

Arkadiusz Kozłowski
Mirosław Szreder

INFORMACJE SPOZA PRÓBY W BADANIACH STATYSTYCZNYCH



Wydawnictwo Uniwersytetu Gdańskiego

INFORMACJE SPOZA PRÓBY W BADANIACH STATYSTYCZNYCH

Arkadiusz Kozłowski
Mirosław Szreder

INFORMACJE SPOZA PRÓBY W BADANIACH STATYSTYCZNYCH

Wydawnictwo Uniwersytetu Gdańskiego
Gdańsk 2020

Recenzja
dr hab. Grażyna Dehnel, prof. UEP

Redakcja wydawnicza
Anna Roman

Projekt okładki i stron tytułowych
Filip Sendal

Skład i łamanie
Mariusz Szewczyk

Publikacja sfinansowana ze środków Dziekana Wydziału Zarządzania Uniwersytetu Gdańskiego

© Copyright by Uniwersytet Gdański
Wydawnictwo Uniwersytetu Gdańskiego

ISBN 978-83-7865-955-6

Wydawnictwo Uniwersytetu Gdańskiego
ul. Armii Krajowej 119/121, 81-824 Sopot
tel./fax 58 523 11 37, tel. 725 991 206
e-mail: wydawnictwo@ug.edu.pl
www.wyd.ug.edu.pl

Księgarnia internetowa: www.kiw.ug.edu.pl

Druk i oprawa
Zakład Poligrafii Uniwersytetu Gdańskiego
ul. Armii Krajowej 119/121, 81-824 Sopot
tel. 58 523 14 49; fax 58 551 05 32

Spis treści

Wprowadzenie	7
Rozdział 1. Badania statystyczne pełne i próbkowe	9
1.1. Badania wyczerpujące (pełne) i ich zastosowania	9
1.2. Specyfika badań próbkowych	11
1.2.1. Określenie celów badawczych	13
1.2.2. Zdefiniowanie populacji	14
1.2.3. Sporządzenie operatu losowania	14
1.2.4. Wybór techniki pobierania próby	15
1.3. Źródła i rodzaje informacji spoza próby	19
1.4. Klasyfikacja błędów w badaniach statystycznych	26
1.5. <i>Big data</i> – szanse i wyzwania	31
Rozdział 2. Rola dodatkowej informacji w próbkowaniu nieprobabilistycznym	35
2.1. Przesłanki i konsekwencje stosowania próbkowania nielosowego	35
2.2. Wygodny dobór próby (wybór przypadkowy)	38
2.3. Próba celowa	40
2.4. Próba kwotowa	42
2.5. Wybór próby metodą kuli śnieżnej	45
Załącznik. Prawdopodobieństwo i jego interpretacje	48
Rozdział 3. Rola informacji spoza próby w badaniach reprezentacyjnych	57
3.1. Znaczenie losowego wyboru próby badawczej	57
3.2. Losowanie proste	59
3.3. Losowanie systematyczne	61
3.4. Losowania warstwowe i zrównoważone	64
3.5. Losowanie zespołowe	71
Rozdział 4. Braki odpowiedzi i skuteczność technik imputacji	75
4.1. Rodzaje i przyczyny powstawania braków odpowiedzi	75
4.2. Zapobieganie brakom odpowiedzi	78
4.3. Wpływ braków odpowiedzi na wnioskowanie z próby	81
4.4. Imputacja danych	88
4.4.1. Metody imputacji jednokrotnej	88
4.4.2. Imputacja wielokrotna	98

Rozdział 5. Zastosowania technik ważenia danych i kalibracji wag	105
5.1. Rodzaje technik ważenia danych i ich wykorzystanie	105
5.1.1. Ogólna idea ważenia danych	105
5.1.2. Wagi wynikające z planu losowania	107
5.1.3. Korekty ze względu na błędy pokrycia	112
5.1.4. Korekty ze względu na braki odpowiedzi	116
5.1.5. Korekty z wykorzystaniem informacji spoza próby	122
5.2. Podejście kalibracyjne	128
5.2.1. Pojęcie kalibracji wag	128
5.2.2. Wyznaczanie wag kalibracyjnych	129
5.2.3. Estymacja regresyjna i szczególne przypadki kalibracji	133
5.2.4. Kalibracja w sytuacji braków odpowiedzi	138
5.2.5. Obciążenie i wariancja estymatora kalibracyjnego	140
Podsumowanie	145
Bibliografia	147
Oznaczenia	153
Indeks	155

Wprowadzenie

Postęp technologiczny, jakiego doświadczamy w gromadzeniu, przetwarzaniu i analizowaniu danych liczbowych, kształtuje nowe tendencje w zakresie projektowania i realizowania badań statystycznych. Można wśród nich dostrzec szczególnie dwie coraz silniej utrwalające się w ostatnich dekadach prawidłowości. Pierwsza polega na stopniowym wypieraniu pełnych badań statystycznych, w których pomiarem objęta jest każda jednostka badanej zbiorowości, przez tańsze i szybsze w realizacji badania częściowe, nazywane inaczej próbkowymi. W badaniach próbkowych obserwacją i pomiarem obejmuje się jedynie niektóre jednostki (wybrane do próby), aby następnie wnioskować o charakterystykach całej populacji. Istotą drugiej tendencji jest zaś dążenie projektantów i realizatorów badań częściowych do wzbogacania informacji pozyskanych z próby o inne źródła wartościowej wiedzy i informacji, mogących się przyczynić do polepszenia jakości wnioskowania. W ten sposób typowe dotąd badanie próbkowe przestaje już być wyłącznie oparte na próbie, a coraz bardziej przekształca się w taki rodzaj badania, w którym kombinacja informacji pochodzących z różnych źródeł tworzy podstawę do wnioskowania o populacji.

