytuacji, gdy operat losowania jest dobrze zaktualizowany i dokładny. Obciążenie B wynosi 6,6% szacowanej średniej. Widzimy, że $MSE(\bar{y}'') > D^2(\bar{y}')$, a więc **zastępstwo okazało** się w rozważanym przykładzie **mniej efektywne od pominięcia wylosowanej jednostki i stosowania estymatora określonego wzorami (2) i (3)**. Chociaż nie zawsze będziemy obserwować zależność $MSE(\bar{y}'') > D^2(\bar{y}')$, to z reguły wystąpi obciążenie estymatora. Uważamy zatem, że metody c) nie należy stosować, lecz pomijać przypadki nie istniejących jednostek, co powoduje zmniejszenie próby, a więc wzrost wariancji estymatora; jednak w przypadkach dużej próby ten wzrost nie będzie tak silnie wpływał na dokładność estymacji, jak to może wystąpić, gdy stosujemy metodę zastępstwa, pociągającą za sobą obciążenie estymatora.

Uwaga 14.1. Jeśli zastępstwo polegałoby na przyjęciu do próby sąsiedniej jednostki (w rozważanej w przykładzie sytuacji byłaby to jednostka $k = 8$ lub $k = 10$, dla których wartość cechy Y wynosi $Y_k = 4$) łatwo obliczyć, że estymator jest również obciążony. Jego obciążenie B wynosi 0,1(1) wariancja $D^2(\bar{y}'') = 0{,}4(4)$ oraz $MSE(\bar{y}'') = 0{,}457$, a więc $MSE(\bar{y}'') < D^2(\bar{y}')$.

14.6. Rozważania, ilustrowane powyższymi przykładami wskazują, jak ważna jest właściwa **aktualizacja** materiałów, stanowiących operat losowania próby. Aktualizację przeprowadzamy w krótkim odstępie czasu przed losowaniem próby. Wykonanie odpowiednich czynności kontrolnych i usunięcie zauważonych niedoskonałości operatu losowania próby przed jego wykorzystaniem przyczynia się do poprawy dokładności ocen badania reprezentacyjnego. Brak odpowiednio wykonanej aktualizacji operatu losowania będzie przyczyną dużych błędów nielosowych szacunków.

14.7. W badaniach reprezentacyjnych GUS losowanie próby odbywa się centralnie, przy zastosowaniu techniki elektronicznej, co zabezpiecza przed popełnianiem błędów przy losowaniu konkretnej próby. Niezmiernie ważna jest kontrola identyfikacji wylosowanych do próby jednostek losowania, aby obserwacja statystyczna dotyczyła jednostek faktycznie do próby wylosowanych, a nie innych, wskutek słabej identyfikowalności pewnych jednostek w terenie. Pomyłki będą powodowały błędy nielosowe, a więc pogorszenie dokładności wyników badania reprezentacyjnego.

14.8 W czasie zbierania informacji statystycznej z wylosowanej próby z różnych przyczyn mogą występować przypadki trudności uzyskania żądanych informacji, prowadzące do deformacji wyników badania. Błędy te mogą wystąpić zarówno w badaniach reprezentacyjnych, jak również w badaniach pełnych. Może występować uzyskanie kontaktu z wylosowaną do próby jednostką badania, odmowy udzielenia żądanych informacji w całości lub w części, niedokładności informacji bądź nawet udzielanie fałszywych danych. Skutki tego rodzaju błędów, zwanych **błędami odpowiedzi**, mogą poważnie zniekształcać oceny różnych parametrów. Nawet na etapie opracowania zebranych informacji mogą wystąpić błędy w czasie: a) kodowania, b) wprowadzania danych na taśmy magnetyczne, c) kontroli uzyskanych zapisów (redagowania), d) tabulacji wyników.

