ZBIGNIEW SAWIŃSKI

ZASTOSOWANIA TABLIC W BADANIACH ZJAWISK SPOŁECZNYCH

ZASTOSOWANIA TABLIC W BADANIACH ZJAWISK SPOŁECZNYCH

Zbigniew Sawiński

ZASTOSOWANIA TABLIC W BADANIACH ZJAWISK SPOŁECZNYCH

Wydawnictwo IFiS PAN
Warszawa 2010

pamięci tych,
z którymi nie mogłem podzielić się radością tworzenia

Wiesławowi Wiśniewskiemu
Wojciechowi Niepokojczyckiemu

Tablice służą do wyrażania myśli,
nie do przechowywania danych

Howard Wainer

Projekt okładki
Andrzej Łubniewski

Redaktor
Elżbieta Morawska

ISBN 978-83-7683-013-1
ISSN 2081-3201

Spis treści

Wprowadzenie

Tabliczka mnożenia zawiera swoistą magię. W najprostszy możliwy sposób wyjaśnia, jak dwie wielkości przekładają się na wspólny efekt. W książce w gruncie rzeczy nie wykroczymy poza tę ideę.

	1	2	3	4	5	6	7	8	9	10
1	1	2	3	4	5	6	7	8	9	10
2	2	4	6	8	10	12	14	16	18	20
3	3	6	9	12	15	18	21	24	27	30
4	4	8	12	16	20	24	28	32	36	40
5	5	10	15	20	25	30	35	40	45	50
6	6	12	18	24	30	36	42	48	54	60
7	7	14	21	28	35	42	49	56	63	70
8	8	16	24	32	40	48	56	64	72	80
9	9	18	27	36	45	54	63	72	81	90
10	10	20	30	40	50	60	70	80	90	100

W tabliczce mnożenia wynik leżący na przecięciu wiersza i kolumny jest funkcją wielkości przypisanych wierszowi i kolumnie. Funkcją tą jest iloczyn dwóch liczb. Mówiąc inaczej, z każdym wierszem i kolumną związany jest pewien potencjał. W wypadku tabliczki mnożenia potencjałem tym są wielkości liczb, które się przez siebie mnoży. Liczebności we wnętrzu tablicy stanowią wypadkowe tych potencjałów. Im wielkości przypisane danemu wierszowi i kolumnie są większe, tym liczebność we wnętrzu tablicy jest odpowiednio większa.

Ten prosty mechanizm warto mieć na uwadze, gdy pragnie się zrozumieć, dlaczego tak niewiarygodną karierę zrobiły tablice w naukach społecznych. Są one bez wątpienia najczęściej stosowanym narzędziem analizy czy opisu zjawisk i procesów społecznych. Stosowane są na wszystkich etapach wnioskowania o rzeczywistym przebiegu zjawisk: począwszy od wstępnej analizy wyników badania (krzyżowanie cech metryczkowych z odpowiedziami na pytania kwestionariusza), a skończywszy na budowie modeli powiązań między wieloma cechami jednocześnie, gdzie tablice pozwalają zilustrować czy zrozumieć elementarne zależności składające się na docelowy model. W zasadzie trudno jest dociec, dlaczego akurat tablica stała się tak uniwersalną formą analizy i prezentacji danych. Pozostańmy więc przy stwierdzeniu, że widocznie umysł ludzki jest tak skonstruowany, że ułożenie danych w postaci dwuwymiarowej tablicy sprzyja zrozumieniu czy odkryciu rządzących nimi powiązań.

Początek kariery tablic sięga dwóch tysięcy lat przed naszą erą, gdy babilończycy na glinianych tabliczkach przedstawiali zależności w wymyślonym właśnie systemie liczbowym. Dwa wieki przed naszą erą Chińczycy zaczęli stosować tabliczkę mnożenia w formie różniącej się od dzisiejszej jedynie tym, że zawierała liczby uporządkowane nie od 1 do 10, a od 9 do 1. Urzędnik średniowiecznego miasta (ang. *exchequer*) układał na stole (ang. *table*) należności podatników w postaci wyższych lub niższych stosów monet. Część etymologów uważa to za źródłosłów terminu tablica (ang. *table*), tak jak rozumiany jest on w erze nowożytnej. W XVIII wieku rozwinęła się statystyka państwowa, zaś tablice stały się podstawowym narzędziem opisu zasobów państwa. Na przełomie XIX i XX wieku zaczęto analizować **formalne** własności tablic. W 1895 roku brytyjski badacz Karl Pearson (1857–1936) nazwał jeden z tych wzorców korelacją (Rodgers i Nicewander 1988). Pojęcie to należy obecnie do najbardziej fundamentalnych narzędzi opisu i interpretacji zjawisk w wielu dyscyplinach nauki. W pierwszych dekadach XX wieku formalny aparat opisu własności tablic rozwijano dalej. W tym czasie opracowano między innymi metody wnioskowania statystycznego znajdujące zastosowanie w sytuacji, gdy tablica skonstruowana była z danych zebranych w wyniku przebadania próby wylosowanej z pełnej zbiorowości. Okres fundamentalnych odkryć zamykają prace amerykańskiego statystyka Williama Deminga (1900–1993) z lat czterdziestych. Deming uświadomił badaczom, że nie muszą ograniczać się do analizy tablic w takiej postaci, jaką uzyskali w wyniku badania. Modelowanie zawartości tablicy pozwala na dodatkowy wgląd w naturę badanych zjawisk.

Po II wojnie światowej najpierw w Stanach Zjednoczonych, a później na całym świecie, rozpoczął się lawinowy rozwój badań sondażowych. Szybko

zauważono, że tablice stanowią najbardziej dogodny sposób analizy i prezentacji ich wyników. Opracowywane w tym czasie raporty z badań zawierały dziesiątki lub setki tablic, w których krzyżowano ze sobą odpowiedzi na pytania kwestionariusza – starając się na tej podstawie formułować wnioski. Zapotrzebowanie na tablice było w tym czasie tak duże, że wywołało konieczność rozwoju wyspecjalizowanych technologii wspomagających ich przygotowanie. Przykład stanowi urządzenie nazwane sorterem, które pozwalało rozdzielać w sposób mechaniczny karty perforowane zawierające zapis odpowiedzi badanych na pytania kwestionariusza. Karty te segregowane były do osobnych przegródek i jednocześnie zliczane. Tą drogą szybko ustalano liczebności w polach nawet dużych tablic. W Polsce sortery stosowano do sporządzania tablic jeszcze na początku lat osiemdziesiątych.

Przechowywanie wyników badań w postaci tablic krzyżujących „wszystko ze wszystkim" stanowiło przez wiele lat konieczność – zwłaszcza w badaniach akademickich. Na ogół nie dawało się bowiem z góry przewidzieć, które tablice okażą się najbardziej przydatne do interpretacji zjawisk stanowiących przedmiot badania. Dlatego lepiej było sporządzić tablice nawet ze znacznym nadmiarem, niż ryzykować, że zabraknie akurat tych, które okażą się najbardziej przydatne w momencie szlifowania raportu przed ostatecznym terminem jego oddania. W rezultacie tak zwane działy tabulacji nawet średniej wielkości ośrodków badawczych zatrudniały po kilkadziesiąt osób, które wykonywały tylko jedno zadanie – przekładały wyniki badań na tablice. Sytuacji tej nie zmieniło wprowadzenie w latach siedemdziesiątych komputerów sieciowych. Ich drukarki wierszowe nadal produkowały tony papieru stanowiące dla badacza namacalną gwarancję, że nic istotnego nie umknie jego oku.

Opisany model wykorzystania tablic do analizy i prezentacji wyników badań można nazwać **ekstensywnym**. Model ten sprowadza się do tego, że w fazie analizy wyników badania przegląda się dość szeroki zakres tablic, dzieląc je na trzy kategorie: (1) tablice umożliwiające sformułowanie i uzasadnienie kluczowych wniosków; (2) tablice, które pozwalają zilustrować wnioski o mniejszym znaczeniu; (3) tablice, które nic nie wnoszą do interpretacji wyników badania. Tablice zaliczone do pierwszej kategorii włącza się do tekstu opracowania czy raportu, natomiast tablice drugiej kategorii dołączane są do niego w formie aneksu tabelarycznego – na ogół bez żadnego komentarza merytorycznego. Tablice zaliczone do trzeciej kategorii są rzecz jasna wykluczane z dalszego opracowania.

Od początku lat dziewięćdziesiątych narzędziem pracy badacza stał się komputer osobisty, a Internet rozszerzył dostęp do źródeł danych. Spowodowało to dość fundamentalne zmiany w sposobach korzystania z wyników badań. Wyniki wielu badań – zwłaszcza realizowanych ze środków publicznych – stały się

dostępne bez ograniczeń. Wystarczało zalogować się do portalu dystrybutora danych i ściągnąć na własny komputer odpowiednie pliki. Analogiczne zmiany nastąpiły w obszarze badań komercyjnych. Coraz większego znaczenia zaczęły nabierać badania syndykatowe, realizowane dla potrzeb wielu klientów. Możliwość skorzystania z wyników badania pozostała w tym wypadku płatna, lecz procedura korzystania z danych jest analogiczna, jak w wypadku badań finansowanych ze środków publicznych. Użytkownik ma możliwość skorzystania z wyników w dowolnym miejscu i dowolnym czasie – poprzez ściągnięcie danych na własny komputer.

Skutkiem omawianych zmian stało się rozdzielenie ról badacza – odpowiedzialnego za realizację badania – oraz badacza, który analizuje wyniki tego badania. Do lat osiemdziesiątych obowiązywała praktyka, zgodnie z którą warunkiem zaistnienia w naukach społecznych było zrealizowanie stosownego badania. Większość publikacji z tego okresu to prace empiryczne, w których opis procedury badawczej zdominował niekiedy zasadnicze wnioski. Nic w tym dziwnego, biorąc pod uwagę fakt, że badacz musiał być wtedy specjalistą zarówno od prowadzenia badań, jak też od interpretacji ich rezultatów. Nie zawsze połączenie obu ról dawało wartościowe efekty w którejkolwiek z tych sfer. Obecnie role dostarczyciela danych i ich użytkownika są rozdzielone. Realizacją badań zajmują się wyspecjalizowane instytuty, zaś duże projekty mają swoje własne komitety metodologiczne grupujące badaczy wyspecjalizowanych w tej dziedzinie. Tym samym metodologia badań stała się odrębną dyscypliną naukową, wydającą własne czasopisma, posiadającą odrębne organizacje profesjonalne[1] i organizującą własne konferencje i kongresy.

Natomiast w wypadku roli użytkownika danych tego rodzaju specjalizacja nie wystąpiła. Wręcz przeciwnie, rola ta uległa „pauperyzacji" ze względu na fakt, że znacznie uprościł się dostęp do danych. Posiadanie szafy pełnej wydruków zawierających tabele nie stanowi obecnie warunku, aby podjąć się interpretacji wyników badania. Z wyników badań korzystają nie tylko badacze akademiccy, lecz również politycy, urzędnicy, publicyści, czy też studenci przygotowujący prace zaliczeniowe i dyplomowe. Gdy dołączymy do tego badania marketingowe, to grono osób korzystających z wyników badań poszerza się o rzeszę badaczy pracujących w agencjach badania rynku, przygotowujących raporty i formułujących rekomendacje. Do grona użytkowników należą również badacze pracujący po stronie zleceniodawców badań, których zada-

[1] Największa organizacja zrzeszająca metodologów w dziedzinie badań społecznych, czyli European Survey Research Association, została utworzona stosunkowo niedawno, bo dopiero w 2005 roku. Wcześniej metodolodzy tworzyli sekcje w ramach organizacji o profilu ogólnym, na przykład w ramach International Sociological Association.

niem jest przełożenie ich wyników na strategię przedsiębiorstwa. Interpretacja wyników badań stanowi też nieodłączny atrybut pracy wszelkich podmiotów rynkowych wspomagających działania marketingowe, jak agencje reklamowe czy domy mediowe. I wreszcie, rozbudowany obecnie sektor mediów również opiera swoje strategie na interpretacji wyników badań. Bez precyzyjnej i aktualnej wiedzy na temat stylów życia i oczekiwań widzów, słuchaczy czy czytelników nie jest możliwe przyciągnięcie audytorium w wypadku stacji telewizyjnej, internetowego portalu, kolorowego tygodnika czy też dowolnego innego kanału komunikacji marketingowej.

Do wszystkich wymienionych grup użytkowników wyników badań adresowana jest ta książka. Cała ta rzesza osób w swojej codziennej pracy posługuje się tablicami jako narzędziem analizy i prezentacji wyników badań.

* * *

Interpretacja danych zapisanych w formie tabelarycznej wydaje się prosta i naturalna. Jest to umiejętność w znacznym stopniu zakulturowana, czyli należy do zestawu metod poznawczych stosowanych w życiu codziennym. Sprzyja temu fakt, że komunikowanie informacji za pomocą tablic jest jednym z elementów kształcenia szkolnego. Podstawowe umiejętności interpretowania danych w tablicy wymagane są od uczniów na szczeblu podstawowym i gimnazjum, zaś bardziej złożone interpretacje stanowią element testów maturalnych. Trudno byłoby bowiem pogodzić się z faktem, że absolwent szkoły średniej nie jest w stanie poprawnie zinterpretować chociażby danych z rocznika statystycznego. Forma tabelarycznej prezentacji jest coraz częściej stosowana również w mediach, a zwłaszcza w prasie. Wyciąganie poprawnych wniosków z rozkładów procentowych umieszczonych w tabeli nie nastręcza więc większych trudności nie tylko współczesnym specjalistom od analizy wyników badań, lecz również osobom obcującym na co dzień z różnorodnymi formatami komunikatów medialnych.

Powyższe uwagi uznałem za konieczne, aby precyzyjnie zdefiniować przedmiot rozważań przedstawionych w tej książce. Nie jest ona poświęcona czytaniu tabel, lecz ich tworzeniu. Różnica jest zasadnicza – co najmniej taka, jak między autorem książki, a jej czytelnikiem. Książkę może przeczytać każdy, interpretując na swój sposób jej treść i lepiej czy gorzej rozumiejąc to, co autor książki chciał czytelnikom przekazać. Sztuką jest natomiast książkę napisać w taki sposób, aby przesłanie autora stało się klarowne dla odbiorcy.

Przedmiotem tej książki jest więc sztuka prezentowania danych za pomocą tablic. Aby komunikaty formułowane w tej postaci pozwalały zrealizować zakładane cele, wymagane jest spełnienie dwóch warunków.

Po pierwsze, twórca tablicy musi dysponować należytą wiedzą na temat prezentowanych zjawisk. Dane, na podstawie których konstruowana jest tablica, pochodzą z badań. Wynik badania stanowi zaś przybliżony, a niekiedy nawet wypaczony obraz rzeczywistości empirycznej. Dlatego tworzenie tablic wymaga świadomości relacji, jakie zachodzą między rzeczywistym zjawiskiem, istniejącym na poziomie empirycznym, a jego odbiciem w wynikach badania.

Po drugie, twórca tablicy musi mieć klarowną wizję tego, co przede wszystkim chciałby zakomunikować odbiorcy. Interpretację dowolnego zjawiska można rozwijać i komplikować w nieograniczony sposób. Naiwnością byłoby oczekiwanie, że odbiorca skoncentruje się na problemie w tym samym stopniu, w jakim uczynił to badacz. Dlatego konieczna jest selekcja przekazywanych treści i wybór takiego ujęcia problemu, które trafnie a zarazem najprościej wyrazi punkt widzenia badacza.

Przedstawienie skutecznych form prezentacji danych z uwzględnieniem dwóch powyższych warunków jest głównym celem książki. Jej zawartość i układ mają dwa źródła. Pierwszym są propozycje statystyków i badaczy dotyczące różnorodnych metod i podejść do analizy wyników badań przedstawianych w postaci tabelarycznej. Uwzględnienie tej wiedzy nie wyróżnia niczym tej książki od innych. Specyficzne wydaje się natomiast drugie źródło inspiracji co do sposobu ujęcia poszczególnych zagadnień. Składają się na nie moje osobiste doświadczenia w komunikowaniu wyników badań różnym audytoriom. W swojej karierze – tak jak wielu badaczy – pisałem artykuły i książki, wygłaszałem referaty na konferencjach, prowadziłem zajęcia ze studentami. Ponadto, przez wiele lat miałem okazję bezpośrednio współpracować z ośrodkami badawczymi – zarówno prowadzącymi badania akademickie, jak i komercyjne. W znacznej części tych przedsięwzięć miałem bezpośredni kontakt z odbiorcą rezultatów badania. Spośród doświadczeń, które zebrałem, dwie kwestie szczególnie zaważyły na zawartości tej książki.

Po pierwsze, za sformułowanie wniosków z badania odpowiedzialny jest zawsze i wyłącznie badacz, który tworzy komunikat. Wyjaśnianie adresatowi wyników badań, że poznanie naukowe jest względne, że badanie objęło jedynie próbę osób, a nie całą populację, że odpowiedzi respondentów mogą być obarczone błędem i tak dalej – rodzi wyłącznie przekonanie, że do prezentowanych wyników nie można mieć zaufania. Na odbiorcę komunikatu w żaden sposób nie uda się przerzucić odpowiedzialności związanej z faktem, że metody w naukach społecznych obarczone są ryzykiem dojścia do błędnych konkluzji. Odpowiedzialność tę musi w całości wziąć na siebie badacz i samodzielnie podjąć decyzję, czy wniosek budzi zaufanie, czy nie. Uwzględniając zarówno ów margines niepewności stosowanych metod badawczych, jak też

swoją dotychczasową wiedzę na temat badanej rzeczywistości – pochodzącą z teorii, z innych badań czy wręcz z codziennych doświadczeń. Jeśli badacz zdecyduje się zakomunikować pewien rezultat, to sam musi mieć do niego pełne zaufanie.

Po drugie, kluczowe wyniki badania muszą być zaprezentowane w sposób syntetyczny. Nawet doświadczeni badacze wpadają niekiedy w pułapkę komunikowania „bogactwa" osiągniętych rezultatów. Zresztą sprzyjają temu pewne okoliczności. Analizując wyniki badania, badacz nabiera do niego pozytywnego stosunku emocjonalnego, przywiązuje się do swojego dzieła. Jest w stanie później godzinami opowiadać o wielu szczegółowych kwestiach, które okazały się inspirujące i z pewnością ciekawe. Cóż z tego, jeśli w tym wszystkim gubi się właściwe cele, dla których badanie zostało przeprowadzone. Odbiorca nie jest w stanie dokonać równie głębokiego wglądu w wyniki badania, jak zrobił to badacz, pracując nad ich interpretacją. Dlatego kluczowe wnioski, które odzwierciedlają podstawowy powód, dla którego w ogóle zdecydowano się przeprowadzić badanie, muszą zostać w szczególny sposób uwypuklone. Wszelkie wątki poboczne – nawet jeśli są wśród nich niespodziewane odkrycia, zmieniające punkt widzenia na wiele spraw – powinny zaś zostać potraktowane jako wartość dodana. Miarą sukcesu badania nigdy nie jest liczba dokonanych ustaleń i zweryfikowanych hipotez, lecz to, na ile skutecznie zakomunikowany został jego kluczowy rezultat.

Aby zrealizować powyższe cele, niezbędna była selekcja metod, które w książce omawiam. Metody analizy tablic, podobnie zresztą jak większość metod analitycznych, podzielić można na dwie kategorie: konfirmacyjne i eksploracyjne.

Metody **konfirmacyjne** służą budowaniu wiedzy w ramach Popperowskiego paradygmatu rozwoju nauki. Mówiąc w pewnym uproszczeniu, na podstawie wspomaganej teorią wiedzy dostępnej przed badaniem formułuje się hipotezy dotyczące prawidłowości na poziomie empirycznym, po czym przeprowadza się badanie weryfikujące te hipotezy. Jeśli wyniki badania nie upoważniają do odrzucenia testowanych hipotez, to są one włączane do zasobów wiedzy. Natomiast odrzucenie hipotezy wymaga modyfikacji lub poszerzenia zasobów dotychczasowej wiedzy, gdyż w świetle wyników badania nie wyjaśnia ona rzeczywistych zjawisk.

Zupełnie inne są cele stosowania metod **eksploracyjnych**. Metody te mają umożliwić wgląd w naturę zjawisk, pozwolić zidentyfikować mechanizmy, które w największym stopniu kształtują istotę danego zjawiska. W podejściu tym jest wiele arbitralności, intuicji czy szukania po omacku. Mniejsza jest kumulacja rezultatów czy systematyczność w budowaniu wiedzy. Korzyści wynikają ze świeżości spojrzenia, z braku skrępowania związanego z wcześ-

niejszymi ustaleniami dotyczącymi badanego problemu. W wielu wypadkach przekłada się to na lepszą skuteczność w realizacji celów, które były powodem przeprowadzenia badania.

Współczesne badania prowadzone są na niespotykaną dotychczas skalę. Chociażby każde wybory parlamentarne czy prezydenckie są szczelnie obudowane badaniami przed, w trakcie i po. Dzięki postępowi w technikach gromadzenia danych cykl badawczy skrócił się, obniżyły się także koszty jednostkowe. W Polsce działa obecnie kilkadziesiąt firm oferujących wykonanie badań o zasięgu ogólnopolskim. Rocznie realizuje się dziesiątki tysięcy badań, w których w sumie uczestniczy ponad 5 milionów osób.

W czasach, gdy liczba komunikatów medialnych dubluje się co dwa lata, możliwość przebicia się przez szum informacyjny ma jedynie **odkrycie**. Stąd też zawrotna kariera pojęcia, jakim jest *insight* badawczy[2]. *Insight* to rezultat dotyczący ukrytej zasady kształtującej ludzkie zachowania, której nie zauważono lub należycie nie opisano we wcześniejszych badaniach. Jest to nowe, odkrywcze spojrzenie na zjawisko w sytuacji, gdy konwencjonalne badania nie były w stanie dostarczyć przekonujących wyjaśnień, dlaczego dzieje się tak, a nie inaczej. Współczesnym celem badań jest poszukiwanie *insightów*. Nowo powstające agencje badawcze w swojej nazwie lub w swojej misji częściej umieszczają słowo „insight" niż tradycyjne „research".

W czasach – gdy tak duży nacisk kładzie się na sukces – uzyskanie klarownych, przekonujących, wyróżniających się czy wręcz spektakularnych wyników nie jest wyłącznie wewnętrzną potrzebą badacza, lecz stanowi warunek konieczny pozytywnej oceny przez środowisko, pracodawcę czy ubiegania się o kolejne granty. Jest to jednym z powodów, dla których badacze wolą stosować metody eksploracyjne niż konfirmacyjne. Metody eksploracyjne stwarzają możliwość dokonania odkrycia, znalezienia *insightu*. Cech tych nie mają metody konfirmacyjne, gdyż z definicji służą weryfikacji hipotez w ramach dotychczasowej wiedzy.

W książce ograniczę się do przedstawienia metod analizy tablic, które służą eksploracji danych. Tym różni się ona od innych prac na ten temat. Książka została poświęcona **praktyce** analizy wyników badań. Nie teorii, nie modelom, lecz praktycznym problemom, jakie napotykają badacze, analizując zawartość tablic. W jaki sposób wydobywają z nich to, co przede wszystkim chcą zakomunikować? Czy mogliby kluczowe wnioski uzyskać w bardziej efektywny

[2] Już w 1946 roku Robert Merton pisał, że jednym z konstytutywnych elementów badań empirycznych jest możliwość uzyskania nieoczekiwanego i nietypowego wyniku. Możliwość ta ma strategiczne znaczenie dla rozwoju teorii wyjaśniających zjawiska społeczne (Merton 2002: 170–174).

sposób? Prezentacja statystycznych teorii czy koncepcji analizy tablic, nawet tych szeroko stosowanych, okrojona zostanie do takich rozmiarów, jakie wystarczą do realizacji zakładanych celów. Na ten temat dostępnych jest wiele wyspecjalizowanych opracowań, do których odsyłał będę Czytelników. W zamian, większą uwagę zwrócę na prace może rzadziej cytowane, lecz zawierające oryginalne idee, pomysły i rozwiązania ułatwiające posługiwanie się tablicami do analizy i prezentacji wyników badań.

Gdy wiemy już, czemu poświęcona została książka i jaki był kontekst jej powstania, spróbujmy w sposób zwięzły określić jej cel. Chciałbym przekonać Czytelników, że tablice stanowią skuteczne narzędzie identyfikacji prawidłowości występujących w wynikach badań. Prawidłowości, które odzwierciedlają logikę badanych zjawisk społecznych. Rozważania zilustruję przykładowymi wynikami badań, które pozwolą zaprezentować metody identyfikacji występujących w tablicach prawidłowości, a także pokazać związki, jakie zachodzą między tymi metodami. Zamierzeniem moim jest również wykazanie, że zebrane w tej książce metody tworzą spójny system, którego elementy obecne są w wielu metodach analizy tablic: zarówno tych najprostszych, jak analiza odsetków, jak też bardziej złożonych. Kluczowa korzyść przedstawionego ujęcia sprowadza się więc do tego, że proponuje ono w jednym zestawie narzędzia o różnym stopniu czułości, które pozwalają wniknąć płycej lub głębiej w analizowane zjawisko, w zależności od potrzeb badacza i celu analizy.