Niniejsza monografia jest poświęcona tej drugiej tendencji, czyli sposobom włączania do badań próbkowych informacji spoza próby na różnych etapach realizacji badania. Aby ułatwić Czytelnikowi zrozumienie zarówno potrzeby wzbogacania informacji zawartej w próbie losowej lub nielosowej i znaczenia wiedzy wstępnej o populacji, jak i sposobów jej inkorporacji do badania, ujęliśmy treść tego opracowania w kilka rozdziałów, uzupełniając ją przykładami liczbowymi, załącznikiem traktującym o prawdopodobieństwie i licznymi odwołaniami do literatury. Głównym celem monografii jest scharakteryzowanie i krytyczne omówienie sposobów włączania informacji spoza próby do badań reprezentacyjnych, a także przedstawienie własnych propozycji udoskonaleń, zarówno analitycznych, jak i praktyki badawczej, pozwalających poprawiać wiarygodność i precyzję wnioskowania.

Wprowadzający do tej problematyki pierwszy rozdział zawiera przede wszystkim charakterystykę badań częściowych – ich projektowanie i specyfikę. Omówienie błędów losowych i nielosowych towarzyszących badaniom próbkowym pozwala uzasadnić potrzebę sięgania we wnioskowaniu poza informacje zawarte w próbie

badawczej. O źródłach tych informacji, a także o znaczeniu dużych zbiorów danych i narzędzi *big data* traktują ostatnie fragmenty tego rozdziału.

W drugim rozdziale scharakteryzowano znaczenie informacji spoza próby w badaniach wykorzystujących próbkowanie nieprobabilistyczne (próby nielosowe). Wskazano w nim na rodzaje informacji wstępnej o populacji, które decydują o użyteczności technik nielosowego wyboru próby w badaniach rynkowych i w niektórych badaniach społecznych. Uzasadniono jednocześnie, że uzyskane tymi technikami próby nie pozwalają na wykorzystanie metod wnioskowania statystycznego w uogólnieniach opartych na ich wynikach. Metody te bowiem zakładają losowy proces generowania obserwacji w próbie i dzięki temu umożliwiają przypisanie określonych prawdopodobieństw wynikom wnioskowania. Załącznik do tego rozdziału prezentuje trzy najważniejsze interpretacje prawdopodobieństwa: klasyczną, częstościową i personalistyczną. Wskazuje się w nim na rosnącą współcześnie rolę tej ostatniej – wykorzystuje ona wiedzę ekspertów i badacza na temat stanowiący przedmiot badania.

W rozdziale trzecim omówiono rodzaje informacji spoza próby, które skłaniają badacza do zastosowania w praktyce innych rodzajów próbkowania niż podstawowego schematu probabilistycznego, jakim jest losowanie proste. W rozdziale tym przedyskutowano znaczenie informacji wstępnej o populacji i sposoby jej włączania w kilku podstawowych technikach próbkowania: systematycznego, warstwowego i zrównoważonego oraz zespołowego.

Czwarty rozdział poświęcono w całości jednemu z największych wyzwań współczesnych badań statystycznych, jakim są braki odpowiedzi. Podstawowym sposobem wypełnienia luki powstałej z powodu odmowy lub braku kontaktu z częścią respondentów jest wykorzystanie informacji o cechach, których wartości są znane z innych źródeł, oraz wszelkiej innej użytecznej wiedzy *a priori*. W rozdziale tym omówiono: możliwości zapobiegania brakom odpowiedzi, konsekwencje braków, a przede wszystkim techniki imputacji i ich skuteczność.

W rozdziale piątym przedstawiono techniki ważenia danych i kalibracji wag jako ważnych kroków w dążeniu do poprawy statystycznych własności próby, która ostatecznie będzie podstawą wnioskowania. Istotą tych technik jest połączenie informacji wstępnej o populacji, informacji o przebiegu badania oraz wyników samego badania w taki sposób, aby przeważona próba stanowiła lepszą od pierwotnej reprezentację populacji. Zwiększa to szanse na lepszej jakości wnioskowanie o charakterystykach populacji.

Ufam, że monografia ta przyczyni się do lepszego zrozumienia wśród Czytelników zagadnień projektowania i realizacji badań reprezentacyjnych, w tym przede wszystkim wykorzystania w nich informacji spoza próby. W dobie rosnących frakcji odmów wśród respondentów z jednej strony, a z drugiej zwiększających się zasobów danych liczbowych, do których badacze mają dostęp, zaprezentowane w tym opracowaniu sposoby wzbogacania informacji próbkowej mogą się okazać użyteczne w praktyce badań statystycznych.

Rozdział 1

Badania statystyczne pełne i próbkowe

1.1. Badania wyczerpujące (pełne) i ich zastosowania

Głównym celem stosowania metod statystycznych w naukowym opisie rzeczywistości jest dążenie do wyodrębnienia i poznania prawidłowości tkwiących w zbiorach liczb. Zwykle chodzi o duże zbiory liczbowe, w których trudno byłoby określić wzorce lub prawidłowości bez zastosowania specjalnych technik i narzędzi, jakie przez wieki wypracowali statystycy. Ludzie zawsze mieli skłonność do poszukiwania wzorców (regularności) w środowisku, w którym żyli, i przypisywania im znaczenia (Młodinow 2008). Chcieli rozumieć prawa rządzące zdarzeniami, jakie obserwowali i jakich doświadczali: choroby i śmierć ludzi, zwierząt, wielkość plonów, a w kolejnych wiekach: błędy pomiarów w fizyce, astronomii, chemii, przyczyny epidemii, wypadków komunikacyjnych. Przeczuli, że w tych danych liczbowych kryje się wartościowa wiedza o człowieku i o świecie.

Współcześnie w celu poznania prawidłowości tkwiących w zjawiskach masowych posługujemy się doskonalonymi przez dziesięciolecia metodami statystycznymi wspartymi coraz bardziej zaawansowaną matematyką i technologią komputerową. W epoce powszechnej komputeryzacji i cyfryzacji, w której pomiar liczbowy staje się powszechniejszy niż kiedykolwiek wcześniej, rośnie rola, a także odpowiedzialność statystyków za wypracowanie skutecznych sposobów wydobywania wiedzy ze zbiorów liczbowych.

Niezależnie od tego, co dane te opisują, jakich elementów rzeczywistości dotyczą, ważne jest określenie już na wstępie, czy poszukiwane w zbiorach liczb prawidłowości chcemy odnieść jedynie do zbioru jednostek, którego liczby te dotyczą, czy też do większej zbiorowości. W drugim wypadku zadanie jest trudniejsze, bo oznacza wnioskowanie o pewnej zbiorowości na podstawie danych dotyczących tylko niektórych jej reprezentantów. Badanie statystyczne, w którym pomiarem obejmuje się wszystkie jednostki zdefiniowanej na wstępie zbiorowości (populacji), nazywa się **badaniem wyczerpującym** lub **pełnym**. Jeżeli zaś obserwacji liczbowych dokonano na pewnej próbie (losowej lub nielosowej) jednostek pobranych z populacji, a celem badania jest poznanie charakterystyk całej tej populacji, to badanie takie nazywa się **badaniem niewyczerpującym** albo **próbkowym**.

W badaniu wyczerpującym zastosowanie znajdują przede wszystkim metody i techniki statystyki opisowej (*descriptive statistics*). W tym dziale statystyki słowo „opis” ma sugerować, że charakterystyka prawidłowości nie wykracza poza pewne ramy, mianowicie: poza zbiorowość, o której zgromadzono dane liczbowe. W zakresie statystyki opisowej mieszczą się wszelkie dociekania o wzorcach i regularnościach populacji opisanej przez zbiór obserwacji liczbowych. Nie zawierają się zaś te metody i techniki, które zaprojektowano do celów wnioskowania na podstawie niepełnej informacji (próby). W statystyce opisowej, dysponując zbiorem liczb, nie pytamy – w przeciwieństwie do wnioskowania statystycznego (*statistical inference*) – jaki stochastyczny mechanizm generujący obserwacje stoi za tymi liczbami. Dociekamy natomiast prawidłowości opisujących badaną zbiorowość, wśród których do najważniejszych zalicza się:

- tendencję centralną (przeciętny poziom wartości badanych cech);
- dyspersję (rozproszenie);
- asymetrię (skośność rozkładów cech);
- koncentrację;
- współzależności (w tym korelację) między badanymi cechami.

W licznych podręcznikach ze statystyki opisowej¹ przedstawiono syntetyczne miary poziomu i natężenia wymienionych właściwości rozkładów cech wraz z ich interpretacją.

Postęp w rozwoju metod wnioskowania statystycznego, w tym rachunku prawdopodobieństwa i technik próbkowania, sprawia, że wiele badań, które w przeszłości wykonywane były jako wyczerpujące, obecnie zastąpiono tańszymi i szybszymi w realizacji badaniami próbkowymi. Inną stosunkowo nową przyczyną malejącej liczby badań wyczerpujących jest zmniejszająca się potrzeba ich realizacji w świetle rosnących i coraz lepiej zorganizowanych zasobów danych administracyjnych. W różnego rodzaju rejestrach urzędowych gromadzone są istotne informacje o zbiorach: osób, podmiotów gospodarczych, instytucji, a także zjawisk. W wielu krajach właśnie te zbiory danych, m.in. rejestry ludności, stanowią alternatywę dla pozyskiwania ważnego zakresu informacji o ludności i mieszkaniach za pomocą spisu powszechnego – najbardziej popularnego badania wyczerpującego². Nie znaczy to jednak, że tzw. statystyka publiczna, czyli system oficjalnej statystyki w danym kraju, nie realizuje już wcale badań wyczerpujących³. Jednak prawie wszystkie spośród tych badań, które przykuwają najwięcej uwagi, np. badania rynku pracy (BAEL – badanie aktywności ekonomicznej ludności), pomiar inflacji

¹ Wśród podręczników można wymienić takie pozycje, jak np. Józwiak i Podgórski (2009), Makać i Urbanek-Krzysztofiak (2008) lub Sobczyk (2010).

² W tym kontekście o końcu ery tradycyjnych spisów ludności obszernie i interesująco pisze Gołata (2018).

³ Przykładami badań wyczerpujących realizowanych przez GUS są: Narodowy Spis Powszechny 2021, Powszechny Spis Rolny 2020, badania z zakresu funkcjonowania oświaty i wychowania oraz szkolnictwa wyższego, działalności służb ratowniczych, partycypacji publicznej na poziomie jednostek samorządu terytorialnego i inne.

(badanie zmian cen towarów i usług konsumpcyjnych), badanie wysokości wynagrodzeń według zawodów, badania budżetów gospodarstw domowych (BBGD), są wykonywane jako badania próbkowe. Jeżeli dodać do tego popularne, zwłaszcza w przekazie medialnym, badania opinii społecznej na różne tematy, szeroki wachlarz badań konsumenckich, a także liczne badania przedwyborcze, łącznie z badaniem *exit poll* w dniu wyborów, to okaże się, że nie tylko w Polsce, ale w wielu innych krajach coraz częściej badanie statystyczne realizowane jest jako badanie niewyczerpujące (próbkowe).

Istnieją jednak sytuacje, w których albo nieduża liczebnie populacja przesądza o tym, że wykonuje się badanie pełne, albo pewne inne czynniki wskazują na celowość poddania pomiarowi wszystkich jednostek populacji. Jednym z takich czynników jest duże rozproszenie wartości badanej cechy lub badanych cech. W wypadku bardzo dużej dyspersji może się okazać konieczne zbadanie wszystkich jednostek populacji, aby z odpowiednią precyzją sformułować oceny dotyczące charakterystyk danej cechy. Inną okolicznością wskazującą na zastosowanie badania wyczerpującego może być żądanie uzyskania wysoce precyzyjnych ocen interesujących badacza parametrów populacji. I mimo że badanie pełne nie jest wolne od błędów, uzyskane na jego podstawie oceny nie są obciążone błędami losowania, obecnymi w każdym badaniu opartym na próbie losowej.

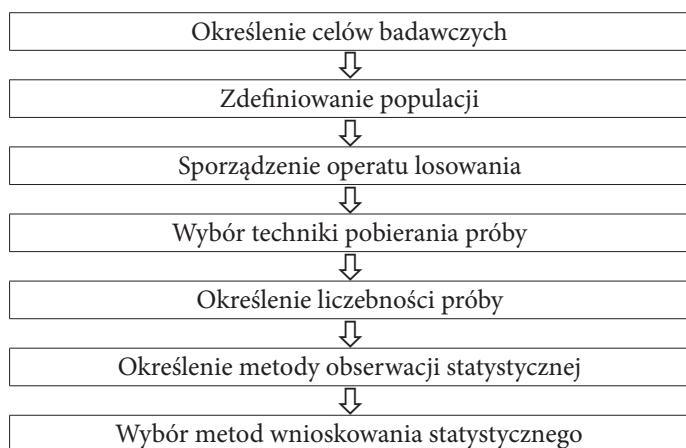
1.2. Specyfika badań próbkowych

Celem badań niewyczerpujących, zwanych inaczej próbkowymi (*sample surveys*), jest statystyczne wnioskowanie o właściwościach lub parametrach określonej na wstępie populacji na podstawie pomiarów jedynie pewnej jej części (próby). Tak jak w potocznym języku mówi się: „Wnioskuje więc, że...”, co oznacza, że wyraża się pewną opinię jedynie na podstawie określonych przesłanek, której prawdziwości nie jest się pewnym, tak w statystyce wnioskuje się o populacji na podstawie próby będącej odpowiednikiem tychże przesłanek. W obu wypadkach jest to działanie w warunkach niepełnej informacji. W konsekwencji wynik tego działania, czyli wnioskowania, zawsze będzie obciążony pewną dozą niepewności. Gdy obserwacje w próbie są generowane losowo, miarą tej niepewności jest prawdopodobieństwo. W badaniach opartych na próbach losowych możliwe jest wyrażenie w kategoriach probabilistycznych zarówno oceny prawdziwości wnioskowania, jak i precyzji uzyskanych rezultatów.

Podstawowe znaczenie dla jakości wnioskowania ma oczywiście wartość wspomnianych przesłanek, czyli we wnioskowaniu statystycznym są to informacje zawarte w próbie badawczej. Sposób wyboru próby oraz jej liczebność w znacznym stopniu determinują wiarygodność i precyzję wnioskowania. W praktyce jednak nawet najstaranniej zaprojektowana próba nie zapewni dobrej jakości wnioskowania, jeżeli znaczny będzie odsetek odmów respondentów, pytania

w kwestionariuszu będą niejasne, rodzaj wywiadu źle dobrany albo zaangażowanie ankietów – niewystarczające. Są to tylko niektóre z przyczyn o charakterze nielosowym mogące negatywnie wpływać na jakość wnioskowania. Z tego m.in. powodu coraz większy nacisk we współczesnych badaniach próbkowych kładzie się na wzbogacenie informacji z próby o inne użyteczne źródła danych, które mogłyby zniwelować skutki działania wymienionych czynników. Przede wszystkim temu zagadnieniu jest poświęcona niniejsza monografia. Trzeba jednak wyraźnie podkreślić, że we wnioskowaniu statystycznym ważne są zarówno wszelkie dostępne źródła informacji – z próby i spoza niej – jak i modele oraz metody wnioskowania, posiadające solidne podstawy matematyczne. Ważne jest bowiem nie tylko to, jakie informacje o populacji badacz posiada, ale także to, w jaki sposób potrafi je przetwarzać i analizować.

Pierwszym etapem każdego badania próbkowego jest jego zaprojektowanie (*research design*). Przez projektowanie badania rozumie się zespół czynności mających na celu zapewnienie efektywnych metod gromadzenia, analizy i udostępniania informacji, tak aby jak najlepiej zaspokoić określone na wstępie potrzeby informacyjne użytkownika (zamawiającego badanie). Projektowanie stanowi jedną z najważniejszych faz badania, w której rozstrzyga się wiele kluczowych kwestii rzutujących na ogólną jego jakość i przydatność wyników. Dość powiedzieć, że właśnie w tej fazie rozstrzygnięciu podlegają tak zasadnicze kwestie jak: co lub kto będzie przedmiotem badania, jaki rodzaj wyboru próby zostanie wykorzystany, jak duża liczebnie będzie próba, jakie formy ankietowania będą zastosowane, jakie techniki wnioskowania zostaną użyte na etapie analizy wyników? Wszystko to sprawia, że badacze poświęcają na ogół dużo uwagi właściwemu zaprojektowaniu badania. Charakteryzując zespół najważniejszych czynności, jakie ten etap obejmuje, można je podzielić na kilka kroków (rys. 1.1).



Rysunek 1.1. Proces projektowania badania próbkowego

Źródło: opracowanie własne.

Scharakteryzowano te kroki w projektowaniu badań próbkowych, do których odwołujemy się w dalszych partiach książki, koncentrując się na głównym zagadnieniu – wykorzystaniu informacji spoza próby badawczej w badaniach niewyczerpujących. Nie rozwijamy natomiast obszernej problematyki metod obserwacji statystycznej (pomiaru sondażowego)⁴, a także zagadnienia określenia właściwej liczebności próby, które szczegółowo omówili w swoich pracach m.in. Szreder (2010), Rószkiewicz i in. (2013), Wywiół (1999).

1.2.1. Określenie celów badawczych

Właściwe określenie celu badania determinuje cały dalszy proces projektowania badania i jest niezbędnym warunkiem zgromadzenia adekwatnych danych liczbowych i informacji. Warunki te najlepiej wyraża stwierdzenie Alberta Einsteina: „sformułowanie problemu jest często ważniejsze niż jego rozwiązanie”⁵, lub znane w języku angielskim powiedzenie: „a problem well defined is a problem half solved”. W zdefiniowaniu celów badawczych pomocne może być wyróżnienie zasadniczego celu badania oraz tzw. celów cząstkowych, precyzujących poszczególne zagadnienia badawcze. Główny cel badania zostaje sformułowany po dokładnym scharakteryzowaniu potrzeb informacyjnych użytkownika i cel ten jest podporządkowany zrealizowaniu tych potrzeb. Potrzeby informacyjne z kolei mogą wynikać z sytuacji decyzyjnej użytkownika, wymagającej posiadania przez niego pewnych dodatkowych informacji, które pozwalałyby zwiększyć prawdopodobieństwo podjęcia dobrej decyzji. Przykładem może być sytuacja wymagająca podjęcia przez menadżera decyzji o nakładzie nowego ilustrowanego czasopisma hobbystycznego. Decyzje takie są zwykle poprzedzone dokładnymi badaniami opinii konsumentów na temat zainteresowania daną problematyką, regularnych zakupów innych czasopism o zbliżonej tematyce oraz chęci prenumeraty nowego pisma. Szczegółowe problemy badawcze natomiast obejmują węższe tematycznie zagadnienia mieszczące się w głównym celu badania, których poznanie prowadzi do realizacji celu głównego. Pełna specyfikacja najważniejszych problemów badawczych zapewnia zaspokojenie potrzeb informacyjnych, które syntetycznie ujęto w głównym celu badania. Pominiecie któregoś z istotnych problemów powoduje lukę w zbiorze informacji gromadzonych w badaniu, czego konsekwencją może być niepełne poznanie badanego zagadnienia i narażenie badacza na zarzut stronnictwa (nie wspominając już o uzasadnionym zarzucie braku fachowości). W większości badań można wyróżnić kilka, rzadziej kilkanaście problemów (zagadnień) badawczych.

⁴ Metody te omówione są m.in. w następujących opracowaniach: Kaczmarczyk (2002), Babbie (2004), Szreder (2010), Rószkiewicz i in. (2013).

⁵ „The formulation of a problem is often more essential than its solution”, cyt. za: Zikmund (1997: 45).

1.2.2. Zdefiniowanie populacji

Populację lub zbiorowość generalną w badaniu próbkowym określa się jako ogół osób, gospodarstw domowych, podmiotów gospodarczych lub zjawisk, których charakterystyki chcemy poznać, dokonując pomiaru (obserwacji) jedynie części jednostek wybranych do próby w wyniku losowania lub w inny sposób. Cel badania przesądza o tym, jakie jednostki i o jakich charakterystykach zostaną włączone do populacji. Jeżeli badacza interesują ogólne informacje o postrzeganiu przez konsumentów danego towaru lub usługi, to badaniem powinna być objęta liczna zbiorowość konsumentów. Jeżeli zaś celem badania jest uzyskanie opinii o dokonanych ostatnio zmianach w składzie produktu lub świadczeniu danej usługi, to populacja objąć powinna mniejszą grupę zaangażowanych konsumentów, częstych użytkowników tych dóbr lub usług, którzy byliby w stanie takie opinie wyrazić. Do badacza należy decyzja, jakie warunki ograniczające należy nałożyć na ogół konsumentów, by uzyskać właściwą operacyjnie definicję populacji.

Poprawne określenie populacji powinno zawierać trzy cechy, nazywane w statystyce **cechami stałymi**:

- rzeczową – wskazującą, kto lub co jest przedmiotem badania (indywidualni konsumenci, dyrektorzy przedsiębiorstw, głowy gospodarstw domowych, wyborcy itp.);
- czasową – informującą, w jakim momencie lub przedziale czasu interesuje badacza dana populacja (każda populacja ma bowiem dynamiczny charakter, a więc zmienia się w czasie, np. populacja studentów szkół wyższych Trójmiasta w roku 2020 różni od analogicznej populacji z 2010);
- przestrzenną – określającą zakres terytorialny (obszar geograficzny) i miejsca, w których są zlokalizowane jednostki tworzące badaną populację (np. właściciele małych i średnich przedsiębiorstw zarejestrowanych na terenie województwa wielkopolskiego).

Dokładne określenie cech stałych pozwala uniknąć dwuznaczności i błędów w sporządzaniu operatu losowania w trakcie obserwacji lub pomiaru jednostek populacji, a także na etapie interpretacji wyników badania.

1.2.3. Sporządzenie operatu losowania

Operat losowania (*sampling frame*) to kompletny wykaz jednostek badanej populacji, którym przydzielono odpowiednie symbole identyfikacyjne (najczęściej liczby) w celu dokonania wyboru próby. Podstawową rolą operatu losowania jest wyszczególnienie wszystkich jednostek populacji w celu umożliwienia każdej z nich dostania się do próby. Posługując się językiem matematyki, można by określić operat losowania jako zbiór nazw lub nazwisk elementów populacji z przyporządkowanymi im symbolami identyfikacyjnymi, który stanowi pełne i jednoznaczne

odwzorowanie rzeczywistych elementów populacji (osób, gospodarstw domowych, przedsiębiorstw itp.). Każda jednostka populacji musi być obecna w operacie losowania, i to obecna dokładnie raz. Operatem losowania mogą być: listy grup studenckich na uczelni, komputerowa baza danych, wyciąg ze statystycznego rejestru przedsiębiorstw i instytucji REGON, krajowy rejestr sądowy, rejestr terytorialny itp.

W każdym badaniu próbkowym obserwacyjny dostęp do populacji jest możliwy wyłącznie przez operat losowania. Najczęściej badacz nie ma możliwości ani organizacyjnych, ani też finansowych, aby przygotować specjalnie dla potrzeb konkretnego badania listy indywidualnych jednostek, które tworzą populację, lub listy inaczej zdefiniowanych jednostek losowania. Jest to utrudnione, tym bardziej że operat losowania powinien zawierać pewne informacje o jednostkach, umożliwiające z jednej strony pełną ich identyfikację, a z drugiej takie operacje, jak warstwowanie jednostek (w technice losowania warstwowego omówioną w rozdz. 3). W wypadku profesjonalnych badań statystycznych funkcję operatów losowania pełnią zwykle rejestry urzędowe.

Najważniejsze właściwości dobrego operatu losowania można ująć w kilku punktach:

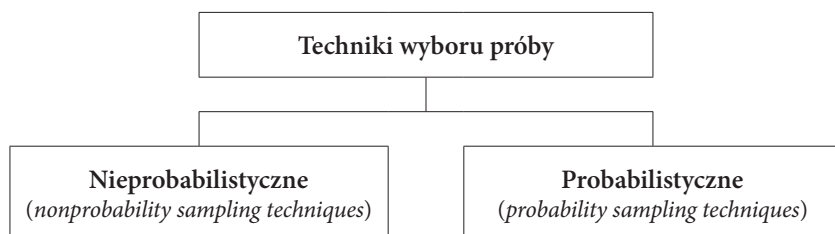
1. Operat musi być kompletny, tzn. powinien dostarczać obserwacyjnego dostępu do wszystkich jednostek w populacji.
2. Operat powinien służyć do uzyskania próby jednostek, które muszą być w jednoznaczny sposób identyfikowalne, tzn. musi zawierać system identyfikacyjny jednostek.
3. Operat powinien być taki, aby można było określić, w jaki sposób jednostki w operacie są powiązane z jednostkami populacji, czyli jaką zastosowano regułę przyporządkowującą poszczególnym jednostkom populacji określone w operacie symbole lub numery.

1.2.4. Wybór techniki pobierania próby

Jedną z najważniejszych decyzji, jakie trzeba podjąć na etapie projektowania badania próbkowego, jest wskazanie techniki wyboru próby badawczej. Najogólniej formułowanym postulatem jest w tym zakresie dążenie do uzyskania **reprezentatywnej** dla danej populacji próby. Termin „próba reprezentatywna” nie przesądza jednak, jaka konkretnie technika wyboru próby zostanie zastosowana w badaniu. **Próba reprezentatywna** oznacza, że jej struktura ze względu na badane cechy (zmienne) jest zbliżona do struktury populacji, z której pochodzi. Stąd słowo „reprezentatywna” może podlegać stopniowaniu, np. bardziej lub mniej reprezentatywna próba, czyli taka, której struktura w większym lub mniejszym stopniu jest zgodna ze strukturą populacji. Reprezentatywną próbę jednostek populacji można otrzymać, stosując zarówno probabilistyczne (losowe), jak i nieprobabilistyczne (nielosowe) techniki wyboru próby. Trzeba jednak dodać, że większą szansę na uzyskanie takiej próby

daje zastosowanie technik wyboru probabilistycznego (losowego). Badanie próbkowe wykorzystujące próbę uzyskaną z populacji w wyniku losowania nazywa się **badaniem reprezentacyjnym**. Termin ten nie rozstrzyga jednak, jaką konkretnie technikę losowania zastosowano do wyboru próby⁶.

W najbardziej podstawowej klasyfikacji technik wyboru próby wyróżnia się techniki probabilistyczne (losowe) i nieprobabilistyczne (nielosowe) (rys. 1.2).



Rysunek 1.2. Klasyfikacja technik wyboru próby

Źródło: opracowanie własne.

Techniki nieprobabilistyczne (nielosowe) to takie techniki wyboru próby, w których nie stosuje się procedur losowania, dających szansę wyboru każdej jednostce populacji, lecz wykorzystuje się inne procedury, oparte przede wszystkim na subiektywnej ocenie badacza, posiadanej przez niego wiedzy wstępnej (*a priori*) lub na wyborze „na chybił trafił”. Ponieważ techniki te nie wykorzystują mechanizmu generowania wartości losowych, nie pozwalają na zastosowanie kategorii związanych z liczbowo wyrażonym prawdopodobieństwem, która to kategoria z definicji dotyczy wyłącznie zdarzeń losowych.

Natomiast **techniki probabilistyczne (losowe)** to procedury oparte na mechanizmach losujących elementy do próby w taki sposób, aby każda jednostka populacji miała szansę dostania się do próby. Losowanie – trzeba podkreślić – to nie to samo co wybór „na chybił trafił”. Ten drugi najczęściej jest obciążony arbitralną decyzją co do czasu i miejsca wyboru jednostek z populacji (np. określonych miejsc w Warszawie w wybrane dni o z góry ustalonej porze w celu wyboru próby mieszkańców stolicy). Takie decyzje powodują, że niektóre jednostki nie mają żadnej możliwości dostania się do próby. W procedurach losowania taka sytuacja nie może zajść (jeżeli oczywiście operat losowania jest kompletny).

Przez **losowy wybór próby** rozumie się taki sposób pobierania próby z badanej populacji, który spełnia dwa następujące warunki:

1. każda jednostka populacji ma dodatnie, znane prawdopodobieństwo dostania się do próby;
2. dla każdego zespołu jednostek populacji można ustalić prawdopodobieństwo tego, że w całości znajdzie się on w próbce⁷.

⁶ Najważniejsze i najczęściej stosowane w praktyce techniki losowania omówiono w rozdziale 3.

⁷ Niektórzy autorzy – np. Zasępa (1972) – dodają w tej definicji jeszcze jeden warunek mówiący o tym, że przed losowaniem nie wiadomo, która jednostka populacji znajdzie się w próbce.

Po pierwsze więc, aby wybór próby można uznać za losowy, każda jednostka populacji musi mieć dodatnie prawdopodobieństwo dostania się do próby. Prawdopodobieństwo – jak wiadomo – jest funkcją przyjmującą wartości z przedziału $[0, 1]$. Postulat ten więc oznacza, że technika wyboru musi zapewniać każdej jednostce niezerowe prawdopodobieństwo znalezienia się w próbie. Druga część pierwszego z postulatów zakłada, że prawdopodobieństwo wylosowania dowolnej jednostki do próby musi być znane. Oznacza to, że badacz musi być w stanie podać (wyliczyć) dla każdej jednostki prawdopodobieństwo tego, że stosując daną technikę wyboru, wylosuje tę jednostkę do próby. Nie można więc uznać za losowy takiego sposobu wyboru jednostek, w którym trzeba by przyznać, że nie ma podstaw do liczbowego określenia prawdopodobieństwa znalezienia się w próbie pewnej części jednostek populacji. Tak jest m.in. w wyborze przypadkowym, gdy ankieterzy w kilku miejscach miasta są w stanie dotrzeć tylko do części populacji i nie ma żadnej możliwości, aby wiarygodnie odpowiedzieć na pytanie, jakie jest prawdopodobieństwo tego, że w próbie znaleźć się może mieszkaniec X tego miasta. Podobnie jest w wypadku radiowej, telewizyjnej lub internetowej sondy, gdzie niemożliwe jest spełnienie wskazanego postulatu. Drugi postulat dotyczy całych prób jednostek populacji. Od technik probabilistycznych wymaga się, aby możliwe było obliczenie prawdopodobieństwa tego, że dowolny zespół jednostek populacji będzie stanowić wylosowaną próbę badawczą. Prawdopodobieństwo to dla niektórych prób może być zerowe.

Podając decyzję o tym, czy do wyboru próby w danym badaniu wykorzystana zostanie któraś z losowych czy też nielosowych technik, bierze się pod uwagę kilka czynników, z których najważniejszy jest cel badania. Jeżeli celem badania jest uzyskanie rozstrzygających wyników (*conclusive results*), np. zeweryfikowanie precyzyjnie określonej hipotezy statystycznej, to preferowane będą probabilistyczne techniki wyboru próby. W sytuacji natomiast gdy celem badania jest jedynie wstępne poznanie danego zagadnienia, a więc badanie ma charakter eksploracyjny (*exploratory research*), wystarczające mogą być mniej pracochłonne i zwykle mniej kosztowne w zastosowaniach techniki nieprobabilistycznego wyboru próby. Decydując się na wybór którejś z technik nieprobabilistycznych, należy być świadomym tego, że co prawda możliwe jest na podstawie takich prób uogólnianie wyników na całą populację, ale nie można tego czynić, stosując metody wnioskowania statystycznego. Wszystkie metody wnioskowania statystycznego – estymacja i weryfikacja hipotez – są oparte na założeniu, że próba będąca źródłem informacji o populacji jest próbą losową. W sytuacjach gdy założenie to nie może być spełnione, trzeba zrezygnować nie tylko ze statystycznych metod określania wartości szacowanych parametrów w populacji, ale także z formuł służących do szacowania precyzji wnioskowania oraz do szacowania wielkości błędów losowych⁸.

⁸ Wielu autorów słusznie podkreśla, że próby nieprobabilistyczne nie upoważniają do stosowania metod wnioskowania statystycznego w celu uogólniania zaobserwowanych prawidłowości w próbie na populację. W.G. Zikmund (1997: 427–428) pisze: „It should be noted that there are

Świadomy tych ograniczeń badacz może się zdecydować na badanie nielosowej próby jednostek populacji. Jednym z powodów może być brak jakiegokolwiek wiarygodnego operatu losowania lub niemożność sporządzenia takiego operatu ze względu na specyfikę populacji. Specyfika ta może się wyrażać tym, że populacja jest niestała w czasie, nie w pełni określona, a nawet quasi-nieskończona. Tak jest na przykład w wypadku populacji osób uczestniczących w masowych imprezach rozrywkowych, klientów dokonujących zakupów w określonym obiekcie handlowym, czytelników jakiegoś czasopisma, słuchaczy stacji radiowej itp. Tego rodzaju praktyczne trudności są najczęściej powodem rezygnacji z probabilistycznych technik wyboru próby. Oczywiście powodem mogą być także ograniczona wielkość budżetu przeznaczanego na badanie lub bardzo krótki czas wyznaczony na realizację badania.

Nie należy jednak postrzegać decyzji o wyborze między losowymi a nielosowymi technikami pobierania próby w kategoriach wyboru między grupą dobrych i złych technik. Niektóre techniki wyboru nieprobabilistycznego pozwalają bowiem uzyskać próby reprezentatywne dla całej populacji ze względu na kilka wyróżnionych, ważnych w badaniu cech (zob. rozdz. 2). Innym powodem stosowania technik nieprobabilistycznych może być bogata wiedza wstępna badacza o populacji będącej celem badania. Dlaczego wiedza o badanej populacji może faworyzować grupę nielosowych technik wyboru próby? Aby dobrze to zrozumieć, warto najpierw rozważyć, czym właściwie jest proces losowania.

Sam akt losowania można rozumieć jako powierzenie ostatecznego składu i kształtu (struktury) próby mechanizmowi czysto losowemu. Mechanizm wyboru jednostek z populacji, w którym każda jednostka populacji ma jednakowe szanse dostania się do próby, ma zapewnić reprezentatywność takiej próby. Ale taki mechanizm wcale nie jest doskonały. Niekiedy losowanie prowadzi do uzyskania mało reprezentatywnych prób tylko dlatego, że taka jest natura losowości, czasami w rzeczywistości realizują się zdarzenia o małym lub bardzo małym prawdopodobieństwie (np. wypadła 4 razy z rzędu 6 w 4 rzutach kostką do gry, albo próba 10-elementowa, która składa się z samych kobiet, mimo że losowano ją z populacji, w której 15% stanowili mężczyźni). Gdyby badacz posiadał szczegółową wiedzę

no appropriate statistical techniques for measuring random sampling error from nonprobability sample. Thus projecting the data beyond the sample is statistically inappropriate” (Należy zauważyć, że nie ma odpowiednich technik statystycznych do pomiaru losowego błędu próbkowania w nieprobabilistycznych technikach doboru próby. Dlatego ekstrapolowanie danych poza próbę jest z punktu widzenia statystyki nieodpowiednie (tłum. własne)). W podobnym duchu Malhotra (1996: 365) stwierdza: „Nonprobability samples do not allow for objective evaluation of the precision of the sample results. Because there is no way of determining the probability of selecting any particular element for inclusion in the sample, the estimates obtained are not statistically projectable to the population” (Nieprobabilistyczne techniki próbkowania nie pozwalają na obiektywną ocenę dokładności wyników w próbie. Ze względu na to, że nie można określić prawdopodobieństwa włączenia danej jednostki populacji do próby, otrzymane oceny próbkowe nie mogą być podstawą uogólnień na całą populację (tłum. własne)).

o strukturze interesującej go populacji, na podstawie tej wiedzy sam wskazałby jednostki, które powinny się znaleźć w próbie, albo przynajmniej ich proporcje, i nie powierzałby tego losowi. W sytuacji gdy badacz nie posiada użytecznej wiedzy o populacji, zdanie się na mechanizm losowania jest tego wyrazem. A takie postępowanie (ostateczne określenie składu próby przez wykorzystanie którejś z technik losowania) trzeba uznać w tych okolicznościach za racjonalne. W praktyce coraz częściej badacz posiada pewną, z reguły ograniczoną, wstępną wiedzę o populacji. Możliwości jej wykorzystania przy wyborze jednostek do próby dają przede wszystkim te techniki (nieprobabilistyczne), w których ciężar odpowiedzialności za ostateczny skład próby nie pozostawia się niezależnemu od badacza mechanizmowi losowania. Wiedza badacza, jeżeli tylko jest wystarczająco bogata i precyzyjna, w niektórych sytuacjach (zwłaszcza w wypadku zróżnicowanej wewnętrznie populacji) może prowadzić do uzyskania prób bardziej reprezentatywnych niż mechanizm losowania. O źródłach tej dodatkowej wiedzy (spoza próby) traktuje następna część tego rozdziału.

1.3. Źródła i rodzaje informacji spoza próby

Informacje służące poprawie dokładności i wiarygodności badania ankietowego są określane mianem informacji dodatkowej, wspomagającej lub pomocniczej (*auxiliary information*). Określenia te odnoszą się zarówno do informacji pochodzących z całkowicie zewnętrznego w stosunku do badania źródła, jak również do informacji uzyskanej z próby lub w jakiś sposób związanej z badaniem, lecz niebędącej bezpośrednim przedmiotem wnioskowania statystycznego.

Źródeł informacji wspomagającej może być wiele – zależą one przede wszystkim od specyfiki danego badania, tj. przedmiotu badania, badanej populacji, celu badania, typu badanych cech i wielu innych, a także od otoczenia, w którym badanie jest realizowane (uwarunkowań prawnych, jakości systemów informacyjnych, rangi podmiotu realizującego badanie itd.). W literaturze statystycznej zazwyczaj ogranicza się do stwierdzenia, że dane określonego typu mogą być dostępne, a następnie przechodzi się do opisu ich pożądanych własności i metod wykorzystania. Wynika to zapewne z różnorodności możliwych źródeł, odmiennych dla każdego typu badania. W tym podrozdziale dokonano ogólnej klasyfikacji źródeł informacji wspomagającej wraz z omówieniem uwarunkowań ich wykorzystania w badaniu statystycznym. Wyróżniono cztery zasadnicze typy źródeł informacji pomocniczej:

- rejestry;
- badania przeszłe;
- wiedza ekspercka;
- inne źródła.

Do grupy **rejestrów** zaliczono wszelkiego rodzaju bazy danych będące rezultatem rejestracji bieżącej dla celów innych niż statystyczne. W większości zastosowań

będą to rejestry o charakterze administracyjnym, tzn. rejestry, które są centralnie regulowane, a ich zapisy mają najczęściej charakter obligatoryjny. Brackstone (1987: 30) wyróżnia sześć głównych typów rejestrów administracyjnych ze względu na cel ich tworzenia – są to rejestry:

- utrzymywane do regulacji przepływu dóbr oraz osób pomiędzy granicami państw (rejestry dotyczące importu, eksportu, imigracji i emigracji);
- wynikające z regulacji prawnych co do rejestracji bieżących zdarzeń, takich jak urodzenia, zgony, małżeństwa, rozwody, zakładanie i zamykanie działalności gospodarczej, udzielanie licencji itp.;
- niezbędne do zarządzania korzyściami lub obowiązkami, takimi jak świadczenia zdrowotne, renty, emerytury, podatki, ubezpieczenia pracownicze itp.;
- potrzebne do administrowania instytucjami publicznymi, takimi jak szkoły, uniwersytety, szpitale, sądy, więzienia itp.;
- wynikające z prawnych regulacji niektórych działów gospodarki, takich jak transport, bankowość, telekomunikacja itp.;
- wynikające z zaopatrywania w wodę, elektryczność, łączność itp.

Poza wymienionymi do rejestrów administracyjnych mogących pełnić funkcję pomocniczą w badaniach ankietowych trzeba również zaliczyć rejestry jednostek podziału terytorialnego kraju (w Polsce jest to rejestr TERYT), a także rejestry wynikające z obowiązku meldunkowego (nie w każdym kraju taki obowiązek występuje), czyli ewidencje ludności. W Polsce istnieje ponad 140 rejestrów administracyjnych prowadzonych przez różne instytucje centralne, jest to więc ogromny potencjał do wykorzystania przy realizacji badań ankietowych (Paradysz 2007: 195).

Poza publicznymi rejestrami administracyjnymi charakter rejestru ma również wiele innych baz danych tworzonych przez prywatne przedsiębiorstwa lub organizacje pozarządowe i one również mogą pełnić funkcję źródła informacji dodatkowej. Można do nich zaliczyć m.in. bazy klientów, dostawców, abonentów, kościelne księgi metrykalne itp.

Szczególnym przykładem masowego wykorzystania rejestrów w badaniu statystycznym jest wspomniane w podrozdziale 1.1 zastąpienie w niektórych krajach spisu powszechnego (*census*) przez badanie próbkowe uzupełniane o dane z rejestrów. Taka sytuacja zaistniała również w Polsce podczas Narodowego Spisu Powszechnego Ludności i Mieszkań w 2011 r., kiedy to posiłowano się danymi z wielu administracyjnych rejestrów centralnych, m.in. zbioru PESEL, Centralnej Ewidencji Pojazdów, Centralnego Wykazu Ubezpieczonych, rejestrów prowadzonych przez urzędy skarbowe, a także z wykazów pozaadministracyjnych, takich jak bazy odbiorców energii elektrycznej oraz abonentów telekomunikacyjnych (GUS 2011).

Badania statystyczne, które w problematyce ekonomicznej i społecznej coraz częściej są badaniami próbkowymi (niewyczerpującymi), wykorzystują nie tylko informacje o cechach jednostek w próbie, ale także wiele innych zasobów informacji związanych z przedmiotem badania. Postępowanie takie z jednej strony jest konieczne ze względu na rosnące znaczenie błędów nielosowych, w tym przede wszystkim tych spowodowanych brakami odpowiedzi. Z drugiej zaś jest to naturalne wykorzystanie możliwości, jakie daje dostęp do coraz większej ilości informacji o wielu populacjach.

W książce tej autorzy uzasadniają potrzebę korzystania we współczesnych badaniach próbkowych z kombinacji różnych źródeł danych statystycznych oraz prezentują sposoby uwzględnienia informacji spoza próby na różnych etapach badania. Szczegółowo przedstawiają źródła i rodzaje danych wykorzystywanych w najbardziej popularnych technikach próbkowania losowego i nielosowego. Omawiają i oceniają także różne metody imputacji wartości brakujących. Ponadto dużo miejsca poświęcają technikom ważenia i kalibracji, których właściwe wykorzystanie może poprawić strukturę próby, a ostatecznie przyczynić się do poprawy jakości wnioskowania. Prezentowane rozwiązania zostały zilustrowane przykładami liczbowymi, które powinny ułatwić Czytelnikom ich zrozumienie.