Od kilkunastu lat coraz więcej miejsca w literaturze naukowej poświęca się problemom redukcji tego rodzaju błędów w badaniach statystycznych. Zakres i ograniczenia czasowe niniejszego wykładu nie pozwalają na szersze omówienie tych spraw.

14.9. **Zadania.** 1. Przedsiębiorstwo uzależnia produkcję nowego typu urządzenia pomocnego w pracy gospodarstwa domowego od rozmiarów jego sprzedaży, co najmniej 5% gospodarstw zakupi dane urządzenie po określonej cenie. W związku z tym wysłano ankietę do wylosowanych (*lpbz*) $n = 10000$ gospodarstw domowych; otrzymało odpowiedź od $n_1 = 8000$ gospodarstw, wśród nich było 10% odpowiedzi pozytywnych. Ocenić „prawdopodobną" dolną granicę frakcji gospodarstw domowych, które mają zamiar zakupić urządzenie planowane do produkcji.

2. Do $n - n_1 = 2000$ gospodarstw, które odmówiły odpowiedzi w powyższej ankiecie wysłano ankietę ponownie i uzyskano $n_2 = 500$ odpowiedzi; wśród nich było $n_3 = 0$ odpowiedzi pozytywnych. Jak dodatkowa informacja wpłynie na ocenę, o którą pytano w zadaniu.

14.10. **LITERATURA UZUPEŁNIAJĄCA:** R. Zasępa [18] str. 275–290, [9] str. 28–44, J. Kordos [5], [6], R. Platek [11] str. 77–97

LITERATURA

[1] Bracha Czesław: *Wykorzystanie informacji o cechach dodatkowych w badaniach reprezentacyjnych*, ZBSE. W-wa 1987.

[2] Greń Jerzy: *Statystyka matematyczna*, PWN, W-wa 1987.

[3] Hartley H.O., Rao J.N.K.: *Sampling with Unequal Probabilities and without Replacement*, AMS. 33 (1962), str. 350—374.

[4] Jóźwiak Janina. Podgórski Jarosław: *Statystyka od podstaw*, skrypt. SGPiS. W-wa 1987.

[5] Kordos Jan: *Dokładność danych w badaniach społecznych*, Bibl. Wiad. Stat., tom 35, W-wa 1987.

[6] Kordos Jan: *Jakość danych statystycznych*. PWE. W-wa 1988.

[7] Kordos Jan: *Metoda rotacyjna w badaniach reprezentacyjnych*. Przegl. Statyst. Nr 14 (1967) str. 373—394.

[8] Lednicki Bronisław: *Badanie efektywności losowania zespołowego na podstawie Narodowego Spisu Powszechnego 1970*. Z Prac ZBSE, z. 78, W-wa 1974.

[9] *Metoda reprezentacyjna w masowych badaniach statystycznych. Teoria i praktyka*. Z Prac ZBSE, z. 122. W-wa 1981.

[10] *Metodologia badań reprezentacyjnych w GUS*. Prace Komisji Matematycznej. Bibl. Wiad. Stat. t. 29, W-wa 1979.

[11] *Problemy badań statystycznych metodą reprezentacyjną*. Bibl. Wiad. Stat. t. 36, W-wa 1989.

[12] *Problemy integracji statystycznych badań gospodarstw domowych*. Bibl. Wiad. Stat. t. 34, W-wa 1987.

[13] Rao J.N.K., Hartley H.O., Cochran W.G.: *On a simple procedure of unequal probability sampling without replacement*. JRSS (1962) seria B. 24.2 str. 482—491.

[14] Steczkowski Jan: *Zastosowanie metody reprezentacyjnej w badaniach społeczno-ekonomicznych*, PWN, W-wa 1988.

[15] *Survey Sampling and Measurement*, ed. K. Namboodiri, New York. 1978.

[16] *Wybrane problemy metodologiczne badań reprezentacyjnych*. Bibl. Wiad. Stat. t. 15, W-wa 1971.

[17] Zasępa Ryszard: *Badania statystyczne metodą reprezentacyjną*, PWN, W-wa 1962.

[18] Zasępa Ryszard: *Metoda reprezentacyjna*. PWE. 1972.

[19] *Zastosowanie metod matematycznych w statystyce*, Bibl. Wiad. Stat. t. 7. W-wa 1969.

[20] *Zastosowanie metody reprezentacyjnej w badaniach statystycznych GUS* (1981—1986), Z Prac ZBSE, z. 166, W-wa 1987.

Dotychczas w serii
BIBLIOTEKA WIADOMOŚCI STATYSTYCZNYCH
ukazały się następujące pozycje:

1967 r.

Tom 1. Wybrane problemy statystyki w Polsce
Tom 2. Zagadnienia statystyki rolniczej
Tom 3. Statystyka regionalna. Aktualny stan i problemy rozwoju
Tom 4. Problemy demograficzne Polski Ludowej

1968 r.

Tom 5. Bilanse gospodarki narodowej

1969 r.

Tom 6. Problemy demograficzne Ziem Zachodnich i Północnych PRL
Tom 7. Zastosowanie metod matematycznych w statystyce
Tom 8. Problemy demograficzne Kielecczyzny
Tom 9. Mierniki rozwoju regionów

1970 r.

Tom 10. Założenia programowe i organizacyjne Narodowego Spisu Powszechnego w 1970 r.
Tom 11. Wybrane problemy prognoz statystycznych
Tom 12. Aktualne problemy statystyki regionalnej w krajach europejskich
Tom 13. Teoretyczne i metodologiczne problemy statystyki społecznej

1971 r.

Tom 14. Badania statystyczne metodą reprezentacyjną w krajach socjalistycznych
Tom 15. Wybrane problemy metodologiczne badań reprezentacyjnych
Tom 16. Rola i zadania statystyki państwowej w planowaniu i zarządzaniu gospodarką narodową w krajach RWPG

1972 r.

Tom 17. Aktualne problemy statystyki państwowej
Tom 18. Eksperymentalne badania budżetów rodzinnych metodą rotacyjną

1973 r.

Tom 19. Problemy demograficzne województwa zielonogórskiego
Tom 20. Podstawowe problemy rozwoju statystyki regionalnej
Tom 21. Stan i perspektywy rozwoju statystyki w Polsce
Tom 22. Rozwój regionalny Polski w świetle badania dochodu narodowego

1974 r.

Tom 23. Problemy mierników poziomu życia ludności
Tom 24. Aktualne problemy demograficzne kraju

1976 r.

Tom 25. Statystyka i ekonometria w Polsce Ludowej
Tom 26. Zagadnienia metodologiczne statystyki społeczno-demograficznej

1978 r.

Tom 27. Rola młodzieży w życiu społeczno-gospodarczym kraju
Tom 28. Narodowy Spis Powszechny 7 grudnia 1978 r. Metodologia i organizacja
Tom 29. Metodologia badań reprezentacyjnych w GUS Prace Komisji Matematycznej

1979 r.

Tom 30. Narodowy Spis Powszechny 1978 jako źródło informacji o migracjach
Tom 31. Statystyka i ekonometria w Polsce Ludowej (wyd. drugie — zmienione)

1981 r.

Tom 32. Tematyka i organizacja spisów powszechnych w Polsce

1983 r.

Tom 33. Problemy demograficzne województwa lubelskiego

1987 r.

Tom 34. Problemy integracji statystycznych badań gospodarstw domowych
Tom 35. Dokładność danych w badaniach społecznych (autor Jan Kordos)

1989 r.

Tom 36. Problemy badań statystycznych metodą reprezentacyjną
Tom 37. Badanie budżetu czasu ludności

1990 r.

Tom 38. Statystyczne metody badania cen

1991 r.

Tom 39. Zarys metody reprezentacyjnej (autor Ryszard Zasępa)