Odrębną kwestią są granice proponowanego podejścia. Zasadniczo obejmuje ono jedynie te badane sytuacje, które dadzą się przedstawić za pomocą konwencjonalnej tablicy, w której skrzyżowane zostały ze sobą dwie cechy. Klasyczny przykład stanowią tablice – występujące w większości raportów badawczych – w których krzyżuje się cechy metryczkowe z odpowiedziami na pytania ankiety. Mimo że proponowane metody omówione zostaną na przykładach tablic obejmujących jedynie dwie cechy, metody te posiadają dość naturalne uogólnienia na wypadek, gdy analizie za pomocą tablicy podlega jednocześnie więcej cech. Możliwości te zasygnalizuję, przedstawiając stosowne przykłady, aczkolwiek nie one stanowią *leitmotiv* książki. Aby dogłębnie wyjaśnić istotę proponowanych metod, rozważania zogniskuję na opisie wyników badań za pomocą konwencjonalnych tablic, obejmujących jedynie dwie cechy.

Układ książki odpowiada jej celom i zakresowi prezentowanych zagadnień. Książka składa się z siedmiu rozdziałów, które umownie podzielić można na dwie części.

Część pierwsza – rozdziały 1 i 2 – stanowi wprowadzenie w problematykę stosowania tablic do analizy i prezentacji zjawisk społecznych. Rozdział 1 dotyczy sposobu reprezentacji rzeczywistości empirycznej za pomocą wyni-

ków badania. Scharakteryzuję najważniejsze pojęcia, które są użyteczne w tym kontekście, a także omówię wybrane problemy związane z interpretacją wyników badań. Rozdział 2 prezentuje „krok po kroku" tworzenie tablicy opisującej związek dwóch cech. Uwagę skoncentruję na środkach pozwalających uwypuklić wybrane aspekty badanego zjawiska, którymi to środkami dysponuje autor tablicy.

Druga część książki – obejmująca rozdziały od 3 do 7 – przedstawia proponowaną metodologię analizy tablic. Tematykę kolejnych rozdziałów starałem się dobrać w taki sposób, aby odpowiadały kolejnym etapom pracy badacza z tablicą zawierającą wyniki badania. Zgodnie z celem książki proponowane ujęcie ma bowiem tworzyć kompletny system, który powinien zaspokoić potrzeby pojawiające się podczas analizy i prezentacji danych za pomocą tablic. Omawiane narzędzia analityczne należą do grupy metod znanych pod nazwą analizy kanonicznej bądź analizy korespondencji. Wybór i ograniczenie zawartości książki akurat do tego podejścia stanowiły świadomą decyzję, podyktowaną przekonaniem, że jest ono dostatecznie elastyczne, aby nie ograniczać wniosków, a zarazem inspiruje do poszukiwania twórczych interpretacji.

Do książki dołączone zostały dwa aneksy. W aneksie A umieściłem matematyczne dowody wybranych własności omawianych metod. Ograniczyłem się jedynie do tych własności, których dowody trudno znaleźć w literaturze. Nie chciałbym bowiem pozostawić Czytelnika w niepewności, że niektóre niestandardowe rozwiązania, o których mówię, mogą być błędne. Drugi z aneksów (aneks B) poświęcony został prezentacji oprogramowania pozwalającego wykorzystać omawiane w książce metody do analizy danych. Kwestie wykonania obliczeń zdecydowałem się umieścić w osobnej części i nie przeplatać ich rozważaniami merytorycznymi. Ułatwia to zogniskowanie uwagi na wyjaśnieniu istoty proponowanych metod.

Chcę jeszcze nadmienić, że prezentacje poszczególnych metod analizy tablic przemieszane są z rozważaniami zahaczającymi o pragmatyczne czy społeczne aspekty naukowego poznania. Z jednej strony chodzi o lepsze osadzenie tego, co robi badacz, w badanej rzeczywistości. Z drugiej zaś – o pokazanie, że badacz nie działa w próżni. Podlega naciskom ze strony innych badaczy, musi uwzględniać branżowe normy postępowania, ograniczają go narzędzia analityczne, które ma w swoim komputerze, dostępna literatura, potrzeby i oczekiwania zleceniodawców badań, a także szereg innych czynników, których nie sposób wymienić. Niekiedy czynniki te decydują o wyborze metod analizy danych w większym stopniu niż treść badanego problemu. Aby jednak sobie to uświadomić, nie należy unikać poruszania tej kwestii.

W nauce poszczególne podejścia związane są na ogół z konkretnymi osobami, które rozwijają je w przekonaniu, że są bardziej użyteczne od innych.

Aby przybliżyć Czytelnikowi twórców metod, o których piszę, w treść książki wplotłem biogramy osób potrafiących przekonać świat badawczy do swoich wizji i swojego spojrzenia na problem analizy danych pochodzących z badań.

Nie będę też ukrywać, że o ostatecznym wyborze metod prezentowanych w książce zadecydowało po prostu to, że do metod tych jestem przywiązany, przez lata przyzwyczaiłem się do nich i osobiście uważam za przydatne i skuteczne. Czy Czytelnik również uzna je za swoje? No cóż, tego nie jestem w stanie zagwarantować.

* * *

Chciałbym podziękować osobom, które zgodziły się przedyskutować projekt książki w fazie jej powstawania, wzbogacając ją o szereg idei i pomysłów, które skrupulatnie wykorzystałem: Tadeuszowi Krauze, Danielowi Krymkowskiemu i Kazimierzowi Maciejowi Słomczyńskiemu. Dziękuję również żonie Monice za inspirujące uwagi do wcześniejszych wersji poszczególnych fragmentów książki.

Szczególne podziękowania winien jestem Henrykowi Domańskiemu. Przez wiele lat miałem możliwość wspólnie z nim eksplorować różnorodne problemy badawcze, stosując między innymi opisane w książce metody. Wnikliwy umysł Henryka nie raz kierował moją uwagę na konieczność krytycznej oceny wyników uzyskiwanych za pomocą tych metod. Nauczyło mnie to pokory wobec faktów empirycznych, a zarazem dystansu wobec narzędzi ich analizy. Dla badacza jest to bezcenny kapitał. Tak się ponadto złożyło, że Wydawnictwo IFiS PAN powierzyło Henrykowi niełatwe zadanie napisania recenzji z pierwszej wersji tej książki. Recenzja uświadomiła mi, że aby przekonać kogokolwiek do proponowanych rozwiązań, trzeba w spójny i klarowny sposób wyjaśnić związane z nimi korzyści. Nie będę ukrywał, że pod wpływem recenzji Henryka dokonałem zmian w układzie i zawartości książki. Sądzę, że w obecnej wersji lepiej zrealizuje ona swoje cele.

Lista podziękowań powinna być dłuższa. Przedstawione ujęcie problematyki tablic zawdzięczam bowiem w dużej mierze osobom stanowiącym audytoria prezentacji wyników badań, zwłaszcza w okresie, gdy pracowałem w agencjach komercyjnych. To ich reakcje pozwoliły mi zrozumieć, że skuteczne komunikowanie wyników badań nie sprowadza się do perfekcyjnego zastosowania określonej metody. Jednakże nawet gdybym bardzo się postarał, to nie potrafiłbym wymienić wszystkich osób, które pomogły mi zrozumieć tę prostą prawdę. Jeśli wezmą do ręki książkę, to przekonają się, że jestem im za to wdzięczny.

Książkę dedykuję dwóm osobom, którym niestety nie mogę podziękować osobiście. Wiesławowi Wiśniewskiemu (1921–1991) zawdzięczam to, że ba-

dania stały się moją pasją. W czasach, gdy życie naukowe przesiąknięte było ideologią, potrafił przekonać współpracowników, że w nauce nie ma ucieczki od prawdy, zaś prowadzenie badań **rzetelnie** dokumentujących społeczną świadomość jest powinnością każdego socjologa. Nie tylko zresztą prowadzenie. Wiesław stanowił wzór badacza, gdy chodzi o kulturę analizy danych. Dociekliwość nie pozwalała mu na prezentację jakiegokolwiek wyniku, co do którego nie miał pewności, że nie jest zależnością pozorną (Sawiński i Zahorska 1991).

Wojciech Niepokojczycki (1955–1994) zrobił wiele dla implementacji metod, które pomogły realizować nakreślone przez Wiesława cele. Pamiętam, jak zimą 1982 roku godzinami wystawaliśmy z Wojciechem pod fabryką imienia Ludwika Waryńskiego, aby dostać się do jedynego dostępnego w Warszawie komputera i zapuścić na nim „joby" perforowanych kart. Żołnierze, grzejący się przy koksowniku obok stojącego nieopodal skota, w sumie nam współczuli. Do ośrodka obliczeniowego prowadził nas zawsze strażnik, gdyż wiązało się to z przejściem przez oddziały produkcyjne. Produkcja na ogół stała, gdyż brakowało surowców. Szare twarze robotników, apatia, poczucie bezsensu. Tak wyglądało społeczeństwo, którego wtedy zabroniono badać. Tak wyglądały czasy, w których udało się nam wspólnie stworzyć pierwsze programy do obliczania korelacji kanonicznych.

ROZDZIAŁ 1

Przejrzystość znaku wobec znaczenia

Celem rozdziału jest omówienie pojęć pozwalających opisać relacje między zjawiskami w realnym świecie społecznym, a ich reprezentacją w postaci danych zgromadzonych w wyniku badania.

Zakres problemów omawianych w tym rozdziale jest dość szeroki i zróżnicowany. Może nawet powstać wrażenie, że jest to zbiór swobodnych skojarzeń. O ile jednak same metody analizy tablic od lat stanowią temat dedykowanych im opracowań, o tyle to, co je otacza, nie stało się przedmiotem systematycznego wykładu. Widocznie punkt styku między metodami analizy danych a żywą empirią – opornie poddaje się refleksji. W rozdziale zebrałem więc to wszystko, co mnie – osobiście – wydaje się w tym kontekście najważniejsze. Czyli kwestie, które mnie również wielokrotnie sprawiały kłopoty. Albowiem żadna, nawet najbardziej dogłębna i finezyjna metoda analizy danych nie pomoże, gdy badacz nie wie, jak mają się dane do badanych zjawisk.

Pierwsze cztery podrozdziały dotyczą rozumienia zjawisk na poziomie empirycznym. W podrozdziale 1.1 przedstawiam sposób wyodrębniania właściwości badanych obiektów, który nazwałem kwantyfikacją. W podrozdziale 1.2 opisuję dwa podstawowe komponenty każdego zjawiska, jakimi są jego mechanizm i zasięg. Identyfikacja mechanizmów i zasięgu zjawisk wiąże się z dwoma podejściami badawczymi, zwanymi ilościowym i jakościowym. W skrócie przedstawiam, jak historycznie kształtowały się oba podejścia oraz jak rozumie się identyfikację mechanizmów zjawisk w badaniach ilościowych. Terminy „ilościowy" i „jakościowy" występują również w podrozdziale 1.3, lecz tym razem w kontekście właściwości badanych obiektów. Wokół kwestii stosowania obu terminów w tym kontekście narosło chyba najwięcej nieporozumień, gdyż nakłada się na nie podział właściwości na ciągłe i skategoryzowane. Odwołując się do przykładów staram się kwestie te uporządkować. Podrozdział 1.4 poświęcony został przestrzennym właściwościom zjawisk. Ich istota jest stosunkowo rzadko omawiana, mimo że są powszechnie stosowane w badaniach. Przykładem jest podział na województwa, które wyodrębniane są w różny sposób, w zależności od przedmiotu badania.

Dalsze podrozdziały poświęcone zostały omówieniu specyficznych problemów związanych z reprezentacją zjawisk za pomocą wyników badań. Część tę otwiera podrozdział 1.5. Omawiam w nim paradoks Simpsona ułatwiający zrozumienie, jak niewłaściwe przełożenie badanego zjawiska na kategorie opisu prowadzić może do fałszywych wniosków. W podrozdziale 1.6 przedstawiam podejście do opisu własności danych oparte na koncepcji skal pomiaru: nominalnej, porządkowej, interwałowej i ilorazowej. Chociaż podejście to traci współcześnie na znaczeniu, w wielu wypadkach nadal decyduje o sposobie wnioskowania na temat badanej rzeczywistości. Podrozdział 1.7 poświęcony został interpretacjom wyników badań w kategoriach probabilistycznych. Ujęcie takie stanowi podstawę wielu metod analizy danych, mimo że w niektórych sytuacjach zamiast upraszczać – komplikuje możliwość wglądu w badane zjawisko. W podrozdziale 1.8 omawiam specyfikę danych uzyskiwanych przy pomocy tak zwanych pytań wieloodpowiedziowych. Kwestii tej poświęca się niewiele uwagi, choć postać danych wieloodpowiedziowych w niektórych badaniach wypiera tradycyjne formaty danych. Kolejny podrozdział (1.9) nawiązuje do problemu wirtualnej reprezentacji danych w formie komputerowych plików. Przedstawię zagrożenia związane z nadinterpretacją danych, czyli przypisaniu obiektom właściwości, których w rzeczywistości nie posiadają.

Ostatni podrozdział (1.10) ma cel odmienny od poprzednich. Określa mianowicie, co rozumieć będziemy jako tablicę w dalszych partiach książki. Zawiera też dyskusję zakresu zastosowań tablic krzyżujących ze sobą dwie cechy oraz relację tego narzędzia do metod analizy danych uwzględniających większą liczbę czynników.

1.1 Właściwości obiektów a cechy

Dziedziną zainteresowań badaczy w naukach społecznych jest pewien zbiór obiektów, to jest jednostek podlegających badaniu bądź analizie. Na ogół obiektami są osoby składające się na pewną zbiorowość, na przykład dorosła ludność Polski, mieszkańcy Warszawy, studenci szkół wyższych. Obiektami mogą też być agregaty osób, na przykład gospodarstwa domowe, pary małżeńskie, czy kategorie zawodowe. Niekiedy obiekty wyodrębnia się na szczeblu instytucji. Tego rodzaju zbiory obiektów stanowią gimnazja czy zakłady pracy. Obiekty można również wyodrębnić w sferze symbolicznej. Przykładami są marki produktów, role zawodowe, czy cenione wartości. Wyodrębnione obiekty charakteryzują się właściwościami, zwanymi też atrybutami. Obiekty daną właściwość mogą posiadać bądź nie. Mogą też różnić się stopniem, czy natężeniem jej posiadania. W proponowanym ujęciu właściwości przynależą

obiektom istniejącym na poziomie rzeczywistości empirycznej, przez co same właściwości są również konstruktami z poziomu empirycznego. Wiedzę na temat właściwości obiektów uzyskuje się drogą badania – niekiedy nazywanego też pomiarem.

Aby pewna właściwość obiektów stała się przedmiotem badania, konieczna jest kwantyfikacja zakresu jej zmienności. **Kwantyfikacja** stanowi pewien sposób interpretacji rzeczywistości przez badacza. Jest zestawem założeń dotyczących tego, które z aspektów danej właściwości należy uznać za istotne dla wyjaśnienia badanego zjawiska, a które niewiele wniosą do jego rozumienia, przez co można je bez szkody pominąć. Jedną z właściwości gospodarstw domowych jest osiągany dochód. Można wyodrębnić go jako właściwość zdychotomizowaną, dzieląc gospodarstwa na takie, które nie uzyskują żadnego dochodu, oraz na takie, które uzyskują dochód. Można zdychotomizować tę właściwość w inny sposób, na przykład dzieląc gospodarstwa na uzyskujące dochód poniżej minimum socjalnego oraz na gospodarstwa uzyskujące dochód powyżej tego minimum. Marketerzy w bankach wyodrębniają potencjalnych klientów według pewnej granicy dochodu, którą uznają za wystarczającą i konieczną, aby zaoferować im konkretny produkt. Klientów banku podzielić również można na kilka grup dochodowych, adresując do każdej z nich inny produkt, dostosowany do potrzeb i możliwości. Są to wszystko przykłady podziału zakresu zmienności danej właściwości na dwie lub więcej rozłącznych kategorii.

W wypadku innych zagadnień podział dochodu na kategorie nie jest potrzebny bądź byłby nieadekwatny. Na przykład, badając zależność postaw konserwatywnych od dochodu, należałoby traktować dochód w wymiarze dwubiegunowej osi, jako wyższy lub niższy. W tej sytuacji podział dochodu na dwie lub więcej kategorii byłby ryzykowny, ponieważ z góry nie wiadomo, w jaki sposób postawy konserwatywne zależą od dochodu i w których fragmentach zróżnicowania dochodowego szukać barier decydujących o tych postawach.

Kwantyfikacja dokonywana jest wobec każdej właściwości obiektów, która staje się przedmiotem badania. Przy czym w różnych analizach może być dokonana w różny sposób. Aby odróżnić właściwość skwantyfikowaną, stanowiącą subiektywne ujęcie rzeczywistości przez badacza, od właściwości istniejącej na poziomie rzeczywistości społecznej, to jest niezależnie od faktu jej badania, przyjmiemy ustalenie, że właściwość skwantyfikowaną nazywać będziemy **cechą**. Cecha odpowiada przyjętemu modelowi istniejącej realnie właściwości obiektów, odzwierciedla aspekty, które badacz uznał za ważne. Tym samym wyznacza sposób operacjonalizacji danej właściwości w badaniu, czyli przełożenia jej na pytania kwestionariusza lub innego typu narzędzie, które posłuży gromadzeniu danych na jej temat.

Użyteczność kwantyfikacji jako osobnego konstruktu sprowadza się do tego, że pozwala każdorazowo sformułować pytanie, czy przyjęta kwantyfikacja jest zasadna. Wynik badania może nie potwierdzić hipotezy nie tylko dlatego, że rzeczywistość kształtuje się w sposób rozbieżny z wcześniejszą wiedzą badacza, lecz również dlatego, że badane właściwości zostały skwantyfikowane w sposób nieadekwatny wobec badanego problemu.

1.2 Mechanizm zjawiska a jego zasięg. Badania jakościowe i ilościowe

Aby przedstawić relacje między mechanizmem zjawiska a jego zasięgiem, rozważmy fikcyjny przykład.

Przyjmijmy, że część ludzi w pewnym wieku zapada na pewne schorzenie, na które dotychczas nie znaleziono środka. Schorzenie to nie stanowi zagrożenia dla życia, lecz jest uciążliwe na co dzień. Grupa lekarzy odkryła preparat, który stwarzał szanse całkowitego wyleczenia. Wniosek taki sformułowano na podstawie testu klinicznego, który objął dwie grupy pacjentów cierpiących na to schorzenie. Obie grupy liczyły po 100 osób. Jednej grupie przez miesiąc podawano nowoodkryty preparat, po czym stwierdzono, że u 70 z tych osób schorzenie całkowicie ustąpiło. Drugiej grupie, pełniącej funkcje grupy kontrolnej, podawano w tym czasie placebo. W grupie tej schorzenie ustąpiło u 20 osób. Wynik testu stanowił wystarczającą podstawę do opatentowania preparatu. Mając w ręku świeży jeszcze patent grupa lekarzy umówiła się z dyrektorem R&D jednego z koncernów farmaceutycznych, aby pozyskać producenta dla nowo odkrytego preparatu. Dyrektor, po wysłuchaniu entuzjastycznej prezentacji na temat przełomowego charakteru odkrycia oraz korzyści, jakie może ono przynieść ludziom gnębionym schorzeniem, zadał pytanie: „Panowie, a jaki odsetek populacji cierpi na to schorzenie?". „No, dokładnie nie wiadomo, gdyż nie istnieje centralna ewidencja chorych – ale z całą pewnością odsetek ten jest znaczny". Dyrektor zadał w związku z tym drugie pytanie: „a jeśli ktoś zapadł już na to schorzenie, to na ile byłby skłonny leczyć się za pomocą nowego preparatu?". „Tego też nie wiemy, ale do naszej przyszpitalnej przychodni wciąż zgłaszają się pacjenci z tym schorzeniem, prosząc o pomoc". Po uzyskaniu tych informacji dyrektor podziękował grupie lekarzy za interesującą prezentację preparatu, który stwarza tak fascynujące możliwości, obiecując, że w ciągu 3 miesięcy, po przeanalizowaniu wszystkich okoliczności, koncern udzieli im odpowiedzi.

Czytelnicy obeznani z marketingową retoryką z łatwością odczytają to jako kompletne *disintéressement*, zapewne będące konsekwencją niepewności całego

przedsięwzięcia. Nie jest jednak wykluczone, że przed podjęciem negatywnej decyzji koncern farmaceutyczny zdecyduje się zrealizować stosowne badania, które pozwoliłyby określić stopień opłacalności produkcji i sprzedaży preparatu. Badanie takie powinno rozstrzygnąć dwie kwestie, o które pytał dyrektor R&D. Po pierwsze, jak kształtuje się rozkład cechy „ma schorzenie–nie ma schorzenia" w populacji osób, w której schorzenie to występuje. Chodzi o oszacowanie wielkości rynku, na którym mógłby zostać zaoferowany nowy preparat. Po drugie, jak wygląda rozkład cechy „jest zainteresowany zakupem preparatu–nie jest zainteresowany" wśród osób, które cierpią z powodu schorzenia.

Grupa lekarzy oferowała informacje dotyczące wyłącznie jednego aspektu zjawiska, które to zjawisko można określić jako leczenie schorzenia za pomocą wynalezionego preparatu. Aspektem tym jest **mechanizm** zjawiska, nazywany też jego istotą. W przedstawionym przykładzie mechanizm zjawiska sprowadza się do tego, że podawanie preparatu powoduje wyleczenie określonego odsetka osób, które poddały się kuracji. Na prezentacji w koncernie farmaceutycznym okazało się jednak, że znajomość mechanizmu zjawiska to nie wszystko. Drugim aspektem każdego zjawiska są bowiem jego rozmiary, które będziemy nazywać **zasięgiem** zjawiska. Bez znajomości zasięgu koncern nie był w stanie podjąć decyzji dotyczącej ewentualnego wejścia w proponowane przedsięwzięcie. Zrealizowane przez lekarzy badanie eksperymentalne umożliwiło wyłącznie odkrycie mechanizmu zjawiska. Ustalenie jego zasięgu wymagało badań innego typu.

W pewnym uproszczeniu można powiedzieć, że odkrywanie mechanizmów zjawisk stanowi domenę badań określanych mianem **jakościowych**. Natomiast ustalanie zasięgu zjawisk jest przedmiotem badań nazywanych **ilościowymi**. Pomimo że oba nurty badań dotyczą **różnych aspektów tych samych zjawisk**, kształtowały się niezależnie od siebie i do chwili obecnej traktowane są jako odrębne.

Trudno ustalić archetyp badania jakościowego, gdyż wywodzi się ono z codziennych refleksji i prób zrozumienia tego, co dzieje się naokoło. Można jednakże uznać, że podstawę współczesnych badań jakościowych stworzyła antropologia funkcjonalistyczna. Jej celem, podobnie jak każdej dziedziny wiedzy, było odkrycie uniwersalnych praw, rządzących zjawiskami społecznymi pod każdą długością i szerokością geograficzną. Jednakże aby cel ten zrealizować stosowano metodę badawczą, która na pierwszy rzut oka stanowi jego zaprzeczenie. Mianowicie badano niewielkie społeczności, odseparowane od głównego nurtu zmian zachodzących na świecie. Społeczności te badano natomiast dogłębnie i wszechstronnie. Bronisław Malinowski (1884–1942) spędził kilka lat na niewielkim archipelagu wysp triobriandzkich położonych na zachodnim Pacyfiku, uczestnicząc w życiu codziennym

badanych społeczności. Pozwoliło mu to zrozumieć funkcje wielu obrzędów, które z pozoru wydawały się być pozbawione jakiejkolwiek społecznej racjonalności. Na podstawie dokonanych obserwacji stworzył teorię kultury, która na lata określiła kanony interpretowania zjawisk zgodnie z duchem funkcjonalizmu.

Współczesne badania jakościowe stanowią w pewnym sensie kontynuację podejścia stosowanego w antropologii czy etnografii. Ich wyróżniającą cechę stanowi to, że badacz jest osobiście zaangażowany w relacje z badanymi. Badanie jakościowe obejmuje swoim zasięgiem wybraną grupę czy sytuację, co pozwala skoncentrować uwagę na mechanizmach odpowiedzialnych za kształt lub przebieg danego zjawiska (Denzin i Lincoln 2005). Pozwala to na wszechstronny i dogłębny opis tych mechanizmów, co niekiedy prowadzi do swoistego odkrycia, określanego mianem *insight'u*. Nie dostarcza natomiast wiedzy na temat zasięgu zjawiska.

Jako pierwowzór badań **ilościowych** przyjmuje się spisy powszechne, które organizowano już w starożytnej Babilonii, to jest około 3800 lat przed naszą erą (van Koppen 2006: 120). Pierwszy spis powszechny w nowożytnej Europie przeprowadzono zaś w 1719 roku w Prusach. Niezależnie jednak od tego, jaką datę przyjmiemy za moment powstania nurtu badań ilościowych, ich podstawowy cel był zawsze jeden: dostarczyć informacji o rozmiarach czy zasięgu zjawisk.

Równie trudno określić jest moment, od którego w obszarze badań ilościowych podjęto świadome próby powiązania zasięgu zjawisk z ich mechanizmami. Mechanizmy w badaniach ilościowych identyfikuje się w sposób dość swoisty. Polega on na zestawianiu i porównywaniu ze sobą wskaźników występowania różnych aspektów danego zjawiska w podzbiorowościach osób badanych. Na przykład, aby ustalić sposób, w jaki wykształcenie rodziców określa szanse kształcenia się ich dzieci, dzieli się ogół badanych osób na kategorie ze względu na poziom wykształcenia ich rodziców, a następnie porównuje się ze sobą osiągnięcia edukacyjne badanych w każdej z tych kategorii. Schemat ten niewiele ma wspólnego z rozumieniem mechanizmów, czy istoty zjawisk, jakie przyjmuje się na gruncie badań jakościowych. Niemniej jednak stanowić może podstawę budowania bardziej pogłębionych wyjaśnień z wykorzystaniem teorii, dotychczas zgromadzonej wiedzy, a także wniosków z badań jakościowych.

Tablice, będące bohaterem książki, służą do identyfikacji mechanizmów zjawisk w takim sensie, w jakim rozumie się je w badaniach ilościowych[1]. Tym

[1] Współcześnie tablice stosuje się również do analizy danych z badań jakościowych. Stanowią one dogodne narzędzie znajdowania podobieństw i różnic między badanymi przypadkami, które wykorzystywane są w budowaniu wyjaśnień (Garcia-Álvarez i López-Sintas 2002).

samym wyniki badań ilościowych stanowią podstawowy obszar stosowania tablic. Należy jednak uwzględnić fakt, że generyczną funkcją badań ilościowych jest określenie zasięgu badanych zjawisk. Dlatego oba aspekty – zarówno mechanizm, jak też zasięg opisanego w tablicy zjawiska – przewijać się będą przez większość metod omawianych w dalszych rozdziałach książki.

1.3 Właściwości ilościowe i jakościowe

Jako **ilościową** (ang. *quantitative*) przyjęto rozumieć właściwość, która charakteryzuje obiekty w większym lub mniejszym stopniu, czy natężeniu. Przykładami ilościowych właściwości osób są wiek, dochód, poziom introwersji, czy liczba posiadanych dzieci. Ze względu na swoją naturę ilościowe właściwości obiektów dzielą się na ciągłe i skategoryzowane.

Jako **ciągłe** (ang. *continuous*) traktuje się między innymi właściwości metryczne, jak na przykład wzrost czy wagę[2]. Właściwości ciągłe przyrównuje się niekiedy do modelu osi liczb rzeczywistych. W modelu tym ciągłość oznacza, że między dwiema dowolnymi liczbami rzeczywistymi zawsze istnieje liczba pośrednia. W wypadku wzrostu można wyobrazić to sobie jako pomiar z coraz większą precyzją. Na przykład, między wzrostem 175 a 176 cm można wskazać wzrost 175,5 cm. Z kolei pomiędzy wzrostem 175 cm a 175,5 cm zawiera się wzrost 175,25 cm. Rozumowanie to można kontynuować przechodząc do coraz mniejszych jednostek pomiaru wzrostu: milimetrów, mikrometrów, nanometrów. Tym samym między dwiema dowolnymi wielkościami wzrostu można wskazać wielkość pośrednią.

Przypisanie właściwości obiektów atrybutu ciągłości nie musi być ściśle zgodne z modelem matematycznym tego pojęcia, który wymaga istnienia pośredniej wartości między dwiema dowolnymi. Jako właściwość ciągłą traktuje się na ogół dochód, mimo że jednostka monetarna ogranicza dokładność, z jaką możemy wyrazić wielkości dochodu. Między dochodem 2345 zł 67 groszy a dochodem 2345 zł 68 groszy nie ma wielkości pośredniej, przez co

[2] Obie wymienione właściwości, tradycyjnie stanowiące domenę zainteresowań antropologów, stają się coraz częściej przedmiotem uwagi badaczy w naukach społecznych. Stanowią bowiem składowe wskaźnika znanego jako *Body Mass Index* (BMI), który określa relację wagi ciała do wzrostu. Oprócz oczywistych konsekwencji fizjologicznych, wskaźnik ten ma wpływ na zachowania i funkcjonowanie jednostek w sferze społecznej (Tovee i inni 1998). *Body Mass Index* jest szeroko stosowany w badaniach marketingowych, gdyż szereg zachowań zakupowych interpretuje się w kontekście społecznych norm w zakresie otyłości. Warto wspomnieć, że wskaźnik BMI wprowadzony został w pierwszej połowie XIX wieku przez Adolphe Quételeta (1796–1874), belgijskiego uczonego, jednego z prekursorów stosowania metod matematycznych i statystycznych w socjologii (ramka 4.1).

matematyczna definicja ciągłości nie jest spełniona. Pomimo tego większość badaczy jest skłonna postrzegać dochód jako pewną wielkość czy też fragment osi liczbowej o stosunkowo gęsto rozmieszczonych kwotach dochodu. Pojęcie właściwości ciągłej dobrze oddaje tego rodzaju intuicje.

Utworzenie cechy reprezentującej właściwość ciągłą zawsze wiąże się z koniecznością dokonania **kwantyfikacji** zakresu jej zmienności. Na przykład wieku – który spełnia kryteria właściwości ciągłej – nie określa się z największą możliwą dokładnością, lecz najczęściej przyjmuje się dokładność co do roku. Opowiada to sposobom komunikowania wieku stosowanych w życiu codziennym: w rozmowach między ludźmi, w publikacjach prasowych czy w dokumentach urzędowych. Wiek najczęściej odpowiada liczbie lat ukończonych, chociaż zdarza się, że ludzie określają swój wiek według liczby lat rozpoczętych. W badaniach obejmujących małe dzieci (na przykład w psychologii rozwojowej) wiek określa się z większą dokładnością. Bezpośrednio po urodzeniu wiek dziecka wyraża się w dniach, następnie w tygodniach, później zaś w miesiącach. Mniej więcej od drugiego roku życia wiek dziecka określa się w latach, jednakże z dokładnością do pół roku. Na przykład mówimy, że dziecko ma trzy i pół roku. Zdecydowana większość prowadzonych w naukach społecznych badań ograniczona jest jednak do populacji osób dorosłych, toteż kwantyfikacja wieku z dokładnością do roku stanowi praktykę dominującą.

Wiek będący wynikiem badania ma każdorazowo skończoną liczbę wartości. Ich zakres rozpoczyna się od wieku najmłodszych osób objętych badaniem (na przykład 18 lat), zaś kończy w zależności od wieku najstarszej badanej osoby. W części badań świadomie ogranicza się od góry zakres wieku badanych osób, między innymi ze względu na obniżające się wraz z wiekiem predyspozycje do komunikowania się, przez co zastosowana procedura gromadzenia informacji – na przykład wywiad kwestionariuszowy – może okazać się nieadekwatna w wypadku najstarszych respondentów. Stąd wiek respondentów będący wynikiem badania ograniczony jest na ogół do kilkudziesięciu wartości. Na przykład, jeśli badanie obejmuje osoby w wieku od 18 do 75 lat, to wiek przyjmuje jedną z 58 różnych wartości: 18, 19, 20, ..., 73, 74 i 75 lat.

W wypadku wielu cech reprezentujących właściwości ciągłe utrata informacji na skutek kwantyfikacji jest niewielka lub nie występuje wcale. Tak się dzieje w wypadku dochodu, gdy badanych prosimy o podanie dokładnej kwoty. Jeśli cechę określimy jako dochód w pełnych złotych, to zaokrągleniu ulegną jedynie końcówki groszowe, a więc niewielka część właściwych kwot. Jeśli zaś cechę stanowi dochód w setkach złotych, to poziom utraty informacji będzie wyższy. Należy się jednak z tym pogodzić, gdyż ludzie myśląc o dochodach mają skłonność do ich zaokrąglania. Odpowiadając na pytanie o dochód

część lub większość badanych od razu poda wielkości zaokrąglone, najczęściej właśnie do setek złotych.

Jako osobną kategorię właściwości ilościowych wyodrębnia się właściwości w naturalny sposób skategoryzowane. W tym wypadku na poziomie empirycznym właściwość przyjmuje tylko pewne stany, pomiędzy którymi nie ma stanów pośrednich. Przykładem mogą być właściwości wyrażane za pomocą liczb naturalnych, jak liczba posiadanych dzieci, liczba odbiorników telewizyjnych w gospodarstwie domowym, czy wielkość gospodarstwa domowego rozumiana jako liczba jego mieszkańców. Posiadanie jednego dziecka i posiadanie dwójki dzieci to sytuacje różniące się „jakościowo". Gospodarstwa jedno- i dwuosobowe to też różne sytuacje, między którymi nie ma stanów pośrednich. Definiując cechę jako odpowiednik właściwości skategoryzowanej badacz na ogół zachowuje większość z kategorii występujących na poziomie empirycznym, łącząc ze sobą jedynie kategorie reprezentowane przez niewielką liczbę jednostek, a przez to rzadko występujące w badanej próbie. Zdefiniowana przez badacza cecha „wielkość gospodarstwa domowego" może mieć przykładowo pięć następujących kategorii: 1-osobowe, 2-osobowe, 3-osobowe, 4-osobowe oraz obejmujące 5 lub więcej osób.

Niezależnie od tego, czy cecha ilościowa ma charakter ciągły czy skategoryzowany, oraz jak wiele informacji utracono, kwantyfikując zakres jej zmienności – cecha ilościowa zawsze zachowuje podstawową własność odpowiadającej jej właściwości obiektów. Wartości cechy ilościowej są zawsze uporządkowane (ułożone w kolejności), a także określony jest kierunek tego uporządkowania. Pozwala to porównywać ze sobą badane obiekty w wymiarze danej cechy. Na przykład można powiedzieć, że jedna osoba jest starsza od drugiej, ma wyższy od niej dochód, w większym stopniu przejawia postawy radykalne, czy też że ma wyższy iloraz inteligencji badany konkretnym testem.

Przejdźmy obecnie do omówienia własności cech **jakościowych** (ang. *qualitative, discrete, nominal, categorical*). Odpowiadają one właściwościom obiektów, które obejmują osobne, odseparowane od siebie stany rzeczy, naturalnie wyodrębniające się na poziomie rzeczywistości empirycznej. Przykładem jakościowej właściwości osób jest płeć. Właściwość ta dzieli populację osób na dwie kategorie, między którymi nie występują kategorie pośrednie. Modelem formalnym właściwości jakościowej jest schemat podziału wyczerpującego, w którym każdy obiekt należy do jednej i tylko jednej kategorii.

Przyjmijmy, że konstytutywnym atrybutem właściwości jakościowych jest **brak naturalnego porządku** kategorii. Porządek taki można rzecz jasna przyjąć, na przykład twierdząc, że w przestrzeni społecznej pozycja mężczyzn jest wyższa niż kobiet, gdyż przeciętnie zarabiają więcej. Nie zmienia to jednak faktu, że wszelkie tego rodzaju uporządkowania mają charakter wtórny wobec

istoty czy natury jakościowej właściwości obiektów. Z faktu bycia mężczyzną czy bycia kobietą nie wynika w konieczny sposób wysokość zarobków. Jakościowa właściwość obiektów to zbiór odrębnych stanów rzeczy, nie zaś ich jednowymiarowe uporządkowanie. Gdyby takie uporządkowanie występowało na poziomie empirycznym, to daną właściwość należałoby uznać za ilościową.

Liczba kategorii właściwości jakościowej może być różna. Płeć stanowi przykład właściwości dychotomicznej. W wyborach parlamentarnych w Polsce wyborca ma możliwość oddania głosu na kandydata jednej z kilkunastu partii. Oglądając telewizję można wybrać jeden z kilkudziesięciu kanałów. Lista wydawanych czasopism obejmuje kilkaset tytułów. W zasadzie trudno tu wskazać górną granicę.

Duża liczba kategorii rozpatrywanej właściwości jakościowej zmusza na ogół badacza do ich pogrupowania. Przykład właściwości o liczbie kategorii wykraczającej poza możliwości ich efektywnej analizy jest model samochodu. Marek samochodów obecnych na polskim rynku jest kilkadziesiąt. Ponadto większość producentów stara się zaoferować klientom swojej marki samochody w różnych segmentach, począwszy od modeli małolitrażowych, poprzez klasę compact, rodzinne kombi, minivan, van, suv, off-road, a skończywszy na modelach klasy sportowej. Wszystko to tworzy mozaikę kilkuset modeli różnych marek, toteż kwestia stworzenia klasyfikacji samochodów, które posiadają respondenci, stanowi dla badacza swoiste wyzwanie.

W takiej sytuacji badacze stosują jedną z dwóch strategii. Pierwsza polega na tym, że tworzy się dwie lub więcej klasyfikacji według różnych kryteriów. W podanym przykładzie podstawą jednej z klasyfikacji byłaby marka samochodu, podstawą innej mógłby być segment, do którego należy dany model. W pierwszej klasyfikacji modele „Suzuki Wagon R+" oraz „Opel Agila" – które mają identyczną karoserię i podobne funkcjonalności – byłyby zaklasyfikowane do marek Suzuki oraz Opel, w drugiej zaś do tego samego segmentu „minivan". Analizy bazujące na równoległych klasyfikacjach mogłyby między innymi udzielić odpowiedzi na pytanie, czy dla respondentów liczy się przede wszystkim marka, czy segment, do którego należy model samochodu.

Druga ze strategii postępowania z właściwością jakościową o zbyt rozdrobnionych kategoriach polega na ich pogrupowaniu drogą analityczną. W omawianym przykładzie lista atrybutów uwzględnionych podczas grupowania mogłaby wykraczać poza markę i segment, obejmując również inne właściwości posiadanego samochodu, jak rok produkcji, to, czy samochód został kupiony po raz pierwszy w kraju, czy też został sprowadzony zza granicy, rodzaj paliwa (benzyna, olej napędowy), kolor nadwozia, a także inne cechy, które są przez ludzi brane pod uwagę przy zakupie i ocenie korzyści z posiadanego samochodu. Utworzona tą drogą klasyfikacja (czy raczej typologia) użyteczna byłaby

w tych analizach, w których samochód jest traktowany jako dobro będące wypadkową potrzeb, oczekiwań i możliwości.

Na zakończenie omawiania cech jakościowych i ilościowych warto nawiązać do faktu, że niekiedy cechy występujące w badaniach stanowią swoistą rekombinację obu typów. Przykład stanowić może dochód, o który pytamy za pomocą pytania zamieszczonego w kwestionariuszu. Odpowiedzi badanych, którzy podali swój dochód, służą do budowy cechy o charakterze ilościowym. Zawsze pojawia się jednak kategoria badanych, którzy odmawiają odpowiedzi na to pytanie, przez co ich dochód pozostaje nieznany. Jest to sytuacja jakościowo odmienna, niemieszcząca się w wymiarze wielkości dochodu. Cecha dochód składa się przez to z dwóch osobnych komponentów: ilościowego – w wymiarze podanych kwot – jak też jakościowego, w wypadku grupy badanych odmawiających podania dochodu.

1.4 Właściwości przestrzenne

Odrębną grupę stanowią właściwości, które mają charakter przestrzenny, czyli takie, do reprezentowania których należałoby posłużyć się przestrzenią dwu- lub więcej wymiarową. Natura tych właściwości powoduje, że kwantyfikacja ich zmienności wymaga większej liczby dodatkowych założeń, niż ma to miejsce w wypadku właściwości jednowymiarowych. Cechuje ją przez to większa arbitralność.

Najprostszym przykładem właściwości przestrzennej jest miejsce zamieszkania. Ściśle biorąc jest to miejsce na kuli ziemskiej wyznaczone przez złożenie dwóch współrzędnych geograficznych, co implikuje dwuwymiarowy charakter tej właściwości. Warto zauważyć, że właściwość ta spełnia kryterium ciągłości, gdyż między dowolnymi dwoma miejscami zawsze można wskazać miejsce pośrednie. W wypadku jednowymiarowej właściwości – takiej jak wiek czy dochód – z ciągłości wynika jej ilościowy charakter. Nawet wtedy, gdy na skutek kwantyfikacji oryginalne wielkości zostaną pogrupowane, nadal jesteśmy w stanie zasadnie operować pojęciami „więcej" lub „mniej". Właściwość dwuwymiarowa zasadzie tej nie podlega, co wynika stąd, że zrzutowanie przestrzeni dwuwymiarowej w jeden wymiar może być dokonane na wiele różnych sposobów.

W badaniu można na przykład rozstrzygnąć, że jedna z badanych osób mieszka bardziej na wschód od innej, co wiąże się z różnicami kontekstu etnograficzno-kulturowego miejsca zamieszkiwania, a przez to może mieć przełożenie na poglądy i zachowania. Jednakże w kontekście wpływu klimatu na życie codzienne bardziej istotne okazałyby się zapewne różnice zamieszkiwania

wzdłuż osi północ–południe. W sumie możliwości jest wiele i wybór dowolnej z nich nie wynika z natury właściwości, którą jest w tym wypadku miejsce określone przez współrzędne geograficzne.

Być może owa niejednoznaczność spowodowała, że właściwości przestrzenne sprowadza się w badaniach do cech jakościowych. Z tym że kryteria wyodrębnienia ich kategorii zarysowują się na ogół dość słabo – aby nie powiedzieć, że są niekiedy zupełnie arbitralne. W badaniach prowadzonych w Polsce miejsce zamieszkania grupuje się na ogół w województwa. „Województwo" spełnia własności cechy jakościowej. Jest podziałem rozłącznym i wyczerpującym, a jednocześnie poszczególne kategorie (województwa) nie mają naturalnego porządku.

Podstawę stosowanego w większości badań podziału na województwa stanowi obowiązujący obecnie podział administracyjny, w którym wyodrębnia się 16 województw. Łatwo jednak wykazać, że podział ten nie pokrywa się z szeregiem cech istotnych dla socjologa, takich jak kraina geograficzna, specyfika kulturowa czy tożsamość etniczna. Co więcej, w niektórych badaniach bardziej użyteczne dla wyjaśnienia badanych zjawisk okazują się inne podziały terytorialne. Na przykład, w badaniach audytoriów stacji radiowych wykorzystuje się podział na 49 województw, obowiązujący w Polsce w latach 1975–1998. Użyteczność tego podziału wynika stąd, że zasięgi techniczne wielu stacji radiowych obejmują obszary zbliżone do dawnych województw. W badaniu czytelnictwa prasy lokalnej używa się jeszcze starszego podziału administracyjnego, na 17 tak zwanych województw gomułkowskich. Wynika to stąd, że lokalne dzienniki wydawane są w miastach, które stanowiły siedziby województw w tamtym okresie. Według tego samego schematu rozmieszczone są po dziś dzień regionalne rozgłośnie Polskiego Radia.

Kryteria lokalizacji terytorialnej nie są jedynymi, które stosuje się przy konstruowaniu cech na podstawie miejsca zamieszkania. Równie często w badaniach używa się podziału miejscowości według statusu administracyjnego: na miasta i wsie. O znaczeniu tej cechy dla wyjaśnienia wielu zjawisk społecznych nie trzeba nikogo przekonywać. Można natomiast dyskutować, na ile ma ona charakter jakościowy, a na ile ilościowy. Cecha ta nie spełnia kryterium ciągłości, gdyż między wsią a miastem nie wyodrębniono jednostki administracyjnej, która miałaby status pośredni. Z drugiej strony miasta często dzieli się na podkategorie, ze względu na liczbę mieszkańców. Jeśli przyjmiemy, że na poziomie empirycznym odpowiada to właściwości „stopień zurbanizowania środowiska zamieszkiwania", to wtedy zasadne wydaje się uporządkowanie w jednym wymiarze zarówno wszystkich kategorii miast według wielkości, jak też kategorii wsi, która w szeregu tym zajmowałaby skrajną pozycję jako środowisko najmniej zurbanizowane.

Miejsce zamieszkania nie jest jedynym przykładem dwuwymiarowej właściwości, która stanowiłaby przedmiot badań. W badaniach percepcji treści prasowych dużą uwagę przywiązuje się do tego, które fragmenty poszczególnych stron gazety czy czasopisma przykuwają największą uwagę czytelnika. Pomiar tej właściwości może być dokonany w miarę precyzyjnie, za pomocą specjalnej kamery śledzącej ruch gałek ocznych badanego. Niemniej jednak rezultaty omawianych badań opisywane są na ogół w języku, w którym dwuwymiarowe współrzędne sprowadza się do kategorii jakościowych, w rodzaju „górna prawa ćwiartka strony", „dolna połowa strony" czy „lewa kolumna". Na podstawie tak skonstruowanej cechy określającej podobszar strony wydawcy decydują o cenach powierzchni reklamowej na poszczególnych fragmentach stron swoich tytułów. Tym samym językiem posługują się biura reklamy, wyjaśniając reklamodawcom sposób czytania czasopisma przez różne kategorie czytelników oraz korzyści, jakie można osiągnąć, aplikując tę wiedzę w kampanii reklamowej. Okazuje się więc, że niekiedy z precyzji pomiaru właściwości ilościowej zrezygnuje się, zastępując ją w komunikacji szerokimi kategoriami o charakterze cechy jakościowej.

O ile dwuwymiarowe właściwości przestrzenne stają się niekiedy przedmiotem badania w naukach społecznych, o tyle trudno wymienić właściwości trójwymiarowe, które byłyby przedmiotem zainteresowań socjologów. Chociaż – niekiedy w rozmowach ludzie używają argumentu, że ktoś miał „pod górkę do szkoły". W socjologii edukacji często zadaje się pytania o odległość do szkoły, czy też o to, czy szkoła znajdowała się w tej samej miejscowości, w której mieszka uczeń. Badana i klasyfikowana jest więc jedynie przestrzeń dwuwymiarowa. Trudno rozstrzygnąć, na ile użyteczne byłoby rozszerzenie w tym wypadku przedmiotu badań do przestrzeni trójwymiarowej:-)

1.5 Paradoks Simpsona

Niekiedy badacz bywa zaskoczony, że zastosowana metoda analizy danych prowadzi do wniosków całkowicie sprzecznych z tym, co się na logikę wydaje. Omówmy to korzystając z danych zgromadzonych przez przez Radeleta i Pierce'a (1991), a dotyczących wyroków kary śmierci w procesach o zabójstwo w stanie Floryda w latach 1976–1987 (tabela 1.1). Przykład jest na tyle spektakularny, że chętnie cytuje się go w podręcznikach metod analizy danych (Agresti 2002: 48–51).

Interesuje nas, czy kolor skóry sprawcy zabójstwa wpływa na orzeczenie kary śmierci. Z odsetków przedstawionych w tabeli 1.1 wynika, że jeśli ofiara zabójstwa była biała, to częściej karę śmierci otrzymywał sprawca czarny

(22,9%) niż biały (11,3%). Analogicznie, gdy ofiarą była osoba o czarnym kolorze skóry, to czarnych sprawców skazano na karę śmierci w 3 procentach przypadków, zaś białych sprawców nie skazano ani razu. A ponieważ dwa dodać dwa równa się cztery, więc płynie stąd wniosek, że czarnych sprawców karano wyrokiem śmierci częściej niż białych – i to niezależnie od koloru skóry ofiary.

Uprośćmy więc strukturę tablicy, pomijając kolor skóry ofiary. Nie to stanowi bowiem przedmiot analizy. W tym celu wystarczy odpowiednio połączyć liczebności w polach tabeli 1.1 a następnie policzyć na nowo odsetki. Dane w zmodyfikowanej postaci przedstawia tabela 1.2. Zaglądamy do jej wnętrza z góry wiedząc, czego można się tam spodziewać, po czym ... doznajemy szoku. Z odsetków wynika bowiem jednoznacznie, że karę śmierci znacząco częściej orzekano w wypadku białych sprawców zabójstw niż w wypadku czarnych!

Tabela 1.1
Odsetki wyroków kary śmierci w procesach o zabójstwo w stanie Floryda w latach 1976–1987 w zależności od koloru skóry sprawcy i ofiary
[*w procentach*]

kolor skóry		czy orzeczono karę śmierci			
sprawcy	ofiary	tak	nie	ogółem	liczba spraw
biały	biały	11,3	88,7	100,0	467
biały	czarny	0,0	100,0	100,0	16
czarny	biały	22,9	77,1	100,0	48
czarny	czarny	2,8	97,2	100,0	143
	ogółem	10,1	89,9	100,0	674

Źródło: Radelet i Pierce 1991.

Przedstawiony efekt nazywany jest w metodologii **paradoksem Simpsona**. Polega on na tym, że agregacja kategorii cechy powoduje odwrócenie się kształtu zjawiska. Prześledzenie liczebności w rubryce „liczba spraw" tabeli 1.1 pozwala zrozumieć, na czym polega istota tego paradoksu. Gdy czarny stał się ofiarą zabójstwa, to karę śmierci orzekano względnie rzadko. Natomiast dużo częściej sąd wydawał taki werdykt w wypadku, gdy ofiarą zabójstwa był biały. Ofiarami sprawców o czarnym kolorze skóry byli około 3 razy częściej czarni (143 sprawy) niż biali (48 spraw). W wypadku białych sprawców było odwrotnie. Biali stawali się prawie 30 razy częściej ofiarami białych sprawców (467 spraw) niż czarni (16 spraw). Czyli ofiarami białych sprawców byli głównie biali, przy czym zabójstwo białego – jak wynika z przedstawionych

danych – było przez sąd częściej karane wyrokiem śmierci. Ofiarami czarnych sprawców byli zaś głównie czarni, w wypadku zabójstwa których sąd rzadziej zasądzał karę śmierci. Tabela 1.2 przedstawia więc faktyczny obraz zjawiska – częściej na karę śmierci skazywano białych sprawców. Natomiast tabela 1.1 wyjaśnia jego mechanizm.

Tabela 1.2
Odsetki wyroków kary śmierci w procesach o zabójstwo w stanie Floryda w latach 1976–1987 w zależności od koloru skóry sprawcy. Dane pogrupowane
[*w procentach*]

	czy orzeczono karę śmierci			
kolor skóry sprawcy	tak	nie	ogółem	liczba spraw
biały	11,0	89,0	100,0	483
czarny	7,9	92,1	100,0	191
ogółem	10,1	89,9	100,0	674

Źródło: Radelet i Pierce 1991. Przedstawiono te same dane co w tabeli 1.1 z pominięciem koloru skóry ofiary zabójstwa.

Wpływ agregacji kategorii cechy na odwrócenie obrazu zjawiska opisał w 1951 roku brytyjski statystyk Edward H. Simpson, przez co jego nazwiskiem przyjęto nazywać ten efekt. Faktycznie zauważono go wiele lat wcześniej. Warto wspomnieć o wkładzie G. Udny Yule'a chociażby z tego powodu, że był on autorem jednego z pierwszych podręczników statystyki przetłumaczonych i wydanych w Polsce (Yule 1921 [1910]). Yule opisał podobny efekt występujący w tablicach, nazywając go **zależnością pozorną** (1921: 51–72). Termin ten kojarzony jest przez współczesnych badaczy głównie ze schematem pozornej zależności między dwiema zmiennymi na skutek tego, że mają wspólną przyczynę (Babbie 2008: 110–112)[3]. Rzadziej zwraca się uwagę na to, że w tablicach krzyżujących ze sobą dwie – a nie trzy cechy – również pojawia się problem wpływu ukrytych podziałów na obserwowany kształt związku w tablicy.

[3] Autor odwołuje się między innymi do dwóch egzemplifikacji znanych bodajże każdemu studentowi socjologii: (1) im więcej wozów strażackich bierze udział w akcji gaszenia pożaru, tym większe są straty oraz (2) regiony, w których żyją bociany, charakteryzują się wyższą stopą urodzeń. W pierwszym wypadku poprzednikiem opisanej zależności jest oczywiście wielkość pożaru, która określa zarówno liczbę wozów strażackich, jak też wielkość strat. W drugim przykładzie wspólną przyczyną jest stopień zurbanizowania regionu. Bociany żyją w obszarach wiejskich, w których przeciętna stopa urodzeń jest wyższa.

1.6 Skale pomiaru

W latach sześćdziesiątych XX wieku w naukach społecznych upowszechniło się klasyfikowanie właściwości obiektów ze względu na poziom pomiaru. Ta elegancka pod względem formalnym koncepcja dotyczy skal liczbowych, na których wyrazić można właściwości badanych obiektów (Brzeziński 2007: 183–215).

Za najsłabszą skalę uznaje się **skalę nominalną**, która pozwala jedynie wnioskować o relacjach równości lub różności między obiektami ze względu na daną właściwość. Odpowiada to tradycyjnemu pojęciu klasyfikacji, czy cechy jakościowej, która każdy z obiektów przyporządkowuje do jednej i tylko jednej kategorii.

Kolejny poziom pomiaru stanowi **skala porządkowa**. Na otrzymanych w badaniu wynikach pomiaru określony jest porządek, który jest odwzorowaniem porządku istniejącego w empirii. Przykładem takiej własności może być wykształcenie. Istnieje zgoda co do tego, że wykształcenie wyższe jest czymś więcej niż wykształcenie średnie, a to ostatnie czymś więcej niż wykształcenie podstawowe. Skala porządkowa posiada przy tym wszystkie własności skali nominalnej – to znaczy klasyfikuje obiekty – a dodatkowo określa między nimi porządek.

Jeśli skalę porządkową wzbogacimy o możliwość określania odległości między obiektami, to otrzymamy **skalę interwałową**. Jako przykład często wymienia się w tym miejscu dochód, gdyż wyraża się on w jednostkach monetarnych spełniających warunek addytywności. 800 złotych minus 700 złotych daje jako różnicę 100 złotych, podobnie jak 5000 złotych minus 4900 złotych. Odległości między kwotami 700 złotych i 800 złotych, a także 4900 złotych i 5000 złotych są więc definiowalne na skali i w tym wypadku jednakowe. W praktyce tego rodzaju skale rodzą jednak wątpliwości, czy nominalna równość różnic przekłada się na jednakową użyteczność dla badanych. Czy wzrost dochodów o 100 złotych u kogoś, kto zarabia 700 złotych, i u kogoś, kto zarabia 4900 złotych jest przez obie osoby traktowane tak samo oraz czy ma porównywalny wpływ na zmiany w strukturze wydatków obu osób? Jeśli nie, to tego rodzaju cechy nie należałoby uznać jako mierzonej na skali interwałowej. Nie wyklucza to możliwości przekształcenia nominalnych kwot przez funkcję malejących różnic (np. logarytmizując dochody), co być może pozwalałoby mówić o skali interwałowej.

Jeśli do własności skali interwałowej dołożymy jeszcze możliwość porównywania stosunków (ilorazów) dowolnych dwóch wielkości, to skala staje się ilorazową, zwaną również stosunkową. **Skala ilorazowa** ma naturalny początek, to jest punkt, w którym dana właściwość nie występuje. Za cechy

mierzone na skalach ilorazowych uznaje się między innymi właściwości, które wyrażają się za pomocą liczb naturalnych: jak liczba rodzeństwa, liczba samochodów w gospodarstwie domowym, czy sama wielkość gospodarstwa domowego. Gospodarstwo czteroosobowe jest 2 razy bardziej liczne od gospodarstwa dwuosobowego, zaś to ostatnie jest 2 razy bardziej liczne od gospodarstwa jednoosobowego. Podobnie jak w wypadku skal interwałowych, również tu rodzą się wątpliwości, czy obliczanie stosunku dwóch liczb znajduje przełożenie na relacje zachodzące w rzeczywistości. Na przykład, gospodarstwo dwuosobowe nie zużywa zapewne aż dwa razy więcej zasobów od gospodarstwa jednoosobowego, lecz relację tę należałoby raczej ocenić jako stan pośredni, między 1 a 2.

Dochód, wymieniany wcześniej jako przykład cechy mierzonej na skali interwałowej, formalnie spełnia również warunki skali ilorazowej. Można bowiem obliczać i porównywać ze sobą ilorazy kwot, chociaż otwartą pozostaje kwestia interpretacji tych ilorazów. W obszarze zjawisk społecznych w zasadzie nie występują właściwości, które formalnie spełniałyby jedynie własności skali interwałowej, a nie spełniały własności skali ilorazowej (Levine 1993: 73). Niemniej jednak skala interwałowa stanowi narzędzie pomocne w wielu interpretacjach wyników metod analizy danych. Dlatego też warto uwzględnić ją wśród omawianych skal pomiaru właściwości obiektów będących przedmiotem zainteresowań w naukach społecznych.

Aby lepiej zrozumieć problemy związane z operacjonalizacją badanych właściwości za pomocą skal przydatna może okazać się refleksja sięgająca kilku dziesięcioleci wstecz. W latach sześćdziesiątych i siedemdziesiątych XX wieku myślenie o symbolicznej reprezentacji badanych zjawisk zdominowane było przez pragmatyzm, wywodzący się właśnie z koncepcji skal pomiaru. Wielu badaczom podejście to wydawało się zbawienne, gdyż pozwalało uniknąć trudności związanych z wyborem narzędzi analizy danych proponowanych przez statystyków. Niekiedy wynikające stąd reguły postępowania przybierały postać wręcz karykaturalną. Jako przykład wskazać można popularne w swoim czasie poradniki, zwane w slangu badawczym „książkami kucharskimi". Tłumaczono w nich, że jeśli jedna cecha jest – przykładowo – porządkowa, zaś druga nominalna, to do analizy zależności między nimi należy posłużyć się taką to a taką metodą (zob. np. Andrews i in. 1981). W propozycjach tych całkowicie abstrahowano od zjawisk, których dotyczy analiza.

Metodologia lat siedemdziesiątych zdominowana została przez dywagacje na temat adekwatności poszczególnych metod statystycznych do cech wyrażonych na różnych skalach pomiaru. Przykład stanowi dyskusja tocząca się na łamach *American Sociological Review* prawie przez 10 lat, a dotycząca w gruncie rzeczy tego, czy współczynnik korelacji Pearsona można stosować dla cech wyrażonych na skalach porządkowych czy też nie jest to uprawnione

(m. in. Labovitz 1970, 1971; Mayer 1970, 1971; Schweitzer i Schweitzer 1971; Grether 1976; O'Brien 1979).

W latach osiemdziesiątych i dziewięćdziesiątych nastąpiła polaryzacja stosowanych metod analitycznych. Z jednej strony znacznie rozbudowano aparat analizy wzajemnych związków między cechami jakościowymi (zob. Goodman 2000), z drugiej zaś intensywnie rozwijano i znajdowano coraz to nowe zastosowania dla metod ekonometrycznych – operujących zasadniczo na cechach ilościowych (Maddala 2006). Cechy mierzone na skalach porządkowych zostały tym samym zmarginalizowane.

Obecnie do skal pomiaru przywiązuje się mniejsze znaczenie. Koncepcja ta okazała się nieefektywna wobec problemów, z którymi na co dzień musi radzić sobie badacz. Rozważmy przykład z polskiego podwórka. Jak zestawić ze sobą 4-letnie liceum ogólnokształcące i 5-letnie technikum. Co prawda szkoły w tej formie obecnie już nie funkcjonują, lecz przez wiele lat będziemy je napotykać w opisach karier edukacyjnych respondentów oraz ich rodziców. Która z tych szkół powinna być umieszczona wyżej na skali? Czy może technikum, gdyż trwa dłużej – co byłoby zresztą zgodne z koncepcjami reprezentacji wykształcenia poprzez liczbę lat spędzonych w systemie szkolnym. Czy może liceum, gdyż większy odsetek jego absolwentów kontynuuje naukę na studiach wyższych. A może oba rodzaje szkół powinny być potraktowane tak samo, gdyż zarówno liceum, jak i technikum umożliwia uzyskanie wykształcenia średniego. Ostatnie z omawianych rozwiązań przyjęto zresztą w międzynarodowej klasyfikacji poziomów wykształcenia ISCED (1997), w której obie szkoły klasyfikowane są do tej samej kategorii oznaczonej symbolem „3A".

Nasuwa się pytanie – czy warto rzeczywistość naginać do typowych schematów analizy danych, czy raczej postępować inaczej – traktując schematy w sposób na tyle elastyczny, aby już na początku nie ograniczyć sobie możliwości interpretacji badanego zjawiska. W przykładzie liceum i technikum rozsądnym wyjściem jest potraktowanie obu szkół jako odrębnych kategorii i nie przyjmowanie z góry żadnych założeń co do relacji między nimi. Jeżeli relacje takie istnieją, to z pewnością ujawnią się jako jeden z wyników analiz. Jeśli zaś kształtują się inaczej niż to sobie wyobrażamy, to nie czyniąc z góry żadnych założeń uchronimy się przed stworzeniem artefaktu.

1.7 Opis wyników badań w kategoriach probabilistycznych

Większość współcześnie prowadzonych badań nie obejmuje całej badanej zbiorowości lecz jedynie niektóre z jednostek wchodzących w jej skład, które w sumie określa się mianem próby. Tego rodzaju badania nazywane są

reprezentacyjnymi w odróżnieniu od badań wyczerpujących, obejmujących wszystkie jednostki w badanej zbiorowości. Powodów realizacji badań reprezentacyjnych zamiast wyczerpujących jest wiele. Najważniejsze z nich to czas i pieniądz. Badanie reprezentacyjne jest tańsze ze względu na ograniczoną liczbę badanych jednostek. Może być też przeprowadzone w krótszym czasie. A czas w niektórych projektach odgrywa kluczową rolę. Przykładami są badania przedwyborcze, badania zauważalności kampanii reklamowych, czy też wiele innych badań, w których na podstawie uzyskanych wyników podejmuje się decyzje operacyjne.

Potrzeba określenia relacji między obrazem zjawiska uzyskanym poprzez przebadanie próby, a jego rzeczywistym kształtem w badanej zbiorowości, zaowocowała rozwojem potężnej gałęzi wiedzy, jaką jest współczesna statystyka. Z jednej strony obejmuje ona zasady doboru próby pozwalające zwiększyć szanse uzyskania stosunkowo niewielkich rozbieżności między uzyskanym w badaniu obrazem zjawiska a jego rzeczywistym kształtem. Z drugiej zaś strony obejmuje aparat opisu wnioskowania o populacji na podstawie przebadanej próby.

Szacunek dla dokonań współczesnej statystyki spowodował, że metodologia analizy danych przynajmniej od kilkudziesięciu lat zdominowana została przez perspektywę, którą można nazwać **probabilistyczną**. Uzyskany w badaniu wynik traktowany jest jako zdarzenie losowe, które zachodzi z pewnym prawdopodobieństwem. Na przykład, respondentka odpowiedziała, że w wyborach parlamentarnych głosowała na PiS. Probabilistyczny charakter tego rezultatu bierze się z dwóch źródeł. Pierwszym jest skłonność do udzielenia danej odpowiedzi. Jeśli w wyborach, o które pytano, PiS osiągnął kiepski rezultat, to respondenci będą raczej skłonni nie przyznawać się do głosowania na tę partię. Gdyby natomiast w rezultacie wyborów PiS uzyskał większość parlamentarną, to do głosowania na PiS byłyby skłonne przyznawać się również osoby, które faktycznie swój głos oddały na kandydatów innych ugrupowań.

Drugim uzasadnieniem probabilistycznej natury uzyskanej odpowiedzi jest to, że owa przykładowa respondentka reprezentuje kobiety z badanej zbiorowości. Do wylosowanej próby trafiła z pewnym prawdopodobieństwem. Równie dobrze do próby w jej miejsce mogła trafić inna kobieta, która – na przykład – powiedziałaby ankieterowi, że głosowała na PO. Uzyskany wynik traktuje się więc jako realizację pewnej zmiennej losowej, której rozkład zależy przede wszystkim od rozkładu danej cechy w populacji. Jeśli każdą osobę w populacji zapytać o to, na kogo oddała głos w wyborach, przy czym większość odpowiedziałaby, że głosowała na PiS, to przy spełnieniu dodatkowych warunków związanych z losowością próby uzyskanie w badaniu odpowiedzi „oddałem

głos na PiS" miałoby większe prawdopodobieństwo wystąpienia niż uzyskanie odpowiedzi świadczącej o głosowaniu na inne ugrupowanie.

Konsekwencją założenia dotyczącego probabilistycznej natury uzyskanych wyników jest aparat pojęciowy stosowany do opisu i analizy danych. Oparty jest on nie na liczebnościach uzyskanych w badaniu, lecz na estymowanych prawdopodobieństwach uzyskania poszczególnych wyników. Przykładowo, współczynnik korelacji traktuje się nie jako konkretną wartość, lecz jako zmienną związaną za pomocą określonej funkcji z prawdopodobieństwami zdarzeń uwzględnianych w formule obliczania współczynnika korelacji. Jako zmienna ma on swój własny rozkład, którego ustalenie stanowi przedmiot badania. Na przykład, w wyniku badania można stwierdzić, że wartość współczynnika korelacji z prawdopodobieństwem 0,95 zawiera się w przedziale od 0,45 do 0,55. Nie przywiązuje się tu wagi do otrzymanej w badaniu konkretnej wartości współczynnika korelacji (na przykład 0,50), gdyż owa konkretna wartość ma charakter przypadkowy. Równie dobrze mogło wyjść coś obok, na przykład 0,47 czy 0,53. Ważny jest rozkład wyników, a nie wynik uzyskany drogą doboru konkretnej próby. Odzwierciedla on bowiem wyłącznie fakt, że do próby trafiły akurat te osoby, a nie inne.

Nie kwestionując trafności ujęcia probabilistycznego jako modelu opisującego mechanizm uzyskiwania wyników w badaniach reprezentatywnych, nie można jednak przejść obojętnie wobec trudności, jakie wprowadza to podejście przy wyjaśnianiu istoty badanych zjawisk. Otóż badacz musi równocześnie skupić uwagę na dwóch czynnikach: na relatywności wyniku związanej z jego probabilistycznym charakterem oraz na właściwym przedmiocie badania. W praktyce jest to trudne. Dlatego w książce zdecydowałem się przyjąć odmienną konwencję. Będę abstrahować od probabilistycznej natury wyników uzyskiwanych w badaniach, traktując je tak, jakby uzyskany w badaniu wynik stanowił trafny obraz badanego zjawiska. Będę więc mówił o liczebnościach czy o odsetkach respondentów, zamiast o prawdopodobieństwach, czy o rozkładzie cechy w badanej populacji.

Przyjęte uproszczenie nie jest zresztą aż tak dotkliwe, jak to mogłoby się wydawać. Gdy brak niezależnej wiedzy na temat badanej zbiorowości, w ramach ujęcia probabilistycznego jako charakterystyki populacji również proponuje się wielkości parametrów otrzymane w wyniku zbadania próby. Na przykład, przedział ufności dla średniej buduje się wokół średniej otrzymanej z próby. Tego rodzaju przykładów można podać wiele. Pomimo że istota obu ujęć jest odmienna, postać komunikowanych wyników jest w wielu wypadkach identyczna.

Zauważmy również, że nie wszystkie badania mają charakter reprezentacyjny. Spis powszechny obejmuje całą ludność Polski, przez co model probabilistyczny nie znajduje zastosowania do analizy jego wyników. Niekiedy

przedmiotem zainteresowań są mikrospołeczności, które można przebadać w całości. Garcia-Álvarez i López-Sintas (2002) badając biznesy zarządzane rodzinnie oparli się na liście 500 największych hiszpańskich przedsiębiorców i wybrali z niej wszystkich, którzy spełniali kryterium prowadzenia rodzinnego biznesu. Przykład z innego podwórka stanowi socjogram sporządzony w klasie szkolnej. Aby miał sens, musi objąć wszystkich uczniów. Co więcej, otrzymany obraz relacji między uczniami ogranicza się tylko i wyłącznie do danej klasy i nie ma możliwości jego uogólnienia na jakąkolwiek inną zbiorowość.

Rezygnacja z ujęcia probabilistycznego nie jest zresztą moim pomysłem. Z podobnych powodów zabieg ten stosuje się w wielu podręcznikach metod statystycznych. Na przykład Lissowski, Haman i Jasiński (2008) utrzymali ponad dwie trzecie swojego podręcznika w konwencji, którą nazywają „opisem statystycznym", opartą na przyjęciu umownego założenia, że „zdołaliśmy zbadać całą populację" (2008: 46). Ujęcie takie ułatwiło autorom wyjaśnienie istoty metod analizy danych stosowanych w badaniach socjologicznych. Dopiero w końcowej części podręcznika autorzy wracają do problemu wnioskowania o zbiorowości na podstawie próby.

Podobne ujęcie odnaleźć można w wielu innych pracach, zwłaszcza starszych. W 1954 roku dwaj amerykańscy statystycy – Leo Goodman i William Kruskal – opublikowali artykuł, który stanowił w pewnym sensie przełom w spojrzeniu na pomiar siły zależności między cechami w tablicach. Autorzy jasno zaznaczyli, że „rozpatrują wyłącznie sytuację, w której cała populacja jest znana i nie trzeba ustosunkowywać się do kwestii doboru próby czy błędów pomiaru" (1954: 733). W zakończeniu artykułu piszą jednak, iż „dotychczasowa dyskusja utrzymana była w konwencji opisu znanych populacji, podczas gdy w praktyce badacze mają do czynienia z próbą dobieraną z nieznanej populacji. W stosunku do proponowanych miar zależności pojawiają się więc pytania jak estymować ich wartości, jak testować hipotezy itp." (1954: 762). Obiecują więc czytelnikowi, że artykuł uzupełnią o teorię rozkładów proponowanych miar. Obietnicę tę udało im się częściowo spełnić 9 lat później (Goodman i Kruskal 1963), natomiast wątek ten w pełni zamknęli dopiero 18 lat później (Goodman i Kruskal 1972). Przez cały ten czas ich pierwotne propozycje żyły własnym życiem, wywołując żywe reakcje wśród badaczy (zob. Goodman i Kruskal 1959). Świadczy to, że kluczową wartość stanowi niekiedy sama idea proponowanych rozwiązań, zaś umieszczenie jej w kontekście modelu probabilistycznego jest sprawą wtórną.

Na zakończenie tego wątku konieczna jest jeszcze jedna uwaga. Proponowane w tej książce zogniskowanie uwagi na wyjaśnieniu idei proponowanych metod nie oznacza, że kwestie wnioskowania statystycznego zostaną pominięte. Wręcz przeciwnie, znaczna część metod analizy tablic – która obejmuje też

metody chętnie przez badaczy stosowane – oparta jest na logice wnioskowania statystycznego. Również i w wypadku metod prezentowanych w konwencji czysto opisowej będę starał się przytoczyć odpowiednie testy statystyczne, a przynajmniej wskazać, gdzie takie testy można znaleźć. Pozwoli to poszerzyć zakres zastosowań proponowanych rozwiązań.

1.8 Dane wieloodpowiedziowe

Jednym z narzędzi uzyskiwania informacji w badaniach kwestionariuszowych są pytania wieloodpowiedziowe. W większości zastosowań pytania te mają postać zamkniętą, to znaczy obejmują kafeterię odpowiedzi zdefiniowanych przez badacza. W pytaniach wieloodpowiedziowych badanemu nie stawia się ograniczeń co do liczby wybranych odpowiedzi. Badany może wybrać jedną odpowiedź, może wybrać ich wiele, jak też może nie wybrać żadnej. W badaniach akademickich pytania wieloodpowiedziowe stosuje się dość rzadko, a jeśli już, to dotyczą na ogół cenionych wartości bądź atrybutów przypisanych postaciom lub sytuacjom. W badaniach marketingowych pytania wieloodpowiedziowe stanowią natomiast jedną z podstawowych form uzyskiwania informacji od respondentów. Stosuje się je w sytuacjach, gdy pyta się konsumenta o rodzaje użytkowanych produktów, marki tych produktów, a także o atrybuty, które można przypisać tym markom. Za pomocą pytań wieloodpowiedziowych ustala się sposoby spędzania czasu wolnego, określa listę czytanych gazet i czasopism, słuchanych stacji radiowych, a także odtwarza wiele innych zachowań.

Dane zgromadzone za pomocą pytania wieloodpowiedziowego posiadają specyficzną strukturę, której warto poświęcić nieco uwagi. Na pierwszy rzut oka wydawać się może, że jedyna komplikacja polega na niejednakowej liczbie atrybutów przynależnych badanym obiektom – w związku z niejednakową liczbą udzielanych odpowiedzi. Tymczasem problem sięga głębiej. Pytanie wieloodpowiedziowe generuje bowiem całkiem nową przestrzeń badanych obiektów, różną od przestrzeni, która stanowi przedmiot badania.

Przyjmijmy, że badamy gospodarstwa domowe, ustalając między innymi marki użytkowanych samochodów. W gospodarstwie nikt może nie mieć samochodu bądź może być użytkowany jeden, dwa lub nawet więcej samochodów. Przyjmijmy dalej, że interesuje nas szczególnie jedna z marek, jaką jest Fiat. W 2006 roku była to marka najczęściej występująca spośród wszystkich marek samochodów w Polsce[4].

[4] Według wyników badania *Target Group Index* (w skrócie TGI) prowadzonego przez SMG/KRC Millward Brown. Próba w 2006 roku liczyła 34 502 osoby. Pytania o markę samochodu

Jeśli liczbę fiatów w Polsce odniesiemy do ogółu gospodarstw, to okaże się, że w 13 procentach gospodarstw jeździ się fiatem. Wynik w tej postaci może być użyteczny dla sieci serwisowej fiata. Jeśli promocję swoich usług chcieliby prowadzić, wysyłając materiały reklamowe pod losowo wybrane prywatne adresy, to 13 procent tych materiałów trafi do gospodarstw, w których użytkowany jest fiat. Analogiczny odsetek obliczyć można przyjmując jako podstawę nie wielkość populacji gospodarstw, lecz sumaryczną liczbę wszystkich samochodów użytkowanych w gospodarstwach. Oszacowany w ten sposób udział samochodów marki Fiat wśród wszystkich samochodów byłby wyższy niż poprzednio i wyniósłby 22 procent. Informacja w tej postaci byłaby cenna dla inwestora, który chciałby budować sieć stacji obsługi nastawionych na klientów indywidualnych. Uzyskany wynik oznacza bowiem, że 22 na 100 zgłaszających się klientów przyjedzie naprawić fiata.

Dane zebrane w pytaniu wieloodpowiedziowym mogą być prezentowane na różne sposoby, przy czym każdy z nich stanowi nieco odmienne spojrzenie na to samo zagadnienie, którym jest w rozpatrywanym wypadku struktura marek samochodów użytkowanych w gospodarstwach różnej wielkości. Sposoby te podzielić można na dwie grupy.

Pierwsza grupa obejmuje sytuacje, w których odpowiedzi na pytanie wieloodpowiedziowe relatywizowane są do wielkości badanej populacji. Suma odsetków odpowiedzi na ogół przekracza wtedy 100 procent, ponieważ respondent może udzielić więcej niż jednej odpowiedzi. Druga grupa obejmuje zaś sytuacje, gdy poszczególne odpowiedzi relatywizowane są do zbiorowości obiektów utworzonych przez odpowiedzi badanych. W omawianym przykładzie obiektami tymi są samochody. Badany, podając jedną markę, dodaje jeden samochód do zbiorowości samochodów, podając dwie marki, dodaje dwa. Również odmowa odpowiedzi dotyczącej marki dodaje samochód, mimo że brakuje informacji o jego marce. Jedynie odpowiedź „żadna z marek" nie dodaje samochodu, gdyż odpowiada sytuacji, w której w gospodarstwie samochodu nie ma.

Relacja między badaną zbiorowością – w omawianym przykładzie zbiorowością gospodarstw w Polsce – a zbiorowością obiektów utworzoną przez odpowiedzi badanych jest na tyle złożona, że nie poddaje się formalnej analizie. Nie można bowiem stworzyć modelu matematycznego, który pozwoliłby ową relację opisać w ogólnym przypadku. Podstawowym powodem jest to, że odpowiedzi nie są od siebie niezależne. W omawianym przykładzie brak niezależności zinterpretować można w ten sposób, że użytkowanie samochodu

zadano w ankiecie samodzielnie wypełnianej przez respondentów. Proszono o podanie maksymalnie dwóch marek użytkowanych samochodów.

kształtuje lojalność czy przywiązanie wobec marki, bądź przeciwnie, zniechęca do danej marki, na przykład ze względu na kosztowną eksploatację lub brak określonych walorów funkcjonalnych. W każdym razie kupując drugi samochód w ramach gospodarstwa domowego, z pewnością bierze się pod uwagę markę dotychczas użytkowanego samochodu.

Brak możliwości skonstruowania formalnego modelu przestrzeni obiektów generowanych przez odpowiedzi badanych nie musi oznaczać, że problemu nie uda się rozwiązać. Niekiedy bowiem badanie zaprojektować można w taki sposób, aby przestrzeń odpowiedzi badanych stała się badaną zbiorowością. W omawianym przykładzie badaną zbiorowość mógłby stanowić ogół samochodów użytkowanych w Polsce. Próbę do tego badania należałoby dobrać w taki sposób, aby szanse wylosowania każdego samochodu były wzajemnie od siebie niezależne. Na przykład losując samochody według odpowiedniego schematu z rejestrów wydziałów komunikacji.

Przykład pytania o markę samochodu jest o tyle specyficzny, że łatwo wskazać niezależny sposób badania zbiorowości utworzonej przez udzielone odpowiedzi. W wypadku wielu pytań wieloodpowiedziowych tak się jednak nie dzieje. W badaniach marketingowych szeroko stosowane jest pytanie o tak zwaną spontaniczną świadomość marek produktów. Na przykład: „jakie marki żelów pod prysznic Pani zna, nawet jeśli ich Pani nie używa?". Odpowiedzi na to pytanie tworzą pewną przestrzeń marek, która zawiera marki aktualnie dostępne na rynku, marki dostępne w przeszłości lub aktualnie dostępne na innych rynkach, marki niedookreślone („takie mydło w płynie w granatowej butelce, to znana marka"), marki obecne w innych segmentach, lecz nieoferujące produktów w kategorii żelów pod prysznic, czy też marki nieistniejące – co do których badanym wydaje się jedynie, że zetknęli się z nimi. Omawiana przestrzeń marek istnieje wyłącznie w świadomości badanych i nie posiada odpowiednika w postaci zbiorowości realnie istniejących obiektów. Nie może więc stać się przedmiotem odrębnego badania.

Mimo tak ulotnej natury zbiorowości marek utworzonej przez odpowiedzi konsumentów na pytanie o spontaniczną świadomość, badania wykorzystujące pytania wieloodpowiedziowe pełnią w marketingu istotne funkcje, służąc między innymi do oceny efektywności wydatkowania środków na kampanie reklamowe czy w podejmowaniu decyzji o dalszej strategii rynkowej przedsiębiorstwa. Użyteczność uzyskiwanych wyników bierze się stąd, że pytania wieloodpowiedziowe zadawane są w określonym cyklu – na przykład przed i po kampanii reklamowej (Sawiński 2007a). Przypuśćmy, że przed kampanią marka miała 10-procentowy udział w świadomości spontanicznej, zaś po kampanii 20-procentowy. Tego rodzaju wynik stanowić może podstawę oceny skuteczności kampanii. Może też być punktem wyjścia dla dalszych wnio-

sków. Przede wszystkim, jak ma się udział w przestrzeni marek do udziału marki w przestrzeni konsumentów. Wynik 20 procent oznaczać bowiem może, że tylko jedna piąta konsumentów zna markę po kampanii, o ile przeciętny konsument przechowuje w świadomości tylko jedną markę produktów z danej kategorii. Ten sam wynik oznaczać również może, że markę znają wszyscy konsumenci, o ile każdy konsument zna ich przynajmniej pięć[5].

Podsumowując ten fragment rozważań, można sformułować wniosek, że cechy przynależne obiektom utworzonym na bazie pytania wieloodpowiedziowego również znajdują zastosowanie w analizach zjawisk społecznych – podobnie jak cechy przynależne obiektom stanowiącym badaną zbiorowość. Zasadne jest więc oczekiwanie, aby metody analizy danych w postaci tabelarycznej uwzględniały również pytania wieloodpowiedziowe.

Tymczasem metodolodzy i statystycy niechętnie podejmują problem analizy danych z pytań wieloodpowiedziowych. Zadanie to jest bowiem niewdzięczne. Struktura danych wieloodpowiedziowych ma na tyle złożoną postać, że utworzenie modelu formalnego jest trudne. W efekcie niewiele jest metod statystycznych dostosowanych do analizy danych z pytań wieloodpowiedziowych, zaś literatura poświęcona tym zagadnieniom jest skąpa. Ogranicza się zresztą do omówienia wyłącznie pewnych aspektów analizy danych wieloodpowiedziowych, natomiast nie oferuje kompleksowych rozwiązań (Umesh i inni 1992; Umesh 1995; Santos 2000; Yancy i Allenby 2003).

1.9 Zapis cech w komputerowych plikach danych

Współcześni badacze analizują związki między cechami, korzystając z komputera. Aplikacje czy pakiety programów dedykowane do obsługi wyników badań socjologicznych wymagają określonego formatu zapisu danych (Niepokojczycki i Sawiński 1984). Format ten oparty jest na pewnych konwencjach dotyczących sposobu, w jaki stosowany zapis reprezentuje wyniki badania, a pośrednio badane zjawiska. Świadomość tych konwencji pozwala uniknąć szeregu mylnych interpretacji, które mogą wziąć się stąd, że badacz przypisuje danym w komputerze własności, których nie mają obiekty na poziomie empirycznym.

[5] Omawiane ustalenie, mimo swojej elegancji, nie ma znaczenia w praktyce. Albowiem konsument, kupując produkt, musi zdecydować się na jedną markę niezależnie od tego, ile przechowuje ich w świadomości. Dlatego na podstawie samego wzrostu udziału reklamowanej marki w puli ogółu marek obecnych w świadomości konsumentów należy spodziewać się wzrostu sprzedaży produktów tej marki. Nie ma tu potrzeby wnikania w kwestie wielkości i struktury przestrzeni marek generowanej przez odpowiedzi respondentów.

Jednostką podziału pamięci komputera jest bajt. Bajt przybierać może jeden z 256 stanów, które obejmują między innymi cyfry od 0 do 9, a także wszystkie litery alfabetu: zarówno małe – od a do z, jak i duże – od A do Z. Z możliwości tych nie korzysta się jednak. Przenosząc wyniki badania na nośnik komputerowy do reprezentowania kategorii cech wykorzystuje się na ogół cyfry lub ciągi cyfr. Na przykład, płeć respondenta bodajże najczęściej zapisywana jest w sposób następujący: mężczyznom przypisuje się symbol „1", zaś kobietom symbol „2". Należy podkreślić, że chodzi tu o symbole, nie o liczby. Równie dobrze można byłoby mężczyznom przypisać literę M, zaś kobietom literę K. Płeć jest cechą wyrażoną na skali nominalnej (jest klasyfikacją), a więc jedyny warunek, który muszą spełnić przypisane symbole sprowadza się do tego, aby obu płciom przypisać odrębne symbole.

Gdy jednak stosuje się symbole cyfrowe, to wtedy łatwo wpaść w pułapkę przypisania badanym obiektom właściwości, które mają wyłącznie liczby. Pakiety statystyczne nie odróżniają bowiem cyfr pełniących wyłącznie funkcję symboli, od cyfr, które pełnią funkcję liczb, czyli wyrażają wielkość bądź natężenie badanego zjawiska. Procedura analityczna wykona więc na cyfrach również te operacje, które miałyby sens jedynie na liczbach. Może na przykład przypisać kobietom – oznaczonym cyfrą „2" – cechę „płeć" w dwa razy większym stopniu niż mężczyznom – oznaczonym cyfrą „1". Równie dobrze może też obliczyć średnią „płci" w badanej próbie.

Kwestia interpretacji symboli reprezentujących kategorie cech prowadzi do nieporozumień również dlatego, że podczas analizy danych badacze powszechnie posługują się terminem „zmienna" (od angielskiego *variable*) dla oznaczenia pojęcia, które przyjęliśmy nazywać cechą, zaś kategorie cechy nazywają „wartościami" (od angielskiego *values*). Prowadzić to może do błędnego przekonania, że wartościami zmiennej są liczby. Samo pojęcie „zmiennej" jest bowiem mocno osadzone nie tylko w statystyce matematycznej, ale również w wiedzy nabytej w szkole średniej, zgodnie z którą zmienna jest kojarzona z jednowymiarowym i ciągłym konstruktem formalnym, odpowiadającym osi liczb rzeczywistych. Odpowiadają temu takie pojęcia jak „zmienna X", „zmienna Y", „wartość zmiennej", „układ współrzędnych". Tymczasem „zmienna" w zbiorze wyników badań ani nie jest z konieczności ciągła, ani też nie musi mieć wartości uporządkowanych wzdłuż jednego wymiaru.

Ważną własność cech reprezentujących wyniki badań sondażowych stanowi fakt, że w każdym z pytań osoba badana ma prawo odmówić odpowiedzi. Może też pytania nie zrozumieć, bądź nie być w stanie wybrać tylko jednej z odpowiedzi spośród proponowanych w kafeterii. Sytuacje te – zwane rezydualnymi – występują praktycznie w każdym zadawanym pytaniu. W zbiorach danych kategorie rezydualne umieszcza się na ogół na krańcach zakresu sym-

boli odpowiadających odpowiedziom badanych. Na przykład, jeśli w pytaniu dochody – oferującym badanemu kafeterię 16 pogrupowanych przedziałów – odpowiedzi są symbolizowane od „01" do „16", to odmowę odpowiedzi oznaczyć można symbolem „-1", zaś „trudno powiedzieć" symbolem „-2". Jeśli dwa ostatnie symbole zinterpretowane zostaną jako liczby, to spełnią relację mniejszości wobec pozostałych kategorii dochodów. Stąd niedaleko do bezsensownego wniosku, że fakt odmowy podania dochodów to **mniej** niż dochody o dowolnej wysokości[6].

Aby uniknąć tego rodzaju błędnych interpretacji – na przykład podczas liczenia średniej czy porządkowania kategorii cechy – pakiety statystyczne umożliwiają zadeklarowanie pewnych symboli jako niemerytorycznych czy rezydualnych (w slangu badawczym zwanych też „missingami"). Osoby, u których wartość cechy zostanie zakodowana za pomocą tak zadeklarowanego symbolu, będą domyślnie usuwane z analiz. Co prawda zabezpiecza to przed błędnymi interpretacjami, lecz w zamian powoduje pominięcie osób, które nie udzieliły odpowiedzi na pytanie. Tymczasem osoby takie stanowić mogą znaczny odsetek badanej próby. Przykładowo, jeśli analizowaną cechą jest dochód, to odsetek osób, które dochodu nie podały, w obecnie prowadzonych badaniach dochodzi nawet do 30–40 procent. Osoby te w rzeczywistości osiągają pewne dochody, lecz w badaniu nie udało się ustalić ich wysokości. Czy automatyczne wykluczenie tych osób z analiz doprowadzi do trafnego opisu sytuacji materialnej ogółu badanych?

Powyższe kwestie przywołane zostały z tego względu, że nieuważne posługiwanie się danymi z badań stwarza niebezpieczeństwo otrzymania mylnych konkluzji. Gdyby zagrożenia czyhające na badacza w fazie analizy danych uporządkować od najmniej do najbardziej poważnych, to kwestia nadinterpretacji symboli zapisu danych znalazłaby się blisko szczytu hierarchii. Omawiane zagrożenie bierze się z dwóch powodów. Po pierwsze, współcześni badacze preferują metody eksploracyjne, w których duże zasoby danych poddawane są obróbce za pomocą procedur analitycznych na tyle złożonych, że sposób ich działania nie zawsze daje się prześledzić. Łatwo więc przeoczyć wiele niuansów w konstrukcji poszczególnych cech oraz w założeniach przyjmowanych w procedurach analitycznych. Spowodować to może, że wyniki obliczeń niewiele mają wspólnego z rzeczywistym obrazem badanego zjawiska. Po drugie,

[6] Niekiedy bezrefleksyjnie przypisuje się znaczące symbole sytuacjom, które odpowiadają brakom danych. Na przykład, w testach psychometrycznych czy w testach osiągnięć używa się symbolu 0 dla oznaczenia sytuacji, w której badany zadanie pominął. W zamierzeniu badacza nie otrzymuje za nie żadnych punktów. Faktycznie natomiast kompetencje czy cechy badanego nie zostały zmierzone. Kwestia ta jest zresztą kontrowersyjna i trudno o jednoznaczną dyrektywę, jak należy postępować (McKnight i in. 2007: 180–181).

coraz częściej role badacza gromadzącego dane oraz badacza zajmującego się ich analizą są rozdzielone – o czym pisałem we wprowadzeniu. Mając kontakt z danymi wyłącznie w fazie analiz trudno na ogół zrekonstruować założenia badania, sposób wyodrębnienia poszczególnych cech czy wpływ warunków realizacji badania na kształt zgromadzonych danych.

1.10 Tablice jako narzędzie analizy danych

W tym miejscu zatrzymamy się na chwilę, aby podjąć próbę doprecyzowania, czym właściwie są tablice stanowiące przedmiot tej książki.

W języku polskim słowo „tablica" ma wiele konotacji. Część z nich odwołuje się do wyodrębnionej w przestrzeni płaszczyzny, na której w określony sposób zostały rozmieszczone pewne elementy (Szymczak 1978: 471). Definiując tablicę w sensie, w którym chcielibyśmy używać tego pojęcia, słownik języka polskiego odwołuje się do terminu „tabela". „Tablica" to „tabela, czyli wykaz, zestawienie, rejestr danych (zwykle liczbowych), rozmieszczonych na arkuszu w określonym porządku według rubryk" (Szymczak 1978: 470). Z definicji tych wynika, że w języku polskim słów „tablica" oraz „tabela" można używać wymiennie.

Na co dzień z „tablicami" czy „tabelami" spotykamy się w publikacjach. Tabela w formacie publikacyjnym zawiera nie tylko samo zestawienie liczb, lecz również pewne objaśnienia. Objaśnienia te obejmują takie elementy, jak tytuł tabeli, opis zawartości poszczególnych wierszy i kolumn, a niekiedy dodatkowe uwagi dotyczące źródła danych bądź wskazówki dotyczące interpretacji liczb przedstawionych w zestawieniu. Można więc powiedzieć, że „tabela" zawsze dotyczy konkretnego zjawiska. Jeśli w jednej tabeli przedstawiony został sposób głosowania mężczyzn i kobiet w wyborach parlamentarnych we wrześniu 2005 roku, zaś w drugiej tabeli sposób głosowania mężczyzn i kobiet w kolejnych wyborach, które odbyły się w październiku 2007 roku, to z całą pewnością w obu wypadkach mamy do czynienia z odrębnymi tabelami. W publikacji otrzymają one osobne oznaczenia, na przykład „tabela 1" i „tabela 2", gdyż obejmują różne dane.

W odróżnieniu od powyższego rozumienia tabeli, można też wskazać inne. Jest nim samo zestawienie liczb ze sobą, w prostokątnym układzie wierszy i kolumn. Zestawienie takie stanowi pośrednią formę przetworzenia danych, w celu poddania ich dalszej analizie. W tym sensie stanowi więc **narzędzie analizy danych**. Narzędzie uniwersalne, którego własności nie zależą od tego, czy umieszczone w nim dane dotyczą płci osób głosujących w wyborach, czy – przykładowo – zbieżności wykształcenia męża i żony. Przyjmijmy ustalenie,

że dla oznaczenia omawianego narzędzia będziemy posługiwać się terminem **tablica**. Ustalenie to ma charakter definicji regulacyjnej, aczkolwiek wydaje się, że niesprzecznej z przyjętymi zwyczajami językowymi. Natomiast termin **tabela** zachowamy dla określenia zestawienia liczb w formacie publikacyjnym. „Tablice" liczb skonfigurowane na podstawie wyników badań prezentować więc będziemy w postaci „tabel", czyli wyodrębnionych fragmentów tekstu, zatytułowanych „Tabela 1", „Tabela 2" i tak dalej.

A więc tablica to zestawienie liczb odzwierciedlających w pewien sposób występowanie każdej z kombinacji kategorii dwóch cech. Zestawienie to ma postać prostokątnego bloku, na który składają się wiersze i kolumny, przy czym poszczególne wiersze tego bloku odpowiadają kategoriom jednej cechy, zaś poszczególne kolumny kategoriom drugiej cechy. Wierszy jest tyle, ile kategorii ma jedna z cech, zaś kolumn tyle, ile kategorii ma druga cecha. W polach leżących na przecięciu poszczególnych wierszy i kolumn podane są wielkości charakteryzujące współwystępowanie danej kombinacji kategorii obu cech. Wielkością taką może być na przykład liczba badanych osób. Może to być też inny wskaźnik liczbowy, na przykład odsetek wyrażony w procentach. W dalszych rozdziałach wprowadzimy wiele tego typu wskaźników.

Warto w tym miejscu wspomnieć, że w tekstach metodologicznych i statystycznych publikowanych w języku polskim na ogół nie dokonuje się terminologicznego rozróżnienia między tabelą – jako formą prezentacji danych oraz tablicą – jako narzędziem ich analizy. W obu wypadkach stosuje się termin „tablica" lub „tabela". Nie podejmę się rozstrzygnięcia, w którym z kontekstów który z tych terminów stosowany jest częściej. Ponieważ jednak mówi się tu o dwóch różnych rzeczach, więc termin odpowiadający narzędziu analizy danych otrzymuje na ogół dookreślenie. Stosuje się między innymi takie dookreślenia, jak tabela czy tablica „liczebności" (Kendall i Buckland 1975: 203), „wielodzielcza" (Góralski 1979: 34), „niezależności" (Steczkowski i Zeliaś 1981: 147), „korelacyjna" (Timofiejuk, Lasek i Pęczkowski 1997: 99) czy „krzyżowa" (Górniak 2000: 117; Perek-Białas i Korzeniecka 2006: 67).

W języku angielskim stosuje się podobne rozwiązania. W opracowaniach akademickich dość wieloznaczny termin „table" najczęściej łączy się z dookreśleniem „contingency"[7]. W nauce stosowanej częściej napotyka się na terminy „cross-tabulation" lub „cross-classification" – co oddaje ideę krzyżowania

[7] Jest rzeczą interesującą, że termin ten, wprowadzony w 1904 roku przez Karla Pearsona, oznaczał pierwotnie „miarę odchylenia klasyfikacji od niezależnych prawdopodobieństw" (Agresti 2002: 621). Czyli w określeniu „contingency table" mieści się już pewna konkretna koncepcja oceny, czy analizy tablic (przedstawimy ją w rozdziale 3). Jest to z pewnością zawężenie uniwersalności tablic jako narzędzia badania związków między cechami. Mimo to termin „contingency table" jest powszechnie używany w tekstach akademickich.

w tablicy cech ze sobą. Należy zaznaczyć, że zarówno określenia „contingency table" jak też „cross-classification" nie ograniczają się do tablic, w których zestawia się ze sobą dwie cechy, lecz mogą być z równym powodzeniem stosowane w wypadku tablic wielowymiarowych, w których zestawia się ze sobą trzy lub więcej cech. Poszukiwanie uniwersalnych metod analizy tablic, niezależnych od liczby jej wymiarów, dominuje zresztą we współczesnych opracowaniach na ten temat (Agresti 2002).

Powstaje w związku z tym pytanie, czy przyjęta w książce perspektywa ograniczenia rozważań do tablic obejmujących wyłącznie dwie cechy nie ograniczy ogólności rozważań i nie odetnie możliwości analizowania za pomocą proponowanych narzędzi również tablic, w których zamieszczono jednocześnie trzy lub więcej cech. Jest bowiem dość oczywiste, że współcześnie prowadzone badania nie ograniczają się do zjawisk opartych na interakcji dwóch cech, lecz obejmują na ogół bardziej skomplikowane powiązania między wieloma cechami.

Rozstrzygnięcie tak sformułowanej kwestii nie jest proste. Gdy wnikniemy w operacje wykonywane podczas analizy danych pochodzących z badań, to łatwo dojść do wniosku, że najczęściej stosowana strategia polega na rozłożeniu splotu badanych zjawisk na zależności elementarne – poprzez zestawianie ze sobą badanych cech parami. Przypuśćmy, że przedmiot analizy stanowiłyby związki zachodzące między płcią, wykształceniem a sposobem głosowania w wyborach. Część badaczy zapewne utworzyłaby dwie tablice, z których jedna przedstawiałaby związek sposobu głosowania z płcią, druga zaś związek sposobu głosowania z wykształceniem. Część badaczy poszłaby inną drogą, tworząc tablice opisujące zależność sposobu głosowania od wykształcenia osobno dla mężczyzn i kobiet.

Powstaje pytanie, jak duża część badaczy byłaby skłonna interpretować dane w postaci tablicy trójwymiarowej – krzyżującej jednocześnie płeć, wykształcenie i sposób głosowania. Przypuszczam, że ta ostatnia frakcja badaczy byłaby niewielka. Większość ludzi nie dysponuje bowiem umiejętnością syntetycznego wglądu w zależności trój- lub więcej wymiarowe w stopniu, w jakim jest w stanie prześledzić zależności w tablicach obejmujących dwie cechy.

Możliwości wykorzystania dwuwymiarowych tablic do analizy większej liczby czynników stwarza też sposób wyodrębnienia obu krzyżowanych cech. Przypuśćmy, że chcemy przedstawić w tablicy związek między wykształceniem a zarobkami. Zauważmy jednak, że ten sam poziom wykształcenie oznaczać może co innego w wypadku mężczyzn i kobiet. Przez wiele lat w polskim systemie kształcenia dominowały szkoły zasadnicze zawodowe. Ich specjalności odzwierciedlały zapotrzebowanie przemysłu ciężkiego na wykwalifikowanych robotników, toteż niewiele było w nich miejsc dla dziewcząt. Po

ukończeniu szkoły podstawowej część dziewcząt zmuszona więc niejako była do kontynuowania nauki w liceach i w technikach. Natomiast gdy liceum lub technikum wybrał chłopiec, to świadczyło to raczej o jego aspiracjach – na przykład do osiągnięcia wykształcenia wyższego.

Jeśli w konstruowanej tablicy nie uwzględnimy płci, to wyodrębnione kategorie wykształcenia będą niejednoznaczne. Na przykład, wykształcenie średnie oznaczałoby co innego w wypadku mężczyzn, a co innego w wypadku kobiet – ze względu na odmienność strukturalnych powodów znalezienia się w szkole średniej. Uzasadnione jest więc utworzenie odrębnych hierarchii poziomów wykształcenia dla mężczyzn i kobiet. Tworzymy w ten sposób cechę, która stanowi skrzyżowanie płci i wykształcenia, po czym cechę tę możemy umieścić w konwencjonalnej tablicy, krzyżując ją z zarobkami.

Tablice krzyżujące ze sobą tylko dwie cechy mogą okazać się więc przydatne również podczas analizy zjawisk, w które uwikłane jest wiele czynników. Kwestiom tym poświęcony został podrozdział 7.10.

1.11 Podsumowanie

Jak sygnalizowałem we wstępie rozdziału, obejmuje on dość zróżnicowane zagadnienia dotyczące kontekstu posługiwania się tablicami w analizach wyników badań. Niekiedy zagadnienia te są odległe od siebie. Dlatego, aby nie powtarzać szczegółowych ustaleń, podsumowanie ograniczę do jednego punktu.

Reprezentacja zjawisk za pomocą kategoryzacji nie jest wolna od zagrożeń prowadzących do mylnych interpretacji. Właściwości, takie jak płeć czy liczba dzieci, mają kategorie wyodrębnione w naturalny sposób. Są jednak i takie cechy – jak dochód, wykształcenie czy miejsce zamieszkania – w wypadku których kształt przyjętej kategoryzacji zależny jest przede wszystkim od badacza. Jeśli okaże się ona nieadekwatna wobec badanego problemu, to uzyskane wnioski nie będą trafne.

Od badacza oczekuje się wniosków dotyczących realnie istniejących zjawisk społecznych, nie zaś opisu relacji między symbolami w komputerowym pliku danych. Dlatego czytając dalsze fragmenty warto mieć na uwadze związki między konstruktem nazwanym cechą a właściwościami obiektów w rzeczywistym świecie. Zawsze też warto szukać dodatkowej wiedzy na temat rozpatrywanego zjawiska, aby móc ocenić trafność sposobu jego ujęcia za pomocą zgromadzonych danych.

ROZDZIAŁ 2

Prezentacja danych za pomocą tabel

Tablice pełnić mogą dwie funkcje: narzędzia analizy danych oraz narzędzia prezentacji wyników. W tym rozdziale omówiona zostanie wyłącznie druga z tych funkcji. Zgodnie z przyjętą wcześniej konwencją tablicę w tym kontekście nazywać będziemy tabelą. Tabela stanowi więc komunikat, którego autorem jest badacz. Za jej pomocą przekazać pragnie odbiorcy określone interpretacje wyników badania i zogniskować jego uwagę na wybranych aspektach badanego zjawiska.

Wyniki współcześnie prowadzonych badań dostępne są w postaci komputerowych plików danych. Na tej podstawie tworzy się modele analizy i opisu badanych zjawisk, w tym tablice opisujące związki dwóch cech. Rozdział rozpoczynam od omówienia sposobu przekształcenia danych zapisanych w formacie komputerowego pliku do postaci tablicy (2.1). W podrozdziale 2.2 omawiam najważniejsze decyzje, jakie badacz musi podjąć w fazie budowy tabeli służącej prezentacji danych. Chodzi o rozstrzygnięcia w takich kwestiach jak wykluczenie z tabeli pewnych kategorii badanych osób, łączenie kategorii o niewielkich liczebnościach czy ustalenie kolejności kategorii cech. Omawiam też wpływ, jaki na treść wniosków ma wybór wielkości prezentowanych w polach tabeli. Osobną uwagę poświęcam zasadom przedstawiania w tabeli danych w postaci ważonej.

Podrozdział 2.3 poświęcony został zasadom edycji tabel. Aby tabela okazała się efektywnym narzędziem prezentacji badanego zjawiska, w fazie jej tworzenia warto wykorzystać wszystkie dostępne środki, w tym środki edycyjne. Podrozdział zamykają wskazówki dotyczące estetyki tabel, która ma również wpływ na perswazyjność komunikatu.

2.1 Od pliku danych do tablicy

Chcąc przedstawić proces budowy tablicy opisującej przykładowe zjawisko społeczne skorzystamy z wyników trzeciej transzy Europejskiego Sondażu

Społecznego (Sztabiński 2004; Sztabiński i Sztabiński 2006) zrealizowanej jesienią 2006 roku. W badaniu tym gromadzono dane metodą wywiadów kwestionariuszowych. Próbę dobrano metodą losową z populacji ludności Polski w wieku 15 lub więcej lat. W sumie zrealizowano 1721 wywiadów. Zebrane dane przechowywane są w postaci komputerowego pliku danych zorganizowanego tak, że możliwy jest dostęp do odpowiedzi na każde z pytań kwestionariusza dla każdej badanej osoby. Strukturę tego pliku przedstawić można jako prostokątną tablicę, w której wiersze odpowiadają badanym osobom, zaś kolumny zawierają symbole odpowiedzi na kolejne pytania (rycina 2.1).

Rycina 2.1
Obraz struktury pliku danych obejmującego wyniki przykładowego badania

identyfikator respondenta	waga	P1	P2	P3	P4	P5	P6	P7	dalsze pytania ...
0001	0,8588961596887979	1	1964	1	10	4	2	5	.....
0002	0,9890803441800305	1	1956	2	66	5	1	5	.....
0003	1,1963709547450171	2	1979	1	06	1	1	4	.....
0004	0,9668964091209586	1	1922	2	66	1	1	5	.....
0005	1,0470887159887524	1	1971	1	05	4	2	8	.....
0006	1,0309645091571681	2	1974	1	03	3	2	5	.....
0007	1,2234325619153087	2	1990	3	66	3	1	5	.....
0008	0,9668964091209586	2	1944	8	66	2	1	4	.....
0009	1,1963709547450171	1	1977	1	06	2	1	5	.....
0010	0,9308354593495910	2	1934	1	88	8	2	2	.....
....	.....	..	....	..	...	..	..	..	.....
....	.....	..	....	..	...	..	..	..	.....
1720	0,9890803441800305	1	1965	2	66	1	1	3	.....
1721	1,0470887159887524	2	1942	1	06	4	2	5	.....

Przyjmijmy, że pole oznaczone jako P1 odpowiada pytaniu kwestionariuszowemu o płeć respondenta, zaś pola oznaczone jako P3 i P4 odpowiadają pytaniom o uczestnictwo w wyborach do sejmu we wrześniu 2005 roku oraz o sposób głosowania w tych wyborach. Treść obu pytań przedstawiona została w ramce 2.1.

Ramka 2.1
Sformułowanie pytań o udział i sposób głosowania w wyborach do Sejmu we wrześniu 2005 roku w badaniu Europejski Sondaż Społeczny 2006

P3. Obecnie wielu ludzi z różnych przyczyn nie bierze udziału w wyborach.
Czy brał/-a P. udział w ostatnich wyborach do sejmu we wrześniu 2005 roku?

Tak	1	***ZADAĆ PYT. P4***
Nie	2	
Nie byłem/-am uprawniony/-a do głosowania	3	***PRZEJŚĆ DO P5***
(Trudno powiedzieć)	8	

ZADAĆ PYT. P4, JEŚLI W PYT. P3 PADŁA ODP. "TAK" (KOD 1)

P4. Na którą partię lub ugrupowanie głosował/-a P. w tamtych wyborach do Sejmu?

Centrum (J. Steinhoff)	01
Dom Ojczysty (B. Pęk)	02
Liga Polskich Rodzin (R. Giertych)	03
Narodowe Odrodzenie Polski (A. Gmurczyk)	04
Ogólnopolska Koalicja Obywatelska (W. Kornowski)	05
Partia Demokratyczna – demokraci.pl (W. Frasyniuk)	06
Partia Inicjatywa RP (Z. Łuczak)	07
Platforma Janusza Korwin-Mikke (J. Korwin-Mikke)	08
Platforma Obywatelska Rzeczypospolitej Polskiej (D. Tusk)	09
Polska Konfederacja – Godność i Praca (A. Słomka)	10
Polska Partia Narodowa (L. Bubel)	11
Polska Partia Pracy (D. Podrzycki)	12
Polskie Stronnictwo Ludowe (W. Pawlak)	13
Prawo i Sprawiedliwość (J. Kaczyński)	14
Ruch Patriotyczny (J. Olszewski)	15
Samoobrona Rzeczpospolitej Polskiej (A. Lepper)	16
Socjaldemokracja Polska (M. Borowski)	17
Sojusz Lewicy Demokratycznej (W. Olejniczak)	18
Inna (WPISAĆ)	19
Inna (WPISAĆ)	20
(Odmowa odpowiedzi)	77
(Nie pamiętam)	88
(BRAK ODPOWIEDZI)	99
(NIE DOTYCZY)	66

Punkt wyjścia budowy każdej tablicy stanowi wybór dwóch kolumn z pliku danych. Wybór taki prowadzi do otrzymania zestawu par symboli odpowiedzi na wybrane pytania. Przyjmijmy, że wybrane zostały kolumny odpowiadające pytaniu P1 o płeć respondenta oraz pytaniu P4 o partię, na kandydata której

badany oddał swój głos. Pary odpowiedzi na wybrane pytania można przedstawić, wypisując je po kolei, jak to zostało przedstawione poniżej. Nazwijmy tę formę prezentacji danych listą par.

1 – 10
1 – 66
2 – 06
1 – 66
1 – 05
2 – 03
2 – 66
2 – 66
1 – 06
2 – 88
. . . . . .
1 – 66
2 – 06

Jak łatwo zauważyć, niektóre pary występują w przedstawionym fragmencie listy więcej niż raz. Na przykład para symboli „2 – 06" (kobieta, która głosowała na kandydata PiS) występuje dwukrotnie. Z kolei para „1 – 66" odpowiadająca mężczyźnie, który nie głosował w wyborach, występuje trzykrotnie. Zamiast więc wypisywać wszystkie pary – czy też mówiąc inaczej – kombinacje odpowiedzi na dwa wybrane pytania, można sporządzić zestawienie częstości występowania poszczególnych par wśród ogółu badanych osób. Zestawienie takie zostało przedstawione w tabeli 2.1.

Przyjrzenie się zawartości tabeli 2.1 prowadzi do dwóch obserwacji. Po pierwsze, poszczególne kombinacje płci i sposobu głosowania mogą różnić się dość znacznie, gdy chodzi o częstość ich występowania wśród badanych osób. Są takie kombinacje, które wystąpiły jedynie raz lub co najwyżej kilka razy. Na przykład, pierwsza z kombinacji w liście oznaczona symbolem „1 – 02" (mężczyzna głosujący na kandydata Domu Ojczystego) wystąpiła tylko raz wśród wszystkich 1721 respondentów. Z drugiej strony są też takie kombinacje obu cech, które wystąpiły kilkaset razy. Na przykład, kombinacja oznaczona symbolem „2 – 66" (kobieta, która nie wzięła udziału w wyborach) pojawiła się wśród badanych osób 339 razy.

Druga z obserwacji dotyczy faktu, że nie wszystkie potencjalnie możliwe kombinacje płci i sposobu głosowania pojawiły się w wynikach badania. W kafeterii pytania kwestionariuszowego (ramka 2.1) wymieniona była między innymi partia Polska Konfederacja – Godność i Praca (A. Słomka), ozna-

Tabela 2.1
Liczba par symboli oznaczających płeć oraz odpowiedź na pytanie o sposób głosowania w wyborach do Sejmu we wrześniu 2005 roku w badaniu Europejski Sondaż Społeczny 2006
Lista posortowana według symboli par

symbol pary	liczba par	płeć – sposób głosowania
1 – 02	1	mężczyzna – Dom Ojczysty (B. Pęk)
1 – 03	8	mężczyzna – Liga Polskich Rodzin (R. Giertych)
1 – 05	1	mężczyzna – Ogólnopolska Koalicja Obywatelska (W. Kornowski)
1 – 06	2	mężczyzna – Partia Demokratyczna demokraci. pl (W. Frasyniuk)
1 – 08	2	mężczyzna – Platforma Janusza Korwin-Mikke
1 – 09	128	mężczyzna – Platforma Obywatelska (D. Tusk)
1 – 13	19	mężczyzna – Polskie Stronnictwo Ludowe (W. Pawlak)
1 – 14	191	mężczyzna – Prawo i Sprawiedliwość (J. Kaczyński)
1 – 16	51	mężczyzna – Samoobrona RP (A. Lepper)
1 – 17	2	mężczyzna – Socjaldemokracja Polska (M. Borowski)
1 – 18	39	mężczyzna – Sojusz Lewicy Demokratycznej (W. Olejniczak)
1 – 19	2	mężczyzna – Inna partia
1 – 66	321	mężczyzna – nie brał udziału w wyborach
1 – 77	11	mężczyzna – odmowa odpowiedzi
1 – 88	37	mężczyzna – nie pamięta na kogo głosował
2 – 03	13	kobieta – Liga Polskich Rodzin (R. Giertych)
2 – 06	6	kobieta – Partia Demokratyczna demokraci.pl (W. Frasyniuk)
2 – 08	1	kobieta – Platforma Janusza Korwin-Mikke
2 – 09	152	kobieta – Platforma Obywatelska (D. Tusk)
2 – 13	9	kobieta – Polskie Stronnictwo Ludowe (W. Pawlak)
2 – 14	216	kobieta – Prawo i Sprawiedliwość (J. Kaczyński)
2 – 16	33	kobieta – Samoobrona RP (A. Lepper)
2 – 17	5	kobieta – Socjaldemokracja Polska (M. Borowski)
2 – 18	46	kobieta – Sojusz Lewicy Demokratycznej (W. Olejniczak)
2 – 19	1	kobieta – Inna partia
2 – 66	339	kobieta – nie brała udziału w wyborach
2 – 77	12	kobieta – odmowa odpowiedzi
2 – 88	72	kobieta – nie pamięta na kogo głosowała
2 – 99	1	kobieta – brak odpowiedzi
	1721	OGÓŁEM

czona symbolem 10. Żadna z osób badanych nie powiedziała ankieterowi, że swój głos oddała na kandydata tej partii. Wynik ten należy uznać za konsekwencję niewielkiej liczby wyborców głosujących w rzeczywistości na tę partię. Według oficjalnego komunikatu Państwowej Komisji Wyborczej wyniosła ona 8353 osoby (ramka 2.3). Stanowi to zaledwie 0,028 procenta osób uprawnionych do głosowania, których w sumie było 30 milionów 229 tysięcy. Wśród respondentów Europejskiego Sondażu Społecznego znalazło się 1618 osób uprawnionych do głosowania w wyborach w 2005 roku. Na podstawie rozkładu dwumianowego (Ferguson i Takane 2007: 111–117) obliczyć można prawdopodobieństwo zdarzenia, że wśród 1618 respondentów nie będzie ani jednej osoby głosującej na Polską Konfederację – Godność i Praca. Prawdopodobieństwo to wynosi 0,64. A więc to, że nikt z badanych nie wymienił tej partii, stanowi wynik najbardziej prawdopodobny z możliwych.

Uwagę badaczy przyciągają zawsze te sytuacje, które występują najczęściej. Przypisuje im się szczególne znaczenie z tego względu, że opisują postawy czy zachowania najbardziej licznych kategorii badanych osób. Identyfikacja najczęściej występujących kombinacji cech może być dokonana poprzez posortowanie zestawienia przedstawionego w tabeli 2.1 według częstości występowania poszczególnych par odpowiedzi wśród badanych osób.

Zestawienie w tej postaci przedstawione zostało w tabeli 2.2. Zarówno wśród kobiet, jak i wśród mężczyzn, najczęściej zdarzały się sytuacje polegające na nie wzięciu udziału w wyborach. Odpowiadają temu dwie najwyższe pozycje w posortowanej liście. Dwie kolejne pozycje zajmują kobiety i mężczyźni głosujący na Prawo i Sprawiedliwość. Dwie następne pozycje odpowiadają kobietom i mężczyznom głosującym na Platformę Obywatelską. Do tego miejsca listy zaznacza się następująca prawidłowość: każde z zachowań wyborczych jest udziałem zarówno mężczyzn, jak i kobiet, przy czym każde z tych zachowań przejawia więcej kobiet niż mężczyzn.

Kolejne miejsce – pod względem częstości występowania – zajmuje sytuacja nie pamiętania przez kobiety, na kogo oddały głos (72 kobiety). Począwszy od tego zdarzenia, łamie się zasada obowiązująca w pierwszej części listy, w myśl której to samo zachowanie występowało kolejno u kobiet i u mężczyzn. Następną kombinację płci i sposobu zachowania przy urnie stanowią mężczyźni, którzy głosowali na Samoobronę (51 mężczyzn). Warto zwrócić uwagę, że jest ich więcej niż kobiet, które oddały swoje głosy na kandydatów tej partii (33 kobiety).

Analiza listy w postaci przedstawionej w tabeli 2.2 prowadzi do wniosku, że w wypadku wielu zachowań wyborczych, zwłaszcza tych występujących najczęściej, kobiety i mężczyźni postępują podobnie. W wypadku innych zachowań, zwłaszcza pojawiających się rzadziej, między kobietami i mężczy-

Tabela 2.2
Liczba par symboli oznaczających płeć oraz odpowiedź na pytanie o sposób głosowania w wyborach do Sejmu we wrześniu 2005 roku w badaniu Europejski Sondaż Społeczny 2006
Lista posortowana według liczby par

symbol pary	liczba par	płeć – sposób głosowania
2 – 66	339	kobieta – nie brała udziału w wyborach
1 – 66	321	mężczyzna – nie brał udziału w wyborach
2 – 14	216	kobieta – Prawo i Sprawiedliwość (J. Kaczyński)
1 – 14	191	mężczyzna – Prawo i Sprawiedliwość (J. Kaczyński)
2 – 09	152	kobieta – Platforma Obywatelska (D. Tusk)
1 – 09	128	mężczyzna – Platforma Obywatelska (D. Tusk)
2 – 88	72	kobieta – nie pamięta na kogo głosowała
1 – 16	51	mężczyzna – Samoobrona RP (A. Lepper)
2 – 18	46	kobieta – Sojusz Lewicy Demokratycznej (W. Olejniczak)
1 – 18	39	mężczyzna – Sojusz Lewicy Demokratycznej (W. Olejniczak)
1 – 88	37	mężczyzna – nie pamięta na kogo głosował
2 – 16	33	kobieta – Samoobrona RP (A. Lepper)
1 – 13	19	mężczyzna – Polskie Stronnictwo Ludowe (W. Pawlak)
2 – 03	13	kobieta – Liga Polskich Rodzin (R. Giertych)
2 – 77	12	kobieta – odmowa odpowiedzi
1 – 77	11	mężczyzna – odmowa odpowiedzi
2 – 13	9	kobieta – Polskie Stronnictwo Ludowe (W. Pawlak)
1 – 03	8	mężczyzna – Liga Polskich Rodzin (R. Giertych)
2 – 06	6	kobieta – Partia Demokratyczna demokraci.pl (W. Frasyniuk)
2 – 17	5	kobieta – Socjaldemokracja Polska (M. Borowski)
1 – 06	2	mężczyzna – Partia Demokratyczna demokraci. pl (W. Frasyniuk)
1 – 08	2	mężczyzna – Platforma Janusza Korwin-Mikke
1 – 17	2	mężczyzna – Socjaldemokracja Polska (M. Borowski)
1 – 19	2	mężczyzna – Inna partia
1 – 02	1	mężczyzna – Dom Ojczysty (B. Pęk)
1 – 05	1	mężczyzna – Ogólnopolska Koalicja Obywatelska (W. Kornowski)
2 – 08	1	kobieta – Platforma Janusza Korwin-Mikke
2 – 19	1	kobieta – Inna partia
2 – 99	1	kobieta – brak odpowiedzi

znami zarysowują się różnice. Prawidłowości te łatwiej byłoby zauważyć, gdyby zamiast sortować wyjściową listę kombinacji płci i sposobu głosowania (tabela 2.1) przekształcić ją w inny sposób. Mianowicie pierwszą część listy, która obejmuje zachowania wyborcze mężczyzn, ułożyć obok drugiej części listy obejmującej zachowania wyborcze kobiet. Powstanie w ten sposób dwuwymiarowa tablica, której wiersze opisują zachowania wyborcze, zaś kolumny płeć osób badanych. Tablicę tę przedstawiono w tabeli 2.3. Po jej prawej stronie dodano kolumnę „ogółem", w której podane zostały sumy mężczyzn i kobiet przejawiających poszczególne zachowania wyborcze. Na dole tablicy dodany został wiersz, w którym podano łączną liczbę badanych mężczyzn, łączną liczbę badanych kobiet, a także liczbę respondentów ogółem.

Prezentacja układu kombinacji obu cech w formie tablicy otwiera nowe możliwości interpretacji wyników badania. Rozważmy przykładowo wiersz oznaczony symbolem „66" odpowiadający sytuacji, w której respondenci Europejskiego Sondażu Społecznego twierdzili, że nie wzięli udziału w wyborach. Jak zauważyliśmy wcześniej, jest to sytuacja najczęściej występująca zarówno wśród mężczyzn jak i wśród kobiet, przy czym zachowania takie przejawia nieco więcej badanych kobiet (339), niż mężczyzn (321). Obecnie – mając dane w postaci tabelarycznej – łatwo zauważyć, że również wśród ogółu respondentów jest więcej kobiet (906) niż mężczyzn (815). Pozwala to sformułować wniosek, że jednym z powodów stwierdzonej w badaniu większej liczby kobiet niegłosujących w wyborach może być po prostu to, że jest ich **więcej niż mężczyzn wśród badanych osób**. Interpretacja ta została w tabeli 2.3 zobrazowana dodatkowo za pomocą symboli graficznych. Liczba 339 kobiet niebiorących udziału w wyborach jest wypadkową dwóch czynników. Pierwszym jest wielkość grupy respondentów, którzy nie głosowali (660 osób), drugim zaś liczba kobiet w przebadanej próbie (906 osób). Podobnie, liczba mężczyzn niebiorących udziału w wyborach (321 osób) zależy od liczności grupy respondentów nie biorących udziału w wyborach (660 osób) oraz od liczby badanych mężczyzn (815).

Warto w tym miejscu przywołać analogię do tabliczki mnożenia, do której nawiązałem we wprowadzeniu. Liczebności we wnętrzu tablicy są wypadkowymi liczebności opisujących wiersze i kolumny. Im większe są obie liczebności brzegowe, tym większej liczebności należy spodziewać we wnętrzu tablicy. Wiadomo, że w tabliczce mnożenia liczebności wnętrza tablicy są iloczynami liczebności brzegowych. Natomiast odtworzenie zasad, które określają zależność liczebności wnętrza od liczebności brzegowych w tablicach skonstruowanych na podstawie wyników badań, stanowi każdorazowo zadanie badacza, który podjął się interpretacji danych.

Na zakończenie tej części rozważań warto jeszcze raz podkreślić, że każda rozpatrywana tablica jest formalnie równoważna innym formom zapisu

danych. Kombinacja dwóch kolumn wybranych z rekordów zbioru danych symbolicznie przedstawionego na rycinie 2.1 jest równoważna tablicy przedstawionej w tabeli 2.3. Między formatem z ryciny 2.1 a docelową tabelą 2.3 można ponadto utworzyć szereg form pośrednich, których przykłady ilustrują tabele 2.1 czy 2.2. Inaczej mówiąc, fizycznej reprezentacji tablicy – tak jak została ona przedstawiona w tabeli 2.3 – odpowiada wirtualna postać tablicy, na którą składają się inne formaty zapisu tych samych danych: związane ze sposobem ich przechowywania i udostępniania, a także z organizacją procesów

Tabela 2.3
Liczba odpowiedzi na pytanie o sposób głosowania w wyborach do Sejmu we wrześniu 2005 roku wśród mężczyzn i kobiet w badaniu Europejski Sondaż Społeczny 2006

symbol	sposób głosowania	mężczyźni	kobiety	ogółem
02	Dom Ojczysty (B. Pęk)	1	0	1
03	Liga Polskich Rodzin (R. Giertych)	8	13	21
05	Ogólnopolska Koalicja Obywatelska (W. Kornowski)	1	0	1
06	Partia Demokratyczna demokraci.pl (W. Frasyniuk)	2	6	8
08	Platforma Janusza Korwin-Mikke	2	1	3
09	Platforma Obywatelska (D. Tusk)	128	152	280
13	Polskie Stronnictwo Ludowe (W. Pawlak)	19	9	28
14	Prawo i Sprawiedliwość (J. Kaczyński)	191	216	407
16	Samoobrona RP (A. Lepper)	51	33	84
17	Socjaldemokracja Polska (M. Borowski)	2	5	7
18	Sojusz Lewicy Demokratycznej (W. Olejniczak)	39	46	85
19	Inna partia	2	1	3
66	nie brał(a) udziału w wyborach	321	339	660
77	odmowa odpowiedzi	11	12	23
88	nie pamięta, na kogo głosował(a)	37	72	109
99	brak odpowiedzi	0	1	1
	ogółem	815	906	1721

Ramka 2.2
Tablice a eksploracja danych

W ostatnich latach karierę zrobiły techniki analityczne określane mianem eksploracji danych (ang. *data mining*). Służą one znajdowaniu pewnych wzorców czy prawidłowości w dużych zasobach danych. Przy czym na ogół nie wymagają żadnych dodatkowych założeń dotyczących tego, na czym owe wzorce mają polegać bądź w których fragmentach danych ich szukać (Hand, Heikki i Smyth 2005; Larose 2006).

Klasyczny przykład zastosowania technik eksploracji danych stanowi tak zwana analiza zawartości koszyka sklepowego. Jej celem jest ustalenie, które produkty klienci najczęściej kupują wraz z innymi produktami. Wyniki formułuje się w postaci tzw. reguł asocjacyjnych. W ogólnym przypadku reguła asocjacyjna ma postać:

jeżeli jedna cecha przybiera określony stan, to druga cecha przybiera określony stan z danym prawdopodobieństwem – zwanym dokładnością reguły.

W hipermarkecie oferującym klientom tysiące produktów znalezienie reguły asocjacyjnej o treści „jeśli klient kupił pieczywo, to spośród wszystkich dostępnych prócz pieczywa asortymentów największe jest prawdopodobieństwo, że kupi jednocześnie produkt do smarowania pieczywa" jest niebłahym wynikiem, który może znaleźć przełożenie na sposób organizacji przestrzeni sklepu, a w konsekwencji przyczynić się do zwiększenia jego obrotów.

Dane będące przedmiotem eksploracji można niekiedy sprowadzić do formatu tablicy. Można tak również zrobić w wypadku analizy zawartości koszyka sklepowego tworząc tablicę, która zarówno w wierszach, jak i w kolumnach miałaby wyszczególnione wszystkie dostępne w sklepie produkty, zaś na przecięciu odpowiedniego wiersza i kolumny notowano by liczbę klientów, którzy kupili daną kombinację produktów. Już jednak średniej wielości sklep może oferować kilka tysięcy produktów, toteż skonstruowana w ten sposób tablica miałaby wiele milionów pól. Jej użyteczność jako narzędzia analizy czy prezentacji danych byłaby przez to niewielka. Z tego powodu, w algorytmach eksploracji danych nie korzysta się z tablic, lecz przeszukanie wykonuje bezpośrednio na danych surowych.

Wobec respondentów Europejskiego Sondażu Społecznego również utworzyć można reguły asocjacyjne. Na przykład, jeśli respondent jest mężczyzną, to z prawdopodobieństwem 0,39 nie wziął udziału w wyborach do Sejmu w 2005 roku. Jeśli respondentka jest kobietą, to z prawdopodobieństwem 0,24 głosowała na „Prawo i Sprawiedliwość". Na pierwszy rzut oka tego rodzaju sformułowania brzmią nienaturalnie, gdyż związek sposobu głosowania z płcią można poddać oglądowi niejako „w całości" prezentując go w postaci tablicy. Bliższe przyjrzenie się problemowi pozwala jednak zauważyć, że zachowanie wyborcy przed urną oraz zachowanie klienta przed półką sklepową mają przynajmniej niektóre mechanizmy wspólne – zaś najbardziej widoczna różnica sprowadza się do liczby alternatyw spośród których dokonywany jest wybór. Tablice i algorytmy eksploracji danych stanowią więc podejścia komplementarne, pomiędzy którymi granica jest płynna.

W badaniach marketingowych stosuje się też pewne elementy strategii eksploracji w sytuacjach, gdy dane przedstawiane są za pomocą tablic. W wielu badaniach uwaga zleceniodawcy koncentruje się na tym, co przede wszystkim różnicuje preferencje konsumentów wobec marek czy wariantów produktów. Przygotowując raport pracownicy agencji badawczej przeglądają więc tysiące tablic krzyżujących cechy metryczkowe bądź segmentacje z postawami i zachowaniami konsumenckimi, starając się wyselekcjonować prawidłowości najbardziej znaczące. W poszukiwaniach tych korzysta się ze wskaźników wspomagających identyfikację najbardziej znaczących prawidłowości w tablicach (przykład tego rodzaju wskaźnika przedstawiony został w podrozdziale 4.10). Oprogramowanie stosowane obecnie nie pozwala jeszcze na „automatyzację" tej czynności, czyli wykonanie jej na danych surowych, bez konieczności produkowania tysięcy tablic. Należy jednak oczekiwać, że luka ta zostanie szybko wypełniona, przez co niektóre z dotychczasowych metod analizy tablic zastąpione zostaną algorytmami eksploracji danych.

przetwarzania danych w pamięci operacyjnej komputera. Z postaci wirtualnej korzysta się w komputerowych procedurach analizy tablic – obliczając na przykład wartość statystyki chi-kwadrat czy współczynnika korelacji. Obliczenie wartości tych i wielu innych parametrów tablicy nie wymaga jej fizycznego skonstruowania, stąd też nie robi się tego, aby w procedurach obliczeniowych uniknąć niepotrzebnych kroków. Z postaci wirtualnej tablic korzystają również coraz częściej stosowane algorytmy eksploracji danych (zob. ramka 2.2).

Gdy jednak celem jest **interpretacja zjawiska**, wtedy należy utworzyć tablicę w docelowym kształcie. Dopóki dane dotyczące zachowań wyborczych kobiet i mężczyzn nie zostały przedstawione w postaci tablicy, tak jak ma to miejsce w tabeli 2.3, dopóty trudno było zauważyć, że częstsze występowanie pewnych zachowań wśród kobiet może wynikać stąd, że było ich więcej wśród badanych osób.

2.2 Utworzenie tabeli prezentującej badane zjawisko

W poprzednim podrozdziale omówiliśmy zasady odpowiedniości między komputerowym zapisem danych a ich reprezentacją w postaci tabelarycznej. Aplikacja komputerowa jest w stanie dostarczyć badaczowi tablicę w formie równoważnej prezentowanej w tabeli 2.3. Postać ta ma dwie cechy charakterystyczne. Po pierwsze, obejmuje wszystkie rekordy zbioru danych (o ile nie zastosowano filtrów). Po drugie, kategorie obu cech ułożone są w kolejności symboli zapisu danych w pliku. Omawiana postać stanowi więc punkt wyjścia, czy też pewien półprodukt, z którego badacz może dopiero utworzyć, czy zbudować tabelę przeznaczoną do prezentacji danego zjawiska. Kolejne kroki prowadzące do tego celu omówione zostaną w kolejnych fragmentach tego podrozdziału.

2.2.1 Zawężenie zbiorowości badanych osób

Aby nie wprowadzać nowych elementów, pozostańmy przy omawianych wynikach badania Europejski Sondaż Społeczny, dotyczących zachowań wyborczych mężczyzn i kobiet w wyborach do sejmu we wrześniu 2005 roku (tabela 2.3). Przyjmijmy, że przedmiotem prezentacji jest porównanie sposobu głosowania mężczyzn i kobiet w tych wyborach.

Tabela 2.3 została utworzona drogą przekształcenia danych obejmujących wszystkich respondentów Europejskiego Sondażu Społecznego z 2006 roku. Obejmuje więc również osoby, które nie brały udziału w wyborach. Ponieważ rozpatrywany problem dotyczy **sposobu głosowania** mężczyzn i kobiet – więc

należy zastanowić się, czy osoby, które nie wzięły udziału w wyborach, powinny być w prezentowanej tabeli uwzględnione czy też pominięte.

Podejmując decyzję co do ewentualnego pominięcia pewnych kategorii respondentów warto odwołać się do istoty analizowanego problemu. Część respondentów nie wzięła udziału w wyborach, gdyż we wrześniu 2005 roku nie była do tego uprawniona ze względu na niepełnoletność (próba w Europejskim Sondażu Społecznym obejmuje osoby w wieku od 15 lat). Co do tej kategorii osób nie ma wątpliwości, że w docelowej tabeli powinny być pominięte. Można powiedzieć, że analizowany problem ich „nie dotyczy". Wyselekcjonowanie tych osób nie stanowi problemu, gdyż kategoria „nie byłem uprawniony do głosowania" była uwzględniona w kafeterii odpowiedzi na pytanie o fakt udziału w wyborach (zob. ramka 2.1).

W wypadku pozostałych respondentów, którzy zadeklarowali, że w wyborach nie wzięli udziału, część mogła nie pójść głosować na skutek niesprzyjającymi okoliczności, jak choroba bądź niespodziewany wyjazd. Osoby te należałoby więc wyłączyć z analizowanej zbiorowości, gdyż nie wzięły udziału w wyborach z przyczyn losowych. Idąc dalej, można domniemywać, że pewna kategoria osób nie poszła głosować na skutek obojętności czy braku zainteresowania wyborami. Wyłączenie tej kategorii również wydaje się zasadne, ponieważ rozpatrywany problem dotyczy w istocie tego, w jakim stopniu społeczne role mężczyzn i kobiet warunkują ich zachowania wyborcze. Jeśli ktoś zachowań tych nie przejawia, to pozostaje poza zakresem rozpatrywanego problemu. To tak, jak osób nie pracujących zawodowo nie pyta się o zarobki z pracy.

Na koniec pozostała kategoria osób, które nie poszły głosować w sposób świadomy, traktując to jako manifestację poglądów politycznych. W grę mogą tu wchodzić różnego typu postawy. Na przykład wyborca mógł uznać, że żaden z kandydatów na liście nie reprezentuje w należyty sposób jego oczekiwań. Bądź mógł kontestować ordynację wyborczą uważając, że każdy wyborca powinien mieć prawo dopisania do listy własnych kandydatów.

Osoby, które nie poszły do urn, manifestując w ten sposób swoje postawy, bez wątpienia mieszczą się w ramach rozpatrywanego problemu. Wyobraźmy sobie bowiem, że postawy takie wykazują wyłącznie mężczyźni. Pominięcie ich w analizie nie pozwoli odkryć, że obok bardziej czy mniej licznych grup mężczyzn głosujących na kandydatów poszczególnych partii istniała również grupa, której formuła wyborów nie stworzyła możliwości wyrażenia swoich preferencji. Nie pozwoli również zauważyć, że wśród kobiet postawy takie nie wystąpiły. Oba wnioski byłyby nie tylko odkrywcze dla socjologa, lecz również miałyby kapitalne znaczenie dla strategów partii wystawiających w wyborach kandydatów.

Aby jednak osoby takie włączyć do analiz, w danych muszą istnieć odpowiednie kryteria ich selekcji. Badanie Europejski Sondaż Społeczny nie koncentrowało swojej tematyki na wyborach parlamentarnych z 2005 roku, przez co nie zawierało pytań pozwalających rozstrzygnąć, czy osoby, które nie głosowały w wyborach, manifestowały w ten sposób swoje postawy polityczne, czy też nie uczestniczyły w nich z przyczyn niemerytorycznych. Ze względu na niemożność dokonania podziału osób nie uczestniczących w wyborach na podkategorie, pozostaje wszystkie te osoby wykluczyć z analiz. A zatem, tabela służąca prezentacji sposobów głosowania mężczyzn i kobiet w wyborach do sejmu we wrześniu 2005 roku powstałaby z tabeli 2.3 drogą pominięcia wiersza „66 – nie brał(a) udziału w wyborach".

2.2.2 Łączenie kategorii o niewielkich liczebnościach

Kolejny krok budowy tabeli stanowi rozstrzygnięcie, czy wszystkie kategorie obu cech powinny być zaprezentowane odbiorcy. Kontrargumentem może być fakt, że uwzględnienie danej kategorii dostarczy nieadekwatnego obrazu prezentowanego zjawiska bądź zmniejszy czytelność komunikatu na skutek pochłonięcia uwagi przez mniej istotne jego aspekty.

Powróćmy do wyników prezentowanych w tabeli 2.3. Tylko jeden respondent stwierdził, że swój głos oddał na kandydata partii 02 – Dom Ojczysty. Respondent ten był mężczyzną. Gdyby wnioskować na tej podstawie o płci osób głosujących na Dom Ojczysty, to należałoby przyjąć, że na kandydatów tej partii głosowali wyłącznie mężczyźni. Wniosek ten jest oczywiście nieuprawniony, gdyż na podstawie jednego przypadku nie można wnioskować o rozkładzie cechy mającej dwie lub więcej wartości. Decyduje to o pominięciu Domu Ojczystego w prezentacji sposobów głosowania mężczyzn i kobiet w wyborach do sejmu w 2005 roku. Z tego samego powodu przesądzone jest również pominięcie Ogólnopolskiej Koalicji Obywatelskiej, na którą również tylko jeden respondent oddał głos.

Przypadek Platformy Janusza Korwina-Mikke nie jest już tak oczywisty. Wydaje się, że trzech respondentów, którzy oddali głos na kandydata tej partii to za mało, aby cokolwiek wnioskować na temat udziału mężczyzn i kobiet wśród głosujących na Janusza Korwin-Mikke. Jeśli jednak zdecydować się na pominięcie tej partii w docelowej tabeli, to jak postąpić w wypadku Socjaldemokracji Polskiej (7 respondentów), czy Partii Demokratycznej demokraci.pl, na którą głos oddało 8 badanych osób. Czy je pominąć, czy pozostawić? W jaki sposób wyznaczyć graniczną liczebność kategorii, którą można byłoby uznać za wystarczającą dla trafnego wnioskowania o rozkładzie w tej kategorii drugiej z rozpatrywanych cech.

Ramka 2.3
Wyniki wyborów do Sejmu Rzeczypospolitej Polskiej w dniu 25 września 2005

Zbiorcze wyniki głosowania

Liczba uprawnionych	30 229 031
Liczba wydanych kart	12 263 640
Liczba wyjętych kart	12 255 875
Liczba ważnych kart	12 244 903
Liczba głosów ważnych	11 804 676
Liczba mandatów	460
Frekwencja	40,57%

Liczba oddanych głosów i podział mandatów

Nr listy	Nazwa komitetu wyborczego	Liczba głosów na listę (ważnych)	Liczba głosów na listę (ważnych) [%]	Liczba mandatów
1	Komitet Wyborczy Ruch Patriotyczny	124 038	1,05	
2	Komitet Wyborczy Polska Partia Pracy	91 266	0,77	
3	Komitet Wyborczy Liga Polskich Rodzin	940 762	7,97	34
4	Komitet Wyborczy Partii Demokratycznej – demokraci.pl	289 276	2,45	
5	Komitet Wyborczy Socjaldemokracji Polskiej	459 380	3,89	
6	Komitet Wyborczy Prawo i Sprawiedliwość	3 185 714	26,99	155
7	Komitet Wyborczy Sojusz Lewicy Demokratycznej	1 335 257	11,31	55
8	Komitet Wyborczy Platforma Obywatelska RP	2 849 259	24,14	133
9	Komitet Wyborczy Polskiej Partii Narodowej	34 127	0,29	
10	Komitet Wyborczy Polskiego Stronnictwa Ludowego	821 656	6,96	25
11	Komitet Wyborczy Centrum	21 893	0,19	
12	Komitet Wyborczy Platformy Janusza Korwin-Mikke	185 885	1,57	
13	Komitet Wyborczy Wyborców – Ogólnopolska Koalicja Obywatelska	16 251	0,14	
14	Komitet Wyborczy Polskiej Konfederacji – Godność i Praca	8 353	0,07	
15	Komitet Wyborczy Samoobrona Rzeczpospolitej Polskiej	1 347 355	11,41	56
16	Komitet Wyborczy Partii Inicjatywa RP	11 914	0,1	
17	Komitet Wyborczy Dom Ojczysty	32 863	0,28	
18	Komitet Wyborczy Narodowego Odrodzenia Polski	7 376	0,06	
19	Komitet Wyborczy Stronnictwa Pracy	1 019	0,01	
19	Komitet Wyborczy Wyborców Społeczni Ratownicy	982	0,01	
19	Komitet Wyborczy Wyborców „Mniejszość Niemiecka”	34 469	0,29	2
19	Komitet Wyborczy Wyborców „Mniejszości Niemieckiej Śląska”	5 581	0,05	
	OGÓŁEM	11 804 676	100,00	460

Źródło: Komunikat Państwowej Komisji Wyborczej (PKB 2005).

Niekiedy do problemu łatwiej podejść z drugiej strony, zastanawiając się nie nad tym, które kategorie należy pominąć, lecz nad tym, które powinny pozostać w docelowej tabeli. Dla dokonania tego rodzaju rozstrzygnięć pomocne mogą być informacje o rozkładzie danej cechy w populacji. Nie zawsze tego rodzaju dane są dostępne, lecz w tym wypadku akurat są. Stanowią je oficjalne wyniki wyborów, podane w komunikacie Państwowej Komisji Wyborczej. Wyniki te zostały przedstawione w ramce 2.3.

Spośród 22 partii, które wystawiły kandydatów na listach wyborczych, jedynie 6 przekroczyło próg 5 procent i uzyskało mandaty[1]. W Europejskim Sondażu Społecznym na partie te głosowało od 21 do 407 respondentów, przy czym najmniej licznie reprezentowany był elektorat Ligi Polskich Rodzin (21 respondentów) oraz Polskiego Stronnictwa Ludowego (28 respondentów). Jest kwestią dyskusyjną, czy tak niewielkie liczebności pozwalają wyciągnąć trafne wnioski na temat udziału mężczyzn i kobiet wśród głosujących na obie partie. Niemniej jednak, ich pozostawienie w tabeli pozwala w klarowny sposób zdefiniować zasadę jej budowy. Zasada ta brzmiałaby następująco: w tabeli przedstawiono płeć respondentów głosujących na partie, które weszły do Sejmu V Kadencji.

Określenie klarownej zasady budowy tabeli jest zawsze czynnikiem sprzyjającym skuteczności komunikacji badacza z odbiorcą opracowania. Dlatego czynnik ten warto uwzględnić nawet wtedy, gdy wymaga to pozostawienia w tabeli kategorii o niewielkich liczebnościach, w wypadku których obraz przedstawianego zjawiska nie budzi zaufania. Ta ostatnia niedogodność nie ma zresztą aż tak dużego znaczenia, gdyż przekaz i tak koncentruje się na tych elementach zjawiska, które są najliczniej reprezentowane. W analizowanym przykładzie były to partie, które zdobyły najwięcej mandatów, gdyż ich rola w Sejmie V Kadencji była największa. Poza tym twórca tabeli ma do dyspozycji szereg środków, za pomocą których może przestrzec odbiorcę przed wyciąganiem ryzykownych wniosków w wypadku kategorii o niewielkich liczebnościach. Zostaną one podane przy omawianiu kolejnych zasad budowy tabel.

Jeśli przyjąć, że w tabeli pozostanie jedynie 6 partii, których kandydaci weszli w skład Sejmu V Kadencji, to powstaje pytanie, co zrobić z partiami, które w docelowej tabeli nie zostaną uwzględnione. Zabiegiem stosowanym w takich sytuacjach najczęściej jest połączenie wszystkich pomijanych kategorii w jedną kategorię zbiorczą, którą w rozważanym przykładzie nazwać można „partie, które brały udział w wyborach, lecz nie weszły do Sejmu".

[1] Bez uwzględnienia Mniejszości Niemieckiej, która ma zagwarantowaną określoną liczbę mandatów niezależnie od przekroczenia progu 5 procent.

Po dokonaniu tej operacji oraz po usunięciu osób nie głosujących wyjściowa tablica przybierze postać przedstawioną w tabeli 2.4.

Warto mieć świadomość, że zabieg łączenia części partii w kategorię zbiorczą nie ma swojego odpowiednika empirycznego. Każda z partii w wyborach występowała osobno, posiadała swój własny program polityczny, wystawiała odrębnych kandydatów. W fazie kampanii wyborczej nie da się wyodrębnić niczego, co stanowiłoby agregat tych partii. Kategoria zbiorcza nie wnosi więc merytorycznych informacji do analizy sposobu głosowania mężczyzn i kobiet.

Tabela 2.4
Liczba mężczyzn i kobiet głosujących w różny sposób w wyborach do Sejmu we wrześniu 2005 roku w badaniu Europejski Sondaż Społeczny 2006

symbol	sposób głosowania	mężczyźni	kobiety	ogółem
03	Liga Polskich Rodzin (R. Giertych)	8	13	21
09	Platforma Obywatelska (D. Tusk)	128	152	280
13	Polskie Stronnictwo Ludowe (W. Pawlak)	19	9	28
14	Prawo i Sprawiedliwość (J. Kaczyński)	191	216	407
16	Samoobrona RP (A. Lepper)	51	33	84
18	Sojusz Lewicy Demokratycznej (W. Olejniczak)	39	46	85
	partie, które brały udział w wyborach, lecz nie weszły do Sejmu	10	13	23
77	odmowa odpowiedzi	11	12	23
88	nie pamięta, na kogo głosował(a)	37	72	109
99	brak odpowiedzi	0	1	1
	ogółem	494	567	1061

Powstaje pytanie, czy utworzonej z pozostałych partii kategorii zbiorczej nie należałoby usunąć z tabeli, tak jak usunięta została kategoria osób, które nie głosowały w wyborach. Jest jednak powód, dla którego decyzja taka przyniosłaby więcej szkody niż pożytku. Od badacza oczekuje się, aby przedstawił w tabeli całościowy obraz danego zjawiska (Marsh 1988: 139–141). Jeśli pewnych kategorii brakuje, to odbiorca może przypuszczać, że badacz z jakiś powodów nie chce ujawnić całości zróżnicowania danej cechy. Zamiast skupić uwagę na zasadniczym przesłaniu komunikatu zaczyna zastanawiać się, co też tam takiego wyszło, że pewne kategorie nie zostały w tabeli przedstawione. Obniża to perswazyjność komunikatu. W rozpatrywanym przykładzie lepiej jest więc pozostawić w tabeli kategorię zbiorczą, mimo że udziały mężczyzn

i kobiet w tej kategorii nie stanowią podstawy do wnioskowania na temat płci osób głosujących na dowolną z partii zgrupowanych w tej kategorii.

2.2.3 Uwzględnienie bądź pominięcie kategorii rezydualnych

Obok kategorii odpowiadających deklaracjom respondentów co do głosowania na kandydatów określonych partii, w tabeli 2.4 znajdują się kategorie opisujące pozostałe reakcje badanych na zadane pytanie. W 23 wywiadach respondent odmówił ujawnienia ankieterowi, na kogo głosował. Reakcje 109 respondentów ankieterzy zaklasyfikowali do kategorii „nie pamięta, na kogo głosował(a)". W jednym wypadku reakcja respondenta nie jest znana, gdyż nie została zakodowana przez ankietera. Powodem mogło być przerwanie wywiadu przed zadaniem pytań o uczestnictwo w wyborach.

Wystąpienie omawianych reakcji jest typowe dla badań realizowanych techniką wywiadu kwestionariuszowego, które zakładają dobrowolność współpracy badanego i jego prawo do nie udzielenia odpowiedzi na dowolne z zadawanych pytań. Ponadto, Europejski Sondaż Społeczny realizowany był rok po wyborach, więc nie można wykluczyć, że część respondentów po prostu nie była w stanie przypomnieć sobie, na kogo oddali głos. Tłumaczyłoby to dużą liczbę reakcji zakodowanych jako „nie pamięta, na kogo głosował(a)". Reakcje należące do tej grupy mogą również stanowić przejawy ukrytych odmów. Przez rok czasu między wyborami a badaniem rzeczywistość polityczna uległa przeobrażeniom. Niektórzy respondenci mogli poczuć dyskomfort, ujawniając ankieterowi swoje decyzje wyborcze sprzed roku, o ile zyskali poczucie, że nie były one trafne w obliczu dokonujących się zmian. Z drugiej strony niektóre z osób, które odmówiły odpowiedzi, mogły rzeczywiście nie pamiętać na kogo głosowały. Jeśli zadane pytanie wywołało zażenowanie, że nie pamięta się tak istotnej sprawy, to odmowa odpowiedzi stwarzała możliwość wyjścia z twarzą z całej sytuacji. Można więc przyjąć, że reakcje respondentów klasyfikowane do różnych kategorii mogą mieć podobne przyczyny, jak niechęć udzielenia odpowiedzi albo nie pamiętanie faktów. Dlatego wydaje się uzasadnione **połączenie wszystkich trzech kategorii w jedną**.

Omawiane kategorie nazywa się też rezydualnymi. Odpowiadają bowiem reakcjom pozostającym na obrzeżach problemu, który stanowił przedmiot zadanego pytania. Nie dostarczają informacji o tym, na kandydata której partii respondent oddał głos. Zarazem jednak wiadomo, że respondent w wyborach uczestniczył, oddał głos w sposób świadomy, lecz z różnych powodów nie chce o tym mówić. Owa dwoistość kategorii rezydualnych powoduje, że decyzje o tym, czy kategorie te pominąć, czy pozostawić w tabeli, należą do najtrudniejszych.

Kategorie rezydualne można pominąć tylko wtedy, gdy nie naruszy to istoty prezentowanego zjawiska. W pierwszej kolejności warto rozważyć częstości ich występowania. Jeśli udział danej kategorii rezydualnej jest niewielki – w porównaniu z częstościami występowania kategorii merytorycznych – to można przyjąć, że ma relatywnie niewielki wpływ na obraz zjawiska jako całości.

Z wielkości przedstawionych w tabeli 2.4 wynika, że kategorie rezydualne obejmują w sumie 133 respondentów. Jest to trzecia co do wielkości liczebność, po liczbie osób, które zadeklarowały, że głosowały na Prawo i Sprawiedliwość (407 osób) oraz po liczbie osób deklarujących fakt głosowania na Platformę Obywatelską (280 osób). Gdyby tak duża liczba respondentów wymieniła partię, na kandydata której oddała głos, to nie można wykluczyć, że rozkład głosów na poszczególne partie uległby znaczącym modyfikacjom. Zaś gdyby kategoria respondentów, którzy nie ujawnili, na którą partię głosowali, była nieliczna, to niebezpieczeństwo pominięcia istotnych aspektów opisywanego zjawiska byłoby odpowiednio mniejsze.

Tabela 2.5
Odsetki respondentów Europejskiego Sondażu Społecznego 2006 deklarujących głosowanie na poszczególne partie w wyborach do Sejmu w 2005 roku oraz odsetki głosów uzyskanych przez poszczególne partie według oficjalnych wyników wyborów
[*w procentach*]

wskaźnik	PiS	PO	SLD	Samo-obrona	PSL	LPR	pozostałe partie	odmowa nie pamięta	ogółem
wyniki wyborów	27	24	11	11	7	8	11	–	100
deklaracje badanych	38	26	8	8	3	2	2	13	100

W wypadku Europejskiego Sondażu Społecznego uwzględniono wyłącznie osoby, które zadeklarowały, że głosowały w wyborach. Oficjalne wyniki wyborów pochodzą z komunikatu Państwowej Komisji Wyborczej (PKW 2005).

Podejmując decyzję o włączeniu lub pominięciu w tabeli kategorii rezydualnych, warto odwołać się do zewnętrznych źródeł wiedzy na temat badanego zjawiska, o ile źródła takie są dostępne. W omawianym przykładzie źródło dodatkowej wiedzy stanowią wyniki wyborów (ramka 2.3). W tabeli 2.5 porównano rozkład odpowiedzi na pytanie o partię, na którą badany głosował w wyborach parlamentarnych w 2005 roku z komunikatem Państwowej Komisji Wyborczej. Z przeprowadzonego porównania wynika, że badani wykazywali skłonność podawania tych partii, które w wyborach zdobyły najwięcej głosów. Fakt głosowania na Prawo i Sprawiedliwość (PiS) zadeklarowało 38 procent badanych, podczas gdy według Państwowej Komisji Wyborczej kandydaci tej

partii uzyskali w wyborach 27 procent głosów. Z drugiej strony badani nie przyznawali się do głosowania na partie, które w wyborach uzyskały mniejsze odsetki głosów. Na przykład głosowanie na Ligę Polskich Rodzin (LPR) zadeklarowało 2 procent badanych, natomiast faktycznie głosy na tę partię oddało 8 procent wyborców.

Uzyskany rezultat pozwala sformułować przypuszczenie, że osoby, które w badaniu nie podały sposobu swojego głosowania, częściej głosowały na kandydatów tych partii, które uzyskały w wyborach mniejsze odsetki głosów. Płynie stąd w każdym razie wniosek, że wobec osób, które nie ujawniły lub nie chciały w badaniu ujawnić, na kandydata której partii oddały głos, nie można przyjąć założenia, że osoby te głosowały w sposób zbliżony do tych, które podały swój sposób głosowania. Aby więc zachować pełny obraz prezentowanego zjawiska kategorię rezydualną należy w tabeli pozostawić.

Na marginesie omawianej kwestii warto wspomnieć, że znaczną część zadawanych w kwestionariuszach pytań stanowią pytania o opinie, w których odpowiedzi respondenci wyrażają za pomocą określeń przedłożonej skali – na przykład „zdecydowanie tak", „raczej tak", „raczej nie", „zdecydowanie nie". Pytania te z zasady dają możliwość uniknięcia przez badanego sformułowania jednoznacznego sądu, w postaci odpowiedzi w rodzaju „ani tak, ani nie", czy „trudno powiedzieć". Tego rodzaju odpowiedzi nie stanowią kategorii rezydualnych w sensie przedstawionym w tym podrozdziale. Odpowiadają bowiem sytuacjom, w których badany nie akceptuje założeń pytania bądź założeń skali. Na przykład uważa, że istnieją zarówno argumenty na tak, jak i na nie. Bądź też nie rozpatruje danej kwestii w sposób proponowany w pytaniu. W badaniach na ogół przyjmuje się, że brak opinii jest również opinią. W badaniach przedwyborczych – oprócz grup respondentów deklarujących gotowość oddania głosu na jednego z kandydatów – zawsze wyodrębnia się kategorię niezdecydowanych. Dla strategów kandydujących partii właśnie ta kategoria stanowi jeden z cenniejszych wyników badania, gdyż do niej należy kierować działania zmierzające do poszerzenia elektoratu danego kandydata czy danej partii.

W każdym razie, tworząc tabele, nie należy pomijać w nich odpowiedzi w rodzaju „nie wiem", czy „trudno powiedzieć". Przynajmniej we wstępnej fazie analiz, gdy jeszcze nie zostało rozstrzygnięte, jaki wpływ na wyniki analiz miałoby ewentualne pominięcie tych kategorii (Domański 2000). Badania respondentów udzielających tego typu odpowiedzi dowodzą ponadto, że stanowią oni specyficzną i zarazem ciekawą grupę (Lutyńska 1999). Jej uważna analiza może niekiedy istotnie wzbogacić obraz badanego zjawiska w porównaniu z sytuacją, w której ograniczono by uwagę do osób mających sprecyzowane sądy.

2.2.4 Prezentacja danych ważonych

W polach dotychczasowych tabel prezentowano liczbę respondentów badania Europejski Sondaż Społeczny o określonej kombinacji płci i sposobu głosowania. Wielkości te wyrażały się liczbami całkowitymi, gdyż dotyczyły osób. W większości współcześnie prowadzonych badań do praktyki należy ważenie badanych osób przez odpowiednio dobrane współczynniki – zwane wagami. Cele wprowadzania wag bywają dwojakie. Z jednej strony ich zadaniem jest kompensacja niejednakowych prawdopodobieństw doboru respondentów przyjętych w schemacie doboru próby. Z drugiej zaś kompensacja niejednakowych ubytków w poszczególnych kategoriach wylosowanych osób powstających w fazie realizacji badania. W konkretnym badaniu wagi mogą pełnić jedną z tych funkcji bądź równocześnie obie (Billet 2007).

W badaniu Europejski Sondaż Społeczny 2006 – podobnie jak to ma miejsce w innych realizowanych obecnie badaniach – odsetki odmów okazały się niejednakowe w miejscowościach różnej wielkości. Na wsiach ankieterom udało się zrealizować ponad 78 procent wylosowanej próby, zaś w dużych miastach (ponad 100 tys. mieszkańców) jedynie 58 procent (Sawiński 2007c: 50). Udostępniając wyniki badania dokonano kompensacji niejednakowych współczynników realizacji (*response rate*) ważąc dane mieszkańców miejscowości różnej wielkości przez odpowiednie współczynniki. Zastosowano przy tym normalizację wag do liczby osób, z którymi zrealizowano wywiady, to jest do 1721. Tym samym suma wag jest równa tej wielkości. Przykłady zastosowanych wag zostały przedstawione na rycinie 2.1. Mają one postać ułamków dziesiętnych. W rezultacie, liczebności w tabelach skonstruowanych po uwzględnieniu ważenia wyrażają się w postaci liczb niecałkowitych.

W tabeli 2.6 przedstawiono liczbę mężczyzn i kobiet głosujących na poszczególne partie obliczoną z uwzględnieniem wag. Porównanie tych liczebności z faktyczną liczbą respondentów reprezentujących poszczególne kombinacje płci i sposobu głosowania (tabela 2.4) prowadzi do wniosku, że różnice są niewielkie i w większości pól tabeli nie przekraczają ±1. Nie należy jednak wyciągać stąd wniosku, że jest rzeczą obojętną, która postać danych wykorzystana zostanie do budowy tabeli. W pracy z danymi przyjmuje się generalną zasadę, że jeśli w zbiorze danych podana została waga, to wagę tę należy uwzględnić. Wynika to stąd, że wiedza użytkownika danych na temat badania jest na ogół uboższa od wiedzy jego autora. Użytkownik może po prostu nie znać pewnych szczegółów dotyczących schematu doboru próby czy poziomu realizacji w poszczególnych warstwach, przez co może nie być świadomy faktu, że dane bez uwzględnienia wag dostarczają skrzywionego (ang. *biased*) obrazu badanej populacji.

Tabela 2.6
Liczba mężczyzn i kobiet głosujących na poszczególne partie wyborach do Sejmu we wrześniu 2005 roku w badaniu Europejski Sondaż Społeczny 2006
Liczebności ważone

symbol	sposób głosowania	mężczyźni	kobiety	ogółem
03	Liga Polskich Rodzin (R. Giertych)	7,761481	13,050642	20,812123
09	Platforma Obywatelska (D. Tusk)	127,649678	153,182506	280,832184
13	Polskie Stronnictwo Ludowe (W. Pawlak)	18,771828	8,746199	27,518027
14	Prawo i Sprawiedliwość (J. Kaczyński)	191,141276	214,710007	405,851283
16	Samoobrona RP (A. Lepper)	51,234603	32,979093	84,213696
18	Sojusz Lewicy Demokratycznej (W. Olejniczak)	38,648941	46,861243	85,510184
	partie, które brały udział w wyborach, lecz nie weszły do Sejmu	9,709097	13,445322	23,154418
77–99	nie pamięta/odmowa odpowiedzi	48,722587	85,219228	133,941815
	ogółem	493,639491	568,194240	1061,833731

Liczebności podano z dokładnością do 6 cyfr po przecinku.

Niebezpieczeństwo to można wyjaśnić na przykładzie badań CATI realizowanych za pośrednictwem telefonów stacjonarnych. Losowany jest numer telefonu, a więc gospodarstwo domowe, w którym następnie przeprowadza się wywiad z jedną z osób. Ważenie wyniku przez łączną liczbę osób w gospodarstwie należących do badanej populacji staje się w tej sytuacji koniecznością. Bez tej operacji opinie osób żyjących samotnie zdominowałyby opinie osób mieszkających w gospodarstwach wieloosobowych. Gdyby zaś to samo badanie prowadzone byłoby nie przez telefony stacjonarne, a przez komórkowe – omawiany problem nie pojawiłby się. Telefon komórkowy jest bowiem dobrem użytkowanym indywidualnie, toteż losując numer telefonu komórkowego losujemy osobę. Podobna różnica występuje między próbą adresową a próbą imienną (Sawiński 2007b: 121–122; Treiman 2009: 204–205).

Poważną niedogodność danych ważonych stanowi fakt, że wszystkie obliczone wielkości i wskaźniki wyrażają się liczbami niecałkowitymi. Ogląd tabeli 2.6 prowadzi do nieuchronnego wniosku, że jej zawartość jest dużo mniej przejrzysta niż prezentowanej wcześniej tabeli 2.4, w której podane wielkości wyrażają się liczbami naturalnymi. Co więcej, język opisu wyników badań nie jest należycie dostosowany do prezentacji wielkości empirycznych wyrażonych liczbami niecałkowitymi. Wygłoszenie podczas wykładu stwierdzenia w rodzaju: „w Europejskim Sondażu Społecznym dwieście czternaście i sie-

demset dziesięć tysięcznych kobiety głosowało na PiS" wywołałoby na sali konsternację, o ile nie salwę śmiechu. Nawyk przekładania wyników badania na policzalnych respondentów jest bardzo silnie zakorzeniony – z czym pozostaje się pogodzić.

Dlatego prezentując w tabeli liczebności ważone, zasadne staje się ich zaokrąglenie do liczb całkowitych. Uniknie się w ten sposób trudności z opisaniem czy wyjaśnieniem prezentowanego zjawiska. Wtedy stwierdzenie „wśród badanych 215 kobiet głosowało na PiS" stanie się komunikatywne, mimo że – dosłownie potraktowane – nie jest prawdziwe. Faktycznie respondentek głosujących na PiS było bowiem 216 (tabela 2.4). Rozbieżność wynika stąd, że 215 nie jest faktycznie otrzymanym wynikiem, lecz jego estymacją. Formalnie poprawne zdanie opisujące dane ważone powinno w tym wypadku brzmieć następująco: „gdyby szanse wejścia do próby były dla każdego jednakowe oraz gdyby odmowy udziału w badaniu były niezależne od badanych cech, to wtedy należałoby się spodziewać, że w pytaniu P5 215 kobiet zadeklaruje oddanie głosu na kandydata PiS". W naukach społecznych przyjmuje się jednak, że odbiorca komunikatu rozumie na czym polega ważenie danych, a przynajmniej jest skłonny zaakceptować, że dane ważone stanowią lepsze przybliżenie rzeczywistości od danych nieważonych. Dlatego dopuszczalna jest konwencja przedstawiania wyników ważonych w języku opisu wyników empirycznych. Tym niemniej, na wszelki wypadek w nocie pod tabelą warto umieścić wyjaśnienie, że prezentowane liczebności zostały zaokrąglone do liczb całkowitych. Podobnie jak warto zaznaczyć, że przedmiot prezentacji stanowią wyniki ważone.

O ile jednak w prezentacji wystarczy przedstawić wyniki ważone, o tyle w fazie budowy tabeli należy przyjrzeć się również danym nieważonym. Wagi bywają niekiedy silnie zróżnicowane, co zależy od schematu doboru próby i jakości realizacji badania. Zdarzają się badania, w których relacja najwyższej wagi do najniższej przekracza 1000. Może to prowadzić do błędnych decyzji w kwestii rezygnacji z pewnych kategorii bądź łączenia kategorii jako mało licznych. Przypuśćmy bowiem, że w pewnej kategorii po przeważeniu i zaokrągleniu do liczby całkowitej jest 8 osób. Na tej podstawie badacz może podjąć decyzję, iż kategoria ta jest zbyt mało liczna, aby przedstawić ją w tabeli. Tymczasem faktyczna liczba respondentów w tej kategorii może być znacznie większa i wynosić na przykład ponad 100 osób – co na ogół uznaje się jako wystarczającą podstawę do uwzględnienia kategorii w tabeli. Zależność w drugą stronę jest także możliwa. Po przeważeniu danych liczebność pewnej kategorii może wynieść ponad 100, podczas gdy faktyczna liczba osób badanych, zaliczonych do tej kategorii, wynosi na przykład dwie.

Oprócz opisanego niebezpieczeństwa warto również uwzględnić fakt, że w niektórych bazach danych jedyna dostępna waga ma postać tak zwanej esty-

macji populacyjnej. W tym wypadku funkcja ważenia przelicza badane osoby, których jest na ogół kilkaset bądź kilka tysięcy, na wielkości populacyjne – czyli na zbiorowość liczącą kilka bądź kilkanaście milionów osób. W rezultacie wielkości ważone nie dają żadnych wskazówek co do liczby respondentów w poszczególnych kategoriach uwzględnionych cech.

Z podanych powodów dokonując łączenia kategorii bądź rezygnując z pewnych kategorii jako mało licznych, najlepiej odwołać się do faktycznej liczby badanych jednostek w tych kategoriach.

2.2.5 Kwestia umieszczenia cechy w boczku lub w główce tabeli

Ważną kwestią jest decyzja, którą z cech umieścić w wierszach, a którą w kolumnach tabeli. W praktyce najczęściej w wierszach umieszcza się cechę, która warunkuje drugą. Regułą tą można posłużyć się, gdy między właściwościami odpowiadającymi cechom zachodzi zależność przyczynowa bądź następstwo w czasie. W analizowanym przykładzie płeć potencjalnie może mieć wpływ na oddanie głosu na kandydata określonej partii, natomiast zależność w drugą stronę jest pozbawiona sensu. Kierując się omawianym kryterium płeć należałoby umieścić w wierszach, czyli w boczku tabeli, zaś w główce tabeli – czyli w kolumnach – należałoby umieścić informację o przynależności partyjnej kandydata, na którego badana osoba oddała głos.

Tablica otrzymana „z komputera" – przedstawiona jako tabela 2.6 – wymaga więc przeformatowania polegającego na zamianie miejscami wierszy i kolumn. Tego rodzaju przeformatowanie nazywane jest **transpozycją** tablicy. Tablica po transpozycji przedstawiona została jako tabela 2.7.

Tabela 2.7
Liczba kobiet i mężczyzn głosujących na kandydatów poszczególnych partii w wyborach do Sejmu we wrześniu 2005 roku w badaniu Europejski Sondaż Społeczny 2006

	sposób głosowania								
płeć	LPR	PO	PSL	PiS	Samo-obrona	SLD	pozostałe partie	odmowa lub nie pamięta	ogółem
mężczyźni	8	128	19	191	51	39	10	49	494
kobiety	13	153	9	215	33	47	13	85	568
ogółem	21	281	28	406	84	86	23	134	1062

Uwzględniono wyłącznie osoby, które zadeklarowały, że głosowały w wyborach. Dane ważone. Liczebności zaokrąglone do liczb całkowitych.

Proponowana konwencja rozmieszczenia cech w tabeli odwołuje się do natury mechanizmów percepcyjnych, zorientowanych na analizę informacji w wymiarze horyzontalnym w kierunku od lewej do prawej. Rozpoczynając przeglądanie tabeli od cechy umieszczonej w boczku, odbiorca przechodzi do wnętrza tabeli, oczekując tam zobrazowania skutków działania tej cechy. Zarazem porównywanie ze sobą wielkości umieszczonych w tym samym wierszu jest bardziej naturalne od porównywania wielkości w tej samej kolumnie (Ehrenberg 1981; Lang i Secic 2006: 332). Decyduje to o kolejności czytania danych w tabeli, co można wykorzystać w celu przekazania określonych interpretacji.

Warto też wspomnieć o sytuacji, gdy na określony układ cech w tabeli autor decyduje się, biorąc pod uwagę format publikacji. Formaty książek i czasopism charakteryzują się tym, że wysokość strony na ogół jest większa od szerokości, toteż bardziej „elegancko" prezentują się w nich tabele, które mają więcej wierszy niż kolumn. W tym miejscu warto odwołać się do zasady – stanowiącej motto książki – że tabela służy do prezentacji obrazu zjawiska, nie zaś do przechowywania danych. W związku z tym wszystkie rozwiązania techniczne powinny być podporządkowane temu celowi. Gdy po wydrukowaniu tabela nie mieści się na szerokości strony, to albo trzeba wydrukować ją na wklejce, albo poświęcić na nią osobną stronę, odwracając w druku o 90 stopni.

2.2.6 Ustalenie kolejności kategorii

Określone rozwiązania przyjmuje się również w odniesieniu do kolejności rozmieszczenia w tabeli kategorii każdej z cech. Porządek ten zależy od tego, na czym badacz pragnie zogniskować uwagę odbiorcy. Najbardziej istotne kategorie cechy w boczku tabeli umieszcza się w pierwszym i w kolejnych wierszach, natomiast najbardziej znaczące kategorie cechy w główce tabeli umieszcza się w kolejnych kolumnach, licząc od lewej do prawej. Omawiany porządek znajduje potwierdzenie w badaniach czytelności stron internetowych prowadzonych techniką *eye-trackingu* (Bojko 2006). Użytkownicy rozpoczynają przeglądanie strony internetowej od informacji zawartych w lewym górnym obszarze, a następnie podążają wzrokiem w prawo i w dół. Przy czym szanse dotarcia do najniższych wierszy ekranu, a także do fragmentów położonych na prawym skraju strony – są niewielkie. Jeśli tabela przekazać ma odbiorcy pewne interpretacje danych, to najważniejsze wyniki muszą być skupione w górnych wierszach i w kolumnach umieszczonych z lewej strony.

Partie wyszczególnione w główce tabeli 2.7 można ułożyć w kolejności ustalonej na podstawie liczby wskazań w badaniu. Tego typu rozwiązanie stosuje się, gdy kategorie cechy nie mają naturalnego porządku (Ehrenberg 1981,

1986). Na pierwszym miejscu – w skrajnej kolumnie z lewej strony – należy więc umieścić PiS, który wskazało 406 osób. Następną w kolejności partią będzie PO, wskazane przez 281 osób. Ostatnią z umieszczonych w tabeli partii jest LPR – wymieniony przez 21 osób. Za poszczególnymi partiami umieścić można kategorię zbiorczą, grupującą pozostałe wymienione przez respondentów partie. Ranga kategorii „pozostałe partie" w wyjaśnianiu sposobu głosowania przez płeć jest bowiem mniejsza. Nie można bowiem określić preferencji dla konkretnych partii w ramach tej kategorii, a poza tym partie te miały niewielkie znaczenie dla rezultatów wyborów, ponieważ nie weszły do sejmu. Przyjęte dotąd ustalenia dotyczące aranżacji porządku kategorii przedstawione zostały w tabeli 2.8.

Ostatnia wyodrębniona kategoria obejmuje respondentów, którzy podczas badania nie ujawnili ankieterowi, na kogo głosowali. Jej rola w wyjaśnianiu różnic w sposobie głosowania kobiet i mężczyzn jest co najwyżej uzupełniająca, gdyż nie wskazano w tym wypadku na żadną partię. Kategorie rezydualne na ogół umieszcza się na końcu listy ze względu na ich odmienny charakter.

Cecha umieszczona w boczku tabeli obejmuje tylko dwie kategorie: kobiety i mężczyzn. Kategorie płci nie mają naturalnego porządku, stąd też kolejność obu płci w wierszach tabeli jest poniekąd arbitralna. W przypadku cechy dychotomicznej porządek kategorii nie jest zresztą aż tak ważny, gdyż odbiorca tabeli jest w stanie łatwo ogarnąć wzrokiem i porównać oba wiersze. W tabeli 2.8 na pierwszym miejscu umieszczono kobiety, ponieważ więcej ich uczestniczyło w wyborach parlamentarnych, przez co w większym stopniu zadecydowały o ich wyniku. Jako kryterium uporządkowania przyjęto tutaj kryterium ilościowe, podobnie jak w wypadku partii politycznych.

Tabela 2.8

Liczba kobiet i mężczyzn głosujących na kandydatów poszczególnych partii w wyborach do Sejmu we wrześniu 2005 roku w badaniu Europejski Sondaż Społeczny 2006

Kategorie obu cech uporządkowane

	sposób głosowania								
płeć	PiS	PO	SLD	Samo-obrona	PSL	LPR	pozo-stałe partie	odmowa lub nie pamięta	ogółem
kobiety	215	153	47	33	9	13	13	85	568
mężczyźni	191	128	39	51	19	8	10	49	494
ogółem	406	281	86	84	28	21	23	134	1062

Uwzględniono wyłącznie osoby, które zadeklarowały, że głosowały w wyborach. Dane ważone. Liczebności zaokrąglone do liczb całkowitych.

Porządek prezentacji kategorii w tabeli bywa niekiedy określony przez normy bądź zwyczaje przyjęte w danej branży. W badaniach marketingowych konsekwentnie przestrzega się zasady, aby markę klienta, szczególnie gdy jest on zleceniodawcą badania, umieścić na pierwszym miejscu, to znaczy – w zależności od układu tabeli – w najwyższym wierszu lub w pierwszej kolumnie licząc od lewej strony. Czyni się tak nawet wtedy, gdy pod względem udziału w rynku marka ta ciągnie się w ogonie wszystkich marek. Umieszczenie marki klienta w innym miejscu zmuszałoby go do poszukiwania w tabelach raportu własnej marki pomiędzy innymi, co utrudniałoby odczytanie jej charakterystyk na tle pozostałych marek. Zostałoby to potraktowane jako brak profesjonalizmu agencji badawczej, która sporządziła raport.

2.2.7 Wybór wielkości w polach tabeli pod kątem celu prezentacji

Wszystkie dotychczasowe tabele prezentują liczebności otrzymane w wyniku badania. Wielkości w tej postaci okazują się dogodne do udzielenia odpowiedzi przynajmniej na niektóre z pytań badawczych. Na przykład, liczebności prezentowane w tabeli 2.3 pozwoliły sformułować wyjaśnienie rezultatu polegającego na tym, że wśród badanych osób więcej kobiet niż mężczyzn nie wzięło udziału w wyborach. Prawdopodobną przyczyną jest to, że kobiet było w ogóle więcej wśród badanych osób. Generalnie jednak, prezentacja w tabeli liczebności nie sprzyja sprawnemu odczytaniu kształtu badanego zjawiska. Do tego celu lepiej posłużyć się odsetkami obliczonymi w wierszach bądź w kolumnach tabeli.

Tabela 2.9
Odsetki głosujących na kandydatów poszczególnych partii w wyborach do Sejmu we wrześniu 2005 roku wśród kobiet i mężczyzn w badaniu Europejski Sondaż Społeczny 2006

[*w procentach*]

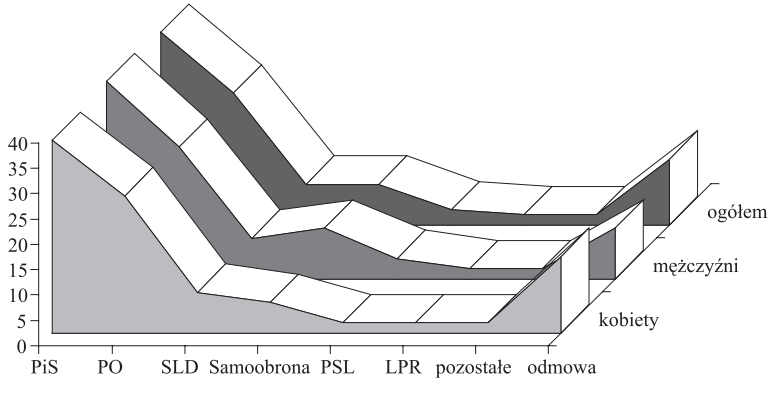
	sposób głosowania									
płeć	PiS	PO	SLD	Samo-obrona	PSL	LPR	pozostałe partie	odmowa lub nie pamięta	ogółem	liczba osób
kobiety	38	27	8	6	2	2	2	15	100	568
mężczyźni	39	26	8	10	4	2	2	10	100	494
ogółem	38	26	8	8	3	2	2	13	100	1062

Uwzględniono wyłącznie osoby, które zadeklarowały, że głosowały w wyborach. Dane ważone (tabela 2.6). Odsetki zaokrąglono do liczb całkowitych.

Odsetki informują, jaka część każdej grupy wyznaczonej przez kategorie jednej cechy przejawia zachowania czy wyraża opinie określone przez kategorie drugiej z cech. Odsetki sposobów głosowania obliczone osobno wśród kobiet, wśród mężczyzn, a także wśród wszystkich badanych, przedstawia tabela 2.9. Powstała ona poprzez obliczenie odsetków dla liczebności z tabeli 2.8 w stosunku do sum brzegowych wierszy (podanych również w ostatniej kolumnie tabeli 2.9). Ten sposób prezentacji danych ułatwia bezpośrednie porównanie sposobów głosowania kobiet i mężczyzn. Ułatwia również określenie specyfiki głosowania kobiet bądź mężczyzn na tle całej zbiorowości.

Rozkłady odsetków w wierszach lub kolumnach tabeli nazywane są też ich **profilami**. Termin ten zaproponowano na gruncie metody zwanej analizą korespondencji (rozdział 7). Chodziło o nadanie interpretacji wizualnej czy przestrzennej odsetkom podawanym w tabelach. Pojęcie profilu kojarzy się z kształtem, czy z przybliżonym obrazem, który zatraca szczegóły. Profile mogą różnić się między sobą, mogą się zmieniać, można je ze sobą porównywać. Wydaje się, że termin ten trafnie oddaje istotę analizowania odsetków w tabeli. Dlatego będziemy posługiwać się nim w dalszych rozważaniach. Graficzny obraz profili głosowania kobiet, mężczyzn oraz ogółu osób przedstawia rycina 2.2.

Rycina 2.2
Profile głosowania kobiet, mężczyzn oraz ogółu osób w wyborach do Sejmu we wrześniu 2005 roku



Europejski Sondaż Społeczny 2006

Wróćmy do profili przedstawionych w tabeli 2.9 i na ich podstawie spróbujmy zrekonstruować typowy sposób odczytania zawartości tabeli oraz wnioski uzyskane na tej podstawie. Większość osób analizować będzie zawartość

tabeli wierszami, od lewej do prawej strony – w taki sposób, jak czyta tekst książki. Z wiersza odpowiadającego kobietom typowy odbiorca wyników badania dowie się, że kobiety najczęściej głosowały na PiS – prawie 40 procent, nieco rzadziej na PO – około jednej czwartej, zaś na kandydatów pozostałych partii głosowało co najwyżej po kilka procent kobiet. Z wiersza tego odczytać też można, że 15 procent kobiet nie potrafiło przypomnieć sobie albo nie chciało powiedzieć, na kogo oddały głos. Następnie osoba analizująca zawartość tabeli przejdzie do jej drugiego wiersza, z którego odczyta jak głosowali mężczyźni. W trzecim kroku zestawi zapewne ze sobą oba rodzaje informacji, dochodząc do wniosku, że w gruncie rzeczy kobiety i mężczyźni głosowali podobnie. W kolejnym kroku postara się ustalić, czy są jakieś różnice, a jeśli tak, to wypadku których partii. Dopiero w tym momencie wielkości w tabeli zacznie porównywać wertykalnie. Bez trudu zauważy, że największa różnica w odsetkach kobiet i mężczyzn ma miejsce w kolumnie odpowiadającej Samoobronie. Zauważy też, że w wypadku PSL-u odsetek mężczyzn jest dwa razy większy od odsetka kobiet. Pozostałe różnice uzna zapewne za mniej ważne, kończąc tym samym ogląd tabeli.

Przypuszczalne wnioski z analizy zawartości tabeli 2.9 jej odbiorca sformułuje następująco: „kobiety i mężczyźni w gruncie rzeczy głosowali podobnie, szczególnie w wypadku partii, które zdobyły najwięcej głosów w wyborach. Jedynie na Samoobronę i na PSL mężczyźni głosowali częściej niż kobiety. Różnice te nie mają jednak dużego znaczenia, ponieważ obie partie uzyskały w wyborach niewielkie odsetki głosów".

Forma prezentacji danych zastosowana w tabeli 2.9 pozwala więc odpowiedzieć na pytanie: na ile sposób głosowania zależy od płci. Najprostsza odpowiedź jest taka, że zależy w niewielkim stopniu. Jeśli badacz chciałby przedstawić wpływ płci na sposób głosowania w wyborach na tle innych cech zróżnicowania społecznego, to może skonstruować analogiczne tabele zastępując płeć wykształceniem, miejscem zamieszkania, dochodem, czy kategorią zawodową. Sposