83524
Biblioteka

WIADOMOŚCI STATYSTYCZNE

WYBRANE PROBLEMY METODOLOGICZNE BADAŃ REPREZENTACYJNYCH

Tom 15 WARSZAWA 1971

EGZEMPLARZ BEZPŁATNY

WYBRANE PROBLEMY METODOLOGICZNE BADAŃ REPREZENTACYJNYCH

•

SEMINARIUM
21—24 IV 1970

GŁÓWNY URZĄD STATYSTYCZNY
ZAKŁAD BADAŃ STATYSTYCZNO-EKONOMICZNYCH

KOMITET REDAKCYJNY BIBLIOTEKI
WIADOMOŚCI STATYSTYCZNYCH

Redaktor naczelny
mgr Henryk Białczyński

Redaktorzy:
mgr Jan Iszkowski
mgr Krystyna Jacek
dr Zbigniew Smoliński
dr Krzysztof Zagórski

Sekretarz redakcji
mgr Halina Cieślik

Prace redakcyjne:
dr Jan Kordos
mgr Jolanta Pawłowska
mgr Jadwiga Rybacka
Henryka Mierowska
(redaktor techniczny)

Okładkę projektował
mgr Krzysztof Dobrowolski

GŁÓWNY URZĄD STATYSTYCZNY
BIBLIOTEKA

83524

Przedruk w całości lub w części oraz wykorzystanie danych statystycznych w druku dozwolone wyłącznie z podaniem źródła.

Wydawca: GŁÓWNY URZĄD STATYSTYCZNY — Departament Wydawnictw, Warszawa, al. Niepodległości 208. Nakład 660 egz., ark. druk. 9,5. Papier IV kl. 70 g, 61 × 86. Przekazano do składu w lipcu 1971 r. Podpisano do druku i druk ukończono w październiku 1971 r. Cena zł 25.—U-76
Druk. Zarząd Drukarni Statystycznej, Warszawa. Zam. 454
Informacje w sprawach zbytu publikacji tel. 25-36-24 25-32-41 w. 285.

SPIS TREŚCI

Str.

Wstęp 5

Przemówienie wstępne 7

Lista uczestników seminarium 8

Część I. Referaty

Jan Kordos — Możliwości szerszego stosowania metody reprezentacyjnej w badaniach statystycznych krajów-członków RWPG (Polska) 11

Johan Carstens — Przybliżona metoda obliczania błędu systematycznego estymatora ilorazowego (NRD) 31

Wacław Czermak — O losowaniu niezależnym ze zwracaniem (Czechosłowacja) 40

Maria Diniewa — Niektóre wstępne badania dotyczące planowania demograficznych mikrospisów w LRB (Bułgaria) 87

Jerzy Greń, Tadeusz Groszek — Program na EMC wyznaczania optymalnej lokalizacji próby ze względu na wiele cech w losowaniu warstwowym (Polska) 93

Erhard König — Problemy zastosowania metody reprezentacyjnej w badaniach warunków pracy i bytu ludności z uwzględnieniem okresu perspektywicznego 1971—1975 (NRD) 102

Jan Kordos — Potrzeba szerszego zastosowania metody rotacyjnej w badaniach reprezentacyjnych (Polska) 124

Iwan Wieniecki — Eksperymentalne badania budżetów rodzinnych metodą obserwacji migawkowych (ZSRR) 134

Część II. Ważniejsze wypowiedzi w dyskusji

Heinrich Schwarz — Porównanie efektywności estymatora prostego i estymatora ilorazowego za pomocą współczynnika regresji . . 140

Wadim Zagradkin — Kilka uwag o eksperymentalnych badaniach budżetów rodzinnych metodą rotacyjną na terenie Polski . . . 143

Ryszard Zasępa — Zastosowanie metody reprezentacyjnej do badania błędów nielosowych 148

Ryszard Zieliński — Losowanie próbek za pomocą emc 150

WSTĘP

W dniach 21—24 kwietnia 1970 r. odbyło się w Warszawie międzynarodowe seminarium na temat zbadania możliwości szerszego niż dotychczas zastosowania metody reprezentacyjnej w badaniach centralnych urzędów statystycznych krajów-członków Rady Wzajemnej Pomocy Gospodarczej. Pierwsza część materiałów z tego seminarium, zawierająca opis metodologii badań reprezentacyjnych prowadzonych przez centralne urzędy statystyczne krajów-członków RWPG oraz Jugosławii, została opublikowana w tomie 14 „Biblioteki Wiadomości Statystycznych". Tom, który oddajemy do rąk czytelnika, obejmuje: referat podstawowy pt.: *Możliwości szerszego stosowania metody reprezentacyjnej w badaniach statystycznych krajów-członków RWPG* (Polska), referaty metodologiczne zgłoszone przez przedstawicieli poszczególnych krajów oraz ważniejsze wypowiedzi uczestników seminarium.

W przeprowadzonej dyskusji wielokrotnie zwracano uwagę na potrzebę prowadzenia prac metodologicznych, badań eksperymentalnych oraz na konieczność systematycznej wymiany doświadczeń między poszczególnymi krajami. Podkreślono konieczność podjęcia specjalnych studiów nad dokładnością wyników badań statystycznych (rzetelnością danych) prowadzonych metodą reprezentacyjną, szerszego niż dotychczas posługiwania się elektronicznymi maszynami cyfrowymi przy losowaniu próbek, opracowywaniu wyników i ocenie ich precyzji. Dużo miejsca poświęcono problemom teoretycznym, szczególnie po przedstawieniu referatu doc. dr W. Czermaka (Czechosłowacja). Wskazano również na potrzebę prowadzenia równoległych prac w zakresie teorii i praktyki metody reprezentacyjnej.

Na podstawie przedłożonych referatów oraz w wyniku dyskusji przyjęto następujące wnioski:

a) uznano za celowe prowadzenie systematycznej wymiany informacji o doświadczeniach (eksperymentach) prowadzonych w poszczególnych krajach w zakresie badań reprezentacyjnych, a zwłaszcza informacji uwzględniającej podstawy teoretyczne tych badań. W tym celu postuluje się szerzej wykorzystać „Informacyjny Biuletyn Statystyczny" Sekretariatu RWPG;
b) w celu rozpracowania teorii metody reprezentacyjnej i znacznego rozszerzenia zakresu jej praktycznego stosowania, a także w celu popularyzacji tej metody należałoby rozważyć możliwości zorgani-

zowania — w centralnych urzędach statystycznych krajów-członków RWPG — stałych grup (wydziałów) specjalistów zajmujących się badaniami reprezentacyjnymi;

c) w celu zwiększenia efektywności badań reprezentacyjnych uznano za niezbędne rozszerzenie badań eksperymentalnych, pogłębienie studiów nad schematami losowania, metodami estymacji, techniką losowania próbek, sprawdzenia jej reprezentatywności, analizą precyzji oszacowań; równocześnie wskazano, iż należy zwrócić szczególną uwagę na problem błędów nielosowych, badając źródła ich powstawania na różnych etapach badań statystycznych;

d) przy planowaniu badań reprezentacyjnych należy uwzględniać — w większym zakresie niż dotychczas — koszt tych badań, co umożliwi wybór bardziej efektywnych schematów losowania;

e) jako szczególnie istotne uznano zwrócenie większej uwagi na przygotowanie kadr statystycznych i ekonomicznych w zakresie zastosowań metody reprezentacyjnej w badaniach społeczno-ekonomicznych;

f) podkreślano potrzebę stosowania w szerszym zakresie w różnych dziedzinach badań statystycznych metody rotacyjnej, a także pogłębiania prac teoretycznych w tej dziedzinie;

g) wskazano na potrzebę popularyzowania metody reprezentacyjnej wśród użytkowników danych statystycznych.

Przebieg seminarium dowiódł, że metoda reprezentacyjna stosowana jest już obecnie w wielu dziedzinach badań statystycznych prowadzonych przez poszczególne kraje należące do RWPG, a w przyszłości należy oczekiwać rozszerzenia zakresu jej zastosowań wraz ze wzrostem zapotrzebowania na wszechstronną informację statystyczną.

Publikacja została przygotowana przez dr Jana Kordosa i mgr Jolantę Pawłowską z Pracowni Metod Matematycznych Zakładu Badań Statystyczno-Ekonomicznych GUS.

ZAKŁAD BADAŃ
STATYSTYCZNO-EKONOMICZNYCH

PRZEMÓWIENIE WSTĘPNE

Otwierając seminarium poświęcone zbadaniu możliwości szerszego stosowania metody reprezentacyjnej w badaniach centralnych urzędów statystycznych krajów-członków RWPG, chciałbym serdecznie powitać wszystkich uczestników seminarium reprezentujących poszczególne kraje należące do RWPG oraz delegację Federalnego Instytutu Statystycznego Jugosławii. Witam także przedstawicieli reprezentujących różne ośrodki naukowe kraju.

Przedmiotem obrad są problemy dość złożone, których przedyskutowanie, a następnie rozwiązanie i stopniowe wdrożenie do praktyki może przyczynić się do ogólnego rozwoju statystyki.

Stały wzrost zapotrzebowania na informacje statystyczne z różnych dziedzin życia gospodarczego i społecznego stwarza konieczność poszukiwania najbardziej efektywnych metod zbierania danych statystycznych. Metody te powinny zapewnić uzyskanie żądanych informacji w krótkim czasie, zebranych niewielkim nakładem pracy oraz zgodnych z rzeczywistością. Wśród tych metod szczególne miejsce zajmuje metoda reprezentacyjna, która znajduje coraz szersze zastosowanie w praktyce badań statystycznych różnych krajów. Istnieje jednakże wiele problemów metodologicznych i organizacyjnych, od których zależy efektywne stosowanie tej metody na szerszą skalę. Problemom tym poświęcone jest dzisiejsze seminarium, które powinno umożliwić po pierwsze — wymianę doświadczeń uzyskanych w tym zakresie przez poszczególne kraje, po drugie — ustalenie form współpracy na przyszłość.

Jestem przekonany, że materiał opracowany na seminarium ułatwi przedyskutowanie węzłowych, istotnych zagadnień dotyczących efektywnego stosowania metody reprezentacyjnej w badaniach statystycznych.

Życzę uczestnikom seminarium owocnych obrad i przyjemnego pobytu w naszym kraju.

dr Stanisław Róg
Wiceprezes GUS

LISTA UCZESTNIKÓW SEMINARIUM

Bułgarska Republika Ludowa

M. Diniewa
W. Żekowa

Czechosłowacka Socjalistyczna Republika

W. Czermak
M. Szimek
J. Wytlaczil

Socjalistyczna Federacyjna Republika Jugosławii

N. Marković
M. Obradović

Mongolska Republika Ludowa

B. Dorżgotow
A. Nijamdulam

Niemiecka Republika Demokratyczna

E. König
G. Weruschek
J. Carstens
H. Schwarz
P. Zaumseil

Polska Rzeczpospolita Ludowa

J. Kordos
Z. Pawłowski
J. Nąć
J. Pawłowska
H. Kowalska

Socjalistyczna Republika Rumunii

S. Munteanu
J. Lehkun

Węgierska Republika Ludowa

K. Szabo
L. Vita
A. Soiom

Związek Socjalistycznych Republik Radzieckich

J. Matiucha
J. Wieniecki

Sekretariat RWPG

D. Georgiewa

Jan KORDOS

MOŻLIWOŚCI SZERSZEGO STOSOWANIA METODY REPREZENTACYJNEJ W BADANIACH STATYSTYCZNYCH KRAJÓW—CZŁONKÓW RWPG

1. Uwagi wstępne

Stały wzrost zapotrzebowania na informacje statystyczne z różnych dziedzin życia gospodarczego i społecznego stwarza konieczność poszukiwania nowych metod zbierania danych statystycznych. Metody te powinny zapewnić uzyskanie żądanych informacji w krótkim czasie, a przy tym zebranych niewielkim nakładem pracy i zgodnych z rzeczywistością. Wśród tych metod szczególne miejsce zajmuje metoda reprezentacyjna, stosowana coraz częściej w praktyce badań statystycznych różnych krajów.

W badaniach statystycznych prowadzonych przez centralne urzędy statystyczne krajów—członków RWPG można zauważyć systematyczny wzrost zainteresowania zastosowaniem metody reprezentacyjnej. Metoda ta coraz częściej stosowana jest w różnych dziedzinach badań statystycznych, takich jak badania rolnicze, badania warunków bytu ludności itp. Jednakże możliwości jej zastosowania są znacznie szersze, a stałe zapotrzebowanie w tych krajach na zwiększenie zakresu informacji statystycznych odpowiednio przetworzonej wymaga, aby tym problemom w przyszłości poświęcić więcej uwagi. Dlatego zagadnieniami tymi zainteresowała się ostatnio Stała Komisja Statystyczna Rady Wzajemnej Pomocy Gospodarczej.

W opracowaniu tym zostały przeanalizowane dotychczasowe kierunki zastosowań metody reprezentacyjnej w badaniach statystycznych krajów—członków RWPG oraz rozważone możliwości jej szerszego zastosowania w przyszłości. Przedyskutowano też podstawowe problemy metodologiczne warunkujące dalszy rozwój badań reprezentacyjnych w statystyce. Nieco miejsca poświęcono również zagadnieniom organizacyjnym związanym z przeprowadzaniem badań statystycznych metodą reprezentacyjną. W zakończeniu podano wnioski wynikłe z przeprowadzonej analizy.

2. Wzrost zapotrzebowania na informacje statystyczne a badania reprezentacyjne

Współczesne metody planowania i zarządzania gospodarką narodową wymagają dostarczania organom zarządzania i organom planującym danych i analiz statystycznych, opracowywanych na zasadach naukowych, prawidłowo i obiektywnie odzwierciedlających daną rzeczywistość. Zakres informacji statystycznych musi być znacznie szerszy niż przy metodach tradycyjnych. Podobnie jak bez zmodernizowania statystyki nie jest możliwy postęp w kierowaniu i zarządzaniu gospodarką narodową, bez podniesienia jakości danych statystycznych nie jest możliwe uzyskanie zamierzonej jakości systemu i procesu planowania. Potrzebne są tu już nie tylko informacje w formie zagregowanej, w postaci prostych zestawień tabelarycznych, ale również informacje znacznie przetworzone przy zastosowaniu odpowiednich metod matematycznych i elektronicznej techniki obliczeniowej, a do tego celu niezbędne są informacje jednostkowe, tj. dotyczące poszczególnych jednostek badania. Informacje takie można wprawdzie uzyskać z badań wyczerpujących, ale często — ze względu na cel analizy — bardziej efektywna okazuje się metoda reprezentacyjna. Jak już wspomniano, efektywne planowanie wymaga szybkiego dostarczenia informacji statystycznych organom zarządzania i planowania, aby wyniki obliczeń mogły być wykorzystane na bieżąco. Dane te powinny być zgodne z rzeczywistością. Z praktyki wiadomo, że każde badanie statystyczne obciążone jest, w mniejszym lub większym stopniu, różnego rodzaju błędami. Rozmiar tych błędów można znacznie zmniejszyć przez zastosowanie metody reprezentacyjnej, pozwalającej zmniejszyć wydatnie liczebność badanej zbiorowości i tym samym poświęcić więcej czasu na jej dokładniejsze zbadanie, a także na bardziej staranne przeszkolenie personelu przeprowadzającego badania. Tym sposobem można również do pewnego stopnia zredukować różnego rodzaju błędy występujące w masowych badaniach statystycznych.

Stosowanie metody reprezentacyjnej pozwoli rozszerzyć zakres informacji statystycznych w porównaniu z zakresem informacji uzyskiwanych przy badaniach wyczerpujących, a zatem uzyskać szereg informacji niezbędnych dla potrzeb planowania i zarządzania gospodarką narodową.

3. Dotychczasowe kierunki zastosowań metody reprezentacyjnej w badaniach statystycznych krajów-członków RWPG

W badaniach statystycznych krajów należących do RWPG metoda reprezentacyjna stosowana jest od wielu lat [6]. Jednakże w początkowym okresie, tj. przed 20 laty, badania prowadzone tą metodą były

dość sporadyczne i ograniczały się w zasadzie do takich dziedzin, jak rolnictwo i budżety rodzinne. Głębsze zainteresowanie zastosowaniami metody reprezentacyjnej w badaniach statystycznych dało się dopiero zauważyć w ciągu ostatniego dziesięciolecia. Rozszerzył się znacznie zakres badań i zwiększyła się ich liczba chociaż do dziś jeszcze w niektórych dziedzinach badań statystycznych metoda reprezentacyjna nie była stosowana. Wydaje się więc celowe przeanalizowanie niżej podanych głównych kierunków zastosowania tej metody badań w krajach należących do RWPG.

Badania z zakresu rolnictwa. W badaniach tych metodę reprezentacyjną stosowano najwcześniej i najczęściej we wszystkich krajach należących do RWPG. Była ona stosowana w badaniach pogłowia zwierząt, w kontrolnych pomiarach spisów rolnych, w badaniach struktury zasiewów oraz wielkości plonów, w badaniach mleczności krów, produkcji rolniczej itd.

Badania warunków bytu ludności. W tej dziedzinie metodę reprezentacyjną najwcześniej stosowano w badaniach budżetów rodzinnych. Badania takie prowadzone są we wszystkich krajach—członków RWPG. Należy tu również wspomnieć o prowadzonych co kilka lat badaniach dochodów ludności i składu gospodarstw domowych oraz ich warunków mieszkaniowych. Badania takie prowadzono w Bułgarii, na Węgrzech, w NRD, Polsce, ZSRR i Czechosłowacji. Uzupełniły one w znacznym stopniu braki informacji z zakresu warunków bytu ludności i były przydatne w polityce gospodarczej i społecznej.

Wstępne opracowanie wyników spisów powszechnych. Metodę reprezentacyjną do opracowania wstępnych wyników powszechnego spisu ludności stosowano już w latach pięćdziesiątych. Zalety tych opracowań są powszechnie znane. Pozwoliły one w krótkim czasie po spisie dostarczyć podstawowych informacji różnym użytkownikom.

Badania kontrolne. Badania kontrolne metodą reprezentacyjną prowadzono w spisach ludności, spisach rolnych oraz innych badaniach statystycznych. Chodziło tu głównie o zbadanie jakości danych uzyskanych z badań podstawowych oraz wykrycie źródeł ważniejszych błędów.

Badania demograficzne. Metodę reprezentacyjną w badaniach demograficznych stosowano tylko w niektórych krajach. Badano wewnętrzny ruch wędrówkowy ludności, przyczyny rozwodów, wpływ czynników demograficznych na wykształcenie, przyczyny zgonów niemowląt itp.

Badania statystyczno-socjologiczne. W badaniach tego typu metodę reprezentacyjną stosowano w wielu krajach. Należy tu wymienić badania budżetu czasu ludności, sposobów spędzania urlopów i wakacji, oddziaływań radia i telewizji, czytelnictwa prasy itd. Badania te

prowadzone w różnych krajach różniły się istotnie między sobą zarówno zakresem, jak i rozwiązaniami metodologicznymi. Na ogół prowadzono je dopiero w ostatnich latach.

Powyżej wymieniono tylko sześć głównych kierunków zastosowań metody reprezentacyjnej, jakie można było wyróżnić z ogółu badań przeprowadzonych w ostatnim okresie przez centralne urzędy statystyczne krajów należących do RWPG. Należy jednak dodać, że w wielu krajach przeprowadzono pojedyncze badania z zakresu statystyki obrotu towarowego, przemysłu i in., zostały one szczegółowo przedstawione w specjalnej publikacji [6]. Badania te różniły się znacznie między sobą przyjętymi rozwiązaniami metodologicznymi, takimi jak schemat losowania, liczebność próbki, metody estymacji, analiza precyzji uzyskanych wyników itp. Ponieważ przyjęte rozwiązania metodologiczne są bardzo ważne dla całości badania reprezentacyjnego, jego efektywności i praktycznego wykorzystania wyników, dalszą część niniejszego opracowania poświęcono rozważaniom na temat poprawności rozwiązań poszczególnych problemów oraz występujących trudności przy ich rozwiązaniu.

4. Wybrane problemy metodologiczne

Rozważając możliwości szerszego zastosowania metody reprezentacyjnej w badaniach statystycznych, należy przede wszystkim ustalić, w jaki sposób rozwiązanie występujących problemów wpływa na powodzenie badania metodą reprezentacyjną. Dlatego zostaną rozpatrzone zagadnienia związane z podjęciem decyzji o zastosowaniu metody reprezentacyjnej, wyborem odpowiedniej metody badania, a w szczególności zagadnienie minimalnej liczebności próbki, operatu losowania, schematu losowania, sposobu losowania, estymacji oraz analizy precyzji wyników.

4.1. Badanie wyczerpujące a badanie reprezentacyjne

Ogólnie panuje pogląd, że badanie reprezentacyjne jest znacznie tańsze i mniej pracochłonne niż badanie wyczerpujące. W zasadzie jest on słuszny, z tym jednak, że udziały nakładów na przygotowanie badania reprezentacyjnego i na przygotowanie badania wyczerpującego różnią się dość znacznie na poszczególnych etapach przygotowania i opracowania obydwu badań. Badanie reprezentacyjne wymaga pewnych dodatkowych nakładów pracy, które nie występują w badaniu wyczerpującym. Chodzi o etap przygotowawczy, który wymaga opracowania schematu losowania, ustalenia minimalnej liczebności próbki, sporządzenia operatów losowania i przeprowadzenia samego losowania jednostek. Proporcje nakładów zmieniają się

na korzyść badania reprezentacyjnego dopiero przy przeprowadzaniu badania w terenie. Łączny koszt badania reprezentacyjnego jest prawie zawsze mniejszy niż badania wyczerpującego. Powstaje więc pytanie, w jakich przypadkach badanie wyczerpujące można zastąpić badaniem reprezentacyjnym. Na to pytanie, nie można udzielić jednoznacznej odpowiedzi. Zależy to bowiem od całego szeregu warunków, takich jak cel badania, dostępne środki finansowe, przedmiot badań itp. Z góry natomiast można powiedzieć, że w niektórych badaniach nie można stosować metody reprezentacyjnej. Dotyczy to tych badań, w których wyniki muszą dotyczyć każdej jednostki lub muszą być opracowane w bardzo szczegółowych przekrojach. W pozostałych przypadkach przed wystąpieniem od badań wyczerpujących należałoby odpowiedzieć na następujące pytania:

a) czy celem badania jest uzyskanie danych dla każdej jednostki badania czy tylko dla całej zbiorowości lub ważniejszych jej grup?
b) czy wyniki uzyskane z badania nie mogą być obarczone żadnymi błędami?
c) czy badanie wyczerpujące dostarczy dokładnych informacji?
d) w jakich przypadkach badanie reprezentacyjne byłoby bardziej wskazane niż badanie wyczerpujące?

Pierwsze pytanie jest bardzo ważne, gdyż — jak już mówiliśmy — konieczność uzyskania danych dla każdej jednostki badania wyklucza możliwość zastosowania metody reprezentacyjnej. W praktyce statystycznej jednak rzadko przeprowadza się badania, których wyniki musiałyby być prezentowane dla każdej jednostki. Zwykle — a zwłaszcza dla planowania i kierowania gospodarką narodową — konieczne są dane charakteryzujące pewną zbiorowość generalną lub takie, które umożliwiają badanie prawidłowości występujących w tej zbiorowości. W tych przypadkach metoda reprezentacyjna jest na ogół bardziej odpowiednia niż metoda badań wyczerpujących.

Dokładne określenie celu badania pozwoli udzielić odpowiedzi na drugie pytanie. Należy zaznaczyć, że zwykle cel badania nie zmusza do zbierania bardzo dokładnych danych. Pewien margines błędu jest dopuszczalny, a zwłaszcza wówczas gdy wyniki badania mają być wykorzystane w planowaniu, które zakłada pewien element niepewności. Żądanie bardzo dokładnych danych z przeszłości nie byłoby zatem uzasadnione, chociaż powinna być określona jakaś granica, której nie należy przekraczać, jeśli się chce, aby dane były należycie użyte.

Odpowiedź na pytanie trzecie jest bardzo ważna z punktu widzenia ostatecznej decyzji dotyczącej wyboru efektywnej metody zbierania informacji. Jak wykazuje praktyka wielu krajów, badanie wyczerpujące nie koniecznie musi dostarczyć dokładnych informacji. Ponieważ błędy występujące w tego rodzaju badaniach tzw. błędy nielosowe, stanowią niejednokrotnie większy problem niż błędy pochodzenia

losowego, postaramy się pokrótce omówić główne źródła występowania tych błędów na tle błędów występujących w badaniach reprezentacyjnych.

Odpowiedź na pytanie czwarte jest bardzo złożona. Zależy ona od wielu czynników, które trzeba uwzględnić w konkretnym przypadku. Szereg problemów ogólnych, związanych z tym pytaniem, omówimy w dalszej części tego opracowania.

4.2. Błędy nielosowe i losowe

Jak już wspomniano, przed podjęciem decyzji dotyczącej przeprowadzenia badania reprezentacyjnego zamiast badania wyczerpującego należy dokładnie rozważyć błędy, które mogą powstać w badaniu wyczerpującym oraz błędy losowe, które mogą wyniknąć na skutek zastosowania metody reprezentacyjnej. Warto przy tej okazji podać główne źródła błędów towarzyszących każdemu badaniu statystycznemu.

Ogólnie można wymienić jedenaście źródeł niedokładności wyników badań statystycznych [6]:

1) niedokładność w materiałach będących podstawą do wyznaczania jednostek badania;
2) trudności praktyczne przy stosowaniu przyjętej definicji jednostki badania;
3) opuszczenie i wielokrotne ujęcie niektórych jednostek badania;
4) niejasne lub zbyt ogólnikowe sformułowanie pytań w formularzu statystycznym;
5) brak koniecznych wyjaśnień w instrukcji statystycznej;
6) niewłaściwa interpretacja pytań formularza przez osoby zbierające materiał statystyczny lub przez osoby udzielające odpowiedzi;
7) świadome lub nieświadome udzielenie fałszywych odpowiedzi;
8) niedokładność pomiarów lub zliczeń, gdy są one podstawą zbierania materiału statystycznego;
9) błędy powstałe w czasie wpisywania danych do formularza statystycznego;
10) pominięcie lub brak odpowiedzi na pewne pytania;
11) błędy powstałe w czasie przenoszenia informacji na karty robocze (maszynowe itd.); mogą to być błędy klasyfikacji i symbolizacji lub segregacji i tabulacji.

Widać, że badanie wyczerpujące może być źródłem poważnych błędów zniekształcających rzeczywistość w sposób istotny. Wszystkie te okoliczności powinny być uwzględnione przy podejmowaniu ostatecznej decyzji o zastąpieniu badań wyczerpujących badaniami reprezentacyjnymi.

W badaniu reprezentacyjnym, oprócz błędów pochodzenia nielosowego, mogą wystąpić błędy losowe. Można powiedzieć, że istnieje

pewna zależność między błędami losowymi i nielosowymi. Błąd losowy zwykle maleje wraz ze wzrostem liczebności próbki, natomiast błąd nielosowy wzrasta. Powstaje więc zagadnienie wyboru optymalnego rozwiązania, tzn. ustalenia takiej liczebności próbki, aby zminimalizować wielkość błędów. Dotychczas w krajach RWPG problemom tym nie poświęcano większej uwagi. Wydaje się jednak, że bez głębszych badań w tym zakresie nie można efektywnie zastępować badań wyczerpujących badaniami reprezentacyjnymi. Powinny być badane nie tylko wielkości błędów losowych, ale również źródła i rozmiary błędów pochodzenia nielosowego. Wtedy w wielu przypadkach, jako konsekwencja takich badań, jedynym rozwiązaniem do przyjęcia w praktyce będą odpowiednio przygotowane badania reprezentacyjne.

4.3. Wybór metody badania

Z chwilą, gdy podjęto decyzję o przeprowadzeniu badania reprezentacyjnego, powstaje do rozwiązania szereg istotnych problemów, od których w dużym stopniu zależy efektywność i powodzenie takiego badania. Chodzi przede wszystkim o wybór najbardziej efektywnego wariantu badania. Sprawa ta nie jest łatwa, gdyż znalezienie rozwiązania optymalnego ze znacznej, stale rosnącej liczby różnych schematów wyboru próby, nastręcza wiele trudności. Wprawdzie ogólnie wiadomo, że chodzi o wybór takiego schematu, który daje żądane informacje o określonym stopniu precyzji przy najmniejszym koszcie badania lub odwrotnie, tj. daje żądane informacje przy określonym koszcie badania o największym stopniu precyzji, a więc o wybór, który dąży do najbardziej efektywnego rozdysponowania środków, czasami poprzez dążenie do najmniejszej liczebności próbki, częściej zaś poprzez minimalizację kosztu badania.

Jednakże bardzo często wiele kłopotów sprawia ustalenie żądanego stopnia precyzji wyników. Decyzje te podejmowane są zwykle przez użytkowników, którzy nie zawsze myślą kategoriami statystycznymi. Dlatego w zasadzie podawane są zbyt wygórowane wymagania odnośnie precyzji wyników.

Wydaje się tu celowe rozważenie wybranych problemów napotykanych przy prowadzeniu badań metodą reprezentacyjną, co rzuci nieco światła na możliwość szerszego zastosowania tej metody w badaniach statystycznych krajów-członków RWPG.

4.4. Operat losowania

Do wylosowania próby potrzebne są odpowiednie wykazy wszystkich jednostek losowania, tzw. o p e r a t l o s o w a n i a [5]. Operat taki powinien być 1) kompletny, tj. obejmować całą populację,

2) aktualny, 3) wolny od błędów, opuszczeń i powtórzeń jednostek losowania. Ponadto powinno być znane prawdopodobieństwo wylosowania każdej jednostki. Dodatkowe informacje, które zawiera operat, mogą być wykorzystane w celu zwiększenia efektywności losowania (przez odpowiednie powarstwowanie jednostek) i estymacji.

Trzeba stwierdzić, że w dotychczasowych badaniach niejednokrotnie brak było dobrych operatów losowania. Za operaty przyjmowano materiały gromadzone w celach administracyjnych w różnym czasie i przestrzeni. Operaty takie okazywały się często niekompletne i przestarzałe. Nie posiadały one na ogół dodatkowych informacji, które umożliwiałyby zwiększenie efektywności losowania. Dotyczyło to zwłaszcza badań warunków bytu ludności. Nieco lepiej przedstawiała się sytuacja w badaniach rolniczych, gdzie często operatem losowania były materiały spisów rolnych.

Wynika z tego, że rozszerzenie badań reprezentacyjnych zależy w dużym stopniu od możliwości skompletowania odpowiednich materiałów, z których można wylosować próbę i wykorzystać je do zwiększenia efektywności badania. Toteż w poszczególnych typach badań reprezentacyjnych należałoby więcej uwagi poświęcić problemom związanym z przygotowaniem odpowiednich operatów losowania, a zatem specjalnym pracom badawczym pozwalającym ustalić, które z dostępnych materiałów i dokumentów gromadzonych dla różnych potrzeb są najbardziej przydatne w badaniach z zakresu rolnictwa, warunków bytu ludności, obrotu towarowego itp.

4.5. Minimalna liczebność próbki

Badania statystyczne prowadzone metodą reprezentacyjną dotyczą zwykle wielu cech charakteryzujących się różnym stopniem rozproszenia. Powstaje więc problem, w jaki sposób ustalić minimalną liczebność próby. Trzeba podkreślić, że problem liczebności próby w badaniach reprezentacyjnych jest o wiele bardziej skomplikowany, niż przedstawia się to w podręcznikach statystyki, gdzie zwykle rozważa się jedną zmienną i liczebność próby ustala się na podstawie znajomości rozkładu tej cechy, minimalizując wielkość błędu losowego przy stałym koszcie lub odwrotnie. Tymczasem, jak już wspomniano, w badaniach reprezentacyjnych bada się wiele cech niezależnych od siebie i różniących się znacznie stopniem rozproszenia. Jeśli w takich przypadkach zastosujemy identyczne postępowanie dla określenia liczebności próby dla każdej pozycji, to otrzymamy różne liczebności prób, (dla żądanej precyzji wyników). Takie postępowanie — naszym zdaniem — nie jest uzasadnione. Lepszym rozwiązaniem będzie wybór pewnej liczby cech, uznanych za podstawowe ze względu na cel badania i określenie dla nich minimalnej liczebności próby. W przypadku gdy tak ustalone liczebności prób różnią się od siebie nieznacz-

nie, można wybrać którąkolwiek z nich jako minimalną liczebność próby i pozostałe parametry zostaną oszacowane z precyzją wyższą lub niższą od parametru przyjętego za podstawę. Gdy wielkości te jednak różnią się znacznie i nie może być przyjęta niższa precyzja wyników, wówczas największą z nich należy przyjąć za liczebność próby. Ten sposób postępowania ma już swoje tradycje w praktyce międzynarodowej i wydaje się uzasadniony, chociaż można przyjąć również i inny sposób określania minimalnej liczebności próby przy równoczesnym badaniu kilku parametrów. Warto tu wspomnieć o metodzie przyjętej w polskim Głównym Urzędzie Statystycznym [1], pozwalającej określić minimalną liczebność próby przy szacowaniu kilku parametrów i jej lokalizację w warstwach. Opracowano tu program standardowy na elektroniczną maszynę cyfrową, dzięki któremu skomplikowane obliczenia przeprowadza się w bardzo krótkim czasie.

Warto jeszcze raz podkreślić, że użytkownicy wyników badań reprezentacyjnych stawiają zwykle zbyt wysokie wymagania, jeśli chodzi o precyzję wyników, co — jak wiadomo — w znacznym stopniu powoduje zwiększenie minimalnej liczebności próby. Od stawiania rozsądnych wymagań dotyczących precyzji wyników zależy w dużym stopniu możliwość rozszerzenia zastosowań metody reprezentacyjnej w badaniach statystycznych.

4.6. Wybór schematu losowania

Z praktyki wiadomo, że wybór odpowiedniego schematu losowania nie jest sprawą łatwą, gdyż proces wypracowania schematu, który wykorzystywałby informacje dostarczone przez operat losowania, nasuwa wiele trudności. Przy opracowywaniu schematu optymalnego wyłania się szereg zagadnień związanych z techniką losowania, a także z kosztami i organizacją badania. Toteż w praktyce statystycznej często zadowalano się schematami będącymi kompromisem między wymaganiami teorii, kosztem i organizacją badania. Zwykle w największym stopniu decydował koszt badania i możliwości organizacyjne.

Przy wyborze optymalnego schematu losowania przyjmuje się zwykle zasadę minimalizacji kosztu jego przeprowadzenia w terenie (w koszt ten włącza się również koszty przygotowania badania) oraz opracowanie przy żądanej precyzji wyników lub odwrotnie. W obydwu przypadkach jest niezwykle ważne, aby podjąć kontrolę błędów nielosowych, które mogłyby wypaczyć wyniki badania. Wybrany schemat powinien być możliwy do wykonania w istniejących warunkach.

W badaniach reprezentacyjnych, prowadzonych na wielką skalę, stosowano schematy jedno lub wielostopniowe (zwykle dwu lub trzy-

stopniowe). Jeśli były dostępne informacje dodatkowe, stosowano losowanie warstwowe. Zwykle tworzono warstwy geograficzne i opierano się na wybranej zmiennej warstwowania. Jednakże nie tworzono warstw w sposób optymalny. Problemom określania liczby potrzebnych warstw ich wielkości oraz sposobom określania granic ze względu na losowania najbardziej efektywne dotychczas nie poświęcono wiele uwagi. Zagadnienia te należałoby w przyszłości dokładnie rozpatrzyć.

Wybór jednostek wewnątrz warstw dokonywany był zwykle w sposób systematyczny lub za pomocą takich schematów, jak losowanie ze zmiennym prawdopodobieństwem i po uprzednim ułożeniu jednostek w odpowiedni sposób. W losowaniu ze zmiennymi prawdopodobieństwami powstał problem wyboru miary wielkości. Zwykle losowano z prawdopodobieństwem proporcjonalnym do liczby jednostek badanych (gospodarstw domowych, osób itp.), wykorzystując do tego informacje z poprzednich badań. Stosowano również losowanie z jednakowym prawdopodobieństwem, gdy jednostki losowania były w przybliżeniu równe lub gdy nie uzyskano odpowiednich danych niezbędnych do innego losowania.

Jak już wspominaliśmy, często stosowano losowanie dwustopniowe, w którym jednostki drugiego stopnia losowano zwykle systematycznie z uprzednio wybranych jednostek pierwszego stopnia z operatów odpowiednio przygotowanych (np. w badaniach budżetów rodzinnych). Losowanie kilkustopniowe przeprowadzano zwykle ze względu na redukcję kosztów badania. Czasami jednak, gdy pozwalała na to organizacja badania, stosowano losowanie jednostopniowe warstwowe, co odpowiadało również wymaganiom zwiększenia efektywności losowania (np. w badaniach pogłowia zwierząt).

W niektórych przypadkach stosowano losowanie podpróbek interpenetracyjnych, co było z pewnością korzystne dla całości badań (badania takie prowadzono w Bułgarii i Polsce). Główną zaletą tego sposobu losowania jest to, że pozwala ono w łatwy sposób obliczyć oceny błędów losowych nawet wówczas, gdy schemat losowania jest skomplikowany. Jest to oczywiście bardzo ważne w przypadkach stosowania takich schematów losowania, w których jednostki pierwszego stopnia były losowane z prawdopodobieństwem zmiennym i gdy stosowano estymatory złożone (estymator regresyjny lub ilorazowy), ponieważ wówczas wyrażenia na wariancję i jej estymator są dość skomplikowane i trudne do numerycznych obliczeń na podstawie pojedynczej próby. Omawiany sposób losowania możemy również zastosować, gdy chcemy sprawdzić jakość pracy poszczególnych zespołów zbierających informacje statystyczne, oraz efekty różnych metod badania itd.

Przy wyborze odpowiedniego schematu losowania, szczególną uwagę należy zwrócić na schematy automatycznie ważone. Ważne jest to

głównie w tych przypadkach, gdy wyniki opracowywane są metodami tradycyjnymi i ważenie w poszczególnych warstwach może stanowić poważną trudność.

Przy wyborze odpowiedniego schematu losowania należy uwzględnić przede wszystkim możliwości przeprowadzenia badań w terenie. Schematy optymalne nie uwzględniające wszystkich elementów, od których zależy powodzenie badania, nie mają większej wartości.

4.7. Technika losowania próbek

Prowadząc badania reprezentacyjne na większą skalę, należy pamiętać o przyjęciu techniki losowania próbek łatwej w wykonaniu i zgodnej z wymaganiami teoretycznymi. Jest to szczególnie ważne, gdy losowanie przeprowadzane jest przez pracowników terenowych.

W praktyce badań prowadzonych w krajach należących do RWPG bardzo często stosowano losowanie systematyczne z losowym początkiem. Stosowano również, ale rzadziej, tablice liczb losowych, a także posługiwano się tablicami pseudolosowymi. Na przykład w Polsce wykorzystywano tablice liczb żelaznych, dające podobne efekty jak losowanie systematyczne i nie mające wady cykliczności [4].

Wydaje się, że na badania w terenie nie zwraca się jeszcze dostatecznej uwagi w pracach na temat zastosowań metody reprezentacyjnej. Problemom związanym z tym etapem badań reprezentacyjnych należałoby poświęcić więcej uwagi, gdyż ma on podstawowe znaczenie dla poprawności uzyskanych wyników i możliwości stosowania wyników na większą skalę. Naszym zdaniem, konieczne jest przeszkolenie większego grona pracowników terenowych służb statystycznych w zakresie przygotowywania operatów losowania i samej techniki losowania.

4.8. Problemy estymacji

Przy szacowaniu wartości parametrów na podstawie wyników próby duże znaczenie dla precyzji szacowania ma rodzaj zastosowanego estymatora. W większości badań reprezentacyjnych prowadzonych przez kraje należące do RWPG stosowano estymatory proste, tj. wyniki uzyskane z próby bezpośrednio uogólniano na całą zbiorowość. Wynikało to z braku dodatkowych informacji niezbędnych przy stosowaniu estymatorów złożonych, takich jak estymator ilorazowy czy regresyjny. Jednakże w wielu badaniach, zwłaszcza z zakresu statystyki rolnictwa, stosowano również estymatory ilorazowe.

Należy zaznaczyć, że w praktyce badań statystycznych iloraz dwóch cech występuje często w różnych obliczeniach i jest niejednokrotnie ważniejszy niż odpowiednie sumy lub przeciętne. Dotyczy to np. badań budżetów rodzinnych, produkcji rolniczej, plonów itd. Przy esty-

macji takich ilorazów brano zwykle stosunek estymatorów nieobciążonych licznika i mianownika.

Zastosowanie estymatora ilorazowego może się okazać pożądane nawet przy estymacji średnich i frakcji. W przypadku gdy znana jest prawdziwa wartość mianownika (np. z badania wyczerpującego) ilorazu, może się wydawać, że wystarczy estymować tylko licznik, a iloraz dla populacji uzyska się przez podzielenie tej wartości estymatora przez znaną wartość mianownika. Jednakże taki estymator nie zawsze jest tak efektywny jak estymator licznika i mianownika. W istocie rzeczy, jeśli estymatory licznika i mianownika są w przybliżeniu proporcjonalne, tj. jeśli istnieje między nimi regresja liniowa i linia regresji przechodzi przez początek układu, estymator ilorazowy jest bardziej efektywny niż estymator prosty opisany wyżej.

Coraz powszechniejsze stosowanie badań reprezentacyjnych wymaga, aby problemom estymacji, stanowiącej bardzo ważne ogniwo w całokształcie tych badań, poświęcić nieco więcej uwagi. Chodzi o badanie możliwości zastosowania różnych metod estymacji w konkretnych badaniach. Dotychczas w bardzo niewielu badaniach stosowano estymator regresyjny, który — jak wiadomo — jest bardziej efektywny niż inne estymatory. Taki stan rzeczy spowodowany był nie tylko trudnościami obliczeniowymi, ale także brakiem szerszych badań w tym zakresie.

Możliwości stosowania coraz efektywniejszych metod estymacji stwarza obecnie elektroniczna technika obliczeniowa, która pozwala wykonać żmudne i uciążliwe obliczenia za pomocą jednorazowo przygotowanych programów standardowych. Ułatwi to dokonywanie wyboru metod w konkretnym przypadku najbardziej efektywnych ze względu na precyzje wyników, a zatem stworzy możliwości szerszego stosowania metody reprezentacyjnej w różnych typach badań statystycznych.

4.9. Analiza precyzji wyników

Zwykle w badaniach statystycznych obserwuje się równocześnie wiele cech na podstawie uzyskanych danych oblicza się dziesiątki parametrów. Powstaje więc problem, w jaki sposób obliczyć uzyskany stopień precyzji wyników przy tak wielkiej liczbie parametrów. Przed podjęciem odpowiednich decyzji, należy jednak rozważyć:

a) dla jakich parametrów trzeba obliczyć uzyskany stopień precyzji wyników; jeśli liczba tych parametrów jest niewielka, obliczenia można przeprowadzić systemem tradycyjnym, przy większej liczbie parametrów konieczne jest użycie elektronicznych maszyn cyfrowych;
b) jaki typ schematu losowania został zastosowany; jest to bardzo ważne, gdyż w przypadku skomplikowanych schematów obliczenia stają się pracochłonne i uciążliwe;

c) w jakim celu prowadzona jest analiza precyzji wyników; jeśli chodzi nam tylko o ogólną orientację co do stopnia precyzji, to można wprowadzić pewne uproszczenia obliczeń; jeśli jednak chodzi o szczegółową analizę, to obliczenia muszą być przeprowadzone bardzo dokładnie;
d) jaki zastosowano estymator; w przypadku losowania prostego dla szacowania struktury obliczenia związane z precyzją wyników nie są skomplikowane, natomiast znacznie się komplikują przy użyciu estymatorów złożonych (np. regresyjnych);
e) jakimi środkami się rozporządza; wykorzystanie elektronicznych maszyn cyfrowych pozwoli w znacznym stopniu zautomatyzować cały proces obliczeń, gdyż można wówczas wykorzystać odpowiednie programy standardowe.

Jak dotychczas, w większości przypadków publikowane są tylko wyniki badania bez błędów losowych. W zasadzie jednak błędy losowe powinny być publikowane razem z wynikami badania, ponieważ umożliwia to użytkownikom budowanie przedziałów ufności i przeprowadzenie wnioskowania według własnych potrzeb. Naszym zdaniem, niepublikowanie błędów losowych jest nieuzasadnione i może doprowadzić do poderwania zaufania użytkowników do wyników uzyskanych z badań reprezentacyjnych. Brak informacji o stopniu precyzji wyników powoduje jednakowe traktowanie wszystkich liczb, a jak wiadomo, jedną z korzyści stosowania metody reprezentacyjnej jest możliwość oceny stopnia precyzji uzyskanych wyników. Brak analizy stopnia precyzji uzyskanych wyników może doprowadzić w wielu przypadkach do błędnych wniosków.

Przy stosowaniu metody reprezentacyjnej na większą skalę należy doprowadzić do tego, aby obok wyników badania były publikowane ich błędy losowe. Wpłynie to niewątpliwie na bardziej odpowiednie wykorzystanie uzyskanych wyników w różnych opracowaniach statystycznych.

Przyjmując, że publikowanie wyników badania łącznie z informacjami o stopniu ich precyzji (publikowanie błędów losowych) jest konieczne, musimy się zastanowić, w jaki sposób należy prezentować te dane. Gdy badanie jest niewielkie i obejmuje nieznaczną liczbę pozycji, wtedy sprawa jest prosta, gdyż błędy losowe można zamieścić obok uzyskanych ocen. Jednakże przy dużych badaniach reprezentacyjnych ten sposób prezentowania błędów losowych nie może być stosowany. Jeśli publikacja wyników z badań reprezentacyjnych obejmuje setki stron, to zamieszczenie obok każdej pozycji jej błędu losowego oznaczałoby podwojenie objętości publikacji. Z tych względów w publikacjach tego rodzaju należy szukać innego rozwiązania. Niektóre kraje, publikując wyniki swych badań reprezentacyjnych, analizę precyzji uzyskanych wyników podają we wstępnym omówieniu i zamieszczają tablice przedstawiające wielkości błędów losowych

jako funkcje odpowiednich wielkości, takich jak liczebność próbki, frakcji szacowanych elementów, współczynników zmienności, itd. Oczywista, że tego rodzaju tablice nie zawsze można łatwo zbudować. Zależy to od zastosowanego schematu losowania i od przyjętych metod estymacji wyników. W zasadzie łatwiejsze jest przygotowanie takiej tablicy przy szacowaniu frakcji elementów wyróżnionych niż dla sum lub wartości przeciętnych cech mierzalnych. Jednakże w praktyce badań statystycznych bardzo często mamy do czynienia z szacowaniem różnych struktur populacji generalnej, więc ten sposób mógłby znaleźć zastosowanie. Można by opracować specjalne tablice, które dla określonych schematów losowania podawałyby wielkości błędów losowych zależne od określonych zmiennych, wymienionych w publikacji, np. od liczebności próbki, od przeciętnych, od współczynników zmienności itp. W Polsce w tym celu opracowano specjalne tablice, w których podano wielkości półprzedziałów ufności w zależności od liczebności próbki, wielkości szacowanej frakcji jednostek wyróżnionych lub współczynników zmienności badanych cech dla losowania indywidualnego [3].

Byłoby wskazane poszukiwanie dalszych sposobów prezentacji błędów losowych w sposób zwięzły, gdyż ograniczenie się tylko do publikowania samych wyników znacznie zmniejsza zakres wnioskowania statystycznego.

5. Możliwości rozszerzenia zastosowań metody reprezentacyjnej

Dotychczas rozważaliśmy główne kierunki zastosowań metody reprezentacyjnej w krajach należących do RWPG oraz ważniejsze problemy metodologiczne związane ze stosowaniem tej metody w praktyce. Problemom metodologicznym celowo poświęciliśmy więcej miejsca, gdyż praktyka statystyczna wskazuje, że nie zawsze były przestrzegane podstawowe zasady badań reprezentacyjnych, co prowadzi ostatecznie do ograniczonego wykorzystania możliwości, jakie stwarza stosowanie tej metody.

Rozważmy teraz możliwości rozszerzenia zakresu zastosowań metody reprezentacyjnej w praktyce. Ograniczymy się do omówienia głównych kierunków zastosowań, pozostawiając na uboczu szczegółowe rozważania możliwości jej zastosowań w konkretnych badaniach statystycznych.

5.1. Badania eksperymentalne

Wiadomo, że każde badanie statystyczne składa się z kilku etapów. Etap wstępny polega na opracowaniu programu badania, organizacji zbierania informacji i opracowywania danych, oraz opracowaniu

szczegółowego harmonogramu czynności. Jednakże przed rozpoczęciem dużych badań statystycznych a zwłaszcza badań nowych, nie mających jeszcze większych tradycji lub prowadzonych poprzednio inną metodą, potrzebne są badania eksperymentalne, które pozwolą sprawdzić w praktyce przyjęte rozwiązania metodologiczne i organizacyjne. Z konieczności muszą być one prowadzone na nieznacznej liczbie jednostek. Jednostki badania powinny być wybierane w sposób losowy, aby uzyskać lepszą reprezentację warunków, jakie mogą wystąpić w badaniu statystycznym. Dane z badań eksperymentalnych mogą być wykorzystane do oceny pewnych parametrów, tkóre mogą być przydatne przy prowadzeniu danego badania metodą reprezentacyjną. Wprawdzie w dotychczasowej praktyce statystyczne badania eksperymentalne prowadzono już nie raz, ale były to badania niewielkie i nie zawsze stosowano do tego celu metodę reprezentacyjną.

5.2. Zastąpienie badań wyczerpujących badaniami reprezentacyjnymi

W części obejmującej rozważania metodologiczne szeroko dyskutowano problemy związane z zastosowaniem metody reprezentacyjnej w badaniach prowadzonych dotychczas metodą wyczerpującą. Wydaje się, że wiele badań statystycznych prowadzonych dotychczas metodą wyczerpującą można by zastąpić badaniami reprezentacyjnymi, lecz każde z nich wymaga osobnych rozważań i nie można tu tych spraw szczegółowo rozpatrzyć. Takie możliwości istnieją niewątpliwie w statystyce rolniczej, w statystyce przemysłu, obrotu towarowego, pracy i płac itd.

5.3. Rozszerzenie zakresu badań

Zakres informacji zbieranych w pełnym badaniu statystycznym musi być z wielu względów ograniczony. Rozszerzenie zakresu informacji może jednak nastąpić, jeśli zastosuje się reprezentacyjną metodę badań. Polega to na tym, że oprócz badania podstawowego, które dotyczy całej zbiorowości, zbiera się dodatkowe informacje od części jednostek wchodzących w skład wylosowanej próbki. Informacje te można zbierać równolegle z badaniem podstawowym lub w pewnym terminie po jego przeprowadzeniu. Metodą wyczerpującą zbierane są wówczas informacje, które muszą być opracowane dla małych jednostek i w szczegółowych przekrojach, natomiast metodą reprezentacyjną — informacje uzupełniające, które nie są potrzebne w tak szczegółowych grupowaniach. Trzeba stwierdzić, że jak dotychczas ta możliwość stosowania metody reprezentacyjnej była wykorzystywana w niewielkim tylko stopniu w niektórych krajach.

5.4. Opracowanie wstępnych wyników

Na ogół metodę reprezentacyjną stosowano tylko do opracowywania wstępnych wyników spisów powszechnych. Chodziło głównie o przyspieszenie uzyskania podstawowych wyników potrzebnych w bieżących pracach. Zastosowanie metody reprezentacyjnej pozwoliło znacznie skrócić czas opracowania i uogólnić uzyskane wyniki na całą zbiorowość. Wydaje się, że opracowania takie można by również stosować w innych badaniach, np. w spisach przemysłowym, handlowym, kadrowym, których opracowanie trwa zwykle dość długo, a wyniki potrzebne są wcześniej do różnych celów. Metodę reprezentacyjną można by wykorzystać również w podobny sposób do opracowania wyników zależnych od występujących określonych prawidłowości (np. badanie współzależności zjawisk), gdy próbka nie musi być zbyt liczna, a do opracowania niezbędne są dane jednostkowe.

5.5. Reprezentacyjne badania kontrolne

Jak już kilkakrotnie podkreślano, wyniki każdego badania statystycznego obarczone są różnego rodzaju błędami, których nie można całkowicie wyeliminować nawet przy najlepszej organizacji badania (o źródłach tych błędów wspomniano przy omawianiu różnych problemów metodologicznych). Konieczne więc jest przeprowadzanie badań kontrolnych, które pozwoliłyby ustalić rozmiary błędów popełnionych w badaniu podstawowym. Badania takie prawie zawsze mogą i powinny być prowadzone metodą reprezentacyjną. W dotychczasowej praktyce ograniczano się jednak głównie do prowadzenia kontrolnych spisów ludności i spisów rolnych, natomiast w innych badaniach statystycznych badania kontrolne stosowano dość rzadko. Istnieją zatem wielkie możliwości stosowania metody reprezentacyjnej w badaniach kontrolnych większości badań statystycznych, a w szczególności tych, które są podstawą ważnych decyzji, a mogą być obarczone różnego rodzaju błędami.

6. Wnioski końcowe

Z naszych rozważań na temat możliwości szerszego stosowania metody reprezentacyjnej w badaniach statystycznych wynika, że metoda ta może w poważnym stopniu przyczynić się do rozszerzenia zakresu informacji, polepszenia ich jakości oraz do przyspieszenia i pogłębienia analiz statystycznych. Jednakże efektywnie jest stosowana w praktyce badań statystycznych wymaga spełnienia wielu warunków. Warunki te przedstawiono tu w formie wniosków wynikłych z przeprowadzonej analizy. Podane wnioski dotyczą ogólnych proble-

mów organizacyjnych i metodologicznych, a zatem w równej mierze związane są z praktyką statystyczną wszystkich krajów-członków RWPG. Nie dotyczą one oczywiście zastosowania metody reprezentacyjnej w konkretnych badaniach statystycznych, gdyż — jak już mówiliśmy — możliwość zastąpienia konkretnego badania pełnego badaniem reprezentacyjnym musi być każdorazowo zanalizowana. Celem podanych tu wniosków ogólnych jest umożliwienie szerszego niż dotychczas stosowania metody reprezentacyjnej oraz zwiększenie jej efektywności.

1. Centralne urzędy statystyczne krajów-członków RWPG powinny posiadać specjalne stałe oddziały zajmujące się badaniami reprezentacyjnymi.

Stosowanie metody reprezentacyjnej na szerszą skalę wymaga utworzenia stałego zespołu zajmującego się całokształtem badań reprezentacyjnych. Zespół ten powinien składać się ze specjalistów z zakresu metody reprezentacyjnej, mających oczywiście odpowiednią dużą praktykę. Zasadniczym zadaniem tego zespołu byłoby planowanie badań reprezentacyjnych, tj. opracowywanie odpowiednich schematów losowania, ustalanie potrzebnej liczebności próby, wybór jednostek do próby, opracowanie metody estymacji i analizy precyzji wyników.

Zespół prowadziłby prace badawcze w zakresie wyboru najbardziej efektywnych metod badań reprezentacyjnych w konkretnych warunkach.

Zespół ten mógłby również podjąć kontrolę błędów nielosowych w badaniach statystycznych. Wydaje się, że do tego rodzaju badań byłby on bardziej odpowiedni niż jakikolwiek inny zespół specjalistów, gdyż wykrycie źródeł błędów nielosowych wymaga znacznego doświadczenia w pracy badawczej. Jak wiemy, do badań błędów pochodzenia nielosowego w wielu krajach dotychczas jeszcze nie przywiązuje się należytej wagi, gdy tymczasem w wielu badaniach błędy te stanowią znacznie poważniejszy problem niż błędy losowe.

Istnienie takiego zespołu umożliwiłoby pełne wykorzystanie doświadczeń zdobytych w poprzednich badaniach i ogólne podniesienie poziomu badań statystycznych.

2. Personel terenowych organów statystyki należałoby drogą specjalnego przeszkolenia zapoznać z podstawami metody reprezentacyjnej.

Część personelu terenowych organów statystyki państwowej należałoby przede wszystkim zapoznać z zasadami przygotowywania dokumentacji do wylosowania próby, przeprowadzania losowania, kontroli przebiegu badania i z podstawowymi metodami zbierania

informacji statystycznej. Od przestrzegania zasad badań reprezentacyjnych na najniższych szczeblach badania zależy w poważnym stopniu ich powodzenie.

3. Do zwiększenia efektywności badań prowadzonych metodą reprezentacyjną konieczne są odpowiednie prace badawcze.

Podjęcie badań reprezentacyjnych na szerszą skalę i efektywne ich prowadzenie wymaga systematycznych prac badawczych. Prace te powinny dotyczyć zarówno zagadnień teoretycznych, jak i praktycznych zastosowań wypracowanych rozwiązań teoretycznych. Należałoby przede wszystkim badać:

a) przydatność różnych operatów losowania,

b) wybór schematów losowania,

c) wybór odpowiednich metod estymacji,

d) metody analizy precyzji wyników,

e) metody zbierania danych,

f) jakość danych statystycznych.

Problem *przydatności operatów losowania* występuje w większości krajów. Często stosowane operaty okazują się niekompletne, nieaktualne i nieodpowiednie do badania. Należałoby badać, do jakich typów badań można wykorzystać istniejącą dokumentację i różne wykazy, takie jak wykazy miast, wsi, bloków, różnego rodzaju mapy oraz materiały z różnego rodzaju spisów. Należałoby badać możliwość wykorzystania z tych wykazów dodatkowych informacji do celów warstwowania, estymacji i wyboru z różnymi prawdopodobieństwami.

Badania *wyboru schematów losowania* umożliwią wybór schematu najbardziej odpowiedniego dla danego badania, czyli optymalnego ze względu na koszt badania lub precyzję wyników.

Badanie *wyboru odpowiednich metod estymacji* pozwoliłoby — w zależności od dostępnych informacji dodatkowych oraz możliwości technicznych (obliczeniowych) — wybrać dla każdego badania najbardziej odpowiednie estymatory.

Badanie *metod analizy precyzji wyników* jest konieczne, ponieważ każde badanie reprezentacyjne powinno dostarczyć informacji o stopniu precyzji uzyskanych wyników. Konieczne to jest do prawidłowego wnioskowania na podstawie danych oraz do efektywnego ustawienia badań przyszłych.

Badania *metod zbierania danych* pozwoliłyby gromadzić informacje o trudnościach i doświadczeniach związanych z różnymi metodami ich zbierania (wywiad, samorejestracja, pomiar, wykorzystanie dokumentów itd.) i ustalać przydatność różnych metod w różnych warunkach.

O konieczności *badania jakości danych statystycznych* na wszystkich etapach badania statystycznego nie musimy nikogo przekonywać. Należałoby badać wpływ różnych czynników na jakość danych, np. wpływ pamięci, metody zbierania informacji, kwalifikacji zespołów prowadzących badania, itd.

4. Konieczność rozszerzenia badań eksperymentalnych.

Przed rozpoczęciem obszernych badań statystycznych, a zwłaszcza badań nowych, nie mających większych tradycji lub gdy chcemy zweryfikować przydatność nowej metody badawczej, konieczne są badania eksperymentalne. Badania takie pozwolą sprawdzić w praktyce przyjęte rozwiązania metodologiczne, zastosowane formularze, definicje, metody zbierania informacji itd. Umożliwią one wybór wariantu badania podstawowego najbardziej odpowiedniego, w istniejących warunkach.

5. Konieczność wymiany doświadczeń w zakresie badań reprezentacyjnych między różnymi krajami.

Wprawdzie możliwość zastosowania metody reprezentacyjnej w danym badaniu statystycznym uzależniona jest od istniejących konkretnych warunków, to jednak ogólne problemy metodologiczne badań reprezentacyjnych są podobne i wymiana doświadczeń w zakresie stosowania różnych metod badawczych, przyjętych rozwiązań metodologicznych itd., byłaby bardzo pożyteczna. I chociaż przystępując do konkretnego badania nie można liczyć na gotowe rozwiązanie, to nieocenioną usługę może oddać znajomość pomysłów i rozwiązań stosowanych w innych badaniach reprezentacyjnych. Dlatego celowe wydaje się międzynarodowe spotkanie seminaryjne specjalistów z zakresu metody reprezentacyjnej.

6. Wśród użytkowników danych statystycznych należy popularyzować badania reprezentacyjne.

Rozszerzenie zastosowań metody reprezentacyjnej wymaga popularyzacji jej zasad wśród użytkowników oraz zapoznania ich ze sposobem wnioskowania na podstawie wyników uzyskanych z badań reprezentacyjnych.

* * *

Wyżej podane wnioski, jak już mówiliśmy, dotyczą tylko rozszerzenia możliwości efektywnego stosowania badań reprezentacyjnych na szerszą skalę. Nie omówiono tu oczywiście wszystkich zagadnień wy-

stępujących w bogatej problematyce zastosowań metody reprezentacyjnej w badaniach statystycznych. Nie poruszono tu problemów związanych ze stosowaniem tej metody w takich konkretnych dziedzinach badań statystycznych, jak statystyka rolnicza, statystyka przemysłu, obrotu towarowego itd. Zagadnienia te wymagają bowiem osobnego omówienia.

Wykaz literatury

[1]. Greń J.: *O pewnym zastosowaniu programowania nieliniowego do metody reprezentacyjnej.* „Przegląd Statystyczny" 1966, z. 3.

[2]. Kordos J.: *Tablice minimalnej liczebności próbki i wielkości przedziałów ufności.* GUS, Warszawa 1967.

[3]. Steinhaus H.: *Liczby złote i żelazne.* „Zastosowania Matematyki" 1958, z. 1.

[4]. Zasępa R.: *Badania statystyczne metodą reprezentacyjną.* PWN, Warszawa 1962.

[5]. Zasępa R.: *Szacowanie niedokładności wyników badań statystycznych.* W: *Zastosowanie metod matematycznych w statystyce.* Biblioteka Wiadomości Statystycznych, tom VII, GUS, Warszawa 1969.

[6]. *Zastosowanie metody reprezentacyjnej w krajach socjalistycznych.* Biblioteka Wiadomości Statystycznych, t. XIV, GUS Warszawa 1971.

Johan CARSTENS

PRZYBLIŻONA METODA OBLICZANIA BŁĘDU SYSTEMATYCZNEGO ESTYMATORA ILORAZOWEGO

1. Estymacja sumy wartości i średniej

Statystyka opisowa zna wiele mniej lub bardziej skomplikowanych wskaźników, które mogą być użyte do scharakteryzowania jedno- albo wielowymiarowej populacji (suma wartości, średnia, miernik dyspersji, współczynniki regresji i korelacji i in.). W tej pracy ograniczymy się jednak do szacowania — na podstawie próby otrzymanej w wyniku losowania nieograniczonego — sumy wartości i średniej, ponieważ dla decyzji dotyczących gospodarki narodowej te dwa wskaźniki mają największe znaczenie.

Sumą wartości będziemy nazywać wskaźnik, który powstaje przez zsumowanie wartości cechy poszczególnych jednostek populacji. Przyporządkujmy każdej z N jednostek populacji skończonej wartość liczbową X_m (wartość określonej cechy), a wtedy sumę wartości cechy będziemy mogli przedstawić w postaci:

$$X = \sum_{m=1}^{N} X_m \tag{1}$$

Sumami wartości są np.: pogłowie trzody chlewnej we wszystkich gminach NRD w określonym momencie czasu, roczne wydatki określonej grupy ludności na bilety wstępu do teatrów, kin i na imprezy kulturalne, liczba zatrudnionych na pół etatu w NRD itd. Suma wartości ma sens tylko dla skończonej liczby jednostek. Dzieląc sumę wartości (1) przez liczbę jednostek populacji, otrzymuje się wartość średnią cechy przypadającą na jednostkę:

$$\bar{X} = \frac{\sum_{m=1}^{N} X_m}{N} = \frac{X}{N} \tag{2}$$

Wzór (2) przedstawia przeciętną wartość cechy wszystkich jednostek populacji.

Suma wartości X i średnia $\bar{X}$ mogą być obliczone dokładnie tylko po zbadaniu wszystkich jednostek populacji. W praktyce statystycznej

często powstaje problem ustalenia sumy wartości X względnie wartości średniej $\bar{X}$, tzn. oszacowania parametrów populacji w oparciu o badania reprezentacyjne.

Niech będzie dana próba losowa n jednostek populacji; frakcję jednostek wylosowanych oznaczamy przez f:

$$f = \frac{n}{N} \tag{3}$$

a czynnik próby — przez φ:

$$\varphi = \frac{N-n}{n} = \frac{N}{n} - 1 = \frac{1}{f} - 1. \tag{4}$$

W wyniku badania reprezentacyjnego każdemu elementowi próby przyporządkujemy wartość liczbową x_i interesującej cechy ($i = 1,2,\ldots, n$ [1]).

Zgodnie z tą regułą wynik badania reprezentacyjnego jest podany w postaci średniej $\bar{x}$:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x}{n}, \tag{5}$$

i odchylenia standardowego s_x:

$$s_x = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}} \tag{6}$$

Gdy na podstawie wyników badania reprezentacyjnego zostanie oszacowana suma wartości X albo średnia $\bar{X}$ wszystkich jednostek populacji generalnej, a więc gdy otrzymany z próby zmniejszony obraz populacji generalnej zostanie ponownie powiększony przez stosowanie pewnych metod rachunkowych, to metody te będziemy nazywać estymacją. Jeśli przy tym do szacowania danych parametrów z populacji generalnej będą używane tylko wartości x_i badanej cechy z próby n — elementowej, to będzie mówiło się o *estymacji prostej*. Przy estymacji prostej estymator sumy wartości X ma postać [2]):

$$\hat{x} = \bar{x}N = x\frac{N}{n}, \tag{7}$$

---

1) W teorii metody reprezentacyjnej używa się dużych liter, gdy symbole dotyczą populacji generalnej. Do scharakteryzowania próby będą stosowane odpowiednie małe litery.

2) Znak $\hat{}$ pokazuje, że chodzi o oszacowanie odpowiedniego parametru populacji generalnej.

a estymator średniej:

$$\hat{\bar{x}} = \bar{x} = \frac{x}{n}. \tag{8}$$

Odchylenia standardowe estymatorów $\hat{x}$ i $\hat{\bar{x}}$ dane są wzorami:

$$s_{\hat{x}} = s_x \sqrt{\varphi N} \tag{9}$$

$$s_{\hat{\bar{x}}} = s_{\bar{x}} = s_x \sqrt{\frac{\varphi}{N}} \tag{10}$$

Estymatory proste (7) i (8) zawierają tylko informacje zaczerpnięte bezpośrednio z próby. Jeśli do oszacowania badanej cechy (X) są używane informacje o innej cesze (dodatkowej), to mówi się wówczas o *estymacji złożonej.*

Zastosowanie estymacji złożonej uzależnione jest od znajomości sumy wartości cechy dodatkowej:

$$Y = \sum_{m=1}^{N} Y_m. \tag{11}$$

Zależnie od sposobu uwzględnienia informacji o cesze dodatkowej przy szacowaniu sumy wartości X albo średniej $\bar{X}$ badanej cechy, rozróżnia się trzy rodzaje estymacji złożonej:

estymacja różnicowa,
estymacja ilorazowa,
estymacja regresyjna.

Nie omawiając poszczególnych metod estymacji (zob. Cochran, Hansen (Hurwitz) Madow, Schwarz, Zasępa i inni) podamy poniżej estymatory sumy wartości X i estymatory średniej $\bar{X}$ dla wymienionych trzech rodzajów estymacji złożonej.

Estymator różnicowy

$$\hat{x}_{\text{diff.}} = Y + (x - y)\frac{N}{n}, \tag{12}$$

$$\hat{\bar{x}}_{\text{diff.}} = \bar{Y} + (\bar{x} - \bar{y}). \tag{13}$$

Estymator ilorazowy

$$\hat{x}_{\text{quot.}} = \frac{x}{y} Y = t \cdot Y, \text{ gdzie } t = \frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} y_i} = \frac{\bar{x}}{\bar{y}}, \tag{14}$$

$$\hat{\bar{x}}_{\text{quot.}} = t \cdot \bar{Y}. \tag{15}$$

Estymator regresyjny

$$\hat{x}_{\text{regr.}} = x\frac{N}{n} + b\left(Y - y\frac{N}{n}\right), \tag{16}$$

$$\hat{\bar{x}}_{\text{regr.}} = \bar{x} + b(\bar{Y} - \bar{y}), \tag{17}$$

gdzie współczynnik regresji b dany jest wzorem:

$$b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y})^2}. \tag{18}$$

Podaliśmy cztery ogólnie znane rodzaje estymacji, stosowane do oszacowania sumy wartości X i średniej $\bar{X}$. Z czterech rodzajów estymatorów jedynie estymator prosty i różnicowy są nieobciążone.

Estymatory obciążone i nieobciążone

Niech ogólnie $\hat{\Theta}$ będzie estymatorem dowolnego parametru Θ populacji generalnej; $\hat{\Theta}$ nazywamy estymatorem nieobciążonym, jeśli wartość oczekiwana estymatora $\hat{\Theta}$, uzyskanego na podstawie próby o dowolnej wielkości, jest równa szacowanemu parametrowi z populacji generalnej:

$$E(\hat{\Theta}) = \Theta. \tag{19}$$

Tego warunku nie będą spełniały estymator ilorazowy $\hat{x}_{\text{quot.}}$ i estymator regresyjny $\hat{x}_{\text{regr.}}$.

Niech $\hat{\Theta}$ będzie obciążonym estymatorem parametry Θ populacji generalnej; wtedy zachodzi

$$E(\hat{\Theta}) = \Theta + B, \tag{20}$$

przy czym wielkość B jest błędem systematycznym oszacowania (błąd ten w literaturze anglo-amerykańskiej określa się słowem: Bias). Pytanie: czy w praktyce powinno się stosować estymatory nieobciążone czy obciążone? — jest w zasadzie pytaniem o dokładność metody estymacji.

Ponieważ ocena precyzji wyników zależy od wielkości odchylenia standardowego estymatora obliczonego na podstawie próby od faktycznej wartości parametru w populacji (a więc pośrednio — od zastosowanej metody estymacji), korzystne jest stosowanie estymatorów

obciążonych zawsze wtedy, gdy to się wiąże z istotnym zmniejszeniem błędu losowego w stosunku do występującego błędu systematycznego. Miarodajny dla zastosowania obciążonego estymatora jest zatem stosunek błędu systematycznego (B) do odchylenia standardowego estymatora ($s_{\hat{\theta}}$), które jest miarą wielkości błędu losowego (por. Cochran s. 8ff; Carstens s. 13ff).

2. Błąd systematyczny estymatora ilorazowego

Ponieważ jak wiemy, odchylenie standardowe estymatora ilorazowego jest z reguły mniejsze od odchyleń standardowych estymatorów prostego i różnicowego, zastosowanie estymatora ilorazowego w praktyce badań reprezentacyjnych prowadzi na ogół do podniesienia efektywności estymacji w porównaniu z nieobciążonymi estymatorami prostym i różnicowym. Jest to, obok takich czynników jak szerokie możliwości zastosowań estymatora ilorazowego (może być stosowany nawet przy różnych zakresach cechy badanej i cechy dodatkowej, może być stosowany nie tylko przy badaniach ciągłych, ale także przy badaniach jednorazowych) i relatywnie małe zwiększenie obliczeń (w przeciwieństwie np. do estymatora regresyjnego), istotną podstawą dla szerszego niż dotychczas zastosowania estymatora ilorazowego przy szacowaniu sumy wartości i średniej. Bardzo mało uwagi poświęcano jednak problemowi oceny precyzji oszacowania, a w szczególności obliczeniom błędu systematycznego estymatora ilorazowego.

Skorzystajmy ze znanego wzoru na obliczenie błędu systematycznego estymatora ilorazowego (por. Cochran, s. 117f; Hansen, Hurwitz Madow, s. 112; Zasępa). Błąd systematyczny przy szacowaniu sumy wartości ustalamy na podstawie wzoru:

$$B_{\hat{x}_{\text{quot.}}} \approx \varphi (V_y^2 - V_{xy}) \bar{X}, \tag{21}$$

lub

$$B_{\hat{x}_{\text{quot.}}} \approx \frac{\varphi}{N} V_y (V_y - \rho V_x) X, \tag{22}$$

a przy szacowaniu średniej $\bar{X}$ — na podstawie wzoru:

$$B_{\hat{\bar{x}}_{\text{quot.}}} \approx \frac{\varphi}{N} (V_y^2 - V_{xy}) \bar{X}. \tag{23}$$

Ponieważ błąd systematyczny obliczany jest w oparciu o wartości próby (w praktyce badań reprezentacyjnych jedynie ta droga jest możliwa), błąd systematyczny estymatora $\hat{\bar{x}}_{\text{quot.}}$ wynosi:

$$B_{\hat{\bar{x}}_{\text{quot.}}} \approx \varphi v_y (v_y - r v_x) \bar{x}. \tag{24}$$

a błąd systematyczny estymatora $\hat{\bar{x}}_{quot.}$:

$$B_{\hat{\bar{x}}_{quot.}} \approx \frac{\varphi}{N}(v_y^2 - v_{xy})\bar{x}, \tag{25}$$

gdzie we wzorach (21), (22), (23):

φ — czynnik próby, $\varphi = \frac{N-n}{n} = \frac{N}{n} - 1 = \frac{1}{f} - 1$,

V_x — współczynnik zmienności cechy badanej w populacji generalnej $V_x = \frac{S_x}{\bar{X}}$,

V_y — współczynnik zmienności cechy dodatkowej w populacji generalnej,

V_{xy} — stosunek kowariancji w populacji generalnej wartości cechy X i cechy Y do iloczynu średnich $\bar{X}$ i $\bar{Y}$ z populacji generalnej

$$V_{xy} = \frac{S_{xy}}{\bar{X}\bar{Y}}, \quad \text{gdzie: } S_{xy} = \frac{1}{N-1}\sum_{m=1}^{N}(X_m - \bar{X})(Y_m - \bar{Y}),$$

ρ — współczynnik korelacji między cechą badaną i cechą dodatkową w populacji generalnej:

$$\rho = \frac{S_{xy}}{S_x S_y} = \frac{\sum_{m=1}^{N}(X_m - \bar{X})(Y_m - \bar{Y})}{\sqrt{\sum_{m=1}^{N}(X_m - \bar{X})^2}\sqrt{\sum_{m=1}^{N}(Y_m - \bar{Y})^2}}$$

a we wzorach (24) i (25)

v_x, v_y, v_{xy} i r — odpowiednie wielkości obliczone z próby.

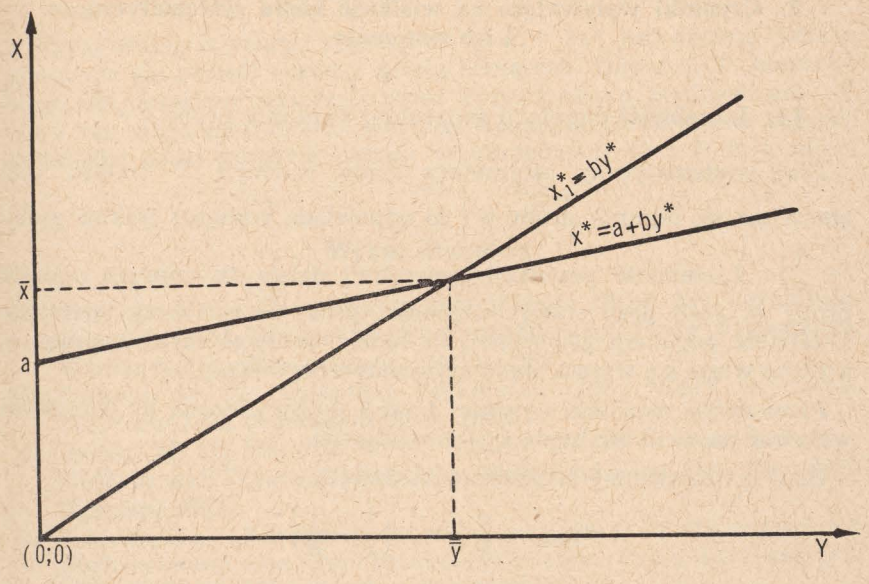
Korzystając z tego, że obliczenia błędu systematycznego oparte zostały na wynikach próby, stwierdzimy natomiast, że zarówno $B_{\hat{\bar{x}}_{quot.}}$, jak i $B_{\hat{\bar{x}}_{quot.}}$ są równe zeru wtedy i tylko wtedy, gdy

$$v_y^2 - v_{xy} = 0.$$

Ten warunek będzie spełniony tylko wtedy, gdy

$$\frac{\bar{x}}{\bar{y}} = b, \tag{26}$$

tzn. gdy prosta regresji określona przez punkty próbowe $(y_i; x_i)$ przechodzi przez początek układu osi współrzędnych (0; 0).



Prosta regresji $x_1^* = by^*$, gdzie $b = \frac{\bar{x}}{\bar{y}}$ (współczynnik proporcjonalności — kątowy) i prosta regresji $x^* = a + by^*$.

Zakładamy, że błąd systematyczny estymatora ilorazowego zależy od przebiegu prostej regresji

$$x^* = a + by^*, \tag{27}$$

a zwłaszcza od jej wyrazu wolnego a. Po przekształceniu równości $a = \bar{x} - b\bar{y}$, otrzymujemy

$$a = \frac{\bar{x}}{v_y}(v_y - r\,v_x). \tag{28}$$

Z wzorów (28), (24) oraz (25) wynika wzór przybliżony na obliczanie błędu systematycznego szacunku sumy wartości X

$$B\hat{}_{\hat{x}\text{quot.}} \approx \varphi\, v_y^2\, a \tag{29}$$

oraz na obliczanie błędu systematycznego szacunku średniej $\bar{X}$

$$B\hat{}_{\hat{\bar{x}}\text{ quot.}} \approx \frac{\varphi}{N} v_y^2\, a. \tag{30}$$

Wzory (29) i (30) pokazują czynniki wpływające na występowanie błędu systematycznego i jego wielkość w przypadku estymatora ilorazowego. Poza tym umożliwiają one stosunkowo szybkie obliczanie błędów systematycznych $B\hat{}_{\hat{x}\text{quot.}}$ i $B\hat{}_{\hat{\bar{x}}\text{quot.}}$.

3. Czynniki wpływające na wielkość błędu systematycznego i ich znaczenie

3.1. Liczebność populacji generalnej i czynnik próby $\varphi = \frac{N}{n} - 1$

Błąd systematyczny estymatora sumy wartości X maleje, jeśli zmniejsza się czynnik próby φ (co odpowiada rosnącej frakcji próby $f = \frac{n}{N}$). Liczebność populacji generalnej działa tu poprzez czynnik próby φ. Jeśli przy stałej liczebności próby n będziemy zwiększać liczebność populacji generalnej, to będzie wzrastać czynnik próby φ, a z tym wiąże się wzrost błędu systematycznego $B_{\hat{x}_{\text{quot.}}}$.

Przez dobór możliwie wysokiej frakcji próby f można by pomyślnie wpływać na wielkość błędu systematycznego.

Błąd systematyczny estymatora średniej $\overline{X}$

$$B_{\hat{\overline{x}}_{\text{quot.}}} \approx \frac{\varphi}{N} v_y^2 a = \left(\frac{1}{n} - \frac{1}{N}\right) v_y^2 a$$

jest tym mniejszy, im większa jest liczebność populacji (N) i im większa jest także liczebność próby (n). Gdy n dąży do N, otrzymujemy:

$$\lim_{n \to N} \varphi = 0 \qquad \text{lub} \qquad \lim_{n \to N} \frac{\varphi}{N} = 0$$

a stąd

$$B_{\hat{x}_{\text{quot.}}} = 0 \qquad \text{lub} \qquad B_{\hat{\overline{x}}_{\text{quot.}}} = 0.$$

Rezultat ten był do przewidzenia; w badaniach wyczerpujących ($N = n$) nie ma obciążenia.

3.2. Współczynnik zmienności $v_y^2 \left(v_y^2 = \frac{s_y^2}{\overline{y}^2}\right)$

Czynnik v_y^2, pokazujący wielkość odchyleń poszczególnych wartości cechy dodatkowej od wartości średniej, wywiera istotny wpływ na rozmiar błędu systematycznego estymatora ilorazowego.

Teoretycznie obciążenie byłoby równe zeru, gdyby wariancja cechy dodatkowej była równa zeru ($s_y^2 = 0$). Cecha dodatkowa posiadająca taką własność jest na pewno nieskorelowana z innymi cechami (a więc i cechą badaną) i jest nieprzydatna jako cecha dodatkowa dla estymatora ilorazowego.

4. Wyraz wolny a prostej regresji

Wyraz wolny a prostej regresji $x^* = a + by^*$ jest odciętą punktu przecięcia się prostej regresji z osią odciętych. Gdy $a \gtrless 0$, kierunek błędu jest określony wyłącznie przez wyraz wolny a, ponieważ ani v_y^2, ani φ nie mogą przyjmować wartości ujemnych. Jeśli prosta regresji przechodzi przez początek układu współrzędnych, to $a = 0$, a zatem

$$B\hat{x}_{\text{quot.}} \approx 0 \qquad \text{lub} \qquad B\hat{\bar{x}}_{\text{quot.}} \approx 0.$$

Wykaz literatury

[1] Carstens J.: — *Der systematische Fehler* (Bias) *bei Verhaltnisschatzung und linearer Regressionschatzung.* Praca dyplomowa, Wyższa Szkoła Ekonomiczna, Berlin 1968.

[2] Cochran W. G.: *Sampling Techniques,* John Wiley and Sons, Inc., New Jork (London) Sydney 1963.

[3] Hansen M. H., Hurwitz W. N., Madow W. G.: — *Sample Survey Methods and Theory.* Vol II, John Wiley and Sons, Inc.. New Jork (London) 1953.

[4] Schwarz H.: — *Die statistische Schatzung von Gesamtwerten durch Hochrechnung von Stichprobenergebnissen.* Praca habilitacyjna, Wyższa Szkoła Ekonomiczna, Berlin 1967.

[5] *Stichproben in der amtlichen Statistik.* Praca zbiorowa. Statistisches Bundesamt Wiesbaden, Verlag W. Kohlhammer, Stuttgart — Mainz 1960.

[6] Zasępa R.: — *Badania Statystyczne metodą reprezentacyjną* PWN, Warszawa 1962.

Wacław CZERMAK

O LOSOWANIU NIEZALEŻNYM ZE ZWRACANIEM

Wstęp

Celem niniejszego referatu jest zapoznanie czytelnika z podstawami teorii estymacji przy stosowaniu losowania niezależnego *ze zwracaniem*. Klasyczne estymatory, $\bar{y}$ oraz s^2, wartości średniej $\bar{Y}$ i wariancji σ^2 populacji generalnej są porównywane z analogicznymi estymatorami zbudowanymi na podstawie różniących się od siebie jednostek pobranych do próbki. Zostały podane wariancje tych estymatorów oraz nieobciążone estymatory tych wariancji. Omawiane estymatory porównywane są parami ze względu na ich efektywność, w szczególności porównania takiego dokonuje się w stosunku do wartości średniej populacji generalnej. W zakończeniu omówione zostało tzw. losowanie inwersyjne. Rozważania teoretyczne są ilustrowane odpowiednimi przykładami, podawane są również tablice wartości funkcji wstępnych we wzorach „nowych" estymatorów.

Referat, będący uzupełnionym i rozszerzonym opracowaniem części pewnej publikacji na temat postępu w teorii badań reprezentacyjnych w ostatnim okresie [1]), podsumowuje nieomal całą wiedzę o danym zagadnieniu, zawartą w rozmaitych artykułach, rozszerzając ją zwłaszcza w zakresie losowania inwersyjnego.

1. Sformułowanie zagadnienia i stosowane symbole

W podręcznikach i monografiach z zakresu teorii badań reprezentacyjnych przy omawianiu losowania niezależnego z reguły podaje się, że przy losowaniu ze zwracaniem estymatorem wartości średniej $\bar{Y}$ z populacji generalnej, określonej wzorem:

$$\bar{Y} = \frac{1}{N} \sum_{k=1}^{N} Y_k, \tag{1.1}$$

[1]) W. Czermak, J. Wytlaczil, W. Jenczowska: *Pokroky teorie wyberovych zjistovani v poslednich 10 az 15 letech.* (Postępy w teorii badań reprezentacyjnych w ciągu ostatnich 10—15 lat). Materiały Naukowe Wydziału Naukowo-Badawczego Instytutu Statystyki i Ewidencji przy Urzędzie Statystycznym w Pradze" 1969, nr 30.

jest wartość średnia $\bar{y}$ z próbki:

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i, \tag{1.2}$$

gdzie y_i oznacza wartość cechy Y, zaobserwowaną w i-tej wylosowanej jednostce. Przy tym nie bierze się pod uwagę faktu, że jednostka została (lub nie została) wylosowana dotychczas. Estymator (1.2) posiada pewne cechy „dodatnie" uzasadniające jego stosowanie: są to przede wszystkim jego nieobciążoność i zgodność. Nie mniej cennym jest również fakt, że jego wariancję $D^2(\bar{y})$ można oszacować na podstawie wartości y_i za pomocą estymatora nieujemnego [2]).

Analogicznie dla wariancji z populacji generalnej, określonej wzorem:

$$\sigma^2 = \frac{1}{N} \sum_{k}^{N} (Y_k - \bar{Y})^2, \tag{1.3}$$

estymatorem może być wariancja obliczona z próbki na podstawie wzoru:

$$s^2 = \frac{1}{n-1} \sum_{i}^{n} (y_i - \bar{y})^2, \tag{1.4}$$

przy czym również tutaj istnieje możliwość powtarzania się pewnych jednostek w próbce. Również ten estymator posiada wspomniane zalety świadczące na jego korzyść.

Z historycznego punktu widzenia estymatory te przyjęto do teorii badań reprezentacyjnych z klasycznej teorii estymacji, z której przeniesiono je zupełnie automatycznie, głównie pod wpływem faktu, iż zwracanie elementów podczas losowania umożliwia traktowanie takiego sposobu losowania jako ciągu powtarzających się zdarzeń niezależnych. Jednakże pominięto jeden istotny szczegół: podczas gdy klasyczna teoria estymacji zakłada istnienie *nieskończonej* populacji generalnej (a więc prawdopodobieństwo ponownego wylosowania tej samej jednostki, albo inaczej — prawdopodobieństwo dokładnego powtórzenia się odpowiedniego zdarzenia jest nieskończenie małe z dowolną dokładnością), to teoria estymacji w zakresie badań reprezentacyjnych może być oparta na istnieniu *skończonej* populacji generalnej. Gdybyśmy chcieli pobrać próbkę ze skończonej populacji generalnej, to populacja ta powinna (przy losowaniu bez zwracania) lub też może (przy losowaniu ze zwracaniem) składać się z indentyfikowalnych (różniących się) jednostek. Z reguły praktycznie mamy do

---

2) Estymatorem nieujemnym nazywa się estymator, przyjmujący jedynie wartości nieujemne. W naszym przypadku np. est $D^2(\bar{y}) = s^2/n \geq 0$.

czynienia z takimi populacjami generalnymi, w których identyfikacja jednostek jest w gruncie rzeczy możliwa i dlatego staje się drugorzędnym problemem technicznym [3]).

Klasyczna teoria estymacji, której produktem są estymatory (1.2) oraz (1.4), naturalnie nie liczyła na możliwość identyfikacji jednostek. A zatem jej wzory w przypadku losowania ze zwracaniem zakładają opracowanie wszystkich n jednostek niezależnie od tego, że pewne wartości y_i tworzą wspólnie jedną i tą samą wartość Y_k. Jest to oczywiście poważny mankament, ponieważ intuicyjnie odczuwa się, że ponowne wylosowanie określonej jednostki nie jest w stanie dostarczyć żadnych nowych informacji o populacji generalnej. Dlatego też powtarzanie się jednostek w próbce jest zjawiskiem negatywnym, którego należy unikać bądź podczas tworzenia próbki (tzn. zastosować losowanie bez zwracania), bądź też scentralizować podczas *opracowania* próbki. To drugie rozwiązanie polega na wykorzystaniu estymatora uwzględniającego tylko raz wartość każdej — nawet kilkakrotnie wylosowanej — jednostki, a więc niezależnie od liczby powtórzeń tej jednostki w próbce. W niniejszym referacie całą naszą uwagę poświęcamy wyłącznie temu drugiemu rozwiązaniu.

A zatem rozpatrujemy populację generalną składającą się z N identyfikowalnych jednostek (jednostek statystycznych), przy czym N jest pewną z góry ustaloną liczbą. Poszczególne jednostki tej zbiorowości przyjmują wartości $Y_1, Y_2, \ldots, Y_k, \ldots, Y_N$ pewnej cechy Y. Suma $\sum_{k=1}^{N} Y_k$, wartość średnia $\overline{Y}$ oraz wariancja σ^2 lub też wariancja S^2, określona wzorem:

$$S^2 = \frac{1}{N-1} \sum_{k}^{N} (Y_k - \overline{Y})^2, \tag{1.5}$$

są szacowanymi parametrami populacji generalnej.

Z populacji generalnej wylosujemy stopniowo n jednostek (z jednakowym prawdopodobieństwem i ze zwracaniem), zapisując szereg wartości cechy Y w postaci $y_1, y_2, \ldots, y_i, \ldots, y_n$. Oczywiście w wyniku takiego losowania nie zawsze otrzymujemy n *różnych* jednostek, jednakże musimy przyjąć, że ich ilość d (od słowa distinct = różny, różniący się; w literaturze stosuje się również wiele innych oznaczeń,

[3]) Jeżeli rodzaj jednostek losowania wymaga, aby były one losowane na podstawie odpowiedniego *operatu losowania* (co w gruncie rzeczy ma miejsce zawsze oprócz poniżej przytoczonego przypadku), to adnotacje o wylosowaniu każdej jednostki dokonuje się bezpośrednio w operacie losowania, a tym samym w sposób bardzo prosty nie dopuszcza się do ponownego losowania tej samej jednostki (przy losowaniu „bez zwracania" lub „bez powtórzeń"), albo też odnotowuje się ilość wylosowań tej samej jednostki (przy losowaniu ze zwracaniem). Jeżeli rodzaj losowanych jednostek umożliwia ich bezpośrednie pobieranie do próbki — dotyczy to jedynie drobnych jednorodnych przedmiotów lub zwierząt — to ponownego ich wylosowania można uniknąć odkładając je, a przy losowaniu „ze zwracaniem" wylosowane jednostki można zaznaczyć w jakiś prosty sposób, np. poprzez nacięcie, farbę itp.

w szczególności ν oraz m) może być dowolną liczbą całkowitą z przedziału $< 1;\ \min\{n, N\} >$. Ciąg wartości cechy Y, ustalonych dla d różnych jednostek próbki (tzn. po wyeliminowaniu wartości przyjmowanych przez jednostki powtórnie wylosowane) zapiszemy w postaci:

$$y_{(1)},\ y_{(2)},\ \ldots,\ y_{(i)},\ \ldots,\ y_{(d)}, \tag{1.6}$$

a ich wartość średnią oznaczymy $\bar{y}_d$:

$$\bar{y}_d = \frac{1}{d} \sum_{(i)}^{d} y_{(i)}. \tag{1.7}$$

Następnie podobnie określimy również wariancję z próbki oraz zbadamy możliwości stosowania obu tych statystyk jako estymatorów odpowiednich parametrów populacji generalnej.

Ponieważ w badaniach reprezentacyjnych nie tyle jest istotna liczba wylosowanych jednostek (tzn. ile wynosi n), ile liczba jednostek, które należy zbadać (tzn. jakie jest d), zwłaszcza z punktu widzenia pracochłonności i kosztów całej operacji, można więc z góry ustalić nie n, lecz d. W takim przypadku istnieje konieczność kontynuowania losowania jednostek do próbki, aż zostanie wylosowanych d różnych jednostek. Sposób ten w literaturze nazywa się „losowaniem inwersyjnym", które szczegółowo rozpatrzymy w końcowym paragrafie niniejszego referatu.

2. Rozkład zmiennej losowej d

Przy zadanej liczbie n (liczba wylosowanych jednostek) wielkość d (liczba różnych jednostek) jest zmienną losową. Przed przystąpieniem do rozważenia naszego zasadniczego problemu, tzn. kwestii konstruowania estymatorów i zbadania ich właściwości, ustalimy rozkład zmiennej losowej d i określimy pewne niezbędne parametry tego rozkładu.

A więc pierwszym zadaniem będzie ustalenie prawdopodobieństw, z którymi zmienna d przyjmuje swoje wartości 1, 2,..., $\min\{n, N\}$. W tym celu przede wszystkim przypomnimy pewną relację rachunku prawdopodobieństwa:

Jeżeli prawdopodobieństwa zdarzeń $A_1, A_2, \ldots, A_k, \ldots, A_N$ oraz prawdopodobieństwa iloczynów tych zdarzeń oznaczymy odpowiednio

$$p_k = P\{A_k\}, \qquad p_{kk'} = P\{A_k A_{k'}\}, \text{ itd.}, \tag{2.1}$$

to dla sumy wszystkich zdarzeń $A_1, \ldots, A_N$ otrzymamy:

$$\begin{aligned} P\{A_1 \cup A_2 \cup \ldots \cup A_N\} &= S_1 - S_2 + S_3 - \ldots + (-1)\, S_N = \\ &= \sum p_k - \sum p_{kk'} + \sum p_{kk'k''} - \ldots \pm \sum p_{kk'\ldots}. \end{aligned} \tag{2.2}$$

Prawdopodobieństwo odwrotnego zdarzenia, tzn. że nie zaistnieje żadne ze zdarzeń $A_1, A_2, \ldots, A_N$, wyniesie wtedy $1 - S_1 + S_2 - S_3 + \ldots \pm S_N$.

Uwaga 2.1. Każdy wynik S_r ma oczywiście $\binom{N}{r}$ składników; wynika stąd m.in., że ostatni wynik $S_N = \sum p_{kk'} \ldots$ można zapisać w postaci $p_{123\ldots N}$.

Po tej dygresji można zająć się zasadniczym tematem, tzn. prawdopodobieństwami $P\{d\}$. W celu łatwiejszego znalezienia rozwiązania wykorzystamy ogólnie znany schemat o przypadkowym rozlokowaniu n przedmiotów w N skrzynkach, za pomocą którego można modelować nasze zagadnienie. Niech każde rozmieszczenie — będzie ich N^n — będzie jednakowo prawdopodobne. Będziemy szukać prawdopodobieństwa tego, że właśnie m skrzynek pozostaje pustych (a więc $N - m = d$ będzie zajętych).

Niech $A_k (k = 1, 2, \ldots, N)$ oznacza zdarzenia, że k-ta skrzynka jest pusta. W takim przypadku wszystkie n przedmiotów wypadnie rozlokować w pozostałych $N - 1$ skrzynkach, czego można dokonać na $(N-1)^n$ sposobów. Podobnie przy dwóch nie zajętych skrzynkach, n przedmiotów można rozlokować na $(N-2)^n$ sposobów itd. Wynika stąd, że prawdopodobieństwa p_k, $p_{kk'}$ itd. wyniosą:

$$p_k = \left(\frac{N-1}{N}\right)^n; \qquad p_{kk'} = \left(\frac{N-2}{N}\right)^n; \qquad p_{kk'k''} = \left(\frac{N-3}{N}\right)^n \text{ itd.}, \tag{2.3}$$

a ponieważ $S_r = \binom{N}{r} \cdot \left(\frac{N-r}{N}\right)^n$, więc prawdopodobieństwo tego, iż co najmniej jedna skrzynka nie jest zajęta, wynosi:

$$S_1 - S_2 + \ldots = \binom{N}{1}\left(\frac{N-1}{N}\right)^n - \binom{N}{2}\left(\frac{N-2}{N}\right)^n + \ldots = \sum_{r=1}^{N} (-1)^{r+1} \binom{N}{r}\left(\frac{N-r}{N}\right)^n. \tag{2.4}$$

Odwrotnie, prawdopodobieństwo wykorzystania wszystkich skrzynek wyniesie:

$$1 - S_1 + S_2 - \ldots = \sum_{r=0}^{N} (-1)^r \binom{N}{r}\left(\frac{N-r}{N}\right). \tag{2.5}$$

Oznaczmy to prawdopodobieństwo $p_0(n, N)$, ponieważ wyrażenie (2.5) jest funkcją wielkości n i N.

Otrzymujemy odpowiedź na pytanie: jakie jest prawdopodobieństwo tego, że akurat n skrzynek pozostaje pustych (a więc d jest zajętych)? Z jednej strony, należy pamiętać, że m skrzynek spośród N można wybrać na $\binom{N}{m}$ sposobów, z drugiej jednak strony, ze wzoru (2.5) wynika, że n przedmiotów można rozlokować w $N - m$ skrzynkach na $(N-m)^n \cdot p_0(n, N-m)$ sposobów. Jeżeli wielkość tę pomnożymy przez współczynnik $\binom{N}{m}$, to otrzymamy liczbę „sprzyjających przypadków". Iloraz będący wynikiem dzielenia tej liczby przez N^n, tzn. przez

wszystkie „możliwe przypadki", określa poszukiwane prawdopodobieństwo:

$$P\{d\} = \frac{1}{N^n}\binom{N}{m}(N-m)^n p_0(n, N-m) =$$
$$= \frac{1}{N^n}\binom{N}{m}\sum_{r=0}^{N-m}(-1)^r\binom{N-m}{r}(N-m-r)^n = \tag{2.6}$$
$$= \frac{1}{N^n}\binom{N}{d}\sum_{r=0}^{d}(-1)^r\binom{d}{r}\cdot(d-r)^n.$$

Wyrażenie

$$\sum_{r=0}^{d}(-1)^r\binom{d}{r}(d-r)^n = d^n - \binom{d}{1}(d-1)^n +$$
$$+\binom{d}{2}(d-2)^n - \ldots(-1)^{d-1}\binom{d}{d-1}1^n \tag{2.7}$$

w dalszej części referatu oznaczymy $C_d(n)$ — zgodnie z oznaczeniami stosowanymi w pracy Pathak (1961) oraz w innych pracach. Wartości tego wyrażenia dla małych d i n zawarte są w tablicy 1.

Uwaga 2.2. Wielkość $C_d(n)$ można również przedstawić w postaci:

$$\sum{}^{*}\frac{n!}{p_1!\,p_2!\ldots p_d!}, \tag{2.8}$$

gdzie $\sum^{*}$ oznacza sumę rozciągającą się na wszystkie liczby naturalne $p_1, p_2, \ldots, p_d$ spełniające warunek: $p_1 + p_2 + \ldots + p_d = n$. Wynika to z następującego rozumowania: przy n losowaniach (ze zwracaniem) liczba *możliwych* wyników wynosi N^n, liczba *trafnych* wyników (tzn. wylosowanie d różnych jednostek) jest równa:

$$\binom{N}{d}\cdot\sum{}^{*}\frac{n!}{p_1!\,p_2!\ldots p_d!},$$

ponieważ mamy wszystkich $\binom{N}{d}$ kombinacji d elementów z N oraz ilość permutacji n elementów, spośród których $p_1, p_2, \ldots, p_d$ jest identycznych, wynosi $n!/p_1!\ldots p_d!$.

Uwaga 2.3. Można łatwo udowodnić, że dla $C_d(n)$ zachodzi stosunek rekurencyjny:

$$C_d(n) = d[C_d(n-1) + C_{d-1}(n-1)]. \tag{2.9}$$

Ustalimy teraz niektóre ważne momenty rozkładu zmiennej losowej d, a mianowicie: wartość oczekiwaną $E(d)$ i wariancję $D^2(d)$. Podamy również pierwszy moment ujemny $E\left(\frac{1}{d}\right)$, z którego później skorzystamy.

Zmienna losowa d jest oczywiście sumą N zmiennych losowych a_k $(k = 1, 2, ..., N)$ przyjmujących jedynie wartość 0 lub 1:

$$d = a_1 + a_2 + ... + a_k + ... + a_N, \tag{2.10}$$

ponieważ każda jednostka populacji generalnej albo została wylosowana do próbki, albo też nie została wylosowana. Każda z tych N zmiennych losowych przyjmuje wartość zero z prawdopodobieństwem $\left(1 - \frac{1}{N}\right)^n$ oraz wartość jeden z prawdopodobieństwem $1 - \left(1 - \frac{1}{N}\right)^n$. A więc

$$E(a_k) = 1 - \left(\frac{N-1}{N}\right)^n \text{ dla } k = 1, 2, ..., N, \tag{2.11}$$

$$D^2(a_k) = \left(\frac{N-1}{N}\right)^n - \left(\frac{N-1}{N}\right)^{2n} \text{ dla } k = 1, 2, ..., N, \tag{2.12}$$

a ponieważ zachodzi

$$P\{a_k = 1;\ a_{k'} = 1\} = 1 - 2\left(\frac{N-1}{N}\right)^n + \left(\frac{N-2}{N}\right)^n, \tag{2.13}$$

to dla kowariancji $C(a_k; a_{k'})$ otrzymujemy wzór:

$$C(a_k; a_{k'}) = \left(\frac{N-2}{N}\right)^n - \left(\frac{N-1}{N}\right)^{2n} \text{ dla } k \neq k' = 1, 2, ..., N. \tag{2.14}$$

Wzory (2.11), (2.12) oraz (2.14) wykorzystamy dla ustalenia wartości oczekiwanej i wariancji zmiennej losowej $d = \sum_{k=1}^{N} a_k$:

$$E(d) = E\left(\sum_k^N a_k\right) = N \cdot E(a_k) = N\left[1 - \left(\frac{N-1}{N}\right)^n\right], \tag{2.15}$$

$$D^2(d) = D^2\left(\sum_k^N a_k\right) = \sum_k^N D^2(a_k) + \sum_{k \neq k'}^N \sum^N C(a_k; a_{k'}) =$$

$$= N\left[\left(\frac{N-1}{N}\right)^n - N\left(\frac{N-1}{N}\right)^{2n} + (N-1)\left(\frac{N-2}{N}\right)^n\right]. \tag{2.16}$$

W tablicach 2 i 3 podajemy wartości $E(d)$ i $D^2(d)$ dla małych i średnich n oraz N.

Podobnie można ustalić wartość oczekiwaną $E\left(\frac{1}{d}\right)$:

$$E\left(\frac{1}{d}\right) = \frac{1^{n-1} + 2^{n-1} + ... + N^{n-1}}{N^n}. \tag{2.17}$$

Ze względu na to, że przy dużych n i N obliczanie $E\left(\frac{1}{d}\right)$ według powyższego wzoru staje się uciążliwe byłoby korzystne posiadanie odpowiednich tablic lub uproszczonego wzoru analogicznego do (2.17).

Pewien względnie dokładny wzór przybliżony zaproponował Pathak (1961); ideę tego postępowania przedstawimy teraz pokrótce. Dla sumy r-ych stopni pierwszych N liczb naturalnych mamy [4]):

$$\sum_{j=1}^{N} j^r = \frac{N^{r+1}}{r+1} + \frac{1}{2} N^r B_0 + \frac{r}{2!} N^{r-1} B_1 - \frac{r(r-1)(r-2)}{4!} N^{r-3} B_2 \pm \ldots, \qquad (2.18)$$

gdzie B_j są to tzw. liczby Bernoulliego ($B_0 = 1, B_1 = \frac{1}{6}, B_2 = \frac{1}{30}$, $B_3 = \frac{1}{42}$ itd.). Jeżeli zamiast r wstawimy n-1, to zamiast wzoru (2.17) otrzymamy:

$$1^{n-1} + 2^{n-1} + \ldots + N^{n-1} = \frac{N^n}{n} + \frac{N^{n-1}}{2} + \frac{n-1}{2} N^{n-2} \frac{1}{6} \mp \ldots, \qquad (2.19)$$

a wtedy dla $E\left(\frac{1}{d}\right)$ po opuszczeniu wyrazów końcowych szeregu (tzn. po pominięciu wyrazów rzędu N^{-4} i niższego) otrzymujemy:

$$E\left(\frac{1}{d}\right) \doteq \frac{1}{n} + \frac{1}{2N} + \frac{n-1}{12N^2}. \qquad (2.20)$$

Można udowodnić, że ten wzór przybliżony daje wartość nie mniejszą od wartości $E\left(\frac{1}{d}\right)$ obliczonej według (2.17), tzn. można dowieść, że zachodzi relacja:

$$\frac{1}{n} + \frac{1}{2N} + \frac{n-1}{12N^2} \geqslant \frac{1^{n-1} + 2^{n-1} + \ldots + N^{n-1}}{N^n}, \qquad (2.21)$$

przy czym równość zachodzi jedynie dla $n = 3$ oraz $n = 4$. Porównania wartości $E\left(\frac{1}{d}\right)$ obliczonych według wzoru (2.20) z odpowiednimi wartościami obliczonymi według (2.17) można dokonać na podstawie tablicy 4 zawierającej wyniki obliczeń według tych wzorów dla małych n i N.

Uwaga 2.4. Z ideą zastąpienia estymatorów, obliczanych na podstawie wszystkich jednostek, estymatorami obliczanymi na podstawie tylko różniących się jednostek spotkamy się po raz pierwszy w roku 1958 od razu w trzech niezależnych pracach: Raj Khamis [10], Basu [1] oraz Hájek [2]. Rozkładowi zmiennej losowej d najwięcej uwagi poświęcono w pracy Basu (1958); ten autor m.in. nieco inną metodą wyprowadził wzory $P\{d\}$, $E(d)$ i $D^2(d)$. Później Pathak [7], o którym już wspominaliśmy, uogólnił te wyniki w tym sensie, że wyprowadził wzór dla ogólnego momentu dowolnego

[4]) Chodzi o szczególny przypadek wzoru Eulera — Maclaurina.

stopnia, tzn. wzory dla $E(d^p)$, $p = 1, 2, \ldots$, oraz ogólny wzór dla pierwszego momentu ujemnego $E\left(\frac{1}{d}\right)$. Wyniki tych uogólnień pozwoliły autorowi określić wariancję $D^2(\bar{y}_d)$ oraz nieobciążony estymator tej wariancji.

3. Estymatory wartości średniej z populacji generalnej

Wartość średnią $\bar{Y}$ z populacji generalnej przy losowaniu ze zwracaniem (oraz z jednakowym prawdopodobieństwem) można oszacować — jak to już zaznaczyliśmy na zakończenie pierwszego paragrafu — nie tylko za pomocą „zwykłego" estymatora $\bar{y}$, lecz również za pomocą estymatora $\bar{y}_d$, tzn. na podstawie *różniących* się jednostek wylosowanych do próbki. Zbadamy teraz dokładnie podstawowe własności tego estymatora, w szczególności jego nieobciążenie, zgodność oraz względną efektywność.

Dla ustalenia wartości średniej $E(\bar{y}_d)$ wykorzystamy pewne, znane z rachunku prawdopodobieństwa twierdzenie, w myśl którego wartość średnią można obliczać jako wartość średnią warunkowych wartości średnich, tzn. w naszym przypadku:

$$E[\bar{y}_d] = E[E(\bar{y}_d \mid d)] = \sum_d E(\bar{y}_d \mid d) \cdot P\{d\} =$$

$$= \sum_d \bar{Y} \cdot P\{d\} = \bar{Y} \cdot \sum_d P\{d\} = \bar{Y} \cdot 1 = \bar{Y}. \tag{3.1}$$

Przy tym wielkość warunkowej wartości średniej można ustalić chociażby w następujący sposób: sumę $\sum_{(i)} y_{(i)}$ zapiszemy w postaci $\sum_k Y_k a_k$, gdzie a_k jest zmienną losową o wartościach 0 i 1 (co oznacza, że wartość Y_k została lub nie została włączona do próbki), a następnie stopniowo otrzymamy:

$$E(\bar{y}_d \mid d) = E\left(\frac{1}{d}\sum_k^N Y_k a_k \mid d\right) = \frac{1}{d}\sum_k^N Y_k E(a_k \mid d) = \frac{1}{d}\sum Y_k \frac{d}{N} = \bar{Y}. \tag{3.2}$$

A więc statystyka $\bar{y}_d$ jest nieobciążonym estymatorem wartości średniej $\bar{Y}$ z populacji generalnej.

Istotna dla dalszych rozważań będzie znajomość wariancji $D^2(\bar{y}_d)$. Wzór na tę wariancję w oparciu o podstawowe parametry (tzn. wzór analogiczny do $D^2(\bar{y}) = \frac{\sigma^2}{n}$ lub $S^2\frac{(1-f)}{n}$) otrzymamy znów za pomocą parametru warunkowego [5]):

[5]) Słowami można to tak wyrazić: wariancja zmiennej losowej jest sumą wartości średniej warunkowych wariancji oraz wariancji warunkowych średnich wartości.

$$D^2(\bar{y}_d) = E[D^2(\bar{y}_d \mid d)] + D^2[E(\bar{y}_d \mid d)]. \tag{3.3}$$

Ponieważ $E(\bar{y}_d \mid d) = \bar{Y}$, tzn. równe jest wartości stałej, wobec tego drugi składnik w poprzednim wzorze równy jest zeru, a więc uznamy, że:

$$\begin{aligned} D^2(\bar{y}_d) &= E\left[D^2(\bar{y}_d \mid d)\right] = E\left[\frac{\sigma^2}{d} \cdot \frac{N-d}{N-1}\right] = \\ &= E\left[S^2\left(\frac{1}{d} - \frac{1}{N}\right)\right] = S^2\left[E\left(\frac{1}{d}\right) - \frac{1}{N}\right]. \end{aligned} \tag{3.4}$$

Już ta postać wyniku umożliwia wyciągnięcie pewnego ciekawego i cennego wniosku. Ponieważ $\lim_{n\to\infty} E\left(\frac{1}{d}\right) = \frac{1}{N}$, więc wariancja $D^2(\bar{y}_d)$ przy rosnącej liczbie losowanych jednostek zbieżna jest do zera. A ponieważ $E(\bar{y}_d) = \bar{Y}$, zatem $\bar{y}_d$ jest nie tylko estymatorem nieobciążonym, lecz również estymatorem zgodnym wartości średniej $\bar{Y}$ z populacji generalnej.

A więc dotychczas widzieliśmy, że oba estymatory $\bar{y}$ i $\bar{y}_d$ są równoważne (oba są nieobciążone i zgodne). Rozstrzygnąć, który z nich jest korzystniejszy, powinna trzecia rozpatrywana przez nas własność względna efektywność. W tym celu do wzoru (3.4) zamiast $E\left(\frac{1}{d}\right)$ podstawimy wyrażenie (2.17) i otrzymamy ostatecznie:

$$D^2(\bar{y}_d) = S^2 \frac{1^{n-1} + 2^{n-1} + \ldots + (N-1)^{n-1}}{N^n}. \tag{3.5}$$

Wzór ten wykorzystamy dla udowodnienia ważnego twierdzenia o relacji pomiędzy $D^2(\bar{y}_d)$ i $D^2(\bar{y})$: dla wszystkich $N \geqslant 2$ oraz dla wszystkich liczb naturalnych n zachodzi relacja:

$$D^2(\bar{y}_d) \leqslant D^2(\bar{y}), \tag{3.6}$$

przy czym równość następuje tylko dla $n = 1$ oraz $n = 2$ (w tych przypadkach estymatory $\bar{y}_d$ i $\bar{y}$ są przecież identyczne), dla $n \geqslant 3$ zachodzi ostra nierówność.

Przypadki równości można łatwo wykazać za pomocą podstawiania, natomiast przypadki nierówności udowodnimy w sposób następujący: przede wszystkim sprowadzimy nierówność

$$S^2 \frac{1^{n-1} + 2^{n-1} + \ldots + (N-1)^{n-1}}{N^n} < \frac{\sigma^2}{n} = \frac{S^2}{n} \cdot \frac{N-1}{N} \tag{3.7}$$

do postaci:

$$n\left[1^{n-1} + 2^{n-1} + \ldots + (N-1)^{n-1}\right] < (N-1)N^{n-1} \tag{3.8}$$

Tę ostatnią nierówność udowodnimy, stosując indukcję matematyczną. Przede wszystkim (3.8) zachodzi dla $N = 2$ (przy $n \geqslant 3$). Następnie należy udowodnić, że jeżeli (3.8) zachodzi dla N, to również jest prawdziwa dla $N + 1$:

$$n[1^{n-1} + 2^{n-1} + \ldots + N^{n-1}] < N(N+1)^{n-1}. \tag{3.9}$$

Jeżeli do obu stron nierówności (3.8) dodamy nN^{n-1}, to wówczas wystarczy wykazać, że

$$(N-1)N^{n-1} + nN^{n-1} \leqslant N(N+1)^{n-1}. \tag{3.10}$$

Jeżeli obie strony tej nierówności podzielimy przez N^{n-1}, to otrzymamy:

$$N - 1 + n \leqslant N\left(1 + \frac{1}{N}\right)^{n-1}. \tag{3.11}$$

Teraz znów zastosujemy indukcję matematyczną, ale tym razem w stosunku do n (tzn. N może być dowolną liczbą naturalną): 1) nierówność (3.11) jest spełniona dla $n = 3$, 2) z prawdziwości (3.11) dla n wynika również prawdziwość tej nierówności dla $n + 1$; nieskomplikowany dowód tego twierdzenia pozostawiamy czytelnikowi.

Z nierówności (3.6) wynika bezpośrednio następujący wniosek: *jeżeli ze skończonej populacji generalnej losuje się n jednostek ze zwracaniem, to wówczas* $\bar{y}_d$ *będzie równomiernie*[6]) *efektywniejszym estymatorem wartości średniej z populacji generalnej niż estymator* $\bar{y}$.

Oczywiście można byłoby udowodnić, że wniosek ten jest tylko szczególnym przypadkiem znacznie bardziej ogólnej zasady, według której estymator bazujący na różniących się jednostkach jest *zawsze* lepszy (w sensie efektywności) od estymatora uwzględniającego możliwość powtarzania się jednostek.

Chwilowo nie będziemy się zajmować zagadnieniem estymatora wariancji $D^2(\bar{y}_d)$; problemem tym zajmiemy się w przedostatnim paragrafie niniejszego referatu po zbudowaniu estymatora podstawowej wariancji S^2. Przed rozważeniem pewnych dalszych estymatorów wartości średniej z populacji generalnej zilustrujemy nasze dotychczasowe wywody teoretyczne pewnym przykładem numerycznym.

Przykład 3.1. Rozpatrzymy miniaturową populację generalną składającą się z pięciu jednostek przyjmujących następujące wartości cechy Y: 1, 2, 3, 4 oraz 10. A więc wartość średnia i wariancja w tej populacji generalnej wyniosą:

$$\bar{Y} = 4, \qquad \sigma^2 = 10, \qquad S^2 = 12{,}5.$$

[6]) Określenie „równomiernie efektywniejszy" oznacza iż $\bar{y}$ będzie estymatorem efektywniejszym w przypadku każdej populacji generalnej.

Na podstawie tej populacji generalnej utworzymy wszystkie możliwe próbki składające się z 3 jednostek (losowanie ze zwracaniem), a następnie obliczymy odpowiednie estymatory $\bar{y}$ i $\bar{y}_d$ oraz ich wartości średnie i wariancje.

$d = 1$

Y_1	Y_2	Y_3	$\bar{Y}$	$\bar{Y}_d$
1	1	1	1	1
2	2	2	2	2
3	3	3	3	3
4	4	4	4	4
10	10	10	10	10
Suma			20	20

$d = 2$

Y_1	Y_2	Y_3	$\bar{Y}$	$\bar{Y}_d$
1	1	2	1,3	1,5
1	1	3	1,6	2
1	1	4	2	2,5
1	1	10	4	5,5
2	2	1	1,6	1,5
2	2	3	2,3	2,5
2	2	4	2,6	3
2	2	10	4,6	6
3	3	1	2,3	2
3	3	2	2,6	2,5
3	3	4	3,3	3,5
3	3	10	5,3	6,5
4	4	1	3	2,5
4	4	2	3,3	3
4	4	3	3,6	3,5
4	4	10	6	7
10	10	1	7	5,5
10	10	2	7,3	6
10	10	3	7,6	6,5
10	10	4	8	7
Suma			80	80

$d = 3$

Y_1	Y_2	Y_3	$\bar{Y}$	$\bar{Y}_d$
1	2	3	2	2
1	2	4	2,3	2,3
1	2	10	4,3	4,3
1	3	4	2,6	2,6
1	3	10	4,6	4,6
1	4	10	5	5
2	3	4	3	3
2	3	10	5	5
2	4	10	5,3	5,3
3	4	10	5,6	5,6
Suma			40	40

$d = 1 : \mathrm{E}(\bar{y} \mid d) = \mathrm{E}(\bar{y}_d \mid d) = \frac{20}{5} = 4;$ $\quad D^2(\bar{y} \mid d) = D^2(\bar{y}_d \mid d) = 10;$

$d = 2 : \mathrm{E}(\bar{y} \mid d) = \mathrm{E}(\bar{y}_d \mid d) = \frac{80}{20} = 4;$ $\quad D^2(\bar{y} \mid d) = 4{,}4;\ D^2(\bar{y}_d \mid d) = 3{,}75$

$d = 3 : \mathrm{E}(\bar{y} \mid d) = \mathrm{E}(\bar{y}_d \mid d) = \frac{40}{10} = 4;$ $\quad D^2(\bar{y} \mid d) = D^2(\bar{y}_d \mid d) = 1{,}6;$

Oczywista, że $E(\bar{y}) = E(\bar{y}_d) = 4$. Ponieważ $P\{d = 1\} = 0{,}04$, $P\{d = 2\} = 0{,}48$ i $P\{d = 3\} = 0{,}48$, to dla wariancji otrzymujemy:

$$D^2(\bar{y}) = 10 \cdot 0{,}04 + 4{,}\dot{4} \cdot 0{,}48 + 1{,}\dot{6} \cdot 0{,}48 = 3{,}\dot{3}$$

$$D^2(\bar{y}_d) = 10 \cdot 0{,}04 + 3{,}75 \cdot 0{,}48 + 1{,}\dot{6} \cdot 0{,}48 = 3{,}0$$

Na zakończenie tego paragrafu w skrócie wspomnimy o pewnych innych estymatorach (wartości średniej z populacji generalnej) opartych na statystyce $\bar{y}_d$. Estymatorem wartości średniej z populacji generalnej, zbudowanym tylko na różniących się jednostkach wylosowanych do próbki, nie musi być akurat średnia $\bar{y}_d$, lecz wiele innych statystyk. Pathak [8] badał klasę estymatorów będącą prostym uogólnieniem średniej $\bar{y}_d$, którą to klasę można zapisać w postaci:

$$\text{est}_P \bar{Y} = f_1(d)\bar{y}_d + f_2(d). \tag{3.12}$$

Autor ten przede wszystkim wykazał, że warunkiem koniecznym i dostatecznym, aby estymatory te nie były obciążone, jest spełnienie równości $E[f_1(d)] = 1$ oraz $E[f_2(d)] = 0$. Następnie znalazł optymalne wartości $f_1(d)$ i $f_2(d)$ — w celu minimalizacji wariancji estymatora (3.12) — i otrzymał następujące wyniki: jeżeli istnieje jakikolwiek estymator *a priori* wartości średniej $\bar{Y}$ — estymator ten oznaczymy $\bar{Y}'$, to optymalnym estymatorem będzie

$$\text{est}_{P(1)} \bar{Y} = \frac{\frac{Nd}{N-d}}{E\left(\frac{Nd}{N-d}\right)} \bar{y}_d + \bar{Y}' \left[1 - \frac{\frac{Nd}{N-d}}{E\left(\frac{Nd}{N-d}\right)} \right]; \tag{3.13}$$

jeżeli natomiast żaden taki estymator $\bar{Y}'$ nie istnieje, to optymalnym estymatorem będzie

$$\text{est}_{P(2)} \bar{Y} = \frac{\frac{Nd}{N-d}}{E\left(\frac{Nd}{N-d}\right)} \bar{y}_d. \tag{3.14}$$

Wartości średnie $E\left(\frac{Nd}{N-d}\right)$ dla małych N i n podane są w tabl. 5. Dla średnich i dużych n obliczanie wartości średnich, a więc i estymatorów (3.13) i (3.14) jest procesem bardzo pracochłonnym. Z tego powodu warto wykorzystać ich granicę przy $\frac{n}{N} \to 0$, co można założyć przy losowaniu niewielkich frakcji:

$$\text{est}_{P(3)} \bar{Y} = \frac{d}{E(d)} \bar{y}_d + \bar{Y}' \left[1 - \frac{d}{E(d)} \right], \tag{3.15}$$

$$\text{est}_{P(4)} \bar{Y} = \frac{d}{E(d)} \bar{y}_d. \tag{3.16}$$

Te dwa ostatnie estymatory Pathaka rozpatrzymy z punktu widzenia ułamka $\frac{n}{N}$ porównując je z „wyjściowym" estymatorem $\bar{y}_d$. Przede wszystkim zauważymy, że oba są estymatorami nieobciążonymi. Jeżeli chodzi o ich efektywność, to Pathak porównał wariancje $D^2(.)$ estymatora (3.15) i estymatora $\bar{y}_d$ i stwierdził, że (3.15) posiada mniejszą wariancję wtedy i tylko wtedy, gdy w przybliżeniu [7]) zachodzi relacja:

$$\frac{S^2}{(\bar{Y} - \bar{Y}')^2} > n - 1. \tag{3.17}$$

Nierówność ta będzie oczywiście spełniona przy doborze „sprzyjającego" wstępnego estymatora $\bar{Y}'$ oraz dla bardzo małych n. Jeżeli natomiast nie istnieje żaden taki estymator $\bar{Y}'$, tzn. w przypadku stosowania estymatora (3.16), warunek $\frac{S^2}{\bar{Y}^2} > n - 1$ będzie spełniony bardzo rzadko.

Przykład 3.2. Do estymatorów $\bar{y}$ i $\bar{y}_d$ z przykładu 3.1. dołączymy estymatory (3.13) oraz (3.16) i obliczymy znów ich wartości średnie i wariancje. Założymy przy tym, że jako wstępny estymator wartości średniej z populacji generalnej mamy $\bar{Y}' = 3$. Wyniki obliczeń, zestawione w poniższej tablicy, potwierdzają korzyści estymatorów (3.13) i (3.15).

Próbka	$\bar{y}_d$	$\text{est}_{P(1)}\bar{Y}$	$\text{est}_{P(2)}\bar{Y}$	$\text{est}_{P(3)}\bar{Y}$	$\text{est}_{P(4)}\bar{Y}$
$d = 1$					
1 1 1	1	2,5238	0,2381	2,1803	0,4098
2 2 2	2	2,7619	0,4762	2,5902	0,8197
3 3 3	3	3,0000	0,7143	3,0000	1,2295
4 4 4	4	3,2381	0,9524	3,4098	1,6393
10 10 10	10	4,6667	2,3810	5,8689	4,0984
$E(\text{est}\,\bar{Y} \mid d = 1)$	4	3,2381	0,9524	3,40984	1,63934
$D^2(\text{est}\,\bar{Y} \mid d = 1)$	10	0,5669	0,5669	1,6797	1,6797

[7]) Przybliżenie wynika z rozwinięcia wzoru na $E(d)$ i $D^2(d)$ w szeregi Taylora z zachowaniem wyrazów do N^{-2} rzędu włącznie.

(dok.)

Próbka	$\bar{y}$	$\text{est}_{P(1)}\bar{Y}$	$\text{est}_{P(2)}\bar{Y}$	$\text{est}_{P(3)}\bar{Y}$	$\text{est}_{P(4)}\bar{Y}$
$d = 2$					
1 1 1	1,5	2,0476	0,9524	1,7705	1,2296
1 1 3	2	2,3651	1,2698	2,1803	1,6393
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
10 10 4	7	5,5397	4,4444	6,2787	5,7377
E (est $\bar{Y} \mid d = 2$)	4	3,6349	2,5397	3,8197	3,2787
D^2 (est $\bar{Y} \mid d = 2$)	3,75	1,5116	1,5116	2,51945	2,51945
$d = 3$					
1 2 3	2	1,5714	2,8571	1,7705	2,4590
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
3 4 10	5,6	6,8095	8,0952	6,2787	6,9672
E (est $\bar{Y} \mid d = 3$)	4	4,4286	5,7143	4,2295	4,9180
D^2 (est $\bar{Y} \mid d = 3$)	1,6	3,4012	3,4012	2,51975	2,51975
E (est $\bar{Y}$)	4	4,0000	4,0000	4,0000	4,0000
D^2 (est $\bar{Y}$)	3	2,5562	5,1867	2,5409	3,3631

4. Estymatory wariancji z populacji generalnej i wariancji $D^2(\bar{y}_d)$ i $D^2(\bar{y})$

W poprzednim paragrafie stwierdzono, że przy losowaniu ze zwracaniem średnia $\bar{y}_d$ z próby, zbudowana na podstawie różniących się jednostek wylosowanych do próbki, jest estymatorem efektywniejszym wartości średniej $\bar{Y}$ z populacji generalnej niż zwykła wartość średnia $\bar{y}$ z próbki. Można więc oczekiwać, że również przy szacowaniu wariancji σ^2 lub S^2 z populacji generalnej korzystniej będzie zamiast wariancji (1.4) z próbki zastosować pewną analogiczną statystykę zbudowaną w oparciu o różniące się jednostki wylosowane. Statystykę tę zdefiniujemy w sposób następujący:

$$s_d^2 = \begin{cases} \dfrac{1}{d-1} \sum_{(i)}^{d} (y_{(i)} - \bar{y}_d)^2, & \text{gdy } d > 1, \quad (4.1a) \\ 0, & \text{gdy } d = 1. \quad (4.1b) \end{cases}$$

Definicja s_d^2 dla $d = 1$ jest niezbędna, gdyż inaczej (4.1a) nie miałaby sensu, jednakże prowadzi do obciążenia s_d^2 jako estymatora wariancji σ^2 i S^2. Dlatego też postaramy się przede wszystkim znaleźć wartość średnią statystyki s_d^2 i na podstawie tej wartości zbudujemy nieobciążony estymator wariancji z populacji generalnej.

Zaczniemy od obliczenia *warunkowej* wartości średniej $E(s_d^2 \mid d)$. Przypadek $d = 1$ jest trywialny, ponieważ $E(s_d^2 \mid d = 1) = E(0 \mid d = 1) = 0$; przypadek $d = 2, 3, \ldots, n$ rozwiążemy przepisując s_d^2 w postaci:

$$s_d^2 = \frac{\sum_{(i)}^{d} \sum_{(i')}^{d} (y_{(i)} - y_{(i')})^2}{2d(d-1)}, \tag{4.2}$$

skąd stopniowo otrzymamy:

$$\begin{aligned} E(s_d^2 \mid d > 1) &= E\left[\frac{\sum_{(i)}^{d} \sum_{(i')}^{d} (y_{(i)} - y_{(i')})^2}{2d(d-1)} \,\middle|\, d > 1 \right] = \\ &= E\left[\frac{\sum_{k}^{N} \sum_{k'}^{N} (Y_k - Y_{k'})^2}{2d(d-1)} a_k a_{k'} \,\middle|\, d > 1 \right] = \\ &= \frac{\sum_{k}^{N} \sum_{k'}^{N} (Y_k - Y_{k'})^2}{2d(d-1)} \cdot E(a_k a_{k'} \mid d > 1) = \\ &= \frac{\sum_{k}^{N} \sum_{k'}^{N} (Y_k - Y_{k'})^2}{2d(d-1)} \cdot \frac{d(d-1)}{N(N-1)} = \frac{\sum_{k}^{N} \sum_{k'}^{N} (Y_k - Y_{k'})^2}{2N(N-1)} = S^2. \end{aligned} \tag{4.3}$$

A więc warunkowa wartość średnia wariancji s_d^2 z próbki zawsze jest równa wariancji S^2 z populacji generalnej, oczywiście tylko dla $d = 2, 3, \ldots, n$. Oznacza to, że również bezwarunkowa wartość średnia *dla tych samych wartości d* równa jest S^2. Warunek $d > 1$ będziemy dalej oznaczać gwiazdką przy E; zatem

$$E^*(s_d^2) = E^*[E(s_d^2 \mid d > 1)] = E^*[S^2] = S^2. \tag{4.4}$$

Wynika stąd, że gdybyśmy przyjęli zasadę szacowania wariancji z populacji generalnej na podstawie próbki pobranej losowo ze zwracaniem i zawierającej co najmniej dwie różne jednostki (w pierwszym przypadku wypada w ogóle zrezygnować z szacowania), to moglibyś-

my wykorzystać wariancję s_d^2 z próbki jako nieobciążony estymator wariancji z populacji generalnej. Symbolicznie można to zapisać tak:

$$\text{est}^* S^2 = s_d^2 \text{ lub } \text{est}^* \sigma^2 = \frac{N-1}{N} \cdot s_d^2. \tag{4.5}$$

Jeżeli jednak zdecydowaliśmy się na szacowanie wariancji z populacji generalnej przy każdym d (przy tym oczywiście dla $d = 1$ estymatorem wariancji z populacji generalnej byłoby 0), to należałoby ustalić wariancję s_d^2 z próbki. Najczęstszym sposobem „korekty" estymatora obciążonego jest mnożenie go przez odpowiedni współczynnik korygujący. W naszym przypadku można zdecydować, czy ten współczynnik ma zależeć — oprócz zależności od rozmiaru N populacji generalnej — od liczby wylosowanych jednostek n czy również od liczby d różnych jednostek.

Rozpatrzymy przede wszystkim pierwszy, prosty przypadek. Współczynnik korygujący K ustalimy z warunku: $E(K \cdot s_d^2) = S^2$:

$$E(s_d^2) = E^*(s_d^2) \cdot P\{d > 1\} + 0 \cdot P\{d = 1\} = S^2\left(1 - \frac{1}{N^{n-1}}\right). \tag{4.6}$$

Stąd

$$K = \frac{N^{n-1}}{N^{n-1} - 1},$$

a więc

$$\text{est } S^2 = \frac{N^{n-1}}{N^{n-1} - 1} \cdot s_d^2 \text{ oraz } \text{est } \sigma^2 = \frac{N^{n-1} - N^{n-2}}{N^{n-1} - 1} \cdot s_d^2. \tag{4.7}$$

Przy dużych N oba współczynniki będą w przybliżeniu równe 1, co pozwala na ich pominięcie w trakcie obliczeń. Wartości tych współczynników dla małych n podane są w tablicach 6 i 7.

Rozpatrzymy teraz drugi przypadek: współczynnik K_d z symbolem d u dołu, co oznacza, iż zakładamy jego zależność od d, obliczymy przyjmując $E(K_d \cdot s_d^2) = \sigma^2$.

Ponieważ $E(K_d s_d^2) = E[E(K_d s_d^2 \mid d)]$, wobec tego należy ustalić warunkowe wartości średnie przy $d = 1$ i $d > 1$. Dla $d = 1$ otrzymujemy

$$E(K_d s_d^2 \mid d = 1) = K_d \cdot E(s_d^2 \mid d = 1) = K_d \cdot E(0) = 0, \tag{4.8}$$

oraz dla $d > 1$ otrzymalibyśmy podobnie $E(K_d s_d^2 \mid d > 0) = K_d S^2$, tak więc

$$E(K_d s_d^2) = E[E(K_d s_d^2 \mid d \geqslant 1)] = 0 \cdot P\{d = 1\} + \sum_{d=2}^{n} K_d S^2 P\{d\} = \tag{4.9}$$

$$= S^2 \sum_{d=2}^{n} K_d \binom{N}{d} C_d(n) N^{-n}.$$

Z warunku, aby (4.9) równe było σ^2 lub $S^2\frac{(N-1)}{N}$, wynika, że powinno zachodzić:

$$\sum_{d=2}^{n} K_d \binom{N}{d} \frac{C_d(n)}{N^n} = 1 - \frac{1}{N}. \tag{4.10}$$

Weźmiemy jeszcze $K_d = 0$ dla $d = 1$, aby z lewej strony wzoru sumę można było napisać już od $d = 1$ (zamiast K_1 można w gruncie rzeczy podstawić dowolną wartość skończoną, ponieważ ze względu na to, iż — zgodnie z definicją — $s_d^2 = 0$ dla $d = 1$, również iloczyn $K_d s_d^2$ wyniesie zero). Jeżeli zamiast 1 z prawej strony wzoru (4.10) podstawimy $\sum_{d=}^{n} P\{d\}$, to otrzymamy:

$$\sum_{d=1}^{n} K_d \binom{N}{d} \frac{C_d(n)}{N^n} = \sum_{d=1}^{n} \binom{N}{d} \frac{C_d(n)}{N^n} - \sum_{d=1}^{n} \binom{N}{d} \frac{C_d(n)}{N^{n+1}} =$$

$$= \sum_{d=1}^{n} \binom{N}{d} \frac{C_d(n)}{N^n} - \sum_{d=1}^{m-1} \binom{N}{d} \frac{C_d(m-1)}{N^m} = \sum_{d=1}^{n} \binom{N}{d} \frac{C_d(n)}{N^n} - \sum_{d=1}^{m} \binom{N}{d} \frac{C_d(m-1)}{N^m}$$

ponieważ $C_m(m-1) = 0$; a więc dalej

$$= \sum_{d=1}^{n} \binom{N}{d} \frac{C_d(n)}{N^n} - \sum_{d=1}^{m} \binom{N}{d} \frac{C_d(m)}{N^m} \cdot \frac{C_d(m-1)}{C_d(m)} =$$

$$= \sum_{d=1}^{n} \binom{N}{d} \frac{C_d(n)}{N^n} \left[1 - \frac{C_d(n-1)}{C_d(n)}\right], \tag{4.11}$$

ponieważ jest wszystko jedno, czy w drugiej sumie będziemy pisać m czy n. Warunkiem dostatecznym (ale oczywiście nie koniecznym) dla spełnienia (4.11) jest, aby

$$K_d = 1 - \frac{C_d(n-1)}{C_d(n)}. \tag{4.12}$$

Ułamek $\frac{C_d(n-1)}{C_d(n)}$, z którym spotkamy się jeszcze niejednokrotnie, oznaczymy krótko $R_d(n)$. Wartości tego ułamka dla małych d i n zawarte są w tablicy 8, dla większych n wartości te można zastąpić wartością graniczną $\frac{1}{d}$.

A więc ostatecznym wynikiem będzie

$$\operatorname{est}\sigma^2 = [1 - R_a(n)] \cdot s_d^2, \qquad \operatorname{est} S^2 = \frac{N}{N-1}[1 - R_a(n)] \cdot s_d^2. \tag{4.13}$$

Należy dodać, że można by udowodnić, iż estymator (4.13) jest estymatorem lepszym (w sensie efektywności) niż estymator (1.4). O wzajemnym powiązaniu między estymatorami (4.13) i 4.7) na razie nic nie wiadomo; prawdopodobnie żaden z nich nie będzie równomiernie efektowniejszym od drugiego.

Przykład 4.1. Zamiast dokładnego porównania skonfrontujemy empirycznie zmienność estymatora wariancji σ^2 na podstawie pewnej konkretnej populacji generalnej. W tym celu powrócimy do naszej miniaturowej populacji generalnej z przykładu 3.1 i dokonamy na jej podstawie wszelkich możliwych losowań po 4 jednostki. Jeżeli teraz na podstawie wyników tych losowań obliczymy wariancje s^2 i s_d^2, a następnie estymatory (1.4), (4.7) i (4.13), to otrzymamy następujące wyniki:

$d = 2$

Próbka	Liczba permutacji	(1.4)	(4.7)	(4.13)
1 1 1 2	4	0,2500	0,4032	0,2857
1 1 2 2	6	0,3333	0,4032	0,2857
1 2 2 2	4	0,2500	0,4032	0,2857
1 1 1 3	4	1,0000	1,6129	1,1429
itd.	...	...	...	...
Przeciętna		7,14286	10,0806	1,14286
Wariancja		69,5629..	133,6498..	67,1020..

$d = 3$

Próbka	Liczba permutacji	(1.4)	(4.7)	(4.13)
1 1 2 3	12	0,9167	0,8065	0,8333
1 1 2 4	12	2,0000	1,8817	1,9444
itd.	...	...	...	...
Przeciętna		10,41667	10,080645	10,41667
Wariancja		62,31944	55,22676	58,96991

$d = 4$

Próbka				Liczba permutacji	(1.4)	(4.7)	(4.13)
1	2	3	4	24	1,6667	1,3441	1,6667
1	2	3	10	24	16,6667	13,4409	16,6667
1	2	4	10	24	16,2500	13,1048	16,2500
1	3	4	10	24	15,0000	12,0968	15,0000
2	3	4	10	24	12,9167	10,4167	12,9167
Przeciętna					12,5000	10,0806	12,5000
Wariacja					31,04167	20,18839	31,04167

Podsumowanie

d	$P(d)$	(1.4)		(4.7)		(4.13)	
		$E(\cdot \mid d)$	$D^2(\cdot \mid d)$	$E(\cdot \mid d)$	$D^2(\cdot \mid d)$	$E(\cdot \mid d)$	$D^2(\cdot \mid d)$
1	1/125	0	0	0	0	0	0
2	28/125	7,14286	69,5629	10,080645	133,64984	7,14286	67,10204
3	72/125	10,41667	62,3194	10,080645	55,22676	10,41667	58,96991
4	24/125	12,5	31,0417	10,080645	20,18839	12,5	31,04167
Przeciętna		10	57,4381	10	65,62435	10	54,95752
Wariacja		3,92857		0,806447		3,92857	

Wariancje bezwarunkowe otrzymamy z wyrażenia:

$$D^2(\cdot) = \mathrm{E}[D^2(\cdot \mid d)] + D^2[\mathrm{E}(\cdot \mid d)]:$$

estymator (1.4) — $D^2(\cdot) = 61{,}367$,
estymator (4.7) — $D^2(\cdot) = 66{,}431$,
estymator (4.13) — $D^2(\cdot) = 58{,}886$.

Jak z powyższego wynika, najefektywniejszym estymatorem dla danej populacji generalnej jest estymator (4.13).

* * *

Obecnie zajmiemy się zbudowaniem nieobciążonych estymatorów wariancji $D^2(\bar{y}_d)$ i $D^2(\bar{y})$. Poprzednie rozważania o estymatorze wariancji σ^2 lub S^2 przygotowały podstawy tej konstrukcji.

Zaczniemy od pewnego szczególnego przypadku dotyczącego $D^2(.)$. Dla warunkowej wariancji wartości średniej $\bar{y}_d$ mamy

$$D^2(\bar{y}_d \mid d) = S^2\left(\frac{1}{d} - \frac{1}{N}\right). \tag{4.14}$$

Ponieważ nieobciążonym estymatorem wariancji S^2 dla $d > 1$ jest wartość średnia s_d^2 z próbki, zatem za podstawę do dalszych badań można przyjąć statystykę $s_d^2\left(\frac{1}{d} - \frac{1}{N}\right)$. Obliczymy jej wartość średnią dla $d > 1$.

$$\begin{aligned} E^*\left[s_d^2\left(\frac{1}{d} - \frac{1}{N}\right)\right] &= E^*\left[E\left(\frac{1}{d} - \frac{1}{N}\right)s_d^2 \mid d > 1\right] = \\ &= E^*\left[\left(\frac{1}{d} - \frac{1}{N}\right)S^2\right] = S^2\left[E^*\left(\frac{1}{d}\right) - \frac{1}{N}\right]. \end{aligned} \tag{4.15}$$

Z tego wynika, że istotnie statystyka $s_d^2\left(\frac{1}{d} - \frac{1}{N}\right)$ jest nieobciążonym estymatorem *warunkowej wariancji* (4.14). Natomiast dla *wariancji bezwarunkowej* [8]), dla której zgodnie z (3.4) otrzymujemy

$$D^2(\bar{y}_d) = S^2\left[E\left(\frac{1}{d}\right) - \frac{1}{N}\right], \tag{4.16}$$

wyrażenie (4.15) należy jeszcze uporządkować, ponieważ $E^*\left(\frac{1}{d}\right) \neq E\left(\frac{1}{d}\right)$. Dlatego też ustalimy odpowiednie obciążenie, które później wyeliminujemy poprzez wyrażenie jego jako średniej wartości E^* pewnej statystyki obliczonej na podstawie próbki. Ponieważ

$$E^*\left[s_d^2\left(\frac{1}{d} - \frac{1}{N}\right)\right] = D^2(\bar{y}_d) + E^*\left[s_d^2 N^{1-n}\left(\frac{1}{d} - 1\right)\right], \tag{4.17}$$

co można udowodnić po niewielu przekształceniach, wobec tego

$$E^*\left[s_d^2\left(\frac{1}{d} - \frac{1}{N}\right) + s_d^2 N^{1-n}\left(1 - \frac{1}{d}\right)\right] = D^2(\bar{y}_d). \tag{4.18}$$

Innymi słowy, dla $d \geqslant 2$ nieobciążony estymator wariancji $D^2(\bar{y}_d)$ ma postać:

$$\text{est } D^2(\bar{y}_d) = s_d^2\left[\frac{1}{d} - \frac{1}{N} + N^{1-n}\left(1 - \frac{1}{d}\right)\right]. \tag{4.19}$$

[8]) Po zakończeniu losowania wartość d jest wielkością stałą, zatem odchylenie standardowe $D^2(\bar{y}_d \mid d)$ miałoby pewien sens (wskazując średnią wielkość błędu estymatora w przypadku, gdy spośród n wylosowanych jednostek jest różnych d), jednakże jest oczywistą ograniczoność tego stwierdzenia. Na przykład nie istniałaby możliwość dokonywania porównań dokładności badań reprezentacyjnych ani w przestrzeni, ani w czasie, a więc nie byłoby możliwości planowania rozmiarów próbki dla przyszłych badań itd.

Nieobciążony estymator dla $d \geqslant 1$ otrzymamy obliczając wartość średnią:

$$E\left[s_d^2\left(\frac{1}{d}-\frac{1}{N}\right)+s_d^2 N^{1-n}\left(1-\frac{1}{d}\right)\right]=E^*\left[\ldots\right]\cdot P\{d\geqslant 2\}+0\cdot P\{d=1\}=$$

$$=E^*\left[\ldots\right]\frac{N^{n-1}-1}{N^{n-1}}=\frac{N^{n-1}-1}{N^{n-1}}D^2(\bar{y}_d); \tag{4.20}$$

stąd

$$\operatorname{est} D^2(\bar{y}_d)=\frac{N^{n-1}}{N^{n-1}-1}s_d^2\left[\frac{1}{d}-\frac{1}{N}+N^{1-n}\left(1-\frac{1}{d}\right)\right]=$$

$$=s_d^2\left[\frac{1}{d}-\frac{1}{N}+\frac{N-1}{N^n-N}\right]. \tag{4.21}$$

Autorem estymatorów (4.19) i (4.12) jest Des Raj, który po raz pierwszy pisze o nich w artykule [10].

Estymator z próby wariancji $D^2(\bar{y}_d)$ można też zbudować za pomocą estymatorów (4.13). Jeżeli podstawimy je do (3.5), to otrzymamy nieobciążony estymator dla $d \geqslant 1$ w postaci:

$$\operatorname{est} D^2(\bar{y}_d)=\frac{1^{n-1}+\ldots+(N-1)^{n-1}}{N^n}\cdot\frac{N}{N-1}\cdot[1-R_d(n)]\,s_d^2. \tag{4.22}$$

Wartości współczynnika przy s_d^2 dla $d=2$ oraz dla małych N i n podane są w tablicy 9. Estymator (4.22) zaproponował Pathak (1962), który porównał go z obydwoma estymatorami Raja oraz dalszymi dwoma własnymi estymatorami, konkludując, że estymator (4.22) najprawdopodobniej będzie najefektywniejszym estymatorem spośród pięciu wymienionych dla zazwyczaj spotykanych populacji generalnych.

5. Losowanie inwersyjne

Na zakończenie pierwszego paragrafu zaznaczyliśmy, że istnieje możliwość określenia z góry, przy losowaniu ze zwracaniem, nie liczby n (ilość losowanych jednostek), lecz liczby d (ilość różnych jednostek). Ten rodzaj losowania nazywa się *losowaniem inwersyjnym* i teraz omówimy je szczegółowiej.

Przede wszystkim należy ustalić schemat losowania. Można wyróżnić dwa różne sposoby:

1) losowanie zostaje zakończone natychmiast po wylosowaniu ustalonej z góry liczby różnych jednostek; tzn. podczas gdy wśród n-1 wylosowanych jednostek znalazło się d-1 różnych, to n-ta wylosowana jednostka okazała się d-tą jednostką różną od poprzednich. Jest to tzw. *pierwszy sposób losowania inwersyjnego;* z historycznego punktu widzenia jest on starszy i w praktyce częściej stosowany;

2) po wylosowaniu d-tej jednostki losowanie jest kontynuowane aż do wylosowania następnej różniącej się jednostki; wówczas ta ostatnia jednostka jest odrzucana, a zachowuje się u poprzednio wylosowanych; tzn. losuje się tyle jednostek (ilość tę oznaczymy $n+1$), aż liczba różniących się wylosowanych jednostek osiągnie wartość $d+1$, wówczas ostatnia z kolei $(n+1)$ — jednostka, którą się okazała różniąca się jednostka $(d+1)$, jest odrzucana i w ten sposób pozostaje d różnych wśród n wylosowanych. Jest to tzw. *drugi sposób losowania inwersyjnego*, który został zastosowany później niż pierwszy i w praktyce jest rzadziej wykorzystywany [9]).

Chociaż drugi sposób losowania inwersyjnego może wydawać się zbyt skomplikowany (prowadzi bowiem „średnio" do większej liczby losowań), to jednak teoretycznie jest prostszym modelem. Dlatego też w niniejszym referacie zajmiemy się głównie tym drugim sposobem, a o pierwszym zaledwie wspomnimy, czyniąc w kilku miejscach odpowiednie uwagi. Upewnimy się przy tym, że sposób pierwszy prowadzi do bardziej złożonych rozważań teoretycznych i złożonych konstrukcji.

Załóżmy, że zastosujemy pierwszy lub drugi sposób losowania inwersyjnego. Wówczas z teoretycznego punktu widzenia losowanie takie będzie odwrotnym przypadkiem losowania zwykłego. Podczas gdy przy z góry ustalonym n (liczba losowanych jednostek) liczba d jest zmienną losową, to w przypadku losowania inwersyjnego d jest wielkością stałą, a zmienną losową jest liczba losowanych jednostek n. Rozkład tej zmiennej losowej należy zbadać.

Wróćmy teraz do drugiego sposobu losowania inwersyjnego. Przy ustalonym d zmienna losowa n przyjmuje wartości: d, $d+1$, ...in. inf., z prawdopodobieństwem

$$P\{n\} = \binom{N-1}{d} N^{-n} C_d(n). \tag{5.1}$$

Wynika to z faktu, że liczba „pozostawionych" lub nie odrzuconych jednostek wyniesie, przy określonym schemacie losowania, dokładnie n pod warunkiem, że:

a) n pierwszych wylosowanych jednostek stanowią d różnych jednostek;

b) $(n+1)$ jednostka wylosowana przyniosła $(d+1)$ jednostkę różną od poprzednich.

---

[9]) Ma to miejsce w tych przypadkach, gdy dla obliczania estymatorów nie potrzebna jest liczba n i wystarczą tylko N i d. W takim przypadku po osiągnięciu liczby d losowanie zostaje przerwane, a reszta losowania jest już tylko fikcją. W konsekwencji zanika więc praktycznie różnica pomiędzy pierwszym i drugim sposobem losowania inwersyjnego.

Prawdopodobieństwo pierwszego zdarzenia określone jest wzorem (2.6), prawdopodobieństwo drugiego zdarzenia równe jest $\left(\frac{N-d}{N}\right)$ a ich iloczyn daje (5.1).

Znając rozkład zmiennej losowej n, nietrudno będzie wyprowadzić wzory jej podstawowych parametrów: wartości średniej $E(n)$, wariancji $D^2(n)$ oraz ewentualnie pierwszego momentu ujemnego $E\left(\frac{1}{n}\right)$. Tak więc przede wszystkim

$$
\begin{aligned}
E(n) &= \sum_{n=d}^{\infty} n.P\{n\} = \sum_{n} n \binom{N-1}{d} N^{-n} C_d(n) = \\
&= \binom{N-1}{d} \sum_{n} \frac{n}{N^n} \left[d^n - \binom{d}{1}(d-1)^n + \binom{d}{2}(d-2)^n \pm \ldots\right] = \\
&= \binom{N-1}{d} \sum_{n} n \left(\frac{d}{N}\right)^n - \binom{N-1}{d}\binom{d}{1} \sum_{n} n \left(\frac{d-1}{N}\right)^n + \\
&\quad + \binom{N-1}{d}\binom{d}{2} \sum_{n} n \left(\frac{d-2}{N}\right)^n \pm \ldots
\end{aligned}
\tag{5.2}
$$

Na przykład dla $d = 1$, jak łatwo się przekonać [10]), otrzymujemy $\frac{N}{(N-1)}$, zaś dla $d = 2$ będziemy mieli

$$\frac{2N^2 - 3N}{(N-1)(N-2)} = \frac{N}{N-1} + \frac{N}{N-2}.$$

Wyniki te pozwalają wnioskować, że dla ogólnej wartości d otrzymamy

$$E(n) = \frac{N}{N-1} + \frac{N}{N-2} + \ldots + \frac{N}{N-d}. \tag{5.3}$$

Można to udowodnić np. za pomocą indukcji matematycznej:

a) prawdziwość (5.3) dla $d = 1$ i $d = 2$ została już wykazana;

b) z prawdziwości (5.3) dla d-1 względnie łatwo można wykazać prawdziwość tego wzoru również dla d.

---

10) Stosuje się wzór dla sumy nieskończonego szeregu

$$\sum_{n=d}^{\infty} n \cdot q^n = \frac{d\,q^d - d\,q^{d+1} + q^{d+1}}{(1-q)^2} \quad \text{(gdzie } q < 1\text{)},$$

przy tym zamiast q należy podstawić $\frac{d}{N}$ lub $\frac{(d-1)}{N}$.

Ponieważ chodzi nam o wariancję $D^2(n)$, więc przede wszystkim należy obliczyć wartość średnią $E(n^2)$, stosując wzór

$$\sum_{n=d}^{\infty} n^2 q^n = \frac{q^{d+1} + q^d [q(d-1) - d]^2}{(1-q)^3}.$$

W ten sposób, np. dla $d = 1$ otrzymujemy $\frac{(N^2+N)}{(N-1)^2}$, tak więc $D^2(n) = \frac{N}{(N-1)^2}$; dla $d = 2$ otrzymujemy $\frac{(5N^4 - 9N^3 + N^2 + 6N)}{(N-1)^2(N-2)^2}$, a więc $D^2(n) = \frac{N}{(N-1)^2} + \frac{2N}{(N-2)^2}$ itd. Stąd wynika ogólny wzór, którego słuszność można dowieść metodą indukcji matematycznej:

$$D^2(n) = \frac{N}{(N-1)^2} + \frac{2N}{(N-2)^2} + \ldots + \frac{dN}{(N-d)^2}. \tag{5.4}$$

Pozostaje jeszcze ustalić wzór na wartość średnią $E\left(\frac{1}{n}\right)$, z którego skorzystamy na końcu tego paragrafu. A więc:

$$E\left(\frac{1}{n}\right) = \sum_{n=d}^{\infty} \frac{1}{n} \cdot P\{n\} = \sum_{n} \frac{1}{n} \binom{N-1}{d} N^{-n} C_d(n) =$$

$$= \binom{N-1}{d} \sum_{n} \frac{1}{nN} \sum_{j=0}^{d-1} (-1)^j \binom{d}{j} (d-j)^n =$$

$$= \binom{N-1}{d} \sum_{j=0}^{d-1} (-1)^j \binom{d}{j} \sum_{n} \frac{1}{n} \left(\frac{d-j}{N}\right)^n.$$

Korzystając ze wzoru

$$\sum_{n=d}^{\infty} \frac{1}{n} Q^n = -\ln(1-Q) - \sum_{n=1}^{d-1} \frac{1}{n} Q^n, \quad \text{gdzie } |Q| < 1$$

po pewnych przekształceniach algebraicznych, których tu nie sposób, a nawet nie potrzeba przytaczać, otrzymamy ostateczny wzór:

$$E\left(\frac{1}{n}\right) = \binom{N-1}{d} \sum_{j=0}^{d-1} (-1)^{j+1} \binom{d}{j} \ln \frac{N-d+j}{N}. \quad \text{gdzie } |Q| < 1$$

Tak na przykład dla $d = 1$, $d = 2$ i $d = 3$ otrzymujemy:

$$E\left(\frac{1}{n}\right) = (N-1)\left[-\ln\left(1 - \frac{1}{N}\right)\right],$$

$$E\left(\frac{1}{n}\right) = \binom{N-1}{2}\left[-\ln\left(1 - \frac{2}{N}\right) + 2\ln\left(1 - \frac{1}{N}\right)\right],$$

$$E\left(\frac{1}{n}\right) = \binom{N-1}{3}\left[-\ln\left(1 - \frac{3}{N}\right) + 3\ln\left(1 - \frac{2}{N}\right) - 3\ln\left(1 - \frac{1}{N}\right)\right].$$

Wartości średnie $E(n)$ i $E\left(\frac{1}{n}\right)$ oraz wariancja $D^2(n)$ dla małych i średnich d i N podane są w tablicach 10, 11, 12. Dla dużych N wartość średnią $E\left(\frac{1}{n}\right)$ można obliczyć z poniższego przybliżonego wzoru, który wyprowadzimy w sposób następujący:

$$E\left(\frac{1}{n}\right) = \sum_{n=d}^{\infty} \frac{1}{n} \binom{N-1}{d} N^{-n} C_d(n) =$$

$$= \frac{(N-1)^{[d]}}{d!} \left[\frac{C_d(d)}{dN^d} + \frac{C_d(d+1)}{(d+1)N^{d+1}} + \frac{C_d(d+2)}{(d+2)N^{d+2}} + \ldots \right] =$$

$$= \frac{N^d - \frac{1}{2}d(d+1)N^{d-1} + \frac{1}{24}d(d-1)(d+1)(3d+2)N^{d-2}}{d!} \cdot \tag{5.6}$$

$$\cdot \left[\frac{d!}{dN^d} + \frac{d(d+1)!}{2(d+1)N^{d+1}} + \frac{d(3d+1)(d+2)}{24(d+2)N^{d+2}} + \ldots \right] =$$

$$= \frac{1}{d} - \frac{1}{2} \cdot \frac{1}{N} - \frac{d+1}{12} \cdot \frac{1}{N^2} - \ldots$$

Uwaga 5.1.. Im większa jest populacja generalna, tym „szybciej" otrzymujemy żądaną liczbę d różnych jednostek, tzn. n jest stochastycznie zbieżne do d. Zgodnie z tym faktem prawdziwe są granice:

$$\lim_{N\to\infty} E(n) = d, \qquad \lim_{N\to\infty} D^2(n) = 0, \qquad \lim_{N\to\infty} E\left(\frac{1}{n}\right) = \frac{1}{d}.$$

Uwaga 5.2. W przypadku tzw. pierwszego sposobu losowania inwersyjnego zmienna losowa n przyjmuje wartości d, $d+1$,... in inf. z prawdopodobieństwem

$$P\{n\} = \binom{N-1}{d-1} N^{1-n} C_{d-1}(n-1). \tag{5.7}$$

Podstawowe parametry rozkładu zmiennej losowej n wyrażają się w tym przypadku wzorami

$$E(n) = 1 + \frac{N}{N-1} + \frac{N}{N-2} + \ldots + \frac{N}{N-d+1} \tag{5.8}$$

$$D^2(n) = \frac{N}{(N-1)^2} + \frac{2N}{(N-2)^2} + \ldots + \frac{(d-1)N}{(N-d+1)^2} \tag{5.9}$$

$$E\left(\frac{1}{n}\right) = \binom{N-1}{d-1} \left[(-1)^{d-1} + \sum_{j=0}^{d-2} (-1)^{j+1} \binom{d-1}{j} \cdot \right. \tag{5.10}$$

$$\left. \cdot \frac{N}{d-1-j} \ln \frac{N-d+1+j}{N} \right]$$

Rozpatrzymy teraz *własności estymatora* $\bar{y}_d$ w przypadku losowania inwersyjnego (sposób drugi). Estymator ten jest przede wszystkim *nieobciążony;* dowód podobny jest tutaj do postępowania przy (3.1) i (3.2):

$$E(\bar{y}_d) = E[E(\bar{y}_d \mid n)] = E[\bar{Y}] = \bar{Y}, \tag{5.11}$$

ponieważ

$$E(\bar{y}_d \mid n) = E\left(\frac{1}{d}\sum_k^N Y_k a_k \mid n\right) = \frac{1}{d}\sum_k Y_k E(a_k \mid n) = \frac{1}{d}\sum Y_k \frac{d}{N} = \bar{Y}, \tag{5.12}$$

gdzie a_k znów oznacza zmienną losową o wartościach 0 i 1.

Jeżeli chodzi o *wariancję* estymatora $\bar{y}_d$, to przy losowaniu inwersyjnym pozostaje w mocy

$$D^2(\bar{y}_d) = E[D^2(\bar{y}_d \mid n)] + D^2[E(\bar{y}_d \mid n)] =$$
$$= E\left[S^2\left(\frac{1}{d} - \frac{1}{N}\right)\right] + D^2[\bar{Y}] = S^2\left(\frac{1}{d} - \frac{1}{N}\right). \tag{5.13}$$

Wynik wskazuje, że badany schemat losowania i estymacji (tzn. losowanie ze zwracaniem przy założonej liczbie różnych jednostek, a następnie skorzystanie z estymatora $\bar{y}_d$ w rezultacie posiada taką samą względną efektywność, jak i klasyczne losowanie d jednostek różnych bez zwracania.

Porównamy jeszcze efektywność estymatora $\bar{y}_d$ z efektywnością estymatora $\bar{y}$ (oczywiście również przy losowaniu inwersyjnym). W tym celu obliczymy wariancję $D^2(\bar{y})$ i porównamy ją z wariancją $D^2(\bar{y}_d)$. Wzór na wariancję $D^2(\bar{y})$ będzie tu oczywiście nieco bardziej skomplikowany. Przede wszystkim ustalimy *warunkową wariancję* $D^2(\bar{y} \mid n)$, a następnie obliczymy jej wartość średnią $E[D^2(\bar{y} \mid n)] = D^2(\bar{y})$.

Ponieważ

$$D^2(\bar{y} \mid n) = E(\bar{y}^2 \mid n) - \bar{Y}^2,$$

wystarczy więc obliczyć wartość średnią $E(\bar{y}^2 \mid n)$:

$$E(\bar{y}^2 \mid n) = E\left[\left(\frac{1}{n}\sum_k^N Y_k a_k\right)^2 \mid n\right] = \frac{1}{n^2}E\left[\left(\sum_k^N Y_k^2 a_k^2 + \right.\right.$$
$$\left.\left. + \sum_{k \neq k'}\sum Y_k Y_{k'} a_k a_{k'} \mid n\right)\right] = \frac{1}{n^2}\left[\sum_k^N Y_k^2 E(a_k^2) \mid n + \sum_{k \neq k'}\sum Y_k Y_{k'} E(a_k a_{k'}) \mid n\right], \tag{5.14}$$

gdzie tym razem a_k oznacza zmienną losową przyjmującą wartości $0, 1, \ldots, n-d+1$. Przy tzw. drugim sposobie losowania inwersyjnego wartości te przyjmowane są z prawdopodobieństwem:

$$P\{0 \mid n\} = \frac{(N-d)}{N}$$

oraz dla pozostałych a_k

$$P\{a_k \mid n\} = \frac{d}{N}\binom{n}{a_k}\frac{C_{d-1}(n-a_k)}{C_d(n)}. \tag{5.15}$$

Wzór ten można wyprowadzić w sposób następujący:

1) ogólna liczba przypadków (możliwych próbek) wynosi $\binom{N}{d} C_d(n)$, ponieważ z N elementów można utworzyć $\binom{N}{d}$ kombinacji z d elementami, a dalej liczba permutacji z powtórzeniami jest równa (por. uwagę 2.2)

$$\sum^{*} \frac{n!}{p_1! \ldots p_d!} = C_d(n);$$

2) liczba sprzyjających przypadków (tzn. gdy k-ta jednostka populacji generalnej została wylosowana do próbki) podobnie wyniesie

$$\binom{N-1}{d-1}\sum^{*}\frac{n!}{p_1!\ldots p_{d-1}!\,a_k!} = \binom{N-1}{d-1}\sum^{*}\frac{n!}{a_k!\,(n-a_k)!}\cdot\frac{(n-a_k)!}{p_1!\ldots p_{d-1}!} =$$
$$= \binom{N-1}{d-1}\binom{n}{a_k} C_{d-1}(n-a_k).$$

Uwaga 5.3. Przy tzw. pierwszym sposobie losowania inwersyjnego wzory na prawdopodobieństwo $P\{a_k\}$ są nieco bardziej złożone:

$$P\{0 \mid n\} = \frac{N-d}{N}, \tag{5.16}$$

$$P\{1 \mid n\} = \frac{1}{N}\left[1 + (d-1)(n-1)\frac{C_{d-2}(n-2)}{C_{d-1}(n-1)}\right], \tag{5.17}$$

a dla pozostałych a_k

$$P\{a_k \mid n\} = \frac{d-1}{N}\binom{n-1}{a_k}\frac{C_{d-2}(n-a_k-1)}{C_{d-1}(n-1)}. \tag{5.18}$$

Uwaga 5.4. Wzór (5.15) wykorzystamy do wyprowadzenia wzoru na wartość średnią $E(a_k \mid n)$ i na podstawie tego wyniku udowodnimy, że estymator $\bar{y}$ przy losowaniu inwersyjnym jest estymatorem nieobciążonym. Przede wszystkim mamy

$$E(a_k \mid n) = \sum_{a_k=0}^{n-d+1} a_k \cdot P\{a_k \mid n\} = \sum_{1}^{n-d+1} a_k \frac{d}{N} \binom{n}{a_k} \frac{C_{d-1}(n-a_k)}{C_d(n)} =$$

$$= \frac{dn}{N C_d(n)} \sum_{1}^{n-d+1} \binom{n-1}{a_k-1} C_{d-1}(n-a_k), \tag{5.19}$$

a ponieważ dla prawdopodobieństwa $P\{a_k \mid n\}$ zachodzi

$$\sum_{a_k=1}^{n-d+1} P\{a_k \mid n\} = \frac{d}{N}, \tag{5.20}$$

więc zgodnie z (5.15) otrzymamy ważny stosunek

$$\sum_{a_k=1}^{n-d+1} \binom{n}{a_k} \cdot C_{d-1}(n-a_k) = C_d(n). \tag{5.21}$$

Zastosowanie tego wzoru do (5.19) prowadzi do wyniku:

$$E(a_k \mid n) = \frac{n}{N}. \tag{5.22}$$

Stąd wynika

$$E(\bar{y} \mid n) = E\left(\frac{1}{n}\sum_{k}^{N} Y_k a_k \mid n\right) = \frac{1}{n}\sum_{k}^{N} Y_k E(a_k \mid n) = \frac{1}{n}\sum_{k} Y_k \frac{n}{N} = \overline{Y}. \tag{5.23}$$

Zwrócimy jeszcze uwagę na ten fakt, że wzory (5.22) i (5.23) pozostałyby w mocy również przy pierwszym sposobie losowania inwersyjnego.

Wróćmy teraz do wzoru (5.14). Dla wartości średniej $E(a_k^2$ po pewnych przekształceniach można uzyskać poniższą postać:

$$E(a_k^2) = \sum_{a_k=1}^{n-d+1} a_k^2 \cdot P\{a_k\} = \frac{n}{N}[1 + (n-1) R_d(n)], \tag{5.24}$$

a dla wartości średniej $E(a_k; a_{k'}) = \sum a_k a_{k'} P\{a_k; a_{k'}\}$ — ze względu na to, iż $P\{a_k; a_{k'}\}$ dla niezerowych a_k i $a_{k'}$ wynosi

$$P\{a_k; a_{k'}\} = \frac{d(d-1)}{N(N-1)} \binom{a}{a_k} \binom{n-a_k}{a_{k'}} \frac{C_{d-2}(n-a_k-a_{k'})}{C_d(n)} \tag{5.25}$$

otrzymujemy

$$E(a_k a_{k'}) = \frac{n(n-1)}{N(N-1)}[1 - R_d(n)]. \tag{5.26}$$

Po podstawieniu tych średnich do (5.14) otrzymamy

$$E(\bar{y}^2 \mid n) = \frac{1}{nN}\left([1 + (n-1)R_d(n)]\sum Y_k^2 + \right.$$
$$\left. + \frac{n-1}{N-1}[1 - R_d(n)]\sum\sum Y_k Y_{k'}\right), \tag{5.27}$$

skąd

$$D^2(\bar{y} \mid n) = \left(\sum Y_k^2 - \frac{1}{N-1}\sum\sum Y_k Y_{k'}\right)\left[\frac{N-n}{nN^2} + \frac{n-1}{nN}R_d(n)\right] =$$
$$= \frac{\sigma^2}{n}\cdot\frac{N-n}{N-1} + \frac{n-1}{n}\cdot\frac{N}{N-1}\sigma^2 R_d(n) = \tag{5.28}$$
$$= S^2\left[\frac{1}{n} - \frac{1}{N} + \frac{n-1}{n}R_d(n)\right].$$

Wariancję $D^2(\bar{y})$ uzyskamy, jak już wspominaliśmy, w postaci wartości średniej $E[D^2(\bar{y} \mid n)]$:

$$D^2(\bar{y}) = \sum_{n=d}^{\infty} D^2(\bar{y} \mid n).P\{n\} = \sum_n S^2\left[\frac{1}{n} - \frac{1}{N} + \frac{n-1}{n}\cdot\right.$$
$$\left.\cdot\frac{C_d(n-1)}{C_d(n)}\right]\cdot P\{n\} = S^2\left[E\left(\frac{1}{n}\right) - \frac{1}{N}E\left(\frac{1}{n+1}\right)\right]. \tag{5.29}$$

Współczynnik przy S^2 (tzn. zawartość nawiasów kwadratowych) dla małych d i N podany jest w tablicy 13. Dla dużych N można go wystarczająco dokładnie obliczyć za pomocą pierwszych trzech składników szeregu wykładniczego otrzymanego podobnie jak (5.6):

$$E\left(\frac{1}{n}\right) - \frac{1}{N}E\left(\frac{1}{n+1}\right) = \frac{1}{d} - \frac{d+3}{2(d+1)}\cdot\frac{1}{N} - \frac{(d-1)(d-2)}{12(d+2)}\cdot\frac{1}{N^2} - \ldots \tag{5.30}$$

Relacje pomiędzy wariancjami $D^2(\bar{y}_d)$ i $D^2(\bar{y})$ oraz względną efektywność estymatorów $\bar{y}_d$ i $\bar{y}$ można ustalić na podstawie wariancji warunkowych (5.13) i (5.28). Z relacji

$$S^2\left(\frac{1}{d} - \frac{1}{N}\right) \leqslant S^2\left[\frac{1}{n} - \frac{1}{N} + \frac{n-1}{n}R_d(n)\right], \tag{5.31}$$

gdzie równość następuje jedynie przy $d = 1$ i $d = n$, gdy $C_d(n-1) = 0$, wynika, że estymator $\bar{y}_d$ nawet przy losowaniu inwersyjnym jest efektywniejszy niż estymator $\bar{y}$.

Uwaga 5.5. W przypadku pierwszego sposobu losowania inwersyjnego dla zmiennej losowej a_k mielibyśmy:

$$E(a_k) = \frac{n}{N} \tag{5.32}$$

$$E(a_k^2) = \frac{n}{N}\left[1 + \frac{(n-1)(n-2)}{n^2} R_d(n-1)\right] \tag{5.33}$$

$$E(a_k a_{k'}) = \frac{n(n-1)}{N(N-1)}\left[1 - \frac{n-2}{n} R_{d-1}(n-1)\right] \tag{5.34}$$

Na podstawie tych parametrów można wyprowadzić wzory na wariancję warunkową estymatora $\bar{y}$, a następnie wzór na wariancję $D^2(\bar{y})$:

$$D^2(\bar{y} \mid n) = S^2\left[\frac{1}{n} - \frac{1}{N} + \frac{(n-1)(n-2)}{n^2} R_{d-1}(n-1)\right] \tag{5.35}$$

$$D^2(\bar{y}) = S^2\left[\frac{N-3}{N} E\left(\frac{1}{n}\right) + \frac{2}{N} E\left(\frac{1}{n^2}\right)\right]. \tag{5.36}$$

Inny sposób wyprowadzenia powyższych wzorów można znaleźć w pracy Chikkagoudara [4].

* * *

Jako ostatnią kwestię związaną z losowaniem inwersyjnym zbadamy problem szacowania wariancji z populacji generalnej oraz wariancji $D^2(.)$. Uzyskane poprzednio wyniki umożliwiają nam krótkie dowiedzenie twierdzenia:

1. Wariancja s_d^2 z próbki dla $d \geqslant 2$ jest nieobciążonym estymatorem wariancji S^2 z populacji generalnej, ponieważ:

$$E(s_d^2 \mid n) = E\left[\frac{\sum\sum_{k \neq k'}(Y_k - Y_{k'})^2}{2d(d-1)} a_k a_{k'} \mid n\right] =$$
$$= \frac{\sum\sum(Y_k - Y_{k'})^2}{2d(d-1)} \cdot \frac{d(d-1)}{N(N-1)} = S^2. \tag{5.37}$$

Stąd

$$E^*(s_d^2) = E^*[E(s_d^2 \mid n)] = E^*[S^2] = S^2.$$

2. Wariancja s^2 z próbki jest nieobciążonym estymatorem wariancji σ^2 z populacji generalnej, ponieważ

$$E(s^2 \mid n) = E\left[\frac{\sum\sum(Y_k - Y_{k'})^2}{2n(n-1)} a_k a_{k'} \mid n\right] =$$
$$= \frac{\sum\sum(Y_k - Y_{k'})^2}{2n(n-1)} \cdot \frac{n(n-1)}{N(N-1)}[1 - R_d(n)] = S^2[1 - R_d(n)] \tag{5.38}$$

Stąd

$$E(s^2) = \sum_{n=d}^{\infty} S^2[1 - R_d(n)] \cdot P\{n\} = \sigma^2.$$

3. Statystyka $s_d^2\left(\frac{1}{d} - \frac{1}{N}\right)$ dla $d \geqslant 2$ jest nieobciążonym estymatorem wariancji $D^2(\bar{y}_d)$, ponieważ

$$E\left[s_d^2\left(\frac{1}{d} - \frac{1}{N}\right)\right] = E\left(E\left[s_d^2\left(\frac{1}{d} - \frac{1}{N}\right) \mid n\right]\right) =$$

$$= E\left[\left(\frac{1}{d} - \frac{1}{N}\right) E(s_d^2 \mid n)\right] = E\left[\left(\frac{1}{d} - \frac{1}{N}\right) S^2\right] =$$

$$= S^2\left(\frac{1}{d} - \frac{1}{N}\right) = D^2(\bar{y}_d) \qquad \text{zgodnie z (5.13).}$$

4. Statystyki

a) $$s^2 \frac{N}{N-1}\left(\frac{1}{n} - \frac{1}{N} + \frac{n-1}{n} R_d(n)\right),$$

b) $$s_d^2\left(\frac{1}{n} - \frac{1}{N} + \frac{n-1}{n} R_d(n)\right),$$

c) $$s_d^2\left[E\left(\frac{1}{n}\right) - \frac{1}{N} E\left(\frac{1}{n+1}\right)\right]$$

są nieobciążonymi estymatorami wariancji $D^2(\bar{y})$, ponieważ

a) $$E\left(\frac{N}{N-1} s^2\left[\frac{1}{n} - \frac{1}{N} + \frac{n-1}{n} R_d(n)\right]\right) =$$

$$= \frac{N}{N-1} E\left(\left[\frac{1}{n} - \frac{1}{N} + \frac{n-1}{n} R_d(n)\right] E[s^2 \mid n]\right) =$$

$$= \frac{N}{N-1} \sigma^2 E\left[\frac{1}{n} - \frac{1}{N} + \frac{n-1}{n} R_d(n)\right] = D^2(\bar{y}) \quad \text{zgodnie z (5.29).}$$

b) $$E\left(s_d^2\left[\frac{1}{n} - \frac{1}{N} + \frac{n-1}{n} R_d(n)\right]\right) =$$

$$= E\left(S^2\left[\frac{1}{n} - \frac{1}{N} + \frac{n-1}{n} R_d(n)\right]\right) =$$

$$= E[D^2(\bar{y} \mid n)] = D^2(\bar{y}) \qquad \text{zgodnie z (5.29).}$$

c) $$E\left(s_d^2\left[E\left(\frac{1}{n}\right) - \frac{1}{N} E\left(\frac{1}{n+1}\right)\right]\right) =$$

$$= \left[E\left(\frac{1}{n}\right) - \frac{1}{N} E\left(\frac{1}{n+1}\right)\right] \cdot E[s_d^2 \mid n] = D^2(\bar{y}) \quad \text{zgodnie z (5.29).}$$

Tablica 1

Wartości $C_d(n)$

n \ d	1	2	3	4	5	6	7	8	9	10
1 . . .	1									
2 . . .	1	2								
3 . . .	1	6	6							
4 . . .	1	14	36	24						
5 . . .	1	30	150	240	120					
6 . . .	1	62	540	1560	1800	720				
7 . . .	1	126	1806	8400	16800	15120	5040			
8 . . .	1	254	5796	40824	126000	191520	141120	40320		
9 . . .	1	510	18150	186480	834120	1905120	2328480	1451520	362880	
10 . . .	1	1022	55980	818520	5103000	16435440	29635200	30240000	16329600	3628800
11 . . .	1	2046	171006	3498000	29607600	129230640	322494480	479001600	419126400	199584000
12 . . .	1	4094	519156	14676024	165528000	953029440	3162075840	6411968640	8083152000	6187104000

Tablica 2

Wartości nadziei matematycznej E (*d*)

n	*N*															
	2	3	4	5	6	7	8	9	10	15	20	25	30	40	50	100
2 . .	1,500	1,667	1,750	1,800	1,833	1,857	1,875	1,889	1,900	1,933	1,950	1,960	1,966	1,975	1,980	1,987
3 . .	1,750	2,111	2,312	2,440	2,528	2,592	2,641	2,679	2,710	2,805	2,852	2,882	2,901	2,925	2,940	2,967
4 . .	1,875	2,407	2,734	2,952	3,165	3,222	3,311	3,381	3,439	3,618	3,710	3,766	3,804	3,852	3,882	3,936
5 . .	1,938	2,605	3,051	3,362	3,589	3,761	3,897	4,006	4,095	4,376	4,524	4,616	4,678	4,756	4,804	4,896
6 . .	1,969	2,737	3,288	3,689	3,991	4,224	4,410	4,561	4,686	5,085	5,298	5,431	5,522	5,637	5,708	5,846
7 . .	1,984	2,824	3,466	3,951	4,326	4,621	4,858	5,054	5,217	5,746	6,033	6,214	6,338	6,496	6,594	6,786
8 . .	1,992	2,883	3,600	4,161	4,605	4,960	5,251	5,492	5,695	6,363	6,732	6,965	7,126	7,334	7,462	7,717
9 . .	1,996	2,922	3,700	4,329	4,837	5,252	5,595	5,882	6,126	6,938	7,395	7,687	7,889	8,151	8,313	8,640
10 . .	1,998	2,948	3,775	4,463	5,031	5,502	5,895	6,228	6,513	7,476	8,025	8,379	8,626	8,947	9,146	9,552
11 . .	1,999	2,965	3,831	4,571	5,192	5,716	6,158	6,536	6,862	7,977	8,624	9,044	9,338	9,723	9,963	10,456
12 . .	2,000	2,977	3,873	4,656	5,327	5,899	6,389	6,810	7,176	8,446	9,193	9,682	10,027	10,480	10,764	11,350
13 . .	2,000	2,985	3,905	4,725	5,439	6,056	6,590	7,053	7,458	8,883	9,733	10,295	10,693	11,218	11,549	12,236
14 . .	2,000	2,990	3,929	4,780	5,533	6,191	6,766	7,270	7,712	9,290	10,247	10,883	11,336	11,938	12,318	13,112
15 . .	2,000	2,993	3,947	4,824	5,611	6,307	6,921	7,462	7,941	9,671	10,734	11,448	11,959	12,639	13,072	13,980
16 . .	2,000	2,995	3,960	4,859	5,675	6,406	7,055	7,633	8,147	10,026	11,197	11,990	12,560	13,323	13,810	14,840
17 . .	2,000	2,997	3,970	4,887	5,730	6,491	7,174	7,785	8,332	10,358	11,638	12,510	13,141	13,990	14,534	15,690
18 . .	2,000	2,998	3,977	4,910	5,775	6,563	7,277	7,920	8,499	10,667	12,056	13,010	13,703	14,640	15,243	16,532
19 . .	2,000	2,999	3,983	4,928	5,812	6,626	7,367	8,040	8,649	10,956	12,453	13,490	14,246	15,274	15,938	17,366
20 . .	2,000	2,999	3,987	4,942	5,843	6,679	7,446	8,147	8,784	11,226	12.830	13,950	14,772	15,892	16,620	18,182
22 . .	2,000	3,000	3,993	4,963	5,891	6,764	7,576	8,326	9,015	11,712	13,529	14,816	15,770	17,083	17,941	19,818
24 . .	2,000	3,000	3,996	4,976	5,925	6,827	7,675	8,467	9,202	12,136	14,160	15,615	16,703	18,214	19,211	21,412
26 . .	2,000	3,000	3,998	4,985	5,948	6,873	7,752	8,579	9,354	12,505	14,730	16,350	17,574	19,290	20,430	22,983
28 . .	2,000	3,000	3,999	4,990	5,964	6,907	7,810	8,667	9,477	12,827	15,243	17,029	18,389	20,313	21,601	24,505
30 . .	2,000	3,000	3,999	4,994	5,975	6,931	7,854	8,737	9,576	13,107	15,707	17,654	19,150	21,285	22,726	26,004
35 . .	2,000	3,000	4,000	4,998	5,990	6,968	7,925	8,854	9,750	13,659	16,678	19,010	20,842	23,510	25,346	29,628
40 . .	2,000	3,000	4,000	4,999	5,996	6,985	7,962	8,919	9,852	14,050	17,430	20,116	22,270	25,471	27,715	33,073
45 . .	2,000	3,000	4,000	5,000	5,998	6,993	7,980	8,955	9,913	14,327	18,011	21,018	23,475	27,198	29,856	36,350
50 . .	2,000	3,000	4,000	5,000	5,999	6,997	7,990	8,975	9,948	14,524	18,461	21,753	24,492	28,720	31,792	39,466

Wartości wariancji $D^2(d)$

Tablica 3

n	N															
	2	3	4	5	6	7	8	9	10	15	20	25	30	40	50	100
2 . .	0,250	0,222	0,188	0,160	0,139	0,122	0,109	0,099	0,090	0,062	0,047	0,038	0,032	0,024	0,020	0,010
3 . .	0,188	0,321	0,340	0,326	0,305	0,283	0,261	0,243	0,226	0,166	0,131	0,108	0,091	0,070	0,057	0,029
4 . .	0,109	0,316	0,414	0,446	0,447	0,435	0,417	0,398	0,378	0,296	0,240	0,201	0,173	0,134	0,110	0,057
5 . .	0,059	0,264	0,423	0,509	0,548	0,559	0,555	0,544	0,528	0,440	0,367	0,313	0,272	0,215	0,177	0,094
6 . .	0,030	0,202	0,393	0,526	0,605	0,648	0,666	0,670	0,664	0,588	0,505	0,438	0,385	0,309	0,257	0,139
7 . .	0,015	0,147	0,343	0,509	0,626	0,702	0,747	0,771	0,781	0,733	0,650	0,573	0,510	0,414	0,348	0,191
8 . .	0,008	0,104	0,287	0,471	0,619	0,726	0,799	0,846	0,874	0,872	0,795	0,714	0,642	0,530	0,448	0,251
9 . .	0,004	0,072	0,234	0,422	0,591	0,715	0,825	0,896	0,944	1,000	0,939	0,859	0,781	0,653	0,557	0,317
10 . .	0,002	0,048	0,186	0,370	0,550	0,705	0,829	0,923	0,993	1,115	1,078	1,002	0,922	0,783	0,674	0,390
11 . .	0,001	0,034	0,146	0,318	0,502	0,672	0,815	0,931	1,021	1,216	1,211	1,145	1,066	0,917	0,796	0,468
12 . .	0,000	0,024	0,114	0,269	0,451	0,630	0,789	0,923	1,032	1,303	1,335	1,284	1,209	1,055	0,924	0,553
13 . .	0,000	0,015	0,087	0,225	0,400	0,582	0,752	0,902	1,029	1,376	1,450	1,419	1,351	1,196	1,056	0,642
14 . .	0,000	0,010	0,067	0,187	0,352	0,533	0,710	0,871	1,012	1,434	1,555	1,548	1,490	1,338	1,191	0,747
15 . .	0,000	0,007	0,051	0,154	0,306	0,483	0,663	0,833	0,986	1,478	1,649	1,670	1,626	1,480	1,329	0,836
16 . .	0,000	0,005	0,039	0,127	0,265	0,434	0,614	0,789	0,953	1,510	1,733	1,784	1,757	1,622	1,469	0,940
17 . .	0,000	0,003	0,029	0,103	0,228	0,388	0,564	0,743	0,913	1,530	1,806	1,891	1,884	1,763	1,610	1,047
18 . .	0,000	0,002	0,022	0,084	0,195	0,344	0,516	0,695	0,869	1,540	1,869	1,990	2,004	1,903	1,752	1,159
19 . .	0,000	0,001	0,017	0,068	0,166	0,304	0,469	0,646	0,823	1,540	1,921	2,080	2,119	2,040	1,894	1,274
20 . .	0,000	0,001	0,013	0,055	0,141	0,268	0,425	0,598	0,775	1,532	1,964	2,162	2,227	2,174	2,035	1,392
22 . .	0,000	0,000	0,007	0,036	0,101	0,206	0,344	0,505	0,679	1,493	2,022	2,301	2,423	2,432	2,315	1,637
24 . .	0,000	0,000	0,004	0,023	0,072	0,156	0,275	0,422	0,586	1,433	2,048	2,408	2,592	2,673	2,588	1,891
26 . .	0,000	0,000	0,002	0,015	0,050	0,118	0,218	0,348	0,501	1,356	2,045	2,484	2,733	2,897	2,851	2,154
28 . .	0,000	0,000	0,001	0,010	0,035	0,088	0,172	0,285	0,424	1,270	2,019	2,533	2,847	3,102	3,104	2,424
30 . .	0,000	0,000	0,001	0,006	0,025	0,066	0,134	0,232	0,356	1,178	1,973	2,556	2,936	3,287	3,343	2,698
35 . .	0,000	0,000	0,000	0,002	0,010	0,031	0,072	0,135	0,224	0,946	1,800	2,522	3,056	3,659	3,876	3,397
40 . .	0,000	0,000	0,000	0,001	0,004	0,015	0,037	0,077	0,138	0,734	1,581	2,392	3,054	3,908	4,310	4,099
45 . .	0,000	0,000	0,000	0,000	0,002	0,007	0,019	0,044	0,084	0,556	1,350	2,203	2,960	4,045	4,643	4,790
50 . .	0,000	0,000	0,000	0,000	0,001	0,005	0,015	0,035	0,068	0,413	1,129	1,983	2,803	4,086	4,880	5,457

Tablica 4

Wartości nadziei matematycznej E (1/*d*)

Wiersz pierwszy zawsze oznacza dokładną wartość obliczoną wg wzoru (2.17), zaś drugi wiersz oznacza wartość przybliżoną obliczoną wg wzoru (2.20)

n	*N*															
	2	3	4	5	6	7	8	9	10	15	20	25	30	40	50	100
2 . .	0,750	0,667	0,625	0,600	0,583	0,571	0,562	0,556	0,550	0,533	0,525	0,520	0,517	0,512	0,510	0,505
	0,771	0,676	0,630	0,603	0,586	0,573	0,564	0,557	0,551	0,534	0,525	0,520	0,517	0,513	0,510	0,505
3 . .	0,625	0,519	0,469	0,440	0,421	0,408	0,398	0,391	0,385	0,367	0,359	0,354	0,350	0,346	0,343	0,338
	0,625	0,519	0,469	0,440	0,421	0,408	0,398	0,391	0,385	0,367	0,359	0,354	0,350	0,346	0,343	0,338
4 . .	0,562	0,444	0,391	0,360	0,340	0,327	0,316	0,309	0,302	0,284	0,276	0,270	0,267	0,263	0,260	0,255
	0,562	0,444	0,391	0,360	0,340	0,327	0,316	0,309	0,302	0,284	0,276	0,270	0,267	0,263	0,260	0,255
5 . .	0,531	0,403	0,346	0,313	0,293	0,278	0,268	0,260	0,253	0,235	0,226	0,221	0,217	0,213	0,210	0,205
	0,533	0,404	0,346	0,313	0,293	0,278	0,268	0,260	0,253	0,235	0,226	0,221	0,217	0,213	0,210	0,205
6 . .	0,516	0,379	0,317	0,283	0,262	0,247	0,236	0,227	0,221	0,202	0,193	0,187	0,184	0,179	0,177	0,172
	0,521	0,380	0,318	0,283	0,262	0,247	0,236	0,227	0,221	0,202	0,193	0,187	0,184	0,179	0,177	0,172
7 . .	0,508	0,363	0,298	0,263	0,240	0,224	0,213	0,205	0,198	0,178	0,169	0,164	0,160	0,156	0,153	0,148
	0,518	0,365	0,299	0,263	0,240	0,224	0,213	0,205	0,198	0,178	0,169	0,164	0,160	0,156	0,153	0,148
8 . .	0,504	0,353	0,285	0,248	0,224	0,208	0,197	0,188	0,181	0,161	0,151	0,146	0,142	0,138	0,135	0,130
	0,521	0,356	0,286	0,248	0,225	0,208	0,197	0,188	0,181	0,161	0,151	0,146	0,142	0,138	0,135	0,130
9 . .	0,502	0,346	0,276	0,237	0,213	0,196	0,184	0,175	0,168	0,147	0,138	0,132	0,129	0,124	0,121	0,116
	0,528	0,352	0,278	0,238	0,213	0,196	0,184	0,175	0,168	0,147	0,138	0,132	0,129	0,124	0,121	0,116
10 . .	0,501	0,342	0,269	0,229	0,204	0,186	0,174	0,165	0,157	0,137	0,127	0,121	0,117	0,113	0,110	0,105
	0,538	0,350	0,272	0,230	0,204	0,187	0,174	0,165	0,158	0,137	0,127	0,121	0,118	0,113	0,110	0,105

Tablica 4 (dok.)

Wartości nadziei matematycznej E (1/*d*)

n	*N*															
	2	3	4	5	6	7	8	9	10	15	20	25	30	40	50	100
11 . .	0,500	0,339	0,264	0,223	0,197	0,179	0,166	0,157	0,149	0,128	0,118	0,112	0,109	0,104	0,101	0,096
	0,549	0,350	0,268	0,224	0,197	0,179	0,166	0,157	0,149	0,128	0,118	0,112	0,109	0,104	0,101	0,096
12 . .	0,500	0,337	0,261	0,218	0,191	0,173	0,160	0,150	0,142	0,121	0,111	0,105	0,101	0,096	0,094	0,088
	0,562	0,352	0,266	0,220	0,192	0,173	0,160	0,150	0,142	0,121	0,111	0,105	0,101	0,096	0,094	0,088
13 . .	0,500	0,336	0,258	0,214	0,187	0,168	0,155	0,145	0,137	0,115	0,104	0,099	0,095	0,090	0,087	0,082
	0,577	0,355	0,264	0,217	0,188	0,169	0,155	0,145	0,137	0,115	0,104	0,099	0,095	0,090	0,087	0,082
14 . .	0,500	0,335	0,256	0,211	0,183	0,164	0,150	0,140	0,132	0,110	0,099	0,093	0,089	0,085	0,082	0,077
	0,592	0,358	0,264	0,215	0,185	0,165	0,151	0,140	0,132	0,110	0,099	0,093	0,089	0,085	0,082	0,077
15 . .	0,500	0,334	0,254	0,209	0,180	0,161	0,147	0,136	0,128	0,105	0,095	0,089	0,085	0,080	0,077	0,072
	0,608	0,363	0,265	0,213	0,182	0,162	0,147	0,137	0,128	0,105	0,095	0,089	0,085	0,080	0,077	0,072
16 . .	0,500	0,334	0,253	0,207	0,178	0,158	0,144	0,133	0,125	0,101	0,091	0,084	0,081	0,076	0,073	0,068
	0,625	0,368	0,266	0,212	0,181	0,159	0,145	0,133	0,125	0,101	0,091	0,084	0,081	0,076	0,073	0,068
17 . .	0,500	0,334	0,253	0,206	0,176	0,156	0,141	0,130	0,122	0,098	0,087	0,081	0,077	0,072	0,069	0,064
	0,642	0,374	0,267	0,212	0,179	0,157	0,142	0,131	0,122	0,098	0,087	0,081	0,077	0,072	0,069	0,064
18 . .	0,500	0,334	0,252	0,205	0,174	0,154	0,139	0,128	0,119	0,095	0,084	0,078	0,074	0,069	0,066	0,061
	0,660	0,380	0,269	0,212	0,178	0,156	0,140	0,129	0,120	0,095	0,084	0,078	0,074	0,069	0,066	0,061
19 . .	0,500	0,334	0,251	0,204	0,173	0,152	0,137	0,126	0,117	0,092	0,081	0,075	0,071	0,066	0,063	0,058
	0,678	0,386	0,271	0,213	0,178	0,155	0,139	0,127	0,118	0,093	0,081	0,075	0,071	0,066	0,063	0,058
20 . .	0,500	0,333	0,251	0,203	0,172	0,151	0,135	0,124	0,115	0,090	0,079	0,073	0,068	0,063	0,061	0,055
	0,696	0,393	0,274	0,213	0,177	0,154	0,137	0,125	0,116	0,090	0,079	0,073	0,068	0,063	0,061	0,055

Wartości nadziei matematycznej $E\left(\frac{Nd}{N-d}\right)$

N \ n	1	2	3	4	5	6	7	8	9	10	11	12
2 . . .	2,000											
3 . . .	1,500	4,500										
4 . . .	1,333	3,333	6,833									
5 . . .	1,250	2,917	5,250	8,917								
6 . . .	1,200	2,700	4,617	7,158	10,823							
7 . . .	1,167	2,567	4,267	6,367	9,034	12,628						
8 . . .	1,143	2,476	4,043	5,901	8,135	10,877	14,382					
9 . . .	1,125	2,411	3,887	5,592	7,575	9,907	12,692	16,113				
10 . . .	1,111	2,361	3,772	5,370	7,189	9,273	11,678	14,487	17,833			
11 . . .	1,100	2,322	3,683	5,203	6,906	8,820	10,984	13,444	16,266	19,547		
12 . . .	1,091	2,291	3,613	5,073	6,688	8,480	10,474	12,703	15,205	18,032	21,263	
13 . . .	1,083	2,265	3,556	4,968	6,515	8,214	10,083	12,145	14,427	16,961	19,793	22,978
14 . . .	1,077	2,244	3,509	4,882	6,375	8,000	9,772	11,708	13,827	16,152	18,714	21,546
15 . . .	1,071	2,225	3,469	4,810	6,259	7,825	9,519	11,356	13,350	15,517	17,881	20,463

Tablica 6

Wartości $\frac{N^{n-1}}{N^{n-1}-1}$

N \ n	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	16
2 . .	2,0000	1,3333	1,1429	1,0667	1,0323	1,0159	1,0079	1,0039	1,0020	1,0010	1,0005	1,0002	1,0001	1,0001	1,0000	1,0000
3 . .	1,5000	1,1250	1,0385	1,0125	1,0041	1,0014	1,0005	1,0002	1,0001	1,0000						1,0000
4 . .	1,3333	1,0667	1,0159	1,0039	1,0010	1,0002	1,0001	1,0000								1,0000
5 . .	1,2500	1,0417	1,0081	1,0016	1,0003	1,0001	1,0000									1,0000
6 . .	1,2000	1,0286	1,0047	1,0008	1,0001	1,0000										1,0000
7 . .	1,1667	1,0208	1,0029	1,0004	1,0001	1,0000										1,0000
8 . .	1,1429	1,0159	1,0020	1,0002	1,0000											1,0000
9 . .	1,1250	1,0125	1,0014	1,0002	1,0000											1,0000
10 . .	1,1111	1,0101	1,0010	1,0001	1,0000											1,0000
11 . .	1,1000	1,0083	1,0008	1,0001	1,0000											1,0000
12 . .	1,0909	1,0070	1,0006	1,0000												1,0000
13 . .	1,0833	1,0060	1,0005	1,0000												1,0000
14 . .	1,0769	1,0051	1,0004	1,0000												1,0000
15 . .	1,0714	1,0045	1,0003	1,0000												1,0000
16 . .	1,0667	1,0039	1,0002	1,0000												1,0000
17 . .	1,0625	1,0035	1,0002	1,0000												1,0000
18 . .	1,0588	1,0031	1,0002	1,0000												1,0000
19 . .	1,0556	1,0028	1,0001	1,0000												1,0000
20 . .	1,0526	1,0025	1,0001	1,0000												1,0000
25 . .	1,0417	1,0016	1,0001	1,0000												1,0000
30 . .	1,0345	1,0011	1,0000													1,0000
40 . .	1,0256	1,0006	1,0000													1,0000
50 . .	1,0204	1,0004	1,0000													1,0000
100. .	1,0101	1,0001	1,0000													1,0000
∞ . .	1,0000															1,0000

Wartości $\frac{N^{n-1} - N^{n-2}}{N^{n-1} - 1}$

N \ n	2	3	4	5	6	7	8	9	10	11	12	13	14	15	>15
2	1,0000	0,6667	0,5714	0,5333	0,5161	0,5079	0,5039	0,5020	0,5010	0,5005	0,5002	0,5001	0,5001	0,5000	0,5000
3	1,0000	0,7500	0,6923	0,6750	0,6694	0,6676	0,6670	0,6668	0,6667						0,6667
4	1,0000	0,8000	0,7619	0,7529	0,7507	0,7502	0,7500								0,7500
5	1,0000	0,8333	0,8065	0,8013	0,8003	0,8001	0,8000								0,8000
6	1,0000	0,8571	0,8372	0,8340	0,8334	0,8334	0,8333								0,8333
7	1,0000	0,8750	0,8596	0,8575	0,8572	0,8572	0,8571								0,8571
8	1,0000	0,8889	0,8767	0,8752	0,8750										0,8750
9	1,0000	0,9000	0,8901	0,8890	0,8889										0,8889
10	1,0000	0,9091	0,9009	0,9001	0,9000										0,9000
11	1,0000	0,9167	0,9098	0,9092	0,9091										0,9091
12	1,0000	0,9231	0,9172	0,9167											0,9167
13	1,0000	0,9286	0,9235	0,9231											0,9231
14	1,0000	0,9333	0,9289	0,9286											0,9286
15	1,0000	0,9375	0,9336	0,9334	0,9333										0,9333
16	1,0000	0,9412	0,9377	0,9375											0,9375
17	1,0000	0,9444	0,9414	0,9412											0,9412
18	1,0000	0,9475	0,9446	0,9445	0,9444										0,9444
19	1,0000	0,9500	0,9475	0,9474											0,9474
20	1,0000	0,9524	0,9501	0,9500											0,9500
25	1,0000	0,9615	0,9601	0,9600											0,9600
30	1,0000	0,9677	0,9667												0,9667
40	1,0000	0,9756	0,9750												0,9750
50	1,0000	0,9804	0,9800												0,9800
100	1,0000	0,9901	0,9900												0,9900
∞	1,0000	1,0000													1,0000

Tablica 8

Wartości $R_d(n) = \dfrac{C_d(n-1)}{C_d(n)}$

n \ d	2	3	4	5	6	7	8	9	10
2 . .	0,00000								
3 . .	0,33333	0,00000							
4 . .	0,42857	0,16667	0,00000						
5 . .	0,46667	0,24000	0,10000	0,00000					
6 . .	0,48387	0,27778	0,15385	0,06667	0,00000				
7 . .	0,49206	0,29900	0,18571	0,10714	0,04762	0,00000			
8 . .	0,49606	0,31159	0,20576	0,13333	0,07895	0,03571	0,00000		
9 . .	0,49804	0,31934	0,21892	0,15106	0,10053	0,06061	0,02778	0,00000	
10 . .	0,49902	0,32422	0,22783	0,16346	0,11598	0,07911	0,04800	0,02222	0,00000
11 . .	0,49951	0,32736	0,23400	0,17235	0,12716	0,09169	0,06334	0,03896	0,01818
12 . .	0,49976	0,32939	0,23835	0,17887	0,13559	0,10187	0,07474	0,05194	0,03226

Tablica 9

Wartości $\dfrac{1^{n-1} + \ldots + (N-1)^{n-1}}{N^n} \cdot \dfrac{N}{N-1} \cdot [1 - R_d(n)]$

(przypadek $d = 2$)

N \ n	2	3	4	5	6	7	8	9	10
2 . .	0,50000	0,16667	0,07143	0,03333	0,01613	0,00794	0,00394	0,00196	0,00098
3 . .	0,50000	0,18519	0,09524	0,05597	0,03505	0,02264	0,01486	0,00983	0,00653
4 . .	0,50000	0,19444	0,10714	0,06806	0,04637	0,03282	0,02375	0,01741	0,01287
5 . .	0,50000	0,20000	0,11429	0,07552	0,05368	0,03974	0,03016	0,02324	0,01811
6 . .	0,50000	0,20370	0,11905	0,08058	0,05874	0,04467	0,03486	0,02767	0,02222
7 . .	0,50000	0,20635	0,12245	0,08422	0,06245	0,04833	0,03842	0,03109	0,02548
8 . .	0,50000	0,20833	0,12500	0,08698	0,06527	0,05116	0,04120	0,03380	0,02808
9 . .	0,50000	0,20988	0,12698	0,08913	0,06750	0,05340	0,04343	0,03598	0,03021
10 . .	0,50000	0,21111	0,12857	0,09086	0,06929	0,05522	0,04524	0,03778	0,03197
15 . .	0,50000	0,21481	0,13333	0,09608	0,07476	0,06081	0,05089	0,04342	0,03757
20 . .	0,50000	0,21667	0,13571	0,09871	0,07753	0,06368	0,05382	0,04638	0,04054
25 . .	0,50000	0,21778	0,13714	0,10030	0,07921	0,06543	0,05561	0,04820	0,04237
30 . .	0,50000	0,21852	0,13810	0,10135	0,08034	0,06660	0,05681	0,04943	0,04362
40 . .	0,50000	0,21944	0,13929	0,10268	0,08175	0,06807	0,05833	0,05098	0,04520
50 . .	0,50000	0,22000	0,14000	0,10347	0,08260	0,06896	0,05926	0,05193	0,04616
70 . .	0,50000	0,22063	0,14082	0,10438	0,08357	0,06999	0,06031	0,05301	0,04727
100 . .	0,50000	0,22111	0,14143	0,10507	0,08431	0,07076	0,06111	0,05384	0,04811
∞ . .	0,50000	0,22222	0,14286	0,10667	0,08602	0,07256	0,06299	0,05577	0,05010

Wartości nadziei matematycznej E (*n*)

Tablica 10

N \ d	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
2	2,000														
3	1,500	4,500													
4	1,333	3,333	7,333												
5	1,250	2,917	5,417	10,417											
6	1,200	2,700	4,700	7,700	13,700										
7	1,167	2,567	4,317	6,650	10,150	17,150									
8	1,143	2,476	4,076	6,076	8,743	12,743	20,743								
9	1,125	2,411	3,911	5,711	7,961	10,961	15,461	24,461							
10	1,111	2,361	3,790	5,456	7,456	9,956	13,290	18,290	28,290						
11	1,100	2,322	3,697	5,269	7,102	9,302	12,052	15,719	21,219	32,219					
12	1,091	2,291	3,624	5,124	6,839	8,839	11,239	14,239	18,239	24,239	36,239				
13	1,083	2,265	3,565	5,010	6,635	8,492	10,658	13,258	16,508	20,842	27,342	40,342			
14	1,077	2,244	3,516	4,916	6,472	8,222	10,222	12,555	15,355	18,855	23,522	30,522	44,522		
15	1,071	2,225	3,475	4,839	6,339	8,006	9,881	12,023	14,523	17,523	21,273	26,273	33,773	48,773	
16	1,067	2,210	3,440	4,774	6,228	7,828	9,606	11,606	13,892	16,558	19,758	23,758	29,092	37,092	53,092
17	1,063	2,196	3,410	4,718	6,134	7,680	9,380	11,269	13,394	15,822	18,656	22,056	26,306	31,972	40,472
18	1,059	2,184	3,384	4,670	6,054	7,554	9,191	10,991	12,991	15,241	17,812	20,812	24,412	28,912	34,912
19	1,056	2,173	3,361	4,627	5,985	7,446	9,029	10,757	12,657	14,768	17,143	19,857	23,024	26,824	31,574
20	1,053	2,164	3,340	4,590	5,924	7,352	8,891	10,557	12,375	14,375	16,598	19,098	21,955	25,288	29,288
25	1,042	2,129	3,265	4,455	5,705	7,021	8,410	9,881	11,443	13,110	14,896	16,819	18,902	21,175	23,675
30	1,034	2,106	3,217	4,371	5,571	6,821	8,125	9,489	10,917	12,417	13,996	15,663	17,428	19,303	21,303
40	1,026	2,078	3,159	4,270	5,413	6,590	7,802	9,052	10,342	11,676	13,055	14,483	15,965	17,503	19,103
50	1,020	2,062	3,126	4,213	5,324	6,460	7,623	8,814	10,033	11,283	12,565	13,881	15,232	16,621	18,050
100	1,010	2,031	3,061	4,103	5,156	6,220	7,295	8,382	9,481	10,592	11,715	12,852	14,001	15,164	16,340

Tablica 11

Wartości wariancji $D^2(n)$

N \ d	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
2 . . .	2,000														
3 . . .	0,750	6,750													
4 . . .	0,444	2,444	14,444												
5 . . .	0,312	1,424	5,174	25,174											
6 . . .	0,240	0,990	2,990	8,990	38,990										
7 . . .	0,194	0,754	2,067	5,178	13,928	55,928									
8 . . .	0,163	0,608	1,568	3,568	8,012	20,012	76,012								
9 . . .	0,141	0,508	1,258	2,698	5,510	11,510	27,260	99,260							
10 . . .	0,123	0,436	1,048	2,159	4,159	7,909	15,687	35,687	125,687						
11 . . .	0,110	0,382	0,897	1,795	3,323	5,963	10,775	20,553	45,303	155,303					
12 . . .	0,099	0,339	0,784	1,534	2,758	4,758	8,118	14,118	26,118	56,118	188,118				
13 . . .	0,090	0,305	0,695	1,337	2,353	3,945	6,472	10,632	17,945	32,389	68,139	224,139			
14 . . .	0,083	0,277	0,624	1,184	2,049	3,361	5,361	8,472	13,512	22,262	39,373	81,373	263,373		
15 . . .	0,077	0,254	0,567	1,062	1,812	2,924	4,564	7,013	10,763	16,763	27,075	47,076	95,826	305,826	
16 . . .	0,071	0,234	0,518	0,963	1,624	2,584	3,967	5,967	8,906	13,350	20,390	32,390	55,501	111,501	351,501
17 . . .	0,066	0,218	0,478	0,880	1,470	2,313	3,503	5,182	7,573	11,042	16,237	24,397	38,209	64,654	128,404
18 . . .	0,062	0,203	0,443	0,810	1,343	2,093	3,134	4,574	6,574	9,387	13,427	19,427	28,787	44,537	74,537
19 . . .	0,059	0,183	0,406	0,744	1,228	1,903	2,827	4,083	5,793	8,138	11,404	16,057	22,918	33,558	51,371
20 . . .	0,055	0,179	0,386	0,699	1,143	1,756	2,584	3,695	5,183	7,183	9,899	13,649	18,955	26,733	38,733
25 . . .	0,043	0,138	0,293	0,520	0,832	1,248	1,788	2,480	3,359	4,470	5,873	7,648	9,905	12,798	16,548
30 . . .	0,036	0,112	0,236	0,413	0,653	0,966	0,363	1,859	2,471	3,221	4,135	5,246	6,595	8,236	10,236
40 . . .	0,026	0,082	0,169	0,293	0,456	0,664	1,921	1,233	1,608	2,052	2,576	3,188	3,901	4,730	5,690
50 . . .	0,021	0,064	0,132	0,227	0,350	0,505	0,694	0,921	1,189	1,501	1,863	2,278	2,753	3,293	3,906
100 . . .	0,010	0,031	0,063	0,106	0,162	0,230	0,311	0,405	0,514	0,637	0,776	0,931	1,103	1,292	1,500

Tablica 12

Wartości nadziei matematycznej E (1/*n*)

N \ *d*	1	2	3	4	5	6	7	8	9
2 . .	0,693								
3 . .	0,811	0,288							
4 . .	0,863	0,353	0,170						
5 . .	0,893	0,387	0,213	0,117					
6 . .	0,912	0,408	0,237	0,148	0,087				
7 . .	0,925	0,423	0,253	0,166	0,111	0,069			
8 . .	0,935	0,433	0,264	0,178	0,125	0,087	0,056		
9 . .	0,942	0,441	0,273	0,188	0,135	0,099	0,072	0,047	
10 . .	0,948	0,447	0,279	0,195	0,143	0,108	0,081	0,060	0,041

Tablica 13

Wartości $E\left(\frac{1}{n}\right) - \frac{1}{N} E\left(\frac{1}{n+1}\right)$

N \ d	2	3	4	5	6	7	8	9	10
3 . .	0,21597								
4 . .	0,28972	0,13448							
5 . .	0,33248	0,17849	0,09611						
6 . .	0,36066	0,20567	0,12648	0,07393					
7 . .	0,38068	0,22452	0,14586	0,09667	0,05957				
8 . .	0,39566	0,23844	0,15973	0,11148	0,07750	0,04957			
9 . .	0,40729	0,24917	0,17023	0,12230	0,08934	0,06422	0,04224		
10 . .	0,41658	0,25771	0,17850	0,13064	0,09812	0,07400	0,05453	0,03666	
11 . .	0,42418	0,26467	0,18520	0,13732	0,10499	0,08133	0,06279	0,04717	0,03228
12 . .	0,43051	0,27044	0,19073	0,14279	0,11055	0,08713	0,06905	0,05429	0,04142
13 . .	0,43586	0,27532	0,19539	0,14737	0,11516	0,09187	0,07404	0,05971	0,04764
14 . .	0,44045	0,27950	0,19936	0,15126	0,11905	0,09583	0,07815	0,06407	0,05240
15 . .	0,44442	0,28311	0,20279	0,15461	0,12238	0,09920	0,08161	0,06768	0,05626
20 . .	0,45832	0,29573	0,21473	0,16618	0,13384	0,11052	0,09317	0,07876	0,06849
25 . .	0,46666	0,30329	0,22184	0,17303	0,14068	0,11640	0,09981	0,08618	0,07524
30 . .	0,47222	0,30828	0,22655	0,17759	0,14507	0,12000	0,10414	0,09050	0,07957
40 . .	0,47916	0,31456	0,23243	0,18342	0,15001	0,12703	0,10946	0,09579	0,08488
50 . .	0,48334	0,31832	0,23595	0,18676	0,15334	0,13023	0,11262	0,09891	0,08795
100 . .	0,49167	0,32583	0,24299	0,19332	0,16022	0,13658	0,11885	0,10494	0,09404
∞ . .	0,50000	0,33333	0,25000	0,20000	0,16667	0,14286	0,12500	0,11111	0,10000

Wykaz literatury

Książki

[1] Murthy M. N.: *Sampling Theory and method.* Statistical Publishing Society, Calcutta (1967), (paragraf 3.4).

[2] Pathak P. K.: *Recent adwances in sampling theory.* Indian Statistical Institute, Research and Training School, Calcutta, 1965, (rozdział 2).

[3] Raj D.: *Sampling theory.* McGraw-Hill, New York, 1968, (paragraf 3.6).

Artykuły

[1] Basu D.: *On sampling with and without replacement.* „Sankhyā", 1958, nr 20, s. 287—294.

[2] Hájek J.: *Some contributions to the theory of probability sampling.* „Bulletin de l' I.I.S", 1958, nr 36, (3), s. 127—134.

[3] Chanda K.: *A note on the comparative efficiencies of selection of sampling units with and without replacement.* „Sciences and Culture", 1952, nr 18, s. 6, 288—289.

[4] Chikkagoudar M. S.: *A note on inverse sampling with equal probabilities.* „Sankhya", 1966, (A), nr 28, s. 1, 93—96.

[5] Kleber R. S.: *The probability of duplication in sampling with replacement.* „Amer. Math. Monthly", 1968, nr 75, s. 10, 1084—1087.

[6] Koźniewska I.: *Porównanie efektywności losowania ze zwracaniem i bez zwracania przy nieznanej wariancji populacji generalnej.* „Zastosowania Matematyki", 1955, nr 2, s. 297—303.

[7] Pathak P. K.: *On the evaluation of moments of distinct units in a sample.* „Sankhyā", 1961, (A) nr 23, s. 415—420.

[8] Pathak P. K.: *On simple random sampling with replacement.* „Sankhyā", 1962, (A), nr 24, s. 287—302.

[9] Pathak P. K., Sethuraman J.: *On the asymptotic distribution of the mean of distinct units in sampling from a finite population.* „Magyar Tud. Akad. Mat. Kutató Int. Közl.", 1964, nr 9, s. 113—116.

[10] Raj D., Khamis S. H.: *Some remarks on sampling with replacement.* Annals Mathem. Statist., vol. 29, s. 550—557.

[11] Rao J. N. K.: *On the comparison of sampling with and without replacement.* „Revue de l'I.I.S.", 1966, nr 34, (2), s. 125—138.

[12] Sampford M. R.: *Methods of cluster sampling with and without replacement for clusters of unequal sizes.* „Biometrika", 1962, nr 49, s. 1, 2, 27—40.

[13] Seth G. R., Rao J. N. K.: *On the comparison between simple random sampling with and without replacement.* „Sankhyā", 1964, (A), nr 26, s. 85—86.

[14] Sukhatme P. V., Narain R. D.:*Sampling with replacement.* „Jour. Ind. Soc. Agric. Statist.", 1952, nr 4, s. 42—49.

Maria DINIEWA

NIEKTÓRE WSTĘPNE BADANIA DOTYCZĄCE PLANOWANIA DEMOGRAFICZNYCH MIKROSPISÓW W LRB

W ostatnich czasach, w krajach socjalistycznych daje się zauważyć znaczny wzrost zainteresowania statystyką demograficzną i społeczną, w wąskim tego słowa znaczeniu. Ten rodzaj informacji statystycznej okazał się bowiem niezbędny dla prawidłowego rozwiązania problemów państwa socjalistycznego.

Odpowiednią formą i sposobem periodycznego zbierania takich danych są mikrospisy i niewątpliwie będą one zajmowały coraz więcej miejsca w praktyce statystycznej krajów socjalistycznych. Mikrospisy pozwalają rozszerzyć program spisów ludności w skali niemożliwej do osiągnięcia przy spisach wyczerpujących, a poza tym realizacja ich jest znacznie tańsza w porównaniu z realizacją spisów wyczerpujących. Przeprowadzenie mikrospisu wymaga jednak przygotowania odpowiednich badań. W Bułgarii, podobnie jak i w innych krajach socjalistycznych, tego rodzaju badania zostały już rozpoczęte.

* * *

Powszechnie wiadomo, że przedmiotem badań mikrospisów są zespoły jednostek. Zespołem w danym przypadku może być gospodarstwo domowe, mieszkanie, zagroda, obwód spisowy, miejscowość itp. Jednakże przy tego rodzaju spisach obserwuje się tzw. efekt losowania zespołowego, tj. pogorszenie precyzji estymatorów w porównaniu z precyzją uzyskiwaną przy niezależnym losowaniu próby. Tłumaczy się to występowaniem korelacji pomiędzy jednostkami w zespole. Niekorzystny efekt losowania zespołowego jest różny dla poszczególnych rodzajów zespołów (gospodarstwo domowe, mieszkanie itp.).

Jeżeli poddamy badaniu wszystkie jednostki należące do badanego rodzaju zespołu, to z reguły pogorszenie będzie tym większe, im większy będzie zespół. Tymczasem przy spisach właśnie większe zespoły stwarzają znaczne ułatwienia organizacyjne (i inne) i obniżają koszt badania przypadający na jednostkę spisową.

Wykazana współzależność zmusza do określenia w drodze badań takiego rodzaju zespołu, przy którym wspomniany „efekt" jak najbardziej kompensowałby się oszczędnością środków finansowych.

Ażeby choć w pewnym stopniu wyjaśnić to zagadnienie, w Bułgarii obliczono wartości korelacji wewnątrzzespołowych (dla około 1400 wskaźników dotyczących kraju jako całości i dla 1500 wskaźników dotyczących okręgów) dla 5 rodzajów zespołów: gospodarstwa domowego, mieszkań, zagrody, obwodu spisowego i miejscowości (dla wskaźników charakteryzujących wsie).

Do badań tych w specjalny sposób pobierano próbę z materiałów powszechnego spisu ludności 1965 r.; stanowiły ją 103 miejscowości wiejskie i 101 miejskich obwodów spisowych. Wybrane miejscowości wiejskie obejmowały 302 obwody spisowe, około 19 000 zagród, około 21 000 mieszkań, około 23 000 gospodarstw domowych i 76 000 osób. Miejskie obwody spisowe zawierały 3700 zagród, 6500 mieszkań, około 10 000 gospodarstw domowych i około 28 000 osób. Na podstawie tej próby określono wpływ korelacji wewnątrzzespołowych na precyzję estymatorów prostych i złożonych.

Dokładne wyniki obliczeń są ciągle jeszcze analizowane. Dotychczasowe rozważania dotyczyły tylko części informacji o korelacjach wewnątrzzespołowych, a mianowicie informacji dotyczących tych cech, dla których badano wpływ nie tylko korelacji wewnątrzzespołowej, ale i niektórych innych czynników.

Oto niektóre wyniki obliczeń i wnioski dotyczące korelacji wewnątrzzespołowych:

1. Najmniejszy efekt wewnątrzzespołowy obserwuje się w przypadku małych jednostek zespołowych: gospodarstwo domowe i mieszkanie. A zatem z tego punktu widzenia wymienione jednostki są najodpowiedniejszymi rodzajami zespołów dla przeprowadzenia demograficznych mikrospisów w LRB.

2. Jeżeli uwzględni się również potrzebne środki finansowe, to jednak się okaże, że korzystniejszym (niż wyżej wymienione) rodzajem zespołu, w warunkach LRB, jest obwód spisowy (o przeciętnej wielkości około 300 osób).

3. Stwierdzono, że zespołem nie może być miejscowość, jeżeli przyjmuje się ją w naturalnych rozmiarach, tj. nie tworzy się sztucznych zespołów, obejmujących po kilka niewielkich miejscowości. W związku z możliwością tworzenia takich „łączonych" zespołów przewiduje się badanie wewnątrzzespołowych korelacji również i w nich. Uzyskane wyniki pozwolą ostatecznie wyjaśnić, w jakim stopniu miejscowość może być jednak wykorzystana jako zespół, chociażby w przypadku spisu ludności wiejskiej.

4. Stwierdzono, że w przypadku niektórych cech niekorzystny wewnątrzzespołowy efekt powoduje znaczne pogorszenie precyzji. Okazało się, że to pogorszenie wymaga zwiększenia liczebności próby

od dwóch do trzynastu razy (najmniejsze konieczne zwiększenie — 2,3 raza) w porównaniu z liczebnością niezbędną przy losowaniu niezależnym próby. Największego pogorszenia precyzji ocen w przyszłym mikrospisie można spodziewać się dla wskaźnika „grupa społeczna" oraz wskaźników cech: wykształcenie, wiek i stan rodzinny. Niektóre odmiany cechy „grupa społeczna" wymagają zwiększenia liczebności próby dla mikrospisów nawet 25—30-krotnie w porównaniu z jej liczebnością przy losowaniu niezależnym próby.

Oprócz korelacji wewnątrzzespołowych zmierzono także wpływ dyspersji zespołów według wielkości, zwłaszcza na precyzję estymatorów prostych. Wpływ ten okazał się równie niekorzystny; wielkość jego mierzono tylko w dwóch rodzajach zespołów: mieszkanie i obwód spisowy i to tylko w stosunku do estymatorów prostych. Ponieważ wpływ dyspersji zespołów okazał się bardzo istotny w planowaniu przyszłych prób, trzeba o nim bezwarunkowo pamiętać.

A oto najbardziej ogólne wnioski dotyczące wpływu tego ostatniego nie sprzyjającego czynnika: dyspersja zespołów według wielkości odbija się w najbardziej niekorzystny sposób na precyzji oceny wskaźników cechy „płeć", za którymi postępują wskaźniki cech: „stan rodzinny", „wykształcenie", „wiek" i „grupa społeczna". Na ogół, ujemny wpływ dyspersji zespołów według wielkości jest większy dla cech demograficznych, a całkiem słaby dla cechy „grupa społeczna". Jeśli chodzi o cechy demograficzne, to można się spodziewać pogorszenia precyzji oceny w 20—40%, w zależności od rodzaju zespołu. Okoliczność ta zmusza do dodatkowego zwiększenia planowanej liczebności próbki dla mikrospisów 2,3-krotnie w zależności od rodzaju zespołu i liczby osób, które zostaną wybrane w drugim stopniu losowania.

Jak widać z przedstawionych dotychczas danych, pogorszenie precyzji ocen przy przejściu od niezależnego do zespołowego losowania próby nie powinno być lekceważone przez projektantów przyszłych mikrospisów. Ponieważ z teorii badań reprezentacyjnych wiadomo, iż istnieją określone możliwości polepszenia precyzji ocen, postanowiliśmy stwierdzić, czy dzięki nim można będzie choć częściowo skompensować utratę precyzji wynikłą na skutek występowania efektu losowania zespołowego. W tym celu badaliśmy, jaki efekt dadzą:

a) rejonizacja losowania próby,
b) systematyczne losowanie próby na podstawie spiralnego uporządkowania gniazd,
c) nieproporcjonalne losowanie próby,
d) uzyskanie ocen w połączeniu z dodatkową informacją z niezależnego źródła lub tzw. metody korygowania (precyzowania, adiustacji).

Należy zaznaczyć, że wpływ wymienionych przedsięwzięć (z wyjątkiem systematycznego sposobu losowania próby zespołów) został zba-

dany na podstawie połowy próby i to tylko dla dwóch rodzajów zespołów: mieszkania i obwodu spisowego.

Poza tym pomiary dotyczyły 56 wskaźników przedstawiających różne cechy: płeć, wiek, stan rodzinny, wykształcenie i grupa społeczna, oddzielnie dla ludności wiejskiej i miejskiej oraz dla dwóch rodzajów estymatorów (prostego i złożonego).

Wpływ, jaki wywiera systematyczny sposób losowania próby na precyzję ocen, został zbadany na podstawie materiału trzyprocentowej próby 1965 r., jednak tylko w ramach dwóch okręgów [1]). Większej liczby okręgów nie można było włączyć z powodu wyjątkowej pracochłonności prac badawczych. Efekt systematycznego losowania próby zmierzono empirycznie poprzez porównanie wielkości błędów losowych ocen próby trzyprocentowej z 1965 r., (autokorelacja dodatnia) z błędami losowymi ocen próby bez wstępnego spiralnego uporządkowania obwodów spisowych.

Biorąc pod uwagę, że ta metoda losowania próby wywrze silny wpływ przede wszystkim na precyzję ocen wskaźników ekonomicznych i socjalnych, odpowiednio poszerzono program badania. Oprócz wspomnianych wskaźników demograficznych i socjalnych zostały włączone również niektóre wskaźniki ekonomiczne: aktywność, gałęzie gospodarki narodowej i sfery produkcji.

Efekt zrealizowania pozostałych przedsięwzięć został zbadany analitycznie, przy czym wykorzystano odpowiednie wzory. Spowodowało to konieczność zebrania informacji o:

— stopniu wewnętrznej jednorodności rejonów,
— dyspersji zespołów według rejonów,
— liniowych korelacjach pomiędzy zagregowanymi wartościami poszczególnych interesujących nas cech a zagregowaną wartością cechy (w danym przypadku ogólnej liczby ludności), która definiuje informację uzupełniającą.

Analiza wyników pozwala na sformułowanie najbardziej ogólnych wniosków dotyczących przyszłych mikrospisów:

1. Rejonizacja próby według okręgów doprowadziłaby do istotnego polepszenia precyzji ocen w przypadku mikrospisów, jeśli jako zespół będzie wykorzystywany obwód spisowy.

2. Rejonizacja odbije się bardziej korzystnie na ocenach wskaźników grup społecznych i mniej korzystnie na ocenach wskaźników demograficznych.

3. Zastosowanie systematycznego losowania próby, z poprzedzającym je spiralnym uporządkowaniem obwodów spisowych, spowoduje zmniejszenie błędu ocen wskaźników demograficznych o 5—10%, a ekonomicznych i socjalno-ekonomicznych o 20%.

---

[1]) Okręg — podstawowa administracyjno-terytorialna jednostka w LRB. Kraj jest podzielony na 28 okręgów, łącznie ze stolicą m. Sofia.

4. Zastosowanie losowania nieproporcjonalnego spowoduje zmniejszenie błędu oceny wskaźników demograficznych co najwyżej o 5%. A zatem zastosowanie tego sposobu pobierania próby w tym przypadku nie jest właściwie uzasadnione. Większego zmniejszenia błędu, bo około 10% dla zespołu „obwód spisowy", można się spodziewać jedynie w stosunku do ocen grup społecznych.

5. Uzyskanie ocen metodą estymacji złożonej powoduje zmniejszenie wielkości błędu o 20% — jeżeli w charakterze zespołu wykorzysta się obwód spisowy oraz o 10% — jeżeli w charakterze zespołu wykorzysta się mieszkanie.

Dla poszczególnych grup wskaźników efekt estymatorów złożonych w stosunku do całej liczby ludności również jest różny. Jest on większy przy wskaźnikach demograficznych (około 25%) i mniejszy dla cechy „grupa społeczna" (ok. 15%).

6. Jeżeli rozpatrzymy ogólny (łączny) efekt zastosowania wszystkich wymienionych przedsięwzięć podwyższających precyzję ocen, to stwierdzimy, że jest on w przybliżeniu jednakowy tak dla prostych, jak i złożonych estymatorów.

7. Należy się spodziewać, iż z wszystkich wymienionych przedsięwzięć największy udział w polepszeniu dokładności mikrospisów może mieć przede wszystkim stosowanie złożonych ocen w stosunku do ogólnej liczby ludności, a następnie rejonizacja próby według okręgów. Jest to ważny wniosek uogólniający.

W podsumowaniu należy podkreślić, że niekorzystny wpływ wewnątrzzespołowych korelacji i dyspersji zespołów jest pod względem wielkości silniejszy od łącznego wpływu wszystkich wymienionych przedsięwzięć, zwłaszcza jeśli chodzi o ocenę wskaźników struktury socjalnej, a to oznacza, że w przyszłych mikrospisach całkowite zniwelowanie efektu losowania zespołowego można będzie osiągnąć tylko przez zwiększenie, i to znaczne, liczebności próbki. Jak już wspomniano, z punktu widzenia środków finansowych względnie najwygodniejszą możliwość skompensowania utraty precyzji stwarza obwód spisowy o przeciętnej liczbie około 300 osób.

Obwód spisowy, jako jednostka zespołowa, zasługuje na największą uwagę. W ten sposób, jeżeli przyjmiemy, że jako zespołową jednostkę wykorzysta się obwód spisowy, przy czym z każdego obwodu wybierze się do obserwacji 60 osób i jeżeli oprócz tego założymy, że zostaną zastosowane wszystkie wymienione środki dla skompensowania utraty precyzji, to jednak trzeba będzie zwiększyć liczebność próbki (w porównaniu z tą, która jest niezbędna przy losowaniu niezależnym próbki) kilkakrotnie, a mianowicie:

1. W przypadku estymatora złożonego dla badanych przez nas wskaźników na ogół konieczne jest zwiększenie liczebności próbki przeciętnie 2,3 raza, jedynie dla wskaźników demograficznych —

przeciętnie o 50% i dla wskaźników socjalnych (grupa społeczna) — przeciętnie 4,3 raza.

2. W przypadku estymatora prostego przeciętnie dla wszystkich badanych wskaźników liczebność próbki powinna być zwiększona około 6,5-krotnie, jedynie dla wskaźników demograficznych około 6-krotnie a dla wskaźników socjalnych (grupa społeczna) — około 7,2 raza.

Z dotychczasowych wypowiedzi wynika, że w przygotowaniu i planowaniu mikrospisów LRB nie należy lekceważyć wpływu efektu wewnątrzzespołowego, nawet w tych przypadkach, gdy przewiduje się środki dla podwyższenia precyzji. Prawdopodobnie ten wniosek jest aktualny, jeśli chodzi o mikrospisy innych krajów socjalistycznych. Należy zdawać sobie sprawę z wzajemnego kompensowania się pogarszających i polepszających efektów i w ostatecznym rachunku przewidzieć niezbędne zwiększanie liczebności próby. Tylko wówczas uzyskiwane z mikrospisów oceny będą niezawodne i spełnią swe zadanie — będą stanowić podstawę dla podejmowania prawidłowych decyzji w zakresie aktualnych problemów kierowania socjalistycznym państwem i społeczeństwem.

Jerzy GREŃ, Tadeusz GROSZEK

PROGRAM NA EMC WYZNACZANIA OPTYMALNEJ LOKALIZACJI PRÓBY ZE WZGLĘDU NA WIELE CECH W LOSOWANIU WARSTWOWYM

1. Wstęp

W praktyce metody reprezentacyjnej dużą rolę odgrywa losowanie warstwowe próby. Jest ono znacznie efektywniejsze od losowania nieograniczonego, jeżeli podział populacji generalnej na warstwy jest taki, że warstwy różnią się znacznie między sobą wartościami badanej cechy. Ważnym problemem przy stosowaniu losowania warstwowego jest odpowiedź na pytanie: ile elementów należy wylosować do próby z poszczególnych warstw populacji generalnej? Jest to tzw. zagadnienie lokalizacji próby między warstwy.

W przypadku jednoparametrowym, tzn. przy szacowaniu jednego tylko parametru populacji (najczęściej średniej wartości cechy badanej), stosuje się powszechnie tzw. losowanie proporcjonalne lub też tzw. losowanie optymalne według wzorów J. Neymana.

W praktyce statystycznej konieczne jest jednak szacowanie wielu naraz parametrów na podstawie jednego badania reprezentacyjnego. Gdy z punktu widzenia wielu badanych cech ustalimy lokalizację próby między warstwy w losowaniu warstwowym, otrzymamy dla każdej cechy inny, optymalny układ liczb losowań do próby z poszczególnych warstw. Istnieje zatem w tym przypadku konieczność wyboru jakiegoś kompromisowego rozwiązania, dającego lokalizację próby optymalnej, z pewnego punktu widzenia, ze względu na wszystkie naraz badane cechy.

Jeżeli liczebność próby nie jest z góry ustalona, to optymalną ze względu na wiele cech lokalizację próby w losowaniu warstwowym można uzyskać drogą rozwiązania pewnego nieliniowego programu.

Idea zastosowania metod programowania matematycznego do rozwiązania tego problemu zrodziła się już w 1952 r. Jej twórcą był szwedzki statystyk T. Dalenius. Jednakże dopiero niedawno, tj. gdy

przeciętnie o 50% i dla wskaźników socjalnych (grupa społeczna) — przeciętnie 4,3 raza.

2. W przypadku estymatora prostego przeciętnie dla wszystkich badanych wskaźników liczebność próbki powinna być zwiększona około 6,5-krotnie, jedynie dla wskaźników demograficznych około 6-krotnie a dla wskaźników socjalnych (grupa społeczna) — około 7,2 raza.

Z dotychczasowych wypowiedzi wynika, że w przygotowaniu i planowaniu mikrospisów LRB nie należy lekceważyć wpływu efektu wewnątrzzespołowego, nawet w tych przypadkach, gdy przewiduje się środki dla podwyższenia precyzji. Prawdopodobnie ten wniosek jest aktualny, jeśli chodzi o mikrospisy innych krajów socjalistycznych. Należy zdawać sobie sprawę z wzajemnego kompensowania się pogarszających i polepszających efektów i w ostatecznym rachunku przewidzieć niezbędne zwiększanie liczebności próby. Tylko wówczas uzyskiwane z mikrospisów oceny będą niezawodne i spełnią swe zadanie — będą stanowić podstawę dla podejmowania prawidłowych decyzji w zakresie aktualnych problemów kierowania socjalistycznym państwem i społeczeństwem.

Jerzy GREŃ, Tadeusz GROSZEK

PROGRAM NA EMC WYZNACZANIA OPTYMALNEJ LOKALIZACJI PRÓBY ZE WZGLĘDU NA WIELE CECH W LOSOWANIU WARSTWOWYM

1. Wstęp

W praktyce metody reprezentacyjnej dużą rolę odgrywa losowanie warstwowe próby. Jest ono znacznie efektywniejsze od losowania nieograniczonego, jeżeli podział populacji generalnej na warstwy jest taki, że warstwy różnią się znacznie między sobą wartościami badanej cechy. Ważnym problemem przy stosowaniu losowania warstwowego jest odpowiedź na pytanie: ile elementów należy wylosować do próby z poszczególnych warstw populacji generalnej? Jest to tzw. zagadnienie lokalizacji próby między warstwy.

W przypadku jednoparametrowym, tzn. przy szacowaniu jednego tylko parametru populacji (najczęściej średniej wartości cechy badanej), stosuje się powszechnie tzw. losowanie proporcjonalne lub też tzw. losowanie optymalne według wzorów J. Neymana.

W praktyce statystycznej konieczne jest jednak szacowanie wielu naraz parametrów na podstawie jednego badania reprezentacyjnego. Gdy z punktu widzenia wielu badanych cech ustalimy lokalizację próby między warstwy w losowaniu warstwowym, otrzymamy dla każdej cechy inny, optymalny układ liczb losowań do próby z poszczególnych warstw. Istnieje zatem w tym przypadku konieczność wyboru jakiegoś kompromisowego rozwiązania, dającego lokalizację próby optymalnej, z pewnego punktu widzenia, ze względu na wszystkie naraz badane cechy.

Jeżeli liczebność próby nie jest z góry ustalona, to optymalną ze względu na wiele cech lokalizację próby w losowaniu warstwowym można uzyskać drogą rozwiązania pewnego nieliniowego programu.

Idea zastosowania metod programowania matematycznego do rozwiązania tego problemu zrodziła się już w 1952 r. Jej twórcą był szwedzki statystyk T. Dalenius. Jednakże dopiero niedawno, tj. gdy

pojawiły się algorytmy programowania nieliniowego, idea ta mogła być zrealizowana w praktyce.

W Polsce w roku 1966 J. Greń opublikował artykuł („Przegląd Statystyczny" 1966, z. 3), w którym zaproponował algorytm, oparty na zmodyfikowanym algorytmie rzutowanego gradientu, J. B. Rosena, umożliwiający wyznaczenie optymalnej lokalizacji próby dla dowolnej ilości cech i warstw. Na podstawie tego algorytmu zespół w składzie: dr J. Greń, mgr T. Groszek, mgr J. Moszczyński i dr R. Zieliński, opracował w języku Algol program dla maszyny ICL-1905, będącej w posiadaniu GUS. Program ten przedstawiony Komisji Matematycznej GUS w czerwcu 1967 r. został przez nią przyjęty do praktycznego wykorzystania w GUS.

W niniejszym referacie podajemy podstawy teoretyczne, krótki opis oraz przykład praktycznego wykorzystania omawianego programu, który znajduje się w bibliotece programów GUS.

2. Podstawy teoretyczne programu

Populacja generalna podzielona jest na L warstw. Na podstawie reprezentacyjnej próby, losowanej według schematu losowania warstwowego, należy oszacować średnie wartości p cech populacji generalnej.

Przyjmijmy następujące oznaczenia:

$n_h \ (h = 1, 2, \ldots, L)$ — szukane liczebności losowań do próby z poszczególnych warstw,

$n = \sum_{h=1}^{L} n_h$ — liczebność całej próby,

$\bar{x}^i \ (i = 1, 2, \ldots, p)$ — średnie arytmetyczne dla poszczególnych cech z próby

$W_h = \frac{N_h}{N}$ — frakcja warstwy h w populacji,

S_{hi}^2 — znana wariancja cechy i w warstwie h w populacji,

$D^2(\bar{x}_i) \approx \sum_{h=1}^{L} \frac{W_h^2 S_{hi}^2}{n_h}$ — wariancja estymatora średniej wartości cechy i.

Optymalną lokalizację próby w losowaniu warstwowym przy szacowaniu średnich wartości p cech badanych można wyznaczyć jako rozwiązanie następującego programu nieliniowego:
znaleźć takie liczby n_h, by

$$n = \sum_{h=1}^{L} n_h = \text{minimum},$$

przy warunkach

$$D^2(\bar{x}) \leqslant c_i \text{ dla } i = 1, 2, \ldots, p,$$

$$n_h > 0 \text{ dla } h = 1, 2, \ldots, L.$$

Ten program nieliniowy (nieliniowość wynika z warunków ograniczających) oznacza, że optymalną ze względu na wszystkie badane cechy lokalizację próby otrzymuje się minimalizując liczebność próby przy warunkach oznaczających, że wariancje estymatorów średnich wartości wszystkich p cech nie mogą przekroczyć z góry danych liczb c_i. A zatem z góry jest zapewniona dokładność szacunku wszystkich badanych cech.

Rozwiązanie podanego wyżej zadania można uzyskać posługując się odpowiednio zmodyfikowanym algorytmem J. B. Rosena, opublikowanym w „Journal of the Society for Industry and Applied Mathematics 1960, vol. 8, a omówionym również w monografii: H. P. Kunzi, W. Krelle, *Nichtlineare Programmierung,* wydanej w Berlinie w 1962 r. Algorytm J. B. Rosena, zwany *metodą rzutowanego gradientu* pozwala na znalezienie optymalnego rozwiązania programu nieliniowego na maksimum wklęsłej funkcji kryterium z liniowymi warunkami ograniczającymi.

Omawiany przykład łatwo jest tak przekształcić, by odpowiadał wymaganym warunkom. W tym celu wprowadzamy zmienne

$$x_h = \frac{1}{n_h} \text{ dla } h = 1, 2, \ldots, L,$$

a iloczyn $W_h^2 S_{hi}^2$ oznaczamy symbolem a_{hi}:

$$W_h^2 S_{hi}^2 = a_{hi}.$$

Funkcja kryterium naszego programu przyjmie teraz postać:

$$n = \sum_{h=1}^{L} \frac{1}{x_h} \text{ (minimum)},$$

warunki zaś postać:

$$\sum_{h=1}^{L} a_{hi} x_h \leqslant c_i \text{ dla } i = 1, 2 \ldots, p,$$

$$x_h > 0 \text{ dla } h = 1, 2 \ldots, L.$$

Żądanie wyznaczenia minimum funkcji wypukłej: $n = \sum_{h=1}^{L} \frac{1}{x_h}$, jest równoważne wyznaczaniu maksimum funkcji wklęsłej $(-n) = -\sum_{h=1}^{L} \frac{1}{x_h}$.

Jedynymi danymi, jakie wymagane są do rozwiązania omawianego programu, są liczby c_i (ograniczenia na wariancje estymatorów wszystkich szacowanych parametrów) oraz wektory $a_i = [a_{1i}, a_{2i} \ldots, a_{Li}]$ dla $i = 1, 2, \ldots, p$, o składowych $a_{hi} = W_h^2 S_{hi}^2$.

Idea algorytmu, na którym oparty jest program dla maszyny cyfrowej, polega na rzutowaniu wektora gradientu (wektora pochodnych cząstkowych) funkcji kryterium na brzeg obszaru rozwiązań dopuszczalnych.

Jeżeli punkt (wektor) x leży na brzegu obszaru rozwiązań dopuszczalnych, utworzonym przez przecięcie się q hiperpłaszczyzn (z warunków ograniczających), to rzut wektora $g = \left[-\frac{1}{x_1^2}, -\frac{1}{x_2^2}, \ldots, -\frac{1}{x_L^2}\right]$ (gradientu funkcji kryterium) na brzeg otrzymuje się jako iloczyn $P_q g$, gdzie P_q jest tzw. macierzą projekcyjną zdefiniowaną jako $P_q = I - A_q \cdot [A_q^T A_q]^{-1} A_q^T$.

Macierz A_q o wymiarach $L \times q$ jest zbudowana z tych wektorów a_i $(i = 1, 2, \ldots, q)$, które odpowiadają przecinającym się hiperpłaszczyznom tworzącym brzeg, na który rzutujemy gradient.

Jeżeli $z = P_q g = 0$ oraz $r = [A_q^T A_q]^{-1} A_q^T g \leqslant 0$, to punkt x, w którym gradient funkcji kryterium wynosi g, jest rozwiązaniem optymalnym programu nieliniowego. Jeżeli nie spełnione są równocześnie te dwa warunki, to punkt x nie jest poszukiwanym optimum i zmieniamy go w kolejnej iteracji. Wzdłuż wektora kierunkowego $z = P_q g$ posuwamy się używając „najdłuższego" możliwego kroku tak, aby nie wyjść poza obszar rozwiązań dopuszczalnych programu i by na tym kierunku osiągnąć możliwie największą wartość funkcji kry-

terium. W ten sposób w kolejnych iteracjach uzyskujemy coraz to lepsze punkty x, aż do osiągnięcia z dowolną dokładnością maksimum funkcji $(-n)$, tj. minimum funkcji (n). Najdłuższy krok określa liczba τ_m będąca odległością, wzdłuż wektora z, między danym punktem x a najbliższą hiperpłaszczyzną brzegową. Jeżeli po tym najdłuższym kroku otrzymalibyśmy mniejszą wartość funkcji kryterium $(-n)$, to stosujemy krok krótszy $\rho\tau_m$, gdzie ułamek ρ jest wyznaczony w odpowiedni sposób drogą interpolacji.

Warunki brzegowe $x_h > 0$, dla $h = 1, 2, ..., L$, wygodnie jest zastąpić warunkami $x_h \geqslant \varepsilon$, gdzie liczby ε są tak dobrane, by $n_h \leqslant N_h$, przy czym N_h jest liczebnością warstwy h w populacji. Oczywista, że w każdej iteracji macierz A_q, a zatem i macierz projekcyjna P_q, muszą być na ogół obliczone na nowo.

Zbieżność tej metody iteracyjnej jest dość szybka. Dokładność z jaką chcemy uzyskać minimalną wielkość próby n, może być z góry przyjęta (na maszynie cyfrowej, znajdującej się w GUS, liczenie odbywało się w zmiennym przecinku, z dokładnością do 11 cyfr znaczących).

3. Opis algorytmu

Program wyznaczania optymalnej lokalizacji próby między warstwy dla wielu cech, sporządzony został dla potrzeb GUS (tj. dla maszyny ICI 1905) w języku ICL — Algol. Opracowano go jedynie dla pamięci wewnętrznej maszyny (bez korzystania z dodatkowych urządzeń do zapamiętywania informacji) i dlatego liczba warstw L i cech p musi być tak ograniczona, by spełniona była nierówność

$$L^2 + (p + 10)L + 2p < 10000.$$

Program zajmuje 9660 słów w pamięci maszyny.

Dane do programu:

L — liczba warstw,

p — liczba cech,

W_h — frakcja warstwy h w populacji dla $(h = 1, 2, ..., L)$,

S_{hi}^2 — wariancja cechy i w warstwie h w populacji (dla $h = 1, 2, ..., L$, oraz $i = 1, 2, ..., p$),

c_i — ograniczenia narzucone z góry na wariancje estymatorów szacowanych średnich (dla $i = 1, 2, ..., p$),

ε — ograniczenie na liczebność próby z warstwy.

Powyższe dane są wczytywane do pamięci maszyny z kart statystycznych 80-kolumnowych. W trakcie liczenia programu, drukowane są wyniki każdej kolejnej iteracji. Po zakończeniu obliczeń drukowane jest rozwiązanie optymalne w postaci zarówno minimalnej liczebności próby n, jak i uzyskanych liczb n_h (dla $h = 1, 2, ..., L$) tworzących optymalną lokalizację próby.

Działania algorytmu przebiega według niżej opisanego bardzo ogólnego schematu.

3.1. Schemat blokowy

Pierwszym etapem, po wczytaniu danych, jest znalezienie rozwiązania początkowego; procedura ta wykorzystuje podaną liczbę ε. Uzyskane rozwiązanie początkowe określa macierz A_q zbudowaną z odpowiednich wektorów a_i. Z macierzy A_q oblicza się następnie macierz projekcyjną P_q. W trakcie tego obliczenia wykorzystuje się odpowiednią, efektywną procedurę odwracania macierzy $[A_q^T A_q]$. Uzyskaną macierz projekcyjną, mnożymy następnie przez wektor g tj. przez gradient funkcji kryterium, który jest aktualnym rozwiązaniem.

Jeżeli po sprawdzeniu, czy uzyskane rozwiązanie jest optymalne, okaże się że tak nie jest, przystępujemy do poprawienia go. Ważnym kolejnym etapem obliczeń jest wyznaczenie właściwego kierunku, tj. wektora z, wzdłuż którego będziemy poszukiwali rozwiązania lepszego. Przy obliczeniu tego wektora używamy (w zależności od pewnych warunków) macierzy projekcyjnej P_q, powstałej z posiadanej aktualnie macierzy A_q bądź też z nowej macierzy A_q różniącej się od poprzedniej zmianą jednego wektora.

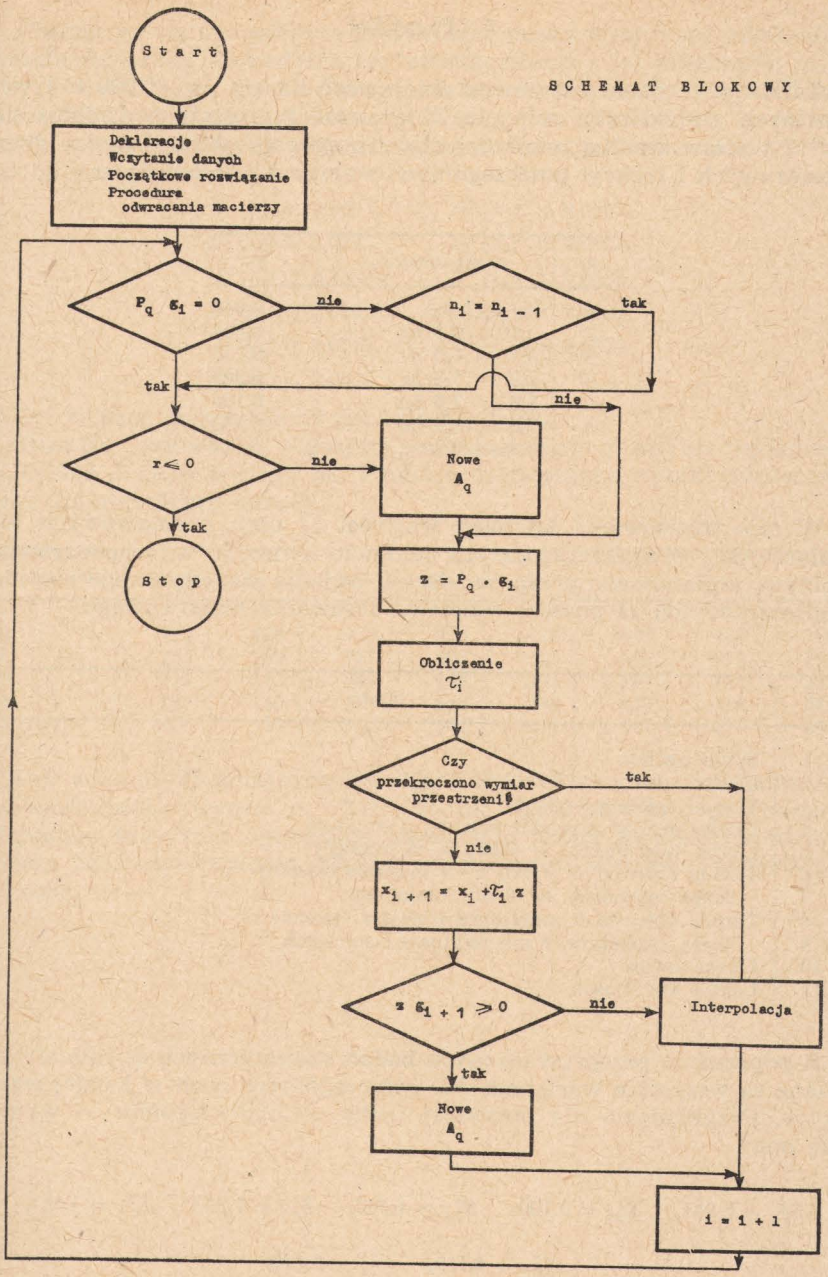
Używając właściwego wektora kierunkowego z, oblicza się następnie krok maksymalny (z użyciem interpolacji — jeżeli potrzeba), jaki można uczynić dla uzyskania lepszego rozwiązania w obszarze rozwiązań dopuszczalnych. Dla nowego rozwiązania wyznaczamy nową macierz projekcyjną P_q oraz nowy gradient funkcji kryterium i po sprawdzeniu optymalności tego nowego rozwiązania, poprawiamy je w kolejnej iteracji, aż do osiągnięcia rozwiązania optymalnego.

Wykorzystanie omówionego programu wyznaczania optymalnej lokalizacji próby w losowaniu warstwowym zilustrujemy następującym przykładem [1]).

---

[1]) J. Pawłowska, *Efektywność metod estymacji i schematów losowania w badaniach pogłowia zwierząt gospodarskich*, Z prac Zakładu Badań Statystyczno-Ekonomicznych GUS 1969, z. 15.

SCHEMAT BLOKOWY



4. Przykład

Załóżmy, że dana jest pewna zbiorowość licząca ok. 60 000 indywidualnych gospodarstw rolnych. Zbiorowość ta została podzielona na $L = 7$ warstw według wielkości obszaru gospodarstwa. Granice grup obszarowych i frakcje poszczególnych warstw są następujące:

h	Grupa obszarowa	W_h
1	do 0,10 *ha* . . .	0,010
2	0,10— 0,5 *ha* . .	0,064
3	0,5 — 2,0 *ha* . .	0,236
4	2,0 — 5,0 *ha* . .	0,309
5	5,0 — 7,0 *ha* . .	0,136
6	7,0 —10,0 *ha* . .	0,133
7	10 *ha* i więcej . .	0,113

W celu oszacowania średnich wartości 11 cech podstawowych tej zbiorowości, charakteryzujących hodowlę zwierząt w gospodarstwie rolnym, zaplanowano przeprowadzenie badania statystycznego metodą reprezentacyjną za pomocą schematu losowania warstwowego.

i	Cecha
1	Bydło ogółem
2	Liczba krów
3	Trzoda chlewna ogółem
4	Liczba prosiąt w wieku powyżej 3 miesięcy
5	Liczba warchlaków w wieku 3—6 miesięcy
6	Trzoda chlewna w wieku 6—9 miesięcy (razem)
7	liczba tuczników od 6 do 9 miesięcy
8	Trzoda chlewna 9 miesięczna i starsza (razem)
9	liczba tuczników 9 miesięcznych i starszych
10	Liczba owiec
11	Liczba kur niosek

Z poprzednio przeprowadzonych badań statystycznych w rolnictwie, znane są wszystkie wariancje S_{hi}^2 poszczególnych cech w każdej warstwie. Przykładowo dla pierwszej cechy — bydło ogółem — wynoszą one:

$$S_{11}^2 = 0{,}964, \quad S_{21}^2 = 0{,}100, \quad S_{31}^2 = 1{,}502, \quad S_{41}^2 = 1{,}353, \quad S_{51}^2 = 1{,}614,$$

$$S_{61}^2 = 4{,}119, \quad S_{71}^2 = 8{,}179.$$

Posługując się omówionym programem należy wyznaczyć minimalną liczebność próby do badania reprezentacyjnego i jej lokalizację między 7 warstw, tak by wariancje estymatorów szacowanych średnich wartości 11 cech nie przekroczyły z góry przyjętych liczb c_i, dobranych w taki sposób, by uzyskać żądaną dokładność szacunku, tzn. by spełnione były warunki:

$$D^2(x_1) \leqslant 0{,}0635, \qquad D^2(x_7) \leqslant 0{,}0061,$$
$$D^2(x_2) \leqslant 0{,}0237, \qquad D^2(x_8) \leqslant 0{,}0121,$$
$$D^2(x_3) \leqslant 0{,}2294, \qquad D^2(x_9) \leqslant 0{,}0036,$$
$$D^2(x_4) \leqslant 0{,}0302, \qquad D^2(x_{10}) \leqslant 0{,}0069,$$
$$D^2(x_5) \leqslant 0{,}0119, \qquad D^2(x_{11}) \leqslant 2{,}6374.$$
$$D^2(x_6) \leqslant 0{,}0074,$$

Do obliczeń przyjmujemy ponadto $\varepsilon = 0{,}0001$.

Stosując omówiony algorytm, uzyskano na maszynie cyfrowej ICT-1905 w GUS w ciągu ok. 1 minuty następujące wyniki ostateczne oraz dla niektórych iteracji:

Nr interacji	n_1	n_2	n_3	n_4	n_5	n_6	n_7	n
0 . . .	428	428	428	428	428	428	428	2996
1 . . .	265	269	517	503	315	352	413	2625
10 . . .	136	147	549	547	296	349	435	2459
20 . . .	127	142	555	548	296	347	436	2451
29 (opt)	121	139	566	542	299	347	437	2451

Tak więc po 29 kolejnych iteracjach stwierdzono, że minimalna liczebność próby wynosi $n = 2451$, tj. ok. 4,1% populacji generalnej. Uzyskana przy tym optymalna lokalizacja próby gwarantuje utrzymanie żądanego poziomu błędów przeciętnych szacunku wszystkich 11 cech.

Erhard KÖNIG

PROBLEMY ZASTOSOWANIA METODY REPREZENTACYJNEJ W BADANIACH WARUNKÓW PRACY I BYTU LUDNOŚCI Z UWZGLĘDNIENIEM OKRESU PERSPEKTYWICZNEGO 1971—1975 w NRD

Statystyka warunków bytu ludności stanowi obszerną dziedzinę stosowania metody reprezentacyjnej. Stosowanie metody reprezentacyjnej jest w tym przypadku nieodzowne.

Zadaniem tej statystyki jest dostarczenie organom planistycznym liczbowych informacji o poziomie, strukturze i kształtowaniu się poziomu warunków bytu ludności i oddziałujących na nie czynnikach. Dane te powinny dotyczyć gospodarstwa domowego i być zróżnicowane według klas społecznych i warstw.

Zasadniczą uwagę przy tym należy skierować na:
— dostarczenie informacji o kształtowaniu się dochodów i spożycia klas i warstw społecznych,
— opracowanie informacji o procesie wychowania i wykształcenia.

Statystyka warunków bytu ludności powinna dostarczać również podstawowych danych:
— do budowy modelu indywidualnego spożycia ludności,
— do budowy modelu tworzenia dochodów i ich wykorzystania w gospodarstwach domowych socjalnych klas i warstw,
— do oceny wpływu stosowania specjalnych środków, w zakresie polityki dochodów i zaopatrzenia, na tworzenie dochodów i ich wykorzystanie,
— do analizy usług bytowych, z których korzysta ludność,
— do analizy wykorzystania budżetu czasu przez ludność,
— do charakterystyki socjalnych sylwetek trybu życia ludności.

Powstaje pytanie dlaczego te informacje muszą być uzyskane, przy określonych nakładach, tylko na drodze badania reprezentacyjnego.

Po pierwsze, niemożliwe jest wykonanie jednego czy kilku badań wyczerpujących, ponieważ, aby dysponować aktualnymi danymi, należałoby stale przeprowadzać wywiady o bardzo obszernym (ze względu na złożoność zagadnień) programie. Wyczerpująca ewidencja wymagałaby olbrzymich nakładów czasu i środków.

Po drugie wywiady dotyczą sfery osobistych spraw badanych osób, a zatem mogą być przeprowadzane tylko na zasadach dobrowolności. A zatem przeprowadzenie wyczerpującego badania byłoby naruszeniem zagwarantowanych konstytucją praw obywatelskich.

Po trzecie, warunki pracy i bytu ludności ulegają tak szybkim zmianom, iż konieczne jest bardzo szybkie uzyskiwanie wyników badań, czego nie można osiągnąć przy badaniu wyczerpującym, w którym ilość danych kilkakrotnie przewyższa ilość informacji otrzymywanych drogą badania reprezentacyjnego.

W okresie lat 1960—1970 organa Centralnego Urzędu Statystycznego prowadząc statystykę warunków bytu ludności stale korzystały z metody reprezentacyjnej. Poniżej podajemy krótki opis zasadniczej treści badań wykonanych tą metodą.

1. Badania reprezentacyjne dochodu

1.1. Badania reprezentacyjne dochodów gospodarstw domowych pracowników fizycznych i umysłowych

Badania reprezentacyjne dochodów gospodarstw domowych pracowników fizycznych i umysłowych po raz pierwszy przeprowadzono w 1959 r. W latach 1960—1965 powtarzano je co roku. Po roku 1965 przeprowadzono już tylko dwa badania: w 1967 i 1970 r.

Celem badań było określenie wysokości i struktury pieniężnych dochodów gospodarstw domowych pracowników fizycznych i umysłowych w zależności od wielkości i składu gospodarstwa domowego, liczby pracujących w każdym gospodarstwie, sytuacji osoby otrzymującej największy zarobek z pracy, oraz podział gospodarstw domowych według różnych kategorii gospodarstw domowych i dochodów.

Jednostkę badania stanowiło gospodarstwo domowe pracowników fizycznych i umysłowych, przedmiot — dochody pieniężne gospodarstw domowych w sierpniu. Do dochodów pieniężnych zaliczano dochody wszystkich członków danego gospodarstwa domowego, a więc zarówno dochody z działalności zawodowej, jak i dochody z funduszów społecznych i socjalnych oraz inne dochody stałe i sporadyczne. Jednocześnie dokonywano spisu niektórych przedmiotów długotrwałego użytkowania.

Badaniem objęto niżej wymienione wskaźniki dochodu poszczególnych członków gospodarstwa domowego:

1) czysty zarobek lub płaca (łącznie ze specjalnymi dodatkami i płacą dodatkową),
2) premie otrzymywane w miejscu pracy,
3) inne zarobki (netto) z pracy,

4) dodatek rodzinny na dzieci lub na małżonka,
5) zasiłki chorobowe,
6) renty (bez zapomóg dla sierot lub z tytułu ubezpieczeń społecznych),
7) stypendia,
8) czysty dochód osoby zarobkującej (ogółem),
9) obowiązujące potrącenia (składki ubezpieczeniowe, podatek dochodowy).

Spośród wskaźników dochodu gospodarstwa domowego (ogółem) badano:
1) inne socjalne zapomogi,
2) inne dochody pieniężne (netto).

Badania obejmowały następujące ogólne dane demograficzne i inne cechy:
— wiek osób,
— płeć,
— sytuacja osób uzyskujących dochód, według rodzaju dochodu,
— stosunek członków gospodarstwa domowego do osoby uzyskującej główny dochód,
— nazwa i adres przedsiębiorstwa, miejsca pracy zatrudnionego,
— forma własności przedsiębiorstwa,
— tygodniowy czas pracy zatrudnionego,
— wiek i płeć osób będących na utrzymaniu gospodarstwa domowego,
— liczba przedmiotów trwałego użytkowania, np. motocykle, skutery, motorowery łącznie z małolitrażowymi skuterami, telewizory, lodówki, pralki.

Jeżeli chodzi o przedmioty trwałego użytkowania ustalono ich liczbę i okres dotychczasowego użytkowania.

Wymienione wskaźniki dochodów i pozostałe informacje o gospodarstwie domowym i osobach z nim związanych uzyskuje się w drodze wywiadu.

1.1.1. Zagadnienia losowania

Liczebność próby, którą należy wylosować wynosiła mniej więcej 30 000 gospodarstw domowych pracowników fizycznych i umysłowych, przy czym osoba uzyskująca główny dochód (pracownik fizyczny lub umysłowy) musiała być zatrudniona w jednym z nierolniczych działów gospodarki uspołecznionej.

Stosowano losowanie dwustopniowe.

Najpierw, liczebność próby dzielono proporcjonalnie na okręgi i gałęzie przemysłu.

W losowaniu pierwszego stopnia losowano przedsiębiorstwa.

Przedsiębiorstwa losowano proporcjonalnie według działów gospodarki i okręgów, uwzględniając liczbę pracowników fizycznych i umysłowych zatrudnionych w przedsiębiorstwie. Operatem losowania był wykaz przedsiębiorstw, instytucji i organizacji działów nierolniczych gospodarki uspołecznionej odnośnego okręgu.

Sprawdzenia poprawności losowania przedsiębiorstw dokonywano za pomocą następującego kontrolnego obliczenia:

$$y_j = \frac{\sum_{k=1}^{n} y_k n_k}{\sum n_k} \text{ lub } y_j = \frac{y_1 n_1 + y_2 n_2 + \ldots + y_n n_n}{n_1 + n_2 + \ldots + n_n},$$

gdzie: y_j — średnia płaca w gałęzi gospodarki j — tego okresu,

y_k — średnia płaca w przedsiębiorstwie k gałęzi gospodarki okręgu,

n_k — liczba badanych jednostek w okręgu k,

$n_k = n_j$ — liczba badanych we wszystkich przedsiębiorstwach objętych próbą.

Jako zasadę przyjęto, że średnia płaca w wybranych przedsiębiorstwach odpowiada średniej płacy w danej gałęzi gospodarki danego okręgu. Przy tym wzięto pod uwagę odchylenie w wysokości ± 5%.

Ażeby zapewnić koncentrację tych osób, którym powierzono przeprowadzenie wywiadu, przy wyborze przedsiębiorstw do próby uwzględniono następujące warunki:

1) w jednym przedsiębiorstwie dopuszcza się badanie minimum 30 i maksimum 150 jednostek; badanie nie powinno obejmować przedsiębiorstw liczących poniżej 300 zatrudnionych;
2) w przedsiębiorstwach zatrudniających więcej niż 2000 osób oddziały (wydziały) przedsiębiorstwa traktuje się jako samodzielne przedsiębiorstwa;
3) w centrum okręgu nie należy badać więcej osób niż 50% ogólnej liczby osób przewidzianych do badania w danym okręgu.

W losowaniu drugiego stopnia, które jest losowaniem systematycznym opartym na kartotece ewidencji płac lub ewidencji zatrudnionych w przedsiębiorstwie, wybierano pracowników fizycznych i umysłowych. Losowanie przeprowadzono dla określonych interwałów z uwzględnieniem 10% rezerwy badanych. Interwał losowania obliczano następująco:

$$\frac{\text{Liczba pracowników fizycznych i umysłowych ogółem w przedsiębiorstwie}}{\text{Liczba badanych łącznie z rezerwą}}$$

Próbą nie byli objęci uczniowie, robotnicy pracujący w domu, urlopowani i chorzy.

Jeżeli w przypadku systematycznego pobierania do próby, natrafiano na robotnika pracującego w domu, wybierano następnego.

Dla większości wyników i zestawień wyników obliczano błędy i dyspersję. Przy czym stwierdzony względny błąd wielkości nie przekraczał 5%.

Przyjęta metoda losowania spośród pracujących prowadzi jednak do systematycznego błędu. Szanse wyboru gospodarstwa domowego, w którego skład wchodzi kilku pracujących, są większe niż tego gospodarstwa, do którego należy tylko jeden pracujący. Szanse wyboru wzrastają wraz z liczbą osób pracujących w gospodarstwie domowym. Systematyczny błąd przy losowaniu prowadzi do zawyżenia niektórych ważnych wskaźników, np. przeciętny dochód gospodarstwa domowego, wielkość rozkładu, stopień wyposażenia w przedmioty trwałego użytkowania itp.

Przypuszcza się, że wskaźniki zostały jednak określone prawidłowo, ponieważ systematyczny błąd powstały na skutek szacunku (zbyt niski szacunek ubocznych dochodów pieniężnych z wykonywania zawodu oraz innych pieniężnych dochodów gospodarstwa domowego) kompensuje systematyczny błąd uwarunkowany losowaniem.

1.2. Badania reprezentacyjne dochodów gospodarstw domowych rencistów

W 1964 r. po raz pierwszy przeprowadzono badanie reprezentacyjne, którego celem było określenie dochodów gospodarstw domowych rencistów. W 1968 r. badanie to zostało powtórzone według szerszego programu (dochody gospodarstwa domowego, socjalna sytuacja gospodarstwa). Bezpośrednim powodem przeprowadzenia tego rodzaju badań było wydanie w 1964 r. zarządzeń zmierzających do podwyższenia renty i wejście w życie w 1968 r. nowej ustawy o rentach.

Celem badania było określenie rozmiaru i struktury dochodów pieniężnych gospodarstw domowych rencistów w zależności od liczby pobierających renty, liczby rencistów jeszcze pracujących, wielkości gospodarstwa domowego oraz podziału gospodarstw domowych rencistów według klas dochodów. Z tym było również związane badanie stopnia wyposażenia gospodarstw rencistów w wybrane artykuły przemysłowe trwałego użytkowania w zależności od rodzaju gospodarstwa domowego, wysokości dochodów gospodarstwa i terminu nabycia tych artykułów.

Jednostkę obserwacji stanowiło całe gospodarstwo domowe rencisty. Pod pojęciem gospodarstwa domowego rencisty rozumiano takie gospodarstwo domowe, w którym uzyskującym główny dochód był

pobierający rentę, niezależnie od tego czy utrzymywał się tylko z renty, czy też zajmował się jeszcze jakąś inną pracą.

Przedmiot obserwacji stanowiły dochody pieniężne gospodarstwa domowego rencisty w maju oraz wybrane artykuły trwałego użytkowania.

Badanie obejmowało niżej podane wskaźniki dochodu poszczególnych odbiorców dochodu w gospodarstwie domowym (w zasadzie otrzymujących renty):

— renta ogółem według rodzajów rent (renty dla osób w wieku starczym, inwalidów, inwalidów w wieku starczym, wdów itp.),
— zapomogi z tytułu opieki,
— płaca, wynagrodzenie łącznie z premią (brutto),
— dochody z przypadkowej pracy (brutto),
— pozostałe dochody z pracy,
— zasiłki chorobowe, zapomogi na dzieci lub na jednego z małżonków, stypendia.

Spośród wskaźników dochodu gospodarstwa domowego ogółem badano:

— zapomogi pieniężne od krewnych i znajomych,
— pozostałe zapomogi socjalne,
— pozostałe pieniężne dochody.

W badaniu uwzględniono również następujące cechy demograficzne i inne:

— wiek i płeć osób,
— stosunek członków gospodarstwa domowego do osoby uzyskującej główny dochód,
— sytuacja otrzymującego dochód w zależności od rodzaju dochodu,
— wiek i płeć członków gospodarstwa domowego nie posiadających własnego dochodu,
— posiadanie artykułów trwałego użytkowania, np. telewizorów, aparatów radiowych, lodówek, pralek, elektroluksów, maszyn do szycia; w badaniu uwzględniano liczbę tych przedmiotów oraz sposób ich nabycia (przed czy po przekroczeniu wieku emerytalnego głównego odbiorcy dochodu, prezent).

Statystyczne wskaźniki dochodów i wyposażenia gospodarstw domowych ustalano w drodze bezpośrednich wywiadów.

1.2.1. Zagadnienia losowania

Liczebność próby wynosiła około 11000 gospodarstw domowych.

W każdym okręgu badano 840 gospodarstw domowych rencistów. Warunkiem objęcia badaniem było otrzymywanie przez głównego odbiorcę dochodu renty przyznawanej osobom w wieku starczym, inwalidom w wieku starczym lub wdowom.

W badaniu stosowano losowanie dwustopniowe.

W losowaniu pierwszego stopnia losowano miejscowość.

Najpierw liczebność próby dzielono proporcjonalnie na okręgi, następnie w granicach okręgów przeprowadzano proporcjonalny dalszy podział na 4 klasy miejscowości:
do 2000 mieszkańców,
2000—20000 mieszkańców,
20000—100000 mieszkańców,
100000 tysięcy i więcej mieszkańców.

W granicach przedziałów wielkości miejscowości stosowano losowanie systematyczne. Operatem losowania był wykaz miejscowości. W celu zapewnienia racjonalnej organizacji pracy, w okręgowych urzędach statystycznych podczas przeprowadzania badań wśród wybranych miejscowości dokonano korektur.

W losowaniu drugiego stopnia wybierano odbiorców rent.

Operat losowania stanowiła kartoteka oddziałów organizacji ubezpieczeń społecznych i dyrekcji rejonów lub kas NTU (Niemieckiego Towarzystwa Ubezpieczeń) znajdujących się w odpowiednich miastach i gminach.

Gospodarstwa domowe losowano systematycznie. Za interwał losowania przyjęto iloraz, gdzie dzielną była ogólna liczba osób otrzymujących, zgodnie z kartoteką, renty wypłacane osobom w wieku starczym, inwalidom w wieku starczym i wdowom, zaś dzielnikiem liczba badanych łącznie z 20-procentową rezerwą.

Przy losowaniu uwzględniano fakt, iż w kartotece figurują także odbiorcy innych rent, np. rent wypłacanych osobom, które uległy wypadkom w czasie pracy, sierotom itp. Liczbę tych osób odejmowano od ogólnej liczby.

Odwiedzając rencistów, w celu uzyskania wywiadu, napotykano tzw. mieszane gospodarstwa domowe, tj. gospodarstwa domowe, w których żyją renciści, ale w których odbiorcą głównego dochodu są robotnicy, pracownicy umysłowi, chłopi spółdzielcy, rzemieślnicy, samodzielni pracownicy lub osoby należące do innej klasy lub warstwy (do tego rodzaju gospodarstw domowych należą również gospodarstwa rencistów zamieszkujących w domach starców). Tego rodzaju gospodarstwa domowe uwzględniano tylko w liczbie, nie poddając je badaniu.

W celu sprawdzenia programu pytań, które miano stosować w badaniu 1964 r., przeprowadzono badanie próbne.

W 1968 r. liczebność próby reprezentacyjnej zredukowano do 9000 gospodarstw domowych, tj. mniej więcej do 600 gospodarstw domowych w każdym okręgu. Podstawą zmniejszenia liczebności prób stanowiły rezultaty dokonanych obliczeń błędów i dyspersji wyników.

W badaniach reprezentacyjnych dochodu prowadzonych w 1968 r. badano ponadto:
— warunki mieszkaniowe (wyposażenie mieszkań, ich wielkość, wysokość czynszu),
— działalność zawodową, ubezpieczenia rencistów, aktywny udział w życiu politycznym i społecznym,
— wykorzystanie czasu wolnego od zajęć.

Metody losowania próby, w porównaniu z 1964 r., nie uległy zmianie.

1.3. Rozległe badania reprezentacyjne dochodów wszystkich klas ludności

W 1966 r. po raz pierwszy zostało przeprowadzone rozległe badanie reprezentacyjne, którego celem było określenie dochodów wszystkich klas.

Jednostkę badania stanowiło gospodarstwo domowe, przedmiotem badania były wszystkie dochody pieniężne gospodarstw domowych oraz wynagrodzenia otrzymywane w naturze w 1965 r., określone ustawowo potrącenia z dochodów członków gospodarstwa domowego, wybrane przemysłowe artykuły trwałego użytkowania, jak również społeczne i indywidualne cechy gospodarstwa domowego i jego członków.

Badane wskaźniki dochodów osób uzyskujących dochód były w tym przypadku bardziej wyczerpujące niż w poszczególnych badaniach reprezentacyjnych:

1) płaca, dochód z pracy, wynagrodzenie za pracę, łącznie z premiami opodatkowanymi,
2) premie nie opodatkowane,
3) uboczny dochód z pracy oraz dochody uzyskiwane w związku z podstawową specjalnością,
4) zasiłki na dzieci i dodatki na jednego z małżonków,
5) zasiłki chorobowe,
6) dochody pieniężne z przedsiębiorstwa rzemieślniczego lub działalności rzemieślniczej,
7) udział w zysku członków spółdzielni (rzemieślnicze spółdzielnie produkcyjne),
8) dochody pieniężne z prywatnej własności (własne domy, działki ziemi itp.),
9) pieniądze wypłacane jako ekwiwalent płacy w naturze (dotyczy w zasadzie chłopów-spółdzielców),
10) płaca uczniów,
11) renty,

12) zapomogi z systemu ubezpieczeń społecznych, łącznie z zapomogami na opiekę,
13) stypendia,
14) wartość artykułów otrzymywanych jako płaca w naturze,
15) składki na ubezpieczenie społeczne,
16) podatki według poz. 1—3, 6—8.

Wskaźniki dochodu gospodarstwa domowego jako całości obejmowały:

1) społeczne zapomogi na dzieci,
2) społeczne zapomogi dla całego gospodarstwa,
3) pieniężne dochody ze sprzedaży rolniczych artykułów,
4) pozostałe dochody pieniężne gospodarstwa domowego.

Jeśli chodzi o cechy demograficzne i inne, w badaniu uwzględniono w zasadzie cechy przyjęte w badaniach reprezentacyjnych dochodów gospodarsw domowych pracowników fizycznych i umysłowych.

Informacje o poszczególnych gospodarstwach domowych uzyskiwano przy pomocy specjalistów dobrze przeszkolonych w zakresie przeprowadzania wywiadów. Trudność w uzyskiwaniu wskaźników polegała na tym, że obliczenie rocznego dochodu gospodarstwa domowego odbywało się wstecz. Rachmistrze odwiedzali gospodarstwa domowe dwukrotnie w ciągu 10 dni. W czasie pierwszej wizyty prosili członków gospodarstwa domowego o wzięcie udziału w badaniu i zapoznawali ich z odpowiednimi materiałami, w czasie drugiej — przeprowadzali właściwy wywiad.

1.3.1. Zagadnienia losowania

Liczebność próby wynosiła około 75000 gospodarstw domowych, a więc w każdym okręgu badano 5000 gospodarstw domowych. Badanie reprezentacyjne objęło 1,2% wszystkich gospodarstw domowych.

Liczebność próby była podzielona na okręgi nie proporcjonalnie, lecz równomiernie.

Wylosowanie tak licznej próby (75000 gospodarstw domowych) wymagało wysokiej techniki losowania.

W losowaniu pierwszego stopnia losowaniem zostały objęte gminy.

Liczbę gospodarstw domowych, jaka miała być zbadana w okręgach dzielono proporcjonalnie na powiaty, odpowiednio do liczby znajdujących się w nich gospodarstw domowych. Gminy każdego powiatu spisywano w porządku alfabetycznym z podaniem liczby gospodarstw (zob. tablica).

Okręg A, powiat 1

Gmina	Liczba gospodarstw domowych	Liczba gospodarstw domowych według obliczenia	Liczba gospodarstw domowych w gminach wybranych	Liczba zbadanych gospodarstw
1				
2				
3				
4				
5				
20				
30				
40				

Interwał losowania ustalano na podstawie wzoru:

$$\frac{\text{Liczba gospodarstw domowych wszystkich gmin każdego okręgu}}{\text{Liczba jednostek badania w okręgu}}$$

gdzie: liczba jednostek = liczba badanych gospodarstw domowych podzielona przez 10.

A zatem w każdej wybranej gminie zbadano co najmniej 10 gospodarstw domowych. Stosowano losowanie systematyczne z losowym początkiem.

W losowaniu drugiego stopnia wylosowano poszczególne gospodarstwa domowe. Operatem losowania gospodarstw domowych były kontrolne wykazy wybranych gmin. Dla drobnych gmin losowanie przeprowadzono mniej więcej tak samo, jak w przypadku badania reprezentacyjnego, mającego na celu określenie dochodu gospodarstw domowych rencistów. Interwał losowania ustalano na podstawie wzoru:

$$\frac{\text{Liczba gospodarstw domowych wybranych gmin}}{\text{Liczba badanych gospodarstw domowych}}$$

Stosowano losowanie systematyczne z losowym początkiem.

Dla większych miast (gminy liczące powyżej 100000 mieszkańców) interwał losowania był nieco uproszczony:

$$\frac{\text{określona funkcja}}{20} \times 3$$

Losowanie odbywało się w dwóch etapach: w pierwszym etapie dokonywano losowania na podstawie list kontrolnych, w drugim — losowano gospodarstwa domowe. Przy czym określono, iż z każdej kontrolnej listy należy wybrać 3 gospodarstwa domowe, a mianowicie: pierwsze, środkowe i ostatnie.

Szczególny sposób pobierania próby w większych miastach spowodowały następujące względy:

1) pobieranie próby w większych miastach wiązało się z dużymi kosztami, toteż wyboru gospodarstw domowych dokonywano nie bezpośrednio, lecz za pomocą list kontrolnych;
2) ażeby ułatwić pracę rachmistrzom, tj. zapewnić zbadanie powiązanych ze sobą gospodarstw domowych, wybór przeprowadzono według określonych odcinków.

W badaniu stosowano estymator regresyjny i prosty:

$$X' = X \cdot \frac{N}{n} = X \cdot \frac{1}{f}$$

gdzie: X' — estymowana wartość cechy X w populacji generalnej,

X — wartości cechy X z próby,

f — frakcja próby,

$\frac{1}{f} = \frac{N}{n}$ — współczynnik uogólniania wyników.

Uogólnianie wyników stosowano do wszystkich wymienionych wyżej wskaźników według wymienionych okręgów, grup ludności, grup gospodarstw domowych i grup dochodów gospodarstw domowych.

1.4. Statystyka budżetów rodzinnych

Celem badania reprezentacyjnego budżetów rodzinnych jest określenie rodzinnych wydatków różnych socjalnych grup ludności ustalonych według ważnych, kształtujących popyt, czynników, np. rozmiarów i struktury dochodu rodzinnego, liczby członków rodziny, ich wieku, płci i działalności zawodowej.

Obecnie statystyka budżetów rodzinnych uwzględnia trzy grupy ludności:

— rodziny pracowników fizycznych i umysłowych (około 3600 rodzin), w których główny dochód uzyskuje pracownik fizyczny lub umysłowy;

— rodziny członków rolniczych spółdzielni produkcyjnych (około 1800 rodzin);

— rodziny rencistów, w których otrzymujący główny dochód pobiera tylko rentę (około 900 rodzin).

Program badanych wskaźników obejmuje:

— ważniejsze wskaźniki dochodu } przeciętne wskaźniki repre-
— ważniejsze cechy rodzinne } zentacyjnych prób dochodu

— około 300 pozycji wydatków pieniężnych według nomenklatury zasadniczego spisu towarów handlu wewnętrznego i według bilansu pieniężnych dochodów i pieniężnych wydatków na pozycje wytwarzania i inne pozycje. W stosunku do rodzin chłopów-spółdzielców obok wspomnianych wskaźników i wielkości pogłowia bydła bierze się pod uwagę również wskaźniki spożycia naturalnego.

Przedmiotem obserwacji jest ogólny obrót pieniężny w prywatnych gospodarstwach domowych, tj. wszystkie przychody i rozchody.

Jednostkę badania stanowi prywatne gospodarstwo domowe. Budżety instytucji społecznych nie są objęte badaniem. Ewidencję przychodów i rozchodów prowadzi się w książeczkach budżetowych przychodów i rozchodów, do których systematycznie bez przerw, wpisuje się odpowiednie dane. Zebrane dane bada się co miesiąc, dla każdego miesiąca przez cały rok. Estymacji danych dokonuje się corocznie dla całego kraju (NRD).

Badanie można wykonać tylko wówczas, gdy zastosuje się stałą obserwację rodzin wypełniających formularze. Obserwację taką wykonuje 750 osób w ramach pracy społecznej. Osoby te udzielają wskazówek i porad rodzinom wypełniającym książeczki budżetowe.

W ciągu ostatniego dziesięciolecia udało się stworzyć stałe grono takich osób, które od szeregu lat zajmują się statystyką budżetów. Osoby te są stale szkolone przez instruktorów i kierowników działów budżetowych okręgowych organów Centralnego Urzędu Statystycznego i otrzymują szczegółowe pisemne instrukcje z kompletnego specjalnego oddziału tegoż urzędu.

1.4.1. Zagadnienie losowania

Przy pobieraniu próby stosowano losowanie dwustopniowe, warstwowe, stanowiące kombinację celowego i losowego pobierania jednostek do próby.

W losowaniu pierwszego stopnia jednostkę losowania stanowiła gmina.

W losowaniu drugiego stopnia jednostką losowania było gospodarstwo domowe.

Pierwszym krokiem losowania gmin był wybór typowych powiatów. Powiaty okręgu zostały podzielone na 3 warstwy:
— powiaty przemysłowe,
— powiaty o strukturze mieszanej,
— powiaty rolnicze.

Powiaty warstwowane są według liczby rodzin. Następnie dokonuje się losowania systematycznego z losowym początkiem.

Kryterium warstwowania gmin stanowiły klasy ich wielkości. Operatem losowania gmin był spis gmin i poszczególnych miejscowości.

Do każdej klasy gmin zaliczano wybrane powiaty w kolejności alfabetycznej i równocześnie dla każdej gminy wybranego powiatu podawano ogólną liczbę rodzin.

Klasy wielkości gmin porządkowano we wzrastającej kolejności i ustalano liczbę rodzin.

Losowanie gmin było losowaniem systematycznym z określonym interwałem losowania. Początek losowania stanowiła liczba losowa.

Dodatkowym warunkiem przy wyborze próby było to, iż w każdej gminie należy przebadać co najmniej 5 rodzin. W ten sposób:
1) zmniejszono rozproszenie gospodarstw domowych na dużą liczbę gmin,
2) zracjonalizowano pracę instruktorów i osób upoważnionych do badania, pracujących społecznie, dzięki skróceniu czasu niezbędnego na przebycie drogi i przestoje.
3) ułatwiono pozyskanie osób do pracy społecznej (zbieranie danych).

Interwał losowania określono na podstawie wzoru:

$$\frac{\text{Ogólna liczba rodzin badanych powiatów} \times 5}{\text{Liczba badanych rodzin w okręgu}}$$

Jeśli badanie którejś z gmin nastręczało trudności (z powodu jej rozmiarów, struktury lub usytuowania terytorialnego) zamieniano ją na inną, równorzędną w granicach powiatu.

Rodziny w wybranych gminach wybierano celowo. Wybór przeprowadzili instruktorzy okręgowych organów statystycznych. Najpierw wybierano przeprowadzającego wywiad (osobę upoważnioną), a następnie razem z nim — rodziny. Przy wyborze rodzin wykorzystywano specjalne instrukcje. Rodziny wybierano z reguły na okres 5 lat.

Opisywaną metodę badania reprezentacyjnego zaczęto stosować w 1967 r. W 1970 r. badanie zakończono.

1.4.2. Rotacja budżetów

Dla statystyki budżetów szczególnie ważna jest wymiana jednostek próby.

Zagadnienie to należy rozpatrywać przede wszystkim z punktu widzenia techniki ewidencji statystycznej. Dlaczego jednak rotacja rodzin jest niezbędna? Oto odpowiedź sformułowana w 3 punktach:

1) rodzina prowadzi książeczkę dochodów i wydatków w ciągu długiego okresu, w następstwie czego powstaje tradycyjne podejście do wpisywania danych i dokładność wyników badań staje się wątpliwa;
2) wpisywanie do książeczki dochodów i wydatków wymaga dużego wysiłku ze strony całej rodziny, co prowadzi do rezygnacji pewnej liczby rodzin z udziału w badaniu, na miejsce rodziny wycofującej się z badania trudno jest znaleźć równorzędną zastępczą rodzinę;
3) wieloletnie wpisywanie do książeczki dochodów i wydatków doprowadza do tego, że część rodzin zaczyna sprawdzać swoje wydatki i kierować nimi.

Wymiana jednostek próby powinna odbywać się w ciągu 5 lat, tzn. że po upływie jednego roku należy wymieniać 1/5 rodzin wypełniających arkusze badań. Nowo pozyskane rodziny wybiera się na 5 lat. Realizacja wymiany rodzin jest jednak jeszcze w stadium początkowym.

2. Badania reprezentacyjne warunków bytu ludności

2.1. Badania warunków mieszkaniowych

Tego rodzaju badanie zostało przeprowadzone w maju 1968 r. Zbadano około 4500 rodzin.

Celem badania było poznanie warunków mieszkaniowych, zwłaszcza rodzin wielodzietnych (rodziny mające 4 i więcej dzieci). Warunki mieszkaniowe (rozmiary i jakość mieszkania) badano uwzględniając, wielkość rodziny, liczbę dzieci, dochód rodziny, wysokość czynszu, sytuację i wielkość gminy.

Jednostkę badania stanowiła rodzina wielodzietna.

W ogólnej liczbie 4500 rodzin zbadano 2500 rodzin wielodzietnych i 2000 innych rodzin (rodziny 1—5 osobowe, biorące udział w badaniach budżetów).

2.1.1. Zagadnienie losowania

Stosowano dwustopniowe losowanie jednostek do próby.

W losowaniu pierwszego stopnia jednostkę losowania stanowiła gmina.

W losowaniu drugiego stopnia dokonywano wyboru rodzin wielodzietnych w wylosowanych gminach.

Niejpierw liczebność próby podzielono, proporcjonalnie do udziału wielodzietnych rodzin, w okręgach. W granicach okręgów przeprowadzono dalszy podział gmin na 4 klasy według ich wielkości.

W losowaniu pierwszego stopnia operatem losowania był spis gmin i poszczególnych miejscowości oraz (jako uzupełnienie) spis współpracowników społecznych uczestniczących w badaniu.

W losowaniu drugiego stopnia operatem losowania były miejscowe spisy rodzin wielodzietnych lub odpowiednie materiały spisu ludności i zawodów. Pobranie jednostek do próby odbywało się przy uwzględnieniu zasady losowania systematycznego.

Interwał losowania ustalono według wzoru:

$$\frac{\text{Liczba wielodzietnych rodzin w gminie}}{\text{Liczba badanych rodzin w gminie}}$$

2.2. Badanie organizacji rocznych urlopów wypoczynkowych ludności

W 1967 r. przeprowadzono po raz pierwszy badanie reprezentacyjne organizacji urlopów w 1966 r. w rodzinach różnych grup ludności.

Próba objęła:

— rodziny pracowników fizycznych i umysłowych	ok. 3500
— rodziny członków rolniczych spółdzielni produkcyjnych	ok. 1600
— rodziny osób samodzielnych (wolnych zawodów)	ok. 1500
Razem ok. 6600 rodzin	

Celem obserwacji było zbadanie organizacji urlopów ludności różnych grup społecznych w zależności od wielkości rodziny, liczby dzieci, dochodu rodziny, wielkości gminy itp.

Badano następujące wskaźniki:
— liczbę osób, które wyjeżdżały na urlop, według wieku, w podziale na wyjazdy wewnątrz kraju i za granicę;
— strukturę osób, które wyjeżdżały, okres urlopu, urlop miesięczny;
— budżet rodzin, które wyjeżdżały;
— osoby wyjeżdżające wewnątrz kraju i za granicę, według wykorzystania podstawowych środków komunikacji, według rodzajów pośrednictwa i rodzajów miejsc zakwaterowania;
— przyczyny nie odbytych wyjazdów na urlop;
— udział dzieci w wyjazdach.

Badanie obejmowało następujące grupy ludności: rodziny pracowników fizycznych i umysłowych, rodziny członków rolniczych spół-

dzielni produkcyjnych typu III, rodziny biorące udział w badaniach budżetów rodzinnych, rodziny osób uprawniających wolne zawody (te ostatnie były badane przez biura podróży).

3. Kierunek rozwoju reprezentacyjnych badań warunków pracy i bytu ludności w okresie planu perspektywicznego 1971—1975

Podstawowym zadaniem w okresie planu perspektywicznego w 1971—1975 jest opracowanie zamkniętego systemu reprezentacyjnego badania rodzin i ludności.

Poszczególne elementy tego systemu stanowią:

— mikrospisy,
— reprezentacyjne badanie dochodów,
— budżety,
— panelowa próba rodzin.

3.1. Mikrospis

Mikrospisy stanowią reprezentacyjne badania ludności, przeprowadzane co 10 lat, w każdej II połowie okresu międzyspisowego. Obejmują one rodziny wszystkich klas i warstw społecznych. Mikrospisy powinny ujawniać podstawowe dane dotyczące:

— społecznej i demograficznej struktury rodzin,
— poziomu i struktury dochodów rodzin.

3.2. Badania reprezentacyjne dochodów

Badania reprezentacyjne dochodów mają za cel zebranie podstawowych danych o rozwoju dochodów ważniejszych klas i warstw ludności.

Społeczna grupa ludności	Rok badania	Liczebność reprezentacyjnej próby
1. Rodziny klasy pracowniczej	1970	30000
2. Rodziny chłopów-spółdzielców	1971	5000
3. Rodziny klasy pracowniczej	1972	15000
4. Rodziny rzemieślników, chałupników i osób samodzielnych	1973	4000
5. Rodziny klasy pracowniczej	1974	15000
6. Rodziny chłopów-spółdzielców	1975	5000

W ciągu 5 lat zostaną przebadane wszystkie podstawowe klasy ludności. Termin badania rodzin rencistów nie został jeszcze ustalony, ponieważ ostatnie rozległe badanie tej grupy ludności przeprowadzono w 1968 r. (badanie zostanie przeprowadzone, jeżeli będzie konieczne z punktu widzenia potrzeb gospodarki narodowej).

Na lata 1971—1975 został opracowany standardowy projekt ewidencji dochodów dla wszystkich ważnych grup ludności. Dane tej ewidencji będą opracowane za pomocą systemu elektronicznego przetwarzania danych JWM 360.

W porównaniu z dotychczas prowadzoną ewidencją dochodów w rodzinach pracowników fizycznych i umysłowych przewidziano zmniejszenie niezbędnej liczebności prób z 30000 do 15000 rodzin. Pozwolił na to wynik obliczeń dyspersji i błędów. Technika ewidencji i pobierania próby pozostanie bez zmiany.

Reprezentacyjne badanie dochodów w rodzinach chłopów-spółdzielców będzie przeprowadzone w rodzinach członków rolniczych spółdzielni produkcyjnych typu I, II i III. Badanie to przygotowuje Centralny Urząd Statystyczny. Ewidencji będą poddane roczne dane, dotyczące dochodów rodzinnych łącznie ze spożyciem naturalnym, oraz przedmioty długotrwałego użytkowania.

Okresem ewidencji będzie nie rok kalendarzowy, lecz rok gospodarczy. Ewidencję danych prowadzić będą z jednej strony — księgowość rolniczych spółdzielni produkcyjnych (wpisy określonych danych) z drugiej strony — rodziny. Takie postępowanie będzie wymagać dużej liczby osób pracujących społecznie, które powinny utrzymywać stałą łączność tak z rodzinami, jak i ze spółdzielniami rolniczymi (jeden badający utrzymuje łączność w jednej spółdzielni przeciętnie z 20 rodzinami).

3.2.1. Zagadnienie losowania

Przewiduje się, iż rodziny do próby będą wybrane w sposób następujący. Liczebność próby zostanie podzielona proporcjonalnie na okręgi, oddzielnie według poszczególnych typów rolniczych spółdzielni produkcyjnych. Najpierw losowane będą powiaty. W tym celu utworzy się w każdym okręgu dwie warstwy:

— warstwa powiatów rolniczych i powiatów o strukturze mieszanej,

— warstwa powiatów przemysłowych i powiatów miejskich.

Wewnątrz warstwy powiaty zostaną podzielone według liczby zatrudnionych w rolnictwie i hodowli zwierząt (na podstawie danych spisu ludności i zawodów z 1964 r.). W każdym okręgu dla obu warstw

ustalać się będzie różną liczbę powiatów do losowania. Interwał losowania ustalony zostanie według wzoru:

$$\frac{\text{Liczba pracujących w okręgu}}{\text{Liczba losowanych powiatów}}$$

Po ustaleniu liczby powiatów w losowaniu pierwszego stopnia będą wybierane rolnicze spółdzielnie produkcyjne, w losowaniu drugiego stopnia — prywatne gospodarstwa członków rolniczych spółdzielni produkcyjnych.

Dla wybranych powiatów w każdym okręgu sporządzany będzie alfabetyczny spis rolniczych spółdzielni produkcyjnych według nazwy gminy, do której odpowiednia spółdzielnia należy. Przy każdej spółdzielni wpisana zostanie liczba rodzin członków spółdzielni. Po podsumowaniu liczby rodzin, obliczany będzie interwał losowania spółdzielni według następującego wzoru:

$$\frac{\text{Liczba rodzin wszystkich spółdzielni wybranych powiatów} \times 20}{\text{Liczba badanych rodzin w okręgu}}$$

W każdej rolniczej spółdzielni produkcyjnej należy uwzględnić co najmniej 20 rodzin.

W losowaniu drugiego stopnia w każdej rolniczej spółdzielni produkcyjnej losowane będą rodziny na podstawie wykazu prywatnych gospodarstw albo kart członkowskich znajdujących się w księgowości spółdzielni. Interwał losowania zostanie obliczony według wzoru:

$$\frac{\text{Liczba prywatnych gospodarstw w wybranej rolniczej spółdzielni produkcyjnej}}{\text{Liczba badanych rodzin}}$$

Po wylosowaniu rodzin, przez instruktorów okręgowych organów, zaangażowane zostaną osoby do prowadzenia wywiadów. Aby przeprowadzić badanie potrzebny jest udział około 300 osób badających.

Dla ewidencji rodzinnych dochodów grupy osób samodzielnych metodą reprezentacyjną istnieje program ewidencji i opracowania danych. Ustalona jest również niezbędna liczebność reprezentacyjnych prób.

Zagadnienia wyboru metody jeszcze nie zostały rozstrzygnięte.

3.3. Statystyka budżetów

W okresie 1971—1975 przewiduje się prowadzenie ciągłego badania następujących grup ludności:

Grupa ludności	Liczebność próby
Rodziny pracowników fizycznych i umysłowych	9500
Rodziny chłopów-spółdzielców członków rolniczych spółdzielni produkcyjnych typu III	1450
Rodziny rencistów	650
mających tylko dochód z rent (415 rodzin)	
mających dochód z rent i pracy (235 rodzin)	
Razem	5600

Dla wszystkich wymienionych grup ludności został opracowany standardowy projekt techniki ewidencji i opracowania danych. Nomenklaturę ewidencji zmniejszono do minimum (z 300 do około 150 pozycji) niezbędnego dla planowania i kierowania indywidualnym spożyciem przez kompetentne organa państwowe.

Nowym elementem, w porównaniu z dotychczasowymi badaniami, jest przestawienie ewidencji z roku kalendarzowego na gospodarczy.

W okresie 1971—1975 zostanie również wprowadzona metoda rotacji rodzin w próbie.

Problemem w dalszym ciągu nie rozwiązanym jest uzyskiwanie, przy określonej liczebności prób wyników reprezentatywnych dla jednostek terytorialnych. Rozwiązanie tego problemu drogą zwiększenia badanych prób jest niemożliwe z powodu wysokich kosztów i braku siły roboczej. Jednakże to zagadnienie ma w dalszym ciągu duże znaczenie dla Centralnego Urzędu Statystycznego, gdzie opracowuje się obecnie „system planowania i kierowania gospodarczym i społecznym rozwojem, zaopatrzenia i obsługi ludności w okręgach, powiatach, miastach i gminach", co nakłada na statystykę nowe, złożone zadania.

3.4. System panelowej próby rodzin

W okresie 1971—1975 system badania reprezentacyjnego zostanie rozszerzony przez wprowadzenie panelowej próby rodzin, tj. próby rodzin, wybieranych z ogólnej zbiorowości rodzin na podstawie określonych założeń [1]).

[1]) W. Lange: *Struktura i stosowanie systemu panelowej próby w badaniu reprezentacyjnym rodzin i ludności.* Lipsk 1969 s. 24.

W ramach takiej próby, której liczebność nie została jeszcze ustalona, rodziny powinny być systematycznie badane 1—2 razy do roku. Stosowanie próby panelowej projektowane jest w badaniach reprezentacyjnych:

— budżetu czasu w rodzinach, według społecznych klas i warstw,
— wydatków i efektywności kwalifikacji dorosłych,
— z zakresu wychowania w rodzinie, w szczególności dzieci i młodzieży,
— przestojów, częstotliwości, jakości prac remontowych i usług oraz kształtowania się cen tych prac i usług,
— usług społecznych i usług leczniczych,
— kulturalno-artystycznej twórczości amatorskiej,
— warunków mieszkaniowych,
— aktywności społecznej,
— specjalnych grup ludności, np. kobiet, młodzieży, rencistów, wielodzietnych rodzin,
— wyposażenia prywatnych gospodarstw w artykuły trwałego użytkowania.

3.4.1. Zastosowanie próby panelowej w badaniu funduszów rodzin w 1969 r.

W ramach ewidencji funduszów stworzono podstawy panelowej próby rodzin. Rodzinom wybranym do omawianego badania, zgodnie z planem badań reprezentacyjnych, przeprowadzający wywiady przekazali materiały reklamowe mające za cel pozyskanie ich do wzięcia udziału w reprezentacyjnych badaniach warunków bytu. W przypadku uzyskania zgody badający wyjaśniał pytania dotyczące:

— nazwiska i adresu rodziny,
— grupy ludności osoby uzyskującej zasadniczy dochód,
— grupy wielkości rodziny,
— grupy dochodu rodzinnego brutto.

Badaniem dotyczącym funduszów objęto około 18000 rodzin wszystkich grup ludności. Około 76%, tj. 13700 rodzin zgłosiło gotowość do wzięcia udziału w pracy w ramach panelowej próby rodzin. Zgłoszenia przewyższały w znacznym stopniu oczekiwanie. Liczono na 60% odpowiedzi twierdzących. Objaw ten wskazuje, że większa część ludności jest zainteresowana w sprawach dalszego rozwoju planowania i kierowania gospodarką narodową.

Porównano socjalno-ekonomiczną strukturę rodzin, które zgłosiły się do udziału w próbie panelowej ze strukturą rodzin NRD. Z porównania wynika, że względny udział w tej próbie rodzin klasy pracowniczej i rodzin rencistów odbiega od udziału tych rodzin w ogólnej liczbie rodzin.

Celem badania funduszów rodzin w 1969 r. było określenie poziomu wyposażenia w artykuły trwałego użytkowania w zależności od wysokości rodzinnego dochodu, wielkości rodziny, czasu nabycia i innych czynników oraz okresu użytkowania tych przedmiotów znajdujących się w gospodarstwie domowym.

Jeśli chodzi o niektóre ważniejsze artykuły trwałego użytkowania badanie miało dostarczyć podstawy do obliczania współczynników ich zużycia.

Przedmiot obserwacji stanowiły 22 techniczne artykuły trwałego użytkowania. Jednostkami obserwacji były prywatne gospodarstwa wszystkich grup ludności NRD.

Badanie odbywało się drogą bezpośrednich wywiadów przeprowadzanych przez specjalne osoby.

Badanie objęło 23000 rodzin (w każdym okręgu badano 1560 rodzin). Około 30% rodzin nie udzieliło wywiadu z powodu nieobecności (rezerwa badania). Większą część tych nieobecności tłumaczy się zmianą miejsca zamieszkania.

3.4.2. Schemat pobierania próby

Zastosowano wielostopniowy, warstwowy schemat pobierania próby, oparty na losowaniu. W zasadzie schemat ten był podobny do schematu losowania obszernej próby reprezentacyjnej, zastosowanego w badaniach dochodu z 1966 r. (por. punkty 1,2,3). W pierwszym etapie losowania dokonano wyboru obwodowych jednostek. Powiaty dzielono na warstwy według liczby osób pracujących w przemyśle, rolnictwie i leśnictwie. Utworzono trzy warstwy:

— powiaty o charakterze przemysłowym,
— powiaty o strukturze mieszanej,
— powiaty o charakterze rolniczym.

Najpierw ustalono liczbę powiatów o różnej wielkości dla poszczególnych okręgów. Następnie w ramach warstw dokonano podziału powiatów. Dla poszczególnych powiatów określono liczbę rodzin mających brać udział w próbie. Powiat, w którym zamieszkiwało najwięcej rodzin, przyjęto jako wielkość wyjściową. Po tym przeprowadzono systematyczne losowanie powiatów według interwału z losowym początkiem.

W drugim etapie wylosowano gminy. W charakterze kryterium podziału na warstwy zastosowano klasę wielkości gminy. Utworzono 4 klasy:

—	—	poniżej 2000 mieszkańców
2000	—	19999 mieszkańców
20000	—	99999 mieszkańców
100000	—	mieszkańców i powyżej.

Najpierw liczbę badanych rodzin w okręgu podzielono proporcjonalnie na klasy wielkości gmin. Gminy wybranych powiatów rozdzielono na klasy wielkości gmin. Jako wyjściową podstawę przyjęto klasę wielkości poniżej 2000 mieszkańców.

W ramach klasy wielkości dokonywano obwodowego uporządkowania, a w ramach powiatów uszeregowano gminy w porządku alfabetycznym. Liczbę rodzin gmin przypisywano w odpowiedni sposób. Losowanie gmin przeprowadzano systematycznie dla określonego interwału losowania z początkową liczbą losową. Jako warunek ustalono, że w każdej gminie należy zbadać co najmniej 10 rodzin.

3.4.3. Niektóre problemy pobierania próby rodzin

W większych miastach znajdowały się duże rejony podlegające wyodrębnieniu np. centrum m. Berlina. W takich podlegających wyodrębnieniu rejonach określono w przybliżeniu liczbę rodzin, które mogły być objęte badaniem. W miejsce eliminowanego wybierano rejon nowych budynków, zamieszkały po 1964 r. leżący w pobliżu rejonu eliminowanego. Równocześnie wykorzystano strukturę rodzin w przybliżeniu podobną do struktury rodzin eliminowanych rejonów. Pod uwagę brano jedną lub kilka wybranych dużych ulic. Opierając się na liczbie mieszkań w nowych budynkach, ustalono interwał losowania.

$$\frac{\text{Liczba mieszkań na ulicy}}{\text{Liczba badanych rodzin}}$$

Następnie dokonywano podziału rodzin według pięter i mieszkań. Często zdarzało się, że wybór przypadał na te rodziny, które już wyprowadziły się z danego mieszkania. W takim przypadku badano rodziny znajdujące się w tych mieszkaniach. Losowanie rodzin nie było związane z konkretnymi nazwiskami. Rodzina znajdująca się w mieszkaniu była badana niezależnie od tego, czy zamieszkiwała w danym mieszkaniu w okresie spisu ludności z 1964 r. Zakładano, że rodzina zajmująca mieszkanie, była z reguły tej samej wielkości jak poprzednia.

4. Badanie budżetu czasu w okresie 1972—1975

Badania budżetu czasu będą prowadzone w latach 1972, 1973, 1974 i 1975 rok. Liczebność badanej próby wyniesie 3500 rodzin osób pracujących, w wieku zdolności do pracy.

Jednostkami obserwacji będą członkowie rodzin w wieku od 16 do 60/65 lat (pracownicy fizyczni i umysłowi, członkowie rolniczych spółdzielni produkcyjnych typu III). Badania będą prowadzone w ciągu jednego tygodnia w maju.

Jan KORDOS

POTRZEBA SZERSZEGO ZASTOSOWANIA METODY ROTACYJNEJ W BADANIACH REPREZENTACYJNYCH

1. Uwagi wstępne

Badania statystyczne prowadzone przez centralne urzędy statystyczne krajów-członków RWPG można ogólnie podzielić na dwie grupy: badania jednorazowe i badania ciągłe prowadzone w różnych okresach. Rozróżnienie to jest szczególnie potrzebne, gdy rozważana jest możliwość szerszego stosowania metody reprezentacyjnej w badaniach statystycznych, ponieważ przy stosowaniu metody reprezentacyjnej w badaniach, w których szacowane parametry mogą ulegać zmianom w czasie, powstaje wiele problemów natury metodologicznej i organizacyjnej nie występujących w badaniach jednorazowych. Dotyczy to przede wszystkim problemów estymacji oraz organizacji badań w terenie. Na przykład przy rozważaniu zagadnień związanych z efektywnością estymacji w badaniach prowadzonych w różnych okresach musimy ustalić następujące problemy:

a) czy we wszystkich okresach powinna być badana ta sama próbka;

b) jeśli nie, to jaka frakcja próbki powinna być wymieniana w poszczególnych okresach badania,

c) w jaki sposób wykorzystać informacje uzyskane z badań w poprzednich okresach, aby zwiększyć efektywność ocen dla okresu badanego.

Należy podkreślić, że w większości podręczników poświęconych teorii i praktyce badań reprezentacyjnych wiele miejsca zajmują badania jednorazowe, a problematyka badań prowadzonych w różnych okresach nie jest dostatecznie opracowana. Wynika to głównie stąd, że rozwiązanie podstawowych problemów występujących przy stosowaniu metody reprezentacyjnej w badaniach prowadzonych w różnych okresach nie jest łatwe i wymaga, w zależności od celu badania, odrębnego potraktowania. Wyłania się tu więc szereg problemów teoretycznych, wymagających bądź to rozwiązania, bądź rozwinięcia, np. optymalna lokalizacja próbek w poszczególnych okresach badania, określenie ich liczebności, zbadanie efektywności różnych schematów

losowania i metod estymacji itd. Są to zagadnienia szczególnie ważne dla praktyki badań reprezentacyjnych prowadzonych w różnych okresach, gdyż od ich rozwiązania zależy efektywność uzyskanych wyników. Dotyczy to w głównej mierze badań prowadzonych przez centralne urzędy statystyczne krajów-członków RWPG, bowiem wiele z tych badań to badania powtarzalne.

Celem tego referatu jest wskazanie na niektóre problemy związane z zastosowaniem metody reprezentacyjnej w badaniach prowadzonych w sposób ciągły lub w pewnych okresach. Podane zostaną podstawowe warianty losowania próbek w poszczególnych okresach oraz ich efektywność w zależności od celu badania. Wykazane zostaną sprzeczności między wyborem najbardziej efektywnego wariantu losowania a głównymi celami stawianymi zwykle przed badaniem statystycznym. Wykazane zostaną również korzyści związane z zastosowaniem określonych wariantów losowania w czasie oraz problemy organizacyjne z nimi związane. W zakończeniu zostanie przedstawiona w skrócie dotychczasowa praktyka badań tego rodzaju w GUS PRL oraz wskazane dziedziny badań statystycznych, w których metody te powinny znaleźć zastosowanie.

2. Dotychczasowy rozwój teorii

Problemy związane z zastosowaniem metody reprezentacyjnej w badaniach prowadzonych w różnych okresach były rozważane przez Jessena [4], Pattersona [10] i Ecklera [2]. W podręcznikach z zakresu metody reprezentacyjnej rozwinięcie tych zagadnień można było znaleźć najpierw w książce Yatesa [15], następnie w wielu innych podręcznikach (por.: Cochran [1], Hansen, Hurwitz, Madow [3], Kish [5] i Murthy [9]).

Należy zaznaczyć, że w literaturze można znaleźć różne nazwy na określenie problemów związanych z zastosowaniem metody reprezentacyjnej w badaniach prowadzonych w różnych okresach. Pierwsza nazwa „losowanie rotacyjne" (rotation sampling) została po raz pierwszy zaproponowana przez Wilksa [13]. Patterson [10] i Yates [15] nazywają ten sposób losowania „losowaniem w kolejnych okresach z częściową wymianą jednostek" (sampling on successive occasions with partial replacement of units), natomiast Hansen, Hurwitz i Madow [3] używają terminu „losowanie dla szeregów czasowych" (sampling from time series). Wyżej wymienione terminy dotyczyły tylko losowej wymiany jednostek w poszczególnych okresach badania.

Wydaje się jednak, że problem badań reprezentacyjnych w czasie, tj. prowadzonych ciągle lub w poszczególnych okresach, powinien być potraktowany kompleksowo, a zatem obejmować nie tylko zagadnienia częściowej wymiany jednostek w poszczególnych okresach ba-

dania. Należałoby wyraźnie rozróżnić problemy występujące w badaniach jednorazowych i badaniach prowadzonych w różnych okresach metodą reprezentacyjną. W ostatnim przypadku można by używać terminu *metoda rotacyjna*, co podkreślałoby możliwość rotacji jednostek badania w poszczególnych okresach badania. Termin ten mógłby obejmować cały system badań reprezentacyjnych dotyczących danej populacji w różnych okresach, a więc zakres jego byłby szerszy niż terminu „losowanie rotacyjne", używanego dotychczas w literaturze. A oto proponowana definicja tego terminu: *Metoda rotacyjna obejmuje całokształt zagadnień związanych z zastosowaniem metody reprezentacyjnej w badaniach populacji podlegających zmianom w czasie, a w szczególności problemy lokalizacji próbki w poszczególnych okresach, ustalania potrzebnej jej liczebności, schematów losowania, metod estymacji oraz metod analizy precyzji wyników.*

Ponieważ problemy te wymagają szerszego, specjalnego opracowania, nie będziemy ich rozważać. Ograniczymy się tylko do podania klasyfikacji losowania rotacyjnego oraz zbadania jego efektywności w zależności od celu badania.

3. Klasyfikacja badania rotacyjnego

Jeśli określona zbiorowość generalna rozważana jest w poszczególnych okresach, a celem badania jest uzyskanie informacji o sumach badanych cech z poszczególnych okresów, o wartości tych cech w okresach badanych lub o różnicy między okresami, to możliwe są trzy warianty losowania próbek. Próbki można losować w każdym okresie niezależnie lub część próbki wylosowanej w pierwszym okresie badać w następnym okresie, dolosowując do niej część nowej podpróbki. Można również tę samą próbkę wylosowaną w pierwszym okresie badać we wszystkich następnych okresach. *Losowanie jednostek w poszczególnych okresach badania nazwaliśmy losowaniem rotacyjnym.* Losowanie to można poklasyfikować ogólnie według kryterium występowania jednostek badania w poszczególnych okresach, a mianowicie:

1) *losowanie o rotacji całkowitej* — gdy w każdym okresie badania losowana jest nowa próbka,
2) *losowanie o rotacji częściowej* — gdy w okresie badanym dolosowuje się część próbki na miejsce eliminowanej części próbki z poprzedniego okresu,
3) *losowanie o rotacji zerowej* — gdy we wszystkich okresach badana jest ta sama próbka.

Jest to oczywiście klasyfikacja ogólna, która jednak okaże się przydatna w dalszych rozważaniach z zakresu metody rotacyjnej.

4. Efektywność losowania rotacyjnego

Analizę efektywności różnych wariantów losowania rotacyjnego przeprowadzimy oddzielnie, uwzględniając różne cele badania.

Najczęściej celem badania statystycznego jest oszacowanie następujących parametrów:

— sumy (lub przeciętnej) ze wszystkich okresów badanych,

— różnicy lub ilorazu między poszczególnymi okresami,

— przeciętnych w poszczególnych okresach badania.

Rozważmy obecnie, który z wariantów losowania rotacyjnego jest najbardziej efektywny przy szacowaniu poszczególnych parametrów, tj. daje najmniejszą wariancję estymatora.

5. Ocena sumy (lub przeciętnej) ze wszystkich okresów

Załóżmy, że badanie przeprowadzono w m okresach. W okresie j ($j = 1, 2, ..., m$) estymatorem przeciętnej jest x_j. Celem badania jest ocena sumy przeciętnych z poszczególnych okresów: $x = \sum_{j=1}^{m} x_j$. Jeśli przez $D^2(x_j)$ oznaczymy wariancję estymatora x_j, a przez r_{jk} — współczynnik korelacji badanej cechy między okresami j i k, wariancja sumy wyrazi się za pomocą następujących wzorów:

a) przy losowaniu o rotacji całkowitej:

$$D^2(x) = \sum_{j=1}^{m} D^2(x_j); \tag{1}$$

b) przy losowaniu o rotacji częściowej:

$$D^2(x) = \sum_{j=1}^{m} D^2(x_j) + \sum_{j \neq k} 2pr_{jk} D(x_j) D(x_k), \tag{2}$$

gdzie p oznacza frakcję próbki, wymienianą w poszczególnych okresach, $0 < p < 1$;

c) przy losowaniu o rotacji zerowej:

$$D^2(x) = \sum_{j=1}^{m} D^2(x_j) + \sum_{j=k} 2r_{jk} D(x_j) D(x_k). \tag{3}$$

Jeśli przyjmujemy, że współczynniki r_{jk} są dodatnie, co potwierdza praktyka, to z wzorów (1), (2) i (3) wynika, iż w tym przypadku najbardziej efektywne jest losowanie o rotacji całkowitej.

6. Ocena zmian między okresami badania

Rozważmy dwa okresy badania, w których estymatorami przeciętnych są odpowiednio x_1 i x_2. Przy badaniu zmian między tymi okresami możemy być zainteresowani w badaniu różnicy, tj. $x_2 - x_1$ lub ilorazu, tj. $w = \frac{x_2}{x_1}$.

Wariancje tych estymatorów wynoszą odpowiednio:

a) przy losowaniu o rotacji całkowitej:

$$D^2(x_2 - x_1) = D^2(x_1) + D^2(x_2), \tag{4}$$

$$D^2(w) = [D^2(x_2) + w^2 D^2(x_1)] \frac{1}{x_1^2}; \tag{5}$$

b) przy losowaniu o rotacji częściowej:

$$D^2(x_2 - x_1) = D^2(x_1) + D^2(x_2) - 2pr_{12} D(x_1) D(x_2) \tag{6}$$

$$D^2(w) = [D^2(x_2) + w^2 D^2(x_1) - 2wpr_{12} D(x_1) D(x_2)] \frac{1}{x_1^2}, \tag{7}$$

gdzie p i r_{12} — jak wyżej;

c) przy losowaniu o rotacji zerowej:

$$D^2(x_2 - x_1) = D^2(x_1) + D^2(x_2) - 2r_{12} D(x_1) D(x_2), \tag{8}$$

$$D^2(w) = [D^2(x_2) + w^2 D^2(x_1) - 2wr_{12} D(x_1) D(x_2)] \frac{1}{x_1^2}. \tag{9}$$

Z porównania wariancji obliczonych na podstawie wzorów (4), (6) i (8) oraz (5), (7) i (9) wynika, że przy badaniu zmian między okresami najbardziej efektywne wyniki daje losowanie o rotacji zerowej, a więc gdy we wszystkich okresach badania występuje ta sama próbka.

7. Ocena przeciętnych wartości w badanych okresach

Jeśli dla bieżących ocen w poszczególnych okresach zastosujemy estymator prosty, to uzyskamy jednakową precyzję ocen wszystkich (trzech) wariantów losowania rotacyjnego. Natomiast jeśli chcielibyśmy informacje uzyskane w poprzednich okresach wykorzystać do oceny przeciętnej w okresie badanym i zastosować estymator złożony (regresyjny), to można wykazać, że najbardziej efektywne okazałoby się losowanie o częściowej rotacji jednostek. Pokażemy to biorąc pod uwagę dwa okresy badane. Znając wartości z pierwszego okresu, chcemy ocenić wartość przeciętnej cechy w drugim okresie, tj. x_2.

Do oceny x_2 zastosujemy estymator złożony w postaci:

$$x'_2 = x_2 + p(1 - p)\, b\,(x_{1a} - x_{1b}), \tag{10}$$

gdzie: b — współczynnik regresji,

x_{1a} — wartość przeciętna obliczona na podstawie próbki z okresu 1, nie występującej już w okresie 2,

x_{1b} — wartość przeciętna obliczona na podstawie próbki z okresu 1, występującej również w okresie 2,

p — jak wyżej.

Wariancje dla estymatora (10) wynoszą

a) przy losowaniu o rotacji całkowitej:

$$D^2(x'_2) = D^2(x_2), \tag{11}$$

b) przy losowaniu o rotacji częściowej:

$$D^2(x'_2) = D^2(x_2)\,[1 - p\,(1 - p)\, r_{12}^2], \tag{12}$$

c) przy losowaniu o rotacji zerowej:

$$D^2(x'_2) = D^2(x_2). \tag{13}$$

Analiza wzorów (11), (12) i (13) potwierdza, że losowanie o rotacji częściowej jest najbardziej efektywne przy szacowaniu wartości parametrów w poszczególnych okresach badania.

Widać więc, że w zależności od celu badania można wybrać najbardziej efektywny schemat losowania rotacyjnego ze względu na stopień precyzji ocen. Istnieją jednak inne korzyści płynące z zastosowania metody rotacyjnej w praktyce statystycznej.

8. Korzyści stosowania metody rotacyjnej w badaniach statystycznych

Stosowanie metody rotacyjnej w badaniach statystycznych korzystne jest nie tylko ze względu na zwiększenie efektywności ocen szacowanych parametrów, ale również z innych powodów (por.: [6]). Zagadnienia te omówimy nieco szerzej.

Równomierne rozłożenie pracy w czasie. Jednorazowe badania reprezentacyjne wymagają organizowania wielkich zespołów osób, przeprowadzających badanie. Zastosowanie metody rotacyjnej pozwala równomiernie rozłożyć pracę w czasie, może ją przy tym przeprowadzić niewielki, odpowiednio przeszkolony zespół.

Zwiększenie jakości wyników. Badając mniejszą próbkę w czasie, można ją zbadać bardziej dokładnie i więcej czasu poświęcić na kontrolę i analizę jakości danych. Masowe, jednorazowe badania reprezentacyjne są w dużym stopniu obciążone błędami właśnie na skutek braku kontroli i odpowiedniego przeszkolenia personelu przeprowadzającego badanie. Metoda rotacyjna stwarza możliwości zminimalizowania tych błędów.

9. Zwiększenie frakcji jednostek uczestniczących w badaniu

Losowanie rotacyjne, szczególnie o rotacji całkowitej lub częściowej pozwala skrócić długość okresu badania. Wiadomo, że w wielu badaniach statystycznych, szczególnie z zakresu statystyki społecznej, znaczna frakcja jednostek zostaje eliminowana z badań i nie dostarcza żądanych informacji. Jedną z głównych przyczyn tej eliminacji jest zbyt długi okres badania tych jednostek. Skrócenie tego okresu zwiększa odpowiednio frakcję uczestników badania. Toteż jeśli z rozważań teoretycznych wynika, że we wszystkich okresach należy badać tę samą próbkę, to jeszcze przed podjęciem decyzji należy zbadać, czy obserwowanie tych samych jednostek przez dłuższy okres nie wpłynie ujemnie na jakość danych. Istnieją bowiem wtedy możliwości powstawania błędów systematycznych (na skutek znużenia badaniem lub nabrania rutyny zniekształcającej wyniki).

10. Możliwość eliminowania błędów zauważonych w czasie badania

Ponieważ badanie rotacyjne rozłożone jest w czasie, uzyskane wyniki można analizować sukcesywnie, co pozwala w następnych okresach badania eliminować zauważone błędy. Takiej możliwości nie ma w badaniach jednorazowych.

10.1. Zwiększenie efektywności estymacji

Ten temat omówiono ogólnie na początku referatu. Warto jednak dodać, że zwykle w jednym badaniu chcemy oszacować różne parametry, zarówno sumy z poszczególnych okresów, jak i zmiany między nimi, oraz uzyskać oceny dla poszczególnych okresów, a w takim przypadku — jak wynika z naszych rozważań — należałoby stosować różne warianty losowania rotacyjnego. W praktyce jednak musimy wybrać jeden z nich. W wielu przypadkach kompromisowym rozwiązaniem będzie wariant losowania o rotacji częściowej, pozwalający odpowiednio zwiększyć efektywność wyników przy szacowaniu różnych parametrów.

10.2. Możliwości objęcia badaniem nowych jednostek

W czasie prowadzenia badania mogą powstawać nowe jednostki, które nie zostały wylosowane do próbki na początku badania. Metoda rotacyjna umożliwia objęcie badaniem takich nowo powstałych jednostek w następnych okresach badań. Jednostki te mogą stanowić odrębną zbiorowość, ważną ze względu na cel badań.

Przedstawiliśmy podstawowe korzyści, jakie daje zastosowanie metody rotacyjnej w badaniach reprezentacyjnych. Należy jednak zaznaczyć, że zastosowanie tej metody na większą skalę w badaniach statystycznych wymaga odpowiedniego przygotowania służb statystycznych w terenie, gdyż etapy przygotowawcze, konieczne do przeprowadzenia tych badań, wymagają znajomości zasad metody reprezentacyjnej i ich praktycznej realizacji.

11. Zastosowanie metody rotacyjnej w badaniach GUS PRL

Szersze zainteresowania GUS PRL problematyką metody rotacyjnej rozpoczęło się 5 lat temu. Prace prowadzone są w trzech kierunkach:

a) śledzenie osiągnięć innych krajów i prowadzenie własnych prac badawczych,
b) prowadzenie badań eksperymentalnych,
c) opracowanie metodologii badań podstawowych.

Wyniki tych prac zostały szerzej omówione w artykule [8]. Należy jednak dodać, że w latach 1967—1968 metodą rotacyjną zbadano zachorowalność i chorobowość ogólną w Polsce [7] oraz budżet czasu. Metodologia ostatniego z tych badań została omówiona w informacji uzupełniającej o badaniach reprezentacyjnych prowadzonych przez GUS PRL w latach 1968—1970. W latach 1968—1969 przeprowadzono dwa eksperymentalne badania budżetów rodzinnych. Wypróbowano wówczas różne schematy rotacyjne oraz zbadano ich wpływ na dokładność danych i procent odmów udziału gospodarstw w badaniach. Badania budżetów rodzinnych dostarczyły bardzo interesujących wyników, które w przyszłości pozwolą zwiększyć efektywność tego rodzaju badań.

Od dnia 1 stycznia 1970 r. przystąpiono do badań sprzedaży pozarynkowej stosując metodę rotacyjną.

Prowadzi się w dalszym ciągu prace badawcze nad celowością i możliwością stosowania metody rotacyjnej w badaniach statystycznych z zakresu rolnictwa, obrotu towarowego i cen, statystyki pracy i płac oraz warunków bytu ludności.

12. Możliwości i warunki szerszego stosowania metody rotacyjnej w badaniach statystycznych

Ponieważ, jak już wspomniano, centralne urzędy statystyczne prowadzą bardzo wiele badań dostarczających informacji o badanych zbiorowościach w poszczególnych okresach należałoby poświęcić więcej uwagi metodologii tych badań. Różnorodność celów badań staty-

stycznych oraz złożona ich problematyka wymagają poszukiwania nowych metod, bardziej efektywnych i tańszych w porównaniu ze stosowanymi dotychczas. Niezwykle pomocną może się okazać w tych przypadkach metoda rotacyjna, która w zależności od celu badania może dostarczyć efektywnego rozwiązania. Można wskazać wiele badań statystycznych, w których metoda rotacyjna da bardziej efektywne wyniki niż metody tradycyjne. Dotyczy to wielu badań z zakresu rolnictwa, warunków bytu, handlu i cen, przemysłu oraz statystyki pracy i płac. Metoda ta szczególnie przydatna może się okazać w badaniach nowych, które wynikły z potrzeb dostarczenia odpowiednich informacji organom planowania i zarządzania gospodarką narodową. Możliwości prowadzenia badań metodą rotacyjną istnieją niemalże w każdej dziedzinie badań statystycznych.

Należy jednak podkreślić, że efektywne stosowanie metody rotacyjnej w badaniach statystycznych na szerszą skalę wymaga spełnienia szeregu warunków. Warunki te dotyczą opracowania tych badań w centralnych urzędach statystycznych oraz ich realizacji w terenie.

Efektywne stosowanie metody rotacyjnej wymaga:

1) prowadzenia prac badawczych (w centralnym urzędzie statystycznym) przez specjalny zespół pracowników,
2) śledzenia osiągnięć innych krajów w tej dziedzinie badań,
3) prowadzenia badań eksperymentalnych przed wdrożeniem ich do praktyki na większą skalę,
4) opracowania całej metodologii badań prowadzonych metodą rotacyjną z uwzględnieniem wymagań teorii i warunków organizacyjnych w terenie,
5) odpowiedniego przygotowania terenowych organów statystyki państwowej do prowadzenia tego rodzaju badań w terenie.

Każdy z tych warunków jest bardzo ważny w całokształcie prowadzonych badań, jednakże ze względu na konieczność spełnienia wszystkich założeń, na których opiera się metoda rotacyjna, specjalne znaczenie ma odpowiednie przygotowanie pracowników terenowych organów statystyki państwowej. Nawet najlepiej — pod względem metodologicznym — przygotowane badanie nie spełni bowiem swego zadania, jeśli na najniższym stopniu nie zostanie właściwie wykonane. Rozwój tych badań powinien iść w parze z podnoszeniem kwalifikacji służb statystycznych. Jakość danych statystycznych zależy od właściwego przygotowania wszystkich etapów badania.

Wykaz literatury

[1] Cochran W. G.: *Sampling Techniques.* New York 1953, s. 287—290.

[2] Eckler A. R.: *Rotation Sampling.* Annals of Mathematical Statistics, vol. 26 (1955).

[3] Hansen M. H., Hurwitz W. N., Madow W. G.: *Sample Survey Methods and Theory.* New York 1953, vol. 1, s. 490—503, vol. 2, s. 268—279.
[4] Jessen R. J.: *Statistical Investigation of a Sample Survey for Obtaining Farm Facts.* Iowa Agr. Exp. Stat. Res. Bull. 304, 1942.
[5] Kish L.: *Survey Sampling.* New York 1965, s. 457—476.
[6] Kordos J.: *Korzyści z zastosowania metody rotacyjnej w badaniach reprezentacyjnych.* „Wiadomości Statystyczne" 1966 nr 11.
[7] Kordos J.: *Zastosowanie metody rotacyjnej w reprezentacyjnym badaniu zachorowalności i chorobowości ogólnej w Polsce.* „Przegląd Statystyczny" 1966 nr 4.
[8] Kordos J.: *Zastosowanie metody rotacyjnej w badaniach statystycznych GUS.* „Wiadomości Statystyczne" 1968, nr 4.
[9] Murthy M. N.: *Sampling Theory and Methods.* Calcutta 1967, s. 415—418.
[10] Patterson H. D.: *Sampling on Successive Occasion with Partial Replacement of Units.* „Journ. Roy. Stat. Soc." 1950, ser. B, vol. 12.
[11] Praca zbiorowa: *Stichprobenverfahren in der amtlichen Statistik.* Wisbadden 1960.
[12] Praca zbiorowa: *Woprosy primienienia matiematiczeskich i statisticzeskich mietodow w izuczenii problemow narodnogo potreblenija.* Akademia Nauk ZSRR, Moskwa 1962 (artykuły: A. G. Wołkowa, A. Ch. Karapetiana i W. A. Wołkońskiego).
[13] Wilks S. S.: *Representative Sampling and Poll Reliabity* „Pub. Opin. Quart." 1940, t. 4.
[14] Woodruff R. S.: *The use of rotating in the Census Bureau's monthly surveys.* „Jour. Am. Stat. Ass." 1963, vol. 58.
[15] Yates F.: *Sampling Methods for Censuses and Surveys.* London 1949, s. 175—182, 233—235, 260—262.

Iwan WIENIECKI

EKSPERYMENTALNE BADANIA BUDŻETÓW RODZINNYCH METODĄ OBSERWACJI MIGAWKOWYCH

W ciągu 50 lat władzy radzieckiej Partia i Rząd poświęcały wiele uwagi sprawie podniesienia stopy życiowej ludności ZSRR.

W celu wyjaśnienia ważnych procesów socjologicznych i bytowych zachodzących w rodzinach osób pracujących oraz zbadania poziomu warunków życia ludności, organy radzieckiej statystyki prowadzą obliczenia wskaźników budżetów rodzinnych. Aby uzyskać odpowiednie informacje i przeprowadzić analizę wskaźników dochodów, wydatków i spożycia ludności, wydział statystyczny budżetów rodzinnych CUS ZSRR prowadzi wiele badań. Źródłem analizowanych danych jest *stała sieć budżetowa, obejmująca różne gałęzie gospodarki i składająca się z różnych grup ludności (pracownicy fizyczni, umysłowi, pracownicy rolni).*

Losowanie próbki z populacji generalnej pracowników fizycznych i umysłowych odbywa się w drodze losowania *przedsiębiorstw.* Rodzina pracownika fizycznego lub umysłowego, która dostała się do próbki, podlega obserwacji przez cały rok. Zbiorowość takich rodzin tworzy właśnie wspomnianą stałą sieć budżetową CUS. Metoda stałej sieci budżetowej pozwala nie tylko badać dynamikę spożycia osób pracujących (spożycie w ciągu wielu lat), ale i jego sezonowość.

Wytyczone w ostatnich latach przez Partię i Rząd zadania z zakresu podwyższania poziomu warunków życia ludności wymagają przede wszystkim stosowania coraz doskonalszych metod matematyczno-statystycznych w celu uzyskiwania i analizowania informacji o budżetach rodzinnych.

Jednym z wariantów metody reprezentacyjnej, stosowanym do obserwacji budżetów rodzinnych, jest metoda *obserwacji migawkowych.*

Realizując w praktyce tę metodę dział badań reprezentacyjnych instytutu naukowo-badawczego CUS oraz wydział statystyki budżetów rodzinnych CUS wraz z filią łotewskiego instytutu naukowo-badawczego zorganizowały jednorazowe, roczne, eksperymentalne badanie budżetów rodzinnych miasta Rygi.

Badanie prowadzono w okresie od października 1967 r. do października 1968 r. W tym czasie rachmistrze odwiedzili ponad 8 tys. rodzin i dla każdej z nich wypełnili specjalnie w tym celu opracowany kwestionariusz zawierający około 160 pytań.

Zaletą metody migawkowych obserwacji budżetów jest po pierwsze to, że próba była losowana w *podziale terytorialnym*, a nie według gałęzi, po drugie nie obserwowano stale tych samych rodzin wchodzących w skład stałej sieci budżetowej (w Rydze 245 rodzin), lecz co dwa dni zmieniano próbę rodzin.

W ten sposób zamiast badania stosunkowo niedużej liczby rodzin podlegających ciągłej obserwacji w ciągu jednego roku obiekt obserwacji rozszerzono do dużej liczby rodzin, z których każdą obserwowano epizodycznie w ciągu 1—2 dni.

W czasie przygotowań do losowania rodzin, mających podlegać obserwacji migawkowej, zwrócono szczególną uwagę na zapewnienie reprezentatywności próbki. W tym celu, w pierwszym stopniu losowano dzielnice (wylosowano 10% dzielnic oddzielnie dla każdego z pięciu rejonów miasta Rygi, proporcjonalnie do liczby rodzin w rejonie). W losowaniu drugiego stopnia losowano budynki. Do badania przyjęto wszystkie rodziny zamieszkałe w wylosowanych budynkach. W liczbach wyglądało to następująco: wylosowano 900 budynków zawierających 6533 mieszkania, w których obserwowano 8784 rodziny. Z otrzymanych kwestionariuszy do opracowań analitycznych nadawało się tylko 8750. Pozostałych 34 kwestionariuszy nie można było opracować ze względu na niepełną liczbę danych.

W badaniu stosowano więc losowanie dwustopniowe, zespołowe, systematyczne (oddzielnie losowano budynki Rady Miejskiej, budynki będące w gestii wydziałów resortowych oraz budynki prywatne).

Na dwa miesiące przed rozpoczęciem badań w jednym z rejonów Rygi przeprowadzono próbną obserwację w celu opracowania programowych zagadnień organizacyjnych. Na przykład ważna była reakcja ludności na pytania dotyczące budżetu rodziny i szansa uzyskania zgodnych z rzeczywistością odpowiedzi na pytania zawarte w kwestionariuszu. Dodatkowo należało określić czas trwania badania jednej rodziny w celu ustalenia liczby rodzin przeznaczonych do zbadania przez jednego rachmistrza.

W wyniku próbnego badania opracowano i ostatecznie zatwierdzono program badania, postać kwestionariusza oraz instrukcję jego wypełniania.

W celu poinformowania ludności, pracownicy łotewskiego oddziału CUS opracowali odezwę do badanych rodzin z prośbą o okazanie pomocy w przeprowadzeniu badania, przygotowali audycje radiowe i telewizyjne.

W badaniu eksperymentalnym metodą obserwacji migawkowych w charakterze rachmistrzów zatrudniono studentów Instytutu Ekono-

micznego Łotewskiego Uniwersytetu im. P. Stuczki. Rachmistrz obserwował codziennie 3 nowe rodziny. Odwiedzał każdą rodzinę 2 razy pod rząd. W związku z tym codziennie odwiedzał 3 nowe rodziny i 3 rodziny, u których był w dniu poprzednim.

Kwestionariusz, opracowany zgodnie z programem badania, zawierał pytania odnoszące się zarówno do każdego członka rodziny oddzielnie, jak i do całej rodziny. Obejmował on wskaźniki charakteryzujące częściowo dochody i wydatki członków rodziny w miesiącu poprzedzającym miesiąc badany oraz w miesiącu bieżącym (badanym).

W 1968 r. ustalono program mechanicznego opracowania (na maszynach analitycznych) danych uzyskanych z tego badania. Program ten przewidywał sporządzenie 36 makiet tablic, z których każda obejmowała określoną grupę wskaźników, usystematyzowaną według różnych cech, np. według grup pochodzenia społecznego (rodziny robotników, pracowników inżynieryjno-technicznych i administracyjnych oraz rodziny mieszane), według gałęzi gospodarki narodowej, w której zatrudniony jest najwięcej zarabiający członek rodziny, według liczby członków rodziny, poziomu dochodu na osobę i poziomu pełnego dochodu na osobę (obejmującego bezpłatne usługi ze społecznego funduszu spożycia), według rodzaju rodziny, rodzaju zatrudnienia, według przynależności państwowej członków rodziny oraz według wielu innych cech. Niektóre makiety dały możliwość uzyskania kombinowanych ugrupowań według dwóch cech, np.

— grup pochodzenia społecznego i dochodu na osobę,
— grup pochodzenia społecznego i rodzaju rodziny,
— liczby członków rodziny i dochodu na osobę itd.

Wszystkie makiety umożliwiały opracowanie danych w kilku przekrojach (kilka niezależnych ugrupowań według różnych cech). Po usystematyzowaniu i zestawieniu danych bezwzględnych przewiduje się obliczenie wartości względnych oraz innych wartości dla każdej grupy.

Opisany projekt opracowania danych uzyskanych z badania przekazano do Centrum Obliczeniowego CUS Łotewskiej Republiki Socjalistycznej, gdzie dane z kwestionariuszy przeniesiono na 8 rodzajów kart perforowanych. Na jeden kwestionariusz zużywano przeciętnie 35—38 kart perforowanych. Na wydziurkowanie informacji zawartych we wszystkich kwestionariuszach zużyto 330 tys. kart perforowanych.

W celu przedwstępnej kontroli reprezentatywności podstawowych wskaźników badania, opracowano kwestionariusze za okres styczeń—czerwiec 1968 r. według następujących wskaźników:

— średnia miesięczna płaca na jednego pracującego członka rodziny,
— podział rodzin według liczby osób w rodzinie.

Kwestionariusze opracowywano zarówno według grup pochodzenia społecznego rodzin (rodziny robotników, pracowników umysłowych,

rodziny mieszane), jak i według podstawowych gałęzi gospodarki narodowej, w których zatrudnieni byli najwięcej zarabiający członkowie rodziny (przemysł, budownictwo, transport, handel i inne gałęzie).

Opracowanie danych wykazało ich stałość w czasie, szczególnie w grupach rodzin o większej liczbie członków rodziny.

Cały okres badania podzielono na 2 podokresy: październik—grudzień 1967 r., styczeń—czerwiec 1968 r.

Uzyskane wyniki świadczą o możliwości zastosowania terytorialnej próbki rodzin do badania budżetów rodzinnych oraz o tym, że rozkład badanych cech w dwóch okresach stosunkowo dobrze odpowiadał rozkładowi tych cech otrzymanemu z innych badań.

W 1968 r. na podstawie danych dotyczących I połowy tego roku opracowano wstępnie niektóre wskaźniki charakteryzujące przeciętny skład rodzin, podział rodzin według liczby członków, wielkość średniego zarobku miesięcznego na jednego pracującego członka rodziny, a także średni dochód na osobę.

Porównanie tych danych z danymi uzyskanymi na podstawie stałej sieci budżetowej pracowników fizycznych i umysłowych w przemyśle wykazało ich nieporównywalność. Tak więc w eksperymentalnym badaniu mieszkańców miasta Rygi metodą obserwacji migawkowych zostały wydzielone trzy grupy pochodzenia społecznego:

1) rodziny, w których wszyscy pracujący byli robotnikami;
2) rodziny, w których wszyscy członkowie pracujący byli zaliczeni do grupy pracowników inżynieryjno-technicznych lub umysłowych;
3) rodziny o składzie mieszanym.

Natomiast w badaniach opartych na danych czerpanych ze stałej sieci budżetów rodzinnych grupę społeczną rodziny określano w zależności od grupy społecznej osoby badanej, tzn. osoby, którą wylosowano do próbki w celu zbadania budżetu rodziny.

W celu usunięcia tej nieporównywalności zbiorcze roczne budżety pracowników fizycznych i umysłowych przemysłu oparte na danych czerpanych ze stałej sieci budżetów rodzinnych miasta Rygi (za rok 1968) przegrupowano według wyżej opisanej zasady i dla takich grup rodzin obliczano odpowiednie wskaźniki.

Wiele wskaźników obliczonych na podstawie danych uzyskanych z badania metodą obserwacji migawkowych nie zgadzało się z danymi uzyskanymi na podstawie stałej sieci budżetowej. Przyczyny tego stanu były następujące:

1) w badaniu metodą obserwacji migawkowych dochód (zarobek) obliczano jako średni miesięczny dochód (zarobek) pierwszej połowy 1968 r., gdy tymczasem w ostatnich miesiącach roku pracownicy otrzymują zwykle większy zarobek i premię, co związane jest z wytężoną pracą w celu wykonania rocznego planu;

2) metodą obserwacji migawkowych badano nie tylko rodziny pracowników fizycznych i umysłowych (których tylko obejmuje stała sieć budżetowa), ale i rodziny emerytów, niższego personelu obsługującego i innych kategorii mieszkańców o niskich dochodach;
3) dane dotyczące średniej liczby członków rodziny, pochodzące ze stałej sieci budżetów, różnią się mało od odpowiednich danych uzyskanych w drodze obserwacji migawkowych; przy bezpośrednim porównaniu obu wybranych zbiorowości widać jednak pewne różnice.

W obu badanych zbiorowościach stosunek liczby pracujących członków rodziny do liczby nie pracujących jest w przybliżeniu jednakowy. W stałej sieci budżetów rodzinnych jest jednak znacznie więcej rodzin, w których pracuje tylko jedna osoba, co powoduje, że dochód na osobę w tej zbiorowości jest niższy niż dochód na osobę uzyskany z badania metodą obserwacji migawkowych.

Aby scharakteryzować stosunek obserwowanych jednostek do zadawanych pytań, opracowano odpowiednie informacje o prawie 600 rodzinach losowo wybranych z obserwowanej zbiorowości.
Stosunek do badania sprawdzano dla cech:
— wielkość dochodu na osobę,
— powierzchnia mieszkania,
— standard mieszkania.

Otrzymane wyniki świadczą, że większość obserwowanych rodzin odpowiada chętnie na zadawane pytania, a obojętny i negatywny stosunek do badania ma raczej charakter losowy i nie jest związany z warunkami życia rodzin i ich dobrobytem materialnym.

Opracowanie materiałów uzyskanych z badania metodą obserwacji migawkowych nie zostało jeszcze zakończone. Oprócz usystematyzowanego opracowania wyników na maszynach analitycznych (średnia mechanizacja) przeprowadzone zostanie specjalne opracowanie mamatematyczne na IBM, obejmujące:
1) obliczenie średnich i maksymalnych błędów szacunków (dla odpowiednich schematów losowania);
2) budowę empirycznych szeregów rozkładu danych dotyczących poziomu dochodu rodzin i poziomu dochodu na osobę, obliczenie statystycznych charakterystyk rozkładu oraz wariancji badanych cech;
3) wyrównanie empirycznych szeregów rozkładu;
4) obliczenie norm spożycia w zależności od wieku;
5) obliczenie elastyczności spożycia względem dochodu;
6) budowę wskaźników różnego rodzaju zależności;
7) budowę matematycznych modeli kształtowania się poziomu dochodu ludności.

Wyniki opracowania materiałów uzyskanych drogą tego badania pozwalają ocenić reprezentatywność metody obserwacji migawkowych dla różnych cech charakteryzujących budżety rodzinne i pozwolą wskazać te cechy, które mogą być z powodzeniem badane tą metodą.

Celem omawianego eksperymentalnego badania jest ocena efektywności metody obserwacji migawkowych w porównaniu z innymi metodami stosowanymi do badań reprezentacyjnych w statystyce budżetów rodzinnych, a także postulowanie stosowania tej metody w statystyce innych gałęzi gospodarki narodowej.

Heinrich SCHWARZ

PORÓWNANIE EFEKTYWNOŚCI ESTYMATORA PROSTEGO I ESTYMATORA ILORAZOWEGO ZA POMOCĄ WSPÓŁCZYNNIKA REGRESJI

Przy szacowaniu sumy wartości

$$X = \sum_{j=1}^{N} X_j, \tag{1}$$

gdzie N oznacza liczebność populacji generalnej lub wartości średniej określonej cechy

$$\bar{X} = \frac{X}{N} \tag{2}$$

można stosować kilka metod estymacji o różnej efektywności. Ograniczymy się tutaj do rozpatrzenia metody estymacji prostej i ilorazowej oraz porównania ich efektywności.

Jak wiadomo, estymatory sumy wartości (1) można wyrazić w postaci wzorów:

$$x' = \bar{x} \cdot N \text{ w przypadku estymacji prostej} \tag{3}$$

a w przypadku estymacji ilorazowej jako

$$x'_Q = \frac{x}{y} \cdot Y = t \cdot Y \tag{4}$$

gdzie: $x = \sum_{i=1}^{n} x_i$ (5)

$$y = \sum_{i=1}^{n} y_i \tag{6}$$

$$Y = \sum_{j=1}^{n} Y_j \text{ — suma wartości cechy dodatkowej} \tag{7}$$

n — oznacza liczebność próbki.

Porównując wariancje s_x^2, i s_{xQ}^2, obu estymatorów otrzymujemy znane zależności:

$$s_{xQ}' \gtrless s_x'$$

jeżeli

$$r \gtrless \frac{1}{2} \frac{v_y}{v_x}. \tag{8}$$

Tutaj r oznacza współczynnik korelacji obu cech, a v_x oraz v_y są współczynnikami zmienności (względne odchylenia standardowe). Z warunku (8) wywodzi się znane, elementarne prawo mówiące, że metoda estymacji ilorazowej jest bardziej efektywna od estymacji prostej gdy $r > \frac{1}{2}$ przy założeniu, że współczynniki zmienności są sobie równe.

W podręcznikach na ogół spotyka się wzór (8), w którym pytanie o efektywność estymatorów wiąże się z wartością liczbową współczynnika korelacji r oraz ilorazem współczynników zmienności. Dla porównania efektywności estymatorów, niezbędna jest znajomość tych wartości.

Obecnie zwrócimy uwagę na możliwość zbadania efektywności estymatorów za pomocą współczynnika regresji b i wartości względnej t

$$t = \frac{x}{y} = \frac{\bar{x}}{\bar{y}}. \tag{9}$$

Jak wiadomo, między współczynnikiem korelacji r oraz współczynnikiem regresji b zachodzi relacja:

$$r = b \frac{s_y}{s_x} \tag{10}$$

lub relacja:

$$r = \frac{b}{t} \cdot \frac{v_y}{v_x}. \tag{11}$$

Podstawiając te wartości do wzoru (8), otrzymujemy

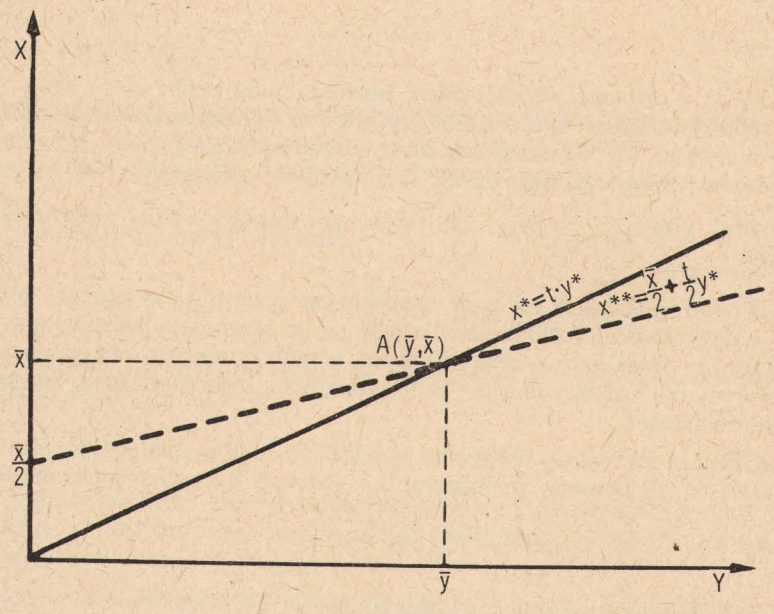
$$s_{xQ}' \lessgtr s_x' \tag{12}$$

jeśli

$$b \gtrless \frac{t}{2}.$$

Interpretacja nierówności (13) jest prosta: estymator ilorazowy jest efektywniejszy od estymatora prostego, jeżeli linia regresji nachylona

jest pod kątem ostrym względem prostej oznaczonej przez x o współrzędnych $\left(0, \frac{\bar{x}}{2}\right)$ i $A\ (\bar{y}, \bar{x})$, tj. jeśli $b > \frac{t}{2}$ (zob. rys.).



Na ogół nietrudno jest rysować na oko (tj. bez obliczania współczynnika regresji) linię regresji dla pewnej liczby punktów. Jeżeli wartości średnie $\bar{x}$ oraz $\bar{y}$ są znane (lub znane w przybliżeniu), to można od razu przedstawić graficznie prostą proporcjonalności x^* i prostą x^{**}.

Jeżeli linia regresji nachylona jest pod kątem ostrym względem prostej x^{**}, to estymator ilorazowy jest efektywniejszy od estymatora prostego.

W ten sposób, metodą graficzną, bez obliczania współczynnika korelacji i współczynników zmienności vx oraz vy można szybko ocenić, czy metoda estymacji ilorazowej jest efektywniejsza od estymacji prostej.

Warunek (13) odnosi się zarówno do estymatora średniej (2) jak i do estymatora sumy wartości (1).

Wadim ZAGRADKIN

KILKA UWAG O EKSPERYMENTALNYCH BADANIACH BUDŻETÓW RODZINNYCH METODĄ ROTACYJNĄ NA TERENIE POLSKI

Efektywne planowanie i zarządzanie gospodarką narodową wymaga szczegółowych informacji z zakresu warunków bytu poszczególnych grup ludności. Jednym z podstawowych źródeł tych informacji są wyniki uzyskane z badań budżetów rodzinnych metodą reprezentacyjną, prowadzonych przez Główny Urząd Statystyczny.

Dotychczasowe badania budżetów rodzinnych w Polsce nie dostarczają pełnych informacji z zakresu warunków bytu ludności, gdyż obejmują tylko określone zbiorowości z danego terenu, np. gospodarstwa domowe [1]) utrzymujące się z pracy w gospodarce uspołecznionej, gospodarstwa pracowników PGR itd. Badania te nie są zadowalające także ze względu na:

— wysoki ich koszt,

— wysoki procent gospodarstw domowych eliminowanych z badania na skutek odmów (około 40%),

— możliwość zniekształcenia struktury wydatków rodziny pod wpływem prowadzenia systematycznych zapisów przez długi okres.

Powstał więc problem opracowania takiej metody badania, która by nie zwiększyła w istotny sposób dotychczasowych środków finansowych przeznaczonych na ich realizację, a równocześnie zapewniłaby uzyskanie reprezentatywnych wyników. Dotyczyło to przede wszystkim metod zbierania informacji, wyboru operatu losowania, długości okresu badania i procentu gospodarstw domowych, odmawiających dalszego udziału w badaniu schematu losowania oraz metod uogólniania danych i oceny ich precyzji. Aby odpowiedzieć na pytanie, czy w badaniach budżetów rodzinnych można zastosować nową metodę

---

1) Pojęcia „gospodarstwo domowe" i „rodzina" w niniejszej publikacji stosowane są zamiennie.

badania — metodę rotacyjną [2]) trzeba było zbadać wszystkie wymienione wyżej problemy.

Na przełomie lat 1968/1969 Główny Urząd Statystyczny przeprowadził dwa eksperymentalne badania budżetów rodzinnych metodą rotacyjną w przekroju gałęziowym i terytorialnym. Eksperyment w przekroju gałęziowym odbywał się na terenie całej Polski, badanie zaś w przekroju terytorialnym odbywało się tylko w m. Wrocławiu.

Przeprowadzenie badań budżetów rodzinnych w dwóch przekrojach podyktowane było między innymi koniecznością zbadania różnych operatów losowania. Dotychczasowe badania były przeprowadzone w przekroju gałęziowym i w związku z tym za podstawę losowania przyjmowano listę płac w wylosowanych zakładach. Przyjęcie takiego operatu losowania przesądza o specyficznym charakterze wybranych rodzin — badaniem obejmowano jedynie rodziny osób zatrudnionych w sektorze uspołecznionym. W eksperymentalnym badaniu w przekroju terytorialnym objęto badaniem nie tylko gospodarstwa domowe osób pracujących w gospodarce uspołecznionej, ale także rodziny osób zatrudnionych w sektorze nieuspołecznionym oraz gospodarstwa domowe rencistów. Rozszerzenie zakresu badania było możliwe dzięki przyjęciu wykazu mieszkań jako operatu losowania. Badanie we Wrocławiu potwierdziło potrzebę zmiany dotychczas stosowanego operatu losowania. W przyszłych badaniach budżetów rodzinnych do budowy operatu posłużą dokumentacje stałych obwodów spisowych.

Po ustaleniu operatu losowania dokonano wyboru próbki. W badaniu przeprowadzonym w przekroju gałęziowym zastosowano losowanie rotacyjne o całkowitej rotacji, tzn. że po zakończeniu badania jednej podpróbki rodzin losowano drugą podpróbkę itd. W tym przypadku okres obserwacji gospodarstw domowych pokrywał się z okresem rotacji rodzin. W drugim eksperymencie zastosowano losowanie rotacyjne o częściowej rotacji podpróbek, tzn. że w każdym okresie badania eliminowano część gospodarstw domowych z poprzedniego okresu i dolosowywano nowe rodziny.

Zastosowanie losowania rotacyjnego umożliwia skrócenie okresu badania wylosowanych gospodarstw. W dotychczasowych badaniach gospodarstw domowych stosowano metodę bieżących zapisów dochodów i wydatków przez okres 1 roku. W eksperymencie przeprowadzonym na terenie całej Polski przyjęto miesięczny okres obserwacji gospodarstwa. W badaniu przeprowadzonym we Wrocławiu wylosowano dwie niezależne równoliczne grupy rodzin. Dla jednej próbki przyjęto dwutygodniowy okres obserwacji, dla drugiej próbki okres czterotygodniowy. Zróżnicowanie okresu badania wprowadzono po to, aby wybrać najodpowiedniejszy okres rotacji do przyszłych badań

[2]) Przez metodę rotacyjną rozumiemy takie badanie reprezentacyjne, które dotyczy populacji w różnych okresach. W badaniu rotacyjnym obserwacja jest dokonywana w ciągu ustalonego wycinka przedziału czasowego.

budżetów rodzinnych. Schemat rotacji podpróbek w badaniu o przekroju gałęziowym można przedstawić w następujący sposób:

Miesięczne badania

podpróbka 1 XXXXX
podpróbka 2 XXXXX
podpróbka 3 XXXXX
itd.

W badaniu przeprowadzonym w przekroju terytorialnym dla próbki rodzin, która była badana przez okres dwóch tygodni, ustalono rotację tygodniową, a dla próbki gospodarstw domowych badanych przez cztery tygodnie okres rotacji wynosił dwa tygodnie. W każdym okresie rotacji wymieniono 50% gospodarstw domowych. Próbka gospodarstw, dla której okres obserwacji wynosił dwa tygodnie, była badana według następującego schematu rotacyjnego:

Tydzień prowadzenia notatek

	1	2		
podpróbka 1	XXXXX	XXXXX		
		1	2	
podpróbka 2		XXXXX	XXXXX	
			1	2
podpróbka 3			XXXXX	XXXXX

itd.

Próbka gospodarstw, dla której okres obserwacji wynosił cztery tygodnie, była badana według następującego schematu rotacyjnego:

Tydzień prowadzenia notatek

podpróbka 1

1	2	3	4
XXXXX	XXXXX	XXXXX	XXXXX

podpróbka 2

	1	2	3	4
	XXXXX	XXXXX	XXXXX	XXXXX

podpróbka 3

		1	2	3	4
		XXXXX	XXXXX	XXXXX	XXXXX

itd.

Przyjęcie skróconego okresu badania rodzin wymagało zmiany dotychczasowej metody zbierania informacji w celu pełnego uchwycenia wszystkich wydatków rodzin (dotyczy głównie artykułów o małej częstotliwości zakupu). Dotychczas głównym źródłem informacji o dochodach i wydatkach była książeczka budżetowa. W eksperymencie zastosowano kombinowaną metodę zbierania informacji, tzn. rejestrację bieżącą i wywiad. Metoda rejestracji bieżącej dochodów i wy-

datków polegała na ewidencjonowaniu przez wylosowane gospodarstwo domowe wszystkich dochodów, jakie otrzymali poszczególni członkowie rodziny w okresie miesiąca do dnia zakończenia badania, oraz wszystkich wydatków na towary i usługi, jakich dokonali członkowie rodziny w okresie dwóch lub czterech tygodni. Niezależnie od zapisów w książeczkach budżetowych informacje o dochodach i wydatkach na wybrane artykuły przemysłowe, dotyczące dłuższego okresu, instruktor uzyskiwał od wszystkich (badanych przez 2 lub 4 tygodnie) drogą wywiadu. Okres, za jaki uzyskiwano informacje tą drogą, był zróżnicowany i wahał się od 1 do 12 miesięcy. Podstawą ustalenia okresów badania wydatków na poszczególne artykuły drogą wywiadu była analiza częstości zakupu danego artykułu, oparta na badaniach budżetów rodzinnych, przeprowadzonych przed rokiem 1968.

Analiza wyników otrzymanych z eksperymentalnych badań budżetów rodzinnych metodą rotacyjną wykazała, że:

1. *Skrócenie okresu obserwacji gospodarstw domowych ma istotny wpływ na zmniejszenie się liczby odmów ze strony wylosowanych rodzin.* I tak w badaniach dwutygodniowych odmówiło prowadzenia notatek 19,2% gospodarstw domowych, w badaniach czterotygodniowych 27,5%, przy miesięcznym zaś okresie obserwacji liczba odmów nie przekraczała 30%, podczas gdy w dotychczasowych badaniach budżetów rodzinnych na terenie m. Wrocławia odmawiało udziału około 60% wylosowanych gospodarstw, w całej zaś Polsce — 40%.

2. *Istnieje jeszcze możliwość zmniejszenia liczby odmów w przyszłych badaniach budżetów rodzinnych, dzięki usprawnieniu organizacji badania,* tj. zastosowanie na szeroką skalę reklamy przy użyciu środków masowego przekazu oraz przez odpowiedni dobór kadry instruktorów itp.

3. *Przeciętne wydatki rodzin uzyskane metodą rotacyjną nie różnią się istotnie od przeciętnych wydatków uzyskanych metodą tradycyjną.* Jedynie dane dotyczące dochodów otrzymane z eksperymentu są niższe od faktycznych. Przyczyną tego stanu rzeczy mógł być m.in. brak anonimowości badania między poszczególnymi członkami rodziny oraz przyjęcie w badaniu jako obowiązującej płacy netto, gdyż większość ankietowanych osób nie orientowała się dokładnie o wysokości tego rodzaju wynagrodzenia.

4. *Rozkład wydatków w poszczególnych tygodniach jest nierównomierny, przy czym różnice te wystąpiły głównie przy porównaniu wydatków między pierwszym i drugim oraz pierwszym i trzecim tygodniem badania, nie wystąpiły natomiast przy porównaniu przeciętnych wydatków w drugim i w trzecim tygodniu.* Nadmienić należy, że przeciętne wydatki osiągały największą wartość w pierwszym tygodniu badań, a najniższą wartość w czwartym tygodniu badań.

5. *Stopień dokładności oszacowanych przeciętnych wydatków żywnościowych nie różni się istotnie w poszczególnych okresach badania.* Natomiast *dokładność ocen wydatków nieżywnościowych była znacznie większa w okresie czterotygodniowym niż dwutygodniowym.*

6. *Należy stosować różne okresy zbierania informacji.* Optymalnym okresem badania dla większości wydatków jest okres miesięczny. Dla wydatków na artykuły charakteryzujące się małą częstotliwością zakupu okres ten należy wydłużyć do pół roku.

7. Przyjęcie w badaniu *kombinowanej metody zbierania informacji* pozwoliło na:

a) odciążenie gospodarstw domowych od obowiązku codziennego notowania wszystkich operacji finansowych przez dłuższy okres,
b) zmniejszenie wpływu, jaki wywiera na wysokość i strukturę wydatków w gospodarstwach domowych notowanie ich przez dłuższy czas.

8. *Nadmierna szczegółowość formularzy zniekształca wyniki i utrudnia instruktorom pracę w terenie.*

9. *W badaniach budżetów rodzinnych metodą rotacyjną, sprawa doboru odpowiedniego personelu ma istotne znaczenie,* gdyż okres badania gospodarstw domowych jest krótki, w związku z tym instruktor musi nawiązywać kontakty i współpracować z coraz to nowymi gospodarstwami domowymi. Zakładano, że osoby młode (studenci) będą lepiej wywiązywać się ze swoich obowiązków. Przypuszczenie takie, jak się okazało w czasie trwania badania, było błędne. Osoby te pracowały gorzej i kilkakrotnie musiano je wymieniać z powodu niesolidności i nieobowiązkowości. Cechą charakterystyczną tej wymiany było to, że studentów zastępowano osobami w wieku 30—40 lat.

W przypadku przyjęcia metody rotacyjnej do badań budżetów rodzinnych należy liczyć się ze zmianą dotychczasowej organizacji badania oraz trudnościami w porównywaniu danych z danymi uzyskanymi innymi metodami. Mimo tych trudności wydaje się celowe zastosowanie tej metody do przyszłych badań budżetów rodzinnych, z tym że wymaga ona jeszcze dopracowania.

Ryszard ZASĘPA

ZASTOSOWANIE METODY REPREZENTACYJNEJ DO BADANIA BŁĘDÓW NIELOSOWYCH

Jedną z przyczyn powodujących, iż stosowanie metody reprezentacyjnej w praktyce napotyka trudności, jest fakt, że wyniki badań przeprowadzanych tą metodą są obarczone błędami losowymi, a to kojarzy się z przeświadczeniem, jakoby badanie metodą reprezentacyjną było zawsze mniej dokładne niż badanie pełne. W rzeczywistości, wyniki masowego badania statystycznego z reguły nie są dokładne, obojętne czy badanie jest pełne, czy nie, zwłaszcza że na wyniki badania mogą mieć poważny wpływ błędy obserwacji. Przeprowadzając badanie pełne jesteśmy często zmuszeni do stosowania tylko niektórych metod obserwacji statystycznej. Jeżeli stosowana metoda obserwacji nie jest doskonała, błędy obserwacji, które z reguły mają charakter tendencyjny, jednokierunkowy, mogą być źródłem poważnych niedokładności wyników badania statystycznego. Tych niedokładności nie byłoby, gdyby zastosować inną lepszą metodę obserwacji (np. zamiast zbierać metodą wywiadu dane o powierzchni mieszkania, które mogą być bardzo niedokładne, można by zmierzyć powierzchnię mieszkania i uzyskać tą metodą wyniki znacznie dokładniejsze). Dokładniejsza metoda obserwacji będzie jednak bardziej pracochłonna i kosztowniejsza i dlatego nie można jej będzie — ze względów technicznych — zastosować w pełnym badaniu statystycznym.

Inaczej sprawa się przedstawia, gdy badanie statystyczne jest prowadzone metodą reprezentacyjną. Wówczas stosowanie dokładniejszej metody obserwacji statystycznej może okazać się wykonalne. Wtedy suma błędów losowych i błędów obserwacji może być mniejsza niż suma błędów obserwacji w badaniu pełnym. Stąd wniosek, że *badanie metodą reprezentacyjną może dostarczać dokładniejszych wyników niż badanie wyczerpujące.* Decyzja czy badanie powinno być pełne, czy prowadzone metodą reprezentacyjną, powinna być podjęta po dokładnym przeanalizowaniu możliwości zastosowania różnych metod obserwacji statystycznej.

Prawdą jest, że przy tej samej metodzie obserwacji statystycznej badanie metodą reprezentacyjną dostarczy wyników mniej dokładnych niż badanie pełne. Jeżeli więc precyzja wymaganych szacunków musi być duża, badanie musi być pełne. Wiadomo, że precyzja szacunku metodą reprezentacyjną jest tym mniejsza, im rzadziej badane zjawisko występuje w populacji generalnej oraz im mniejsza jest badana zbiorowość. Nie ulega wątpliwości, że statystyka państwowa musi dostarczyć niektórych danych, np. z wykonania planu gospodarczego, w bardzo wąskich przekrojach terytorialnych bądź organizacyjnych i wówczas badanie statystyczne musi być pełne. Powstaje jednak pytanie, które wskaźniki statystyczne należy opracować z dużą szczegółowością terytorialną czy organizacyjną, a które wystarczyłoby opracować np. dla całego kraju. Zwykle, dla opracowania danych w pewnym zakresie, zbiera się sprawozdania okresowe, często bardzo szczegółowe, ze wszystkich jednostek sprawozdawczych. Sądzę, że analiza rzeczywiście potrzebnej szczegółowości opracowania pozwoliłaby podzielić zbierane informacje na:

a) takie, które są niezbędne do obliczenia wskaźników w bardzo szczegółowym podziale terytorialnym bądź organizacyjnym,

b) takie, które są potrzebne do obliczenia wskaźników w szerokim ujęciu.

Wówczas informacje kategorii a) należałoby zbierać od wszystkich jednostek sprawozdawczych, natomiast informacje kategorii b) tylko od jednostek wylosowanych do próby. Łatwo zauważyć, że taka kombinacja badania pełnego z badaniem metodą reprezentacyjną znacznie zmniejszy pracę sprawozdawczą, uzyskane wyniki zaś mogą być wystarczająco dokładne. Możliwość takiego ustawienia sprawozdawczości nie jest na ogół wykorzystywana w praktyce. Wprawdzie organizacja takiej sprawozdawczości jest bardziej skomplikowana niż w przypadku sprawozdawczości pełnej, to jednak społecznie uzyskałoby się w ten sposób znaczne oszczędności w pracy sprawozdawczej.

II/3505/71

Ryszard ZIELIŃSKI

LOSOWANIE PRÓBEK ZA POMOCĄ EMC

Podstawowym problemem w praktyce badań reprezentacyjnych jest wylosowanie próbki, na podstawie której będą estymowane interesujące nas parametry populacji lub weryfikowane hipotezy o tej populacji. Aktualna technika pobierania próbek za pomocą drukowanych tablic liczb losowych jest bardzo kłopotliwa, a wylosowanie dużej próby (rzędu kilku lub kilkunastu tysięcy jednostek) staje się zadaniem bardzo trudnym lub nawet niemożliwym do wykonania. Istnieją różne sposoby ominięcia tych trudności, np. losowanie systematyczne, ale stwarzają one trudności związane z obliczaniem efektywności takiego losowania i wariancji uzyskiwanych estymatorów.

Technika obliczeniowa wykorzystująca maszyny cyfrowe pozwala na ominięcie wszystkich trudności i na losowanie dowolnie dużych próbek z dowolnie uporządkowanych populacji w dowolnym schemacie losowania, przy czym numery lub nazwy wylosowanych jednostek mogą być zestawiane przez maszynę cyfrową w dowolnym porządku.

Podstawowe zagadnienia, związane z otrzymywaniem liczb losowych w maszynie cyfrowej oraz z wykorzystaniem programów generujących te liczby w praktyce, były wielokrotnie dyskutowane w Komisji Matematycznej GUS, a w ośrodku obliczeniowym GUS opracowano odpowiedni program dla maszyny ICL 1905. Cała procedura losowania odbywać się może w maszynie cyfrowej. Na przykład w najprostszym przypadku, gdy z populacji N-elementowej należy wylosować n-elementową próbkę prostą, statystyk podaje maszynie liczby N oraz n, a otrzymuje z maszyny ciąg numerów wylosowanych jednostek, uporządkowanych zgodnie z życzeniem statystyka, od najmniejszego do największego. W przypadku gdy materiał jest uporządkowany w bardziej złożony sposób — np. jeżeli arkusze spisowe uporządkowane są według rejonów, można otrzymać z maszyny numery wylosowanych elementów w rozbiciu na te rejony, uporządkowane w żądany sposób (np. w ciągi monotoniczne) wewnątrz każdego rejonu.

Generatory liczb losowych stają się szczególnie przydatne, gdy wszystkie informacje o populacji przeniesione są na taśmy magnetyczne i mogą być wprowadzone do pamięci maszyny. Elementy populacji wylosowane do próbki są natychmiast przez maszynę opracowywane i statystyk otrzymuje gotową odpowiedź na pytania, które sformułował. Możliwe jest przy tym stosowanie złożonych schematów losowania, np. losowania warstwowego. Na przykład przy statystycznym opracowaniu budżetów rodzinnych możemy żądać, żeby maszyna wylosowała taką próbkę, w której będzie nr „formularzy" dotyczących rodzin n-osobowych. Można oczywiście nie podawać maszynie liczby nr, lecz żądać, żeby liczby te zostały przez maszynę obliczone według odpowiedniego programu, a następnie wykorzystane do wylosowania próbki w schemacie żądanego losowania warstwowego.

Możliwe jest oczywiście również losowanie prób według schematów losowania z różnymi prawdopodobieństwami wyboru. Takie schematy są szczególnie efektywne, jeżeli prawdopodobieństwa wyboru są dobrze ustalone.

Prace nad generatorami liczb losowych zostały w GUS stosunkowo niedawno zapoczątkowane, ale już teraz istnieją możliwości rozwiązywania podstawowych problemów wymagających użycia takich liczb.

Dotychczas w serii

BIBLIOTEKA WIADOMOŚCI STATYSTYCZNYCH

ukazały się następujące pozycje:

1967 r.

Tom 1. Wybrane problemy statystyki w Polsce . . . . . 25,—
Tom 2. Zagadnienia statystyki rolniczej . . . . . 20,—
Tom 3. Statystyka regionalna. Aktualny stan i problemy rozwoju 30,—
Tom 4. Problemy demograficzne Polski Ludowej . . . . . 30,—

1968 r.

Tom 5. Bilanse gospodarki narodowej . . . . . 30,—

1969 r.

Tom 6. Problemy demograficzne Ziem Zachodnich i Północnych PRL 25,—
Tom 7. Zastosowanie metod matematycznych w statystyce . . . 20,—
Tom 8. Problemy demograficzne Kielecczyzny . . . . . 15,—
Tom 9. Mierniki rozwoju regionów . . . . . 30,—

1970 r.

Tom 10. Założenia programowe i organizacyjne Narodowego Spisu Powszechnego w 1970 r. . . . . 15,—
Tom 11. Wybrane problemy prognoz statystycznych . . . . . 24,—
Tom 12. Aktualne problemy statystyki regionalnej w krajach europejskich . . . . . 22,—
Tom 13. Teoretyczne i metodologiczne problemy statystyki społecznej 30,—

1971 r.

Tom 14. Badania statystyczne metodą reprezentacyjną w krajach socjalistycznych . . . . . 27,—

GŁÓWNY URZĄD STATYSTYCZNY BIBLIOTEKA

INFORMACJA

o sieci księgarni prowadzących działy sprzedaży publikacji Głównego Urzędu Statystycznego

Publikacje Głównego Urzędu Statystycznego można zamawiać i nabywać we wszystkich miastach wojewódzkich w następujących księgarniach:

w Warszawie	— Główna Księgarnia Naukowa im. B. PRUSA ul. Krakowskie Przedmieście 7
	— Księgarnia MDM ul. Piękna 31/37
	— Księgarnia Prawno-Ekonomiczna ul. Nowy Świat 1
	— Kiosk PP „Dom Książki" w Komisji Planowania przy Radzie Ministrów Plac Trzech Krzyży 3/5
	— Kiosk SPUPiK „Ruch" w gmachu Głównego Urzędu Statystycznego al. Niepodległości 208
w Białymstoku	— Księgarnia Naukowa ul. Próchnika 5
w Bydgoszczy	— Księgarnia Naukowo-Techniczna Stary Rynek 15
w Gdańsku-Wrzeszczu	— Księgarnia Naukowo-Techniczna ul. Grunwaldzka 111/113
w Katowicach	— Księgarnia Naukowo-Techniczna ul. Warszawska 11
w Kielcach	— Księgarnia przy ul. Sienkiewicza 37
w Koszalinie	— Księgarnia Naukowo-Techniczna Pl. Bojowników PPR 2
w Krakowie	— Księgarnia przy ul. Podwale 6
w Lublinie	— Księgarnia Naukowo-Techniczna ul. Krakowskie Przedmieście 39
w Łodzi	— Księgarnia Naukowa i Medyczna ul. Piotrkowska 102a
w Olsztynie	— Księgarnia przy Pl. Wolności 2/3
w Opolu	— Księgarnia Naukowa Rynek 19/23
w Poznaniu	— Księgarnia Naukowa ul. Armii Czerwonej 69
w Rzeszowie	— Księgarnia Naukowo-Techniczna ul. 3 Maja 2
w Szczecinie	— Księgarnia Naukowa Al. Wojska Polskiego 39
we Wrocławiu	— Księgarnia przy ul. Rynek 60
w Zielonej Górze	— Księgarnia Naukowo-Techniczna ul. Karola Marksa 3

Wymienione księgarnie przyjmują również zamówienia na sprzedaż wysyłkową egzemplarzową i na całe serie wydawnicze oraz udzielają informacji o wydanych publikacjach GUS i terenowych organów statystycznych.

A/ 48193

2/3505/71

Cena zł 25.–

CENTRALNA BIBLIOTEKA
STATYSTYCZNA - GUS

108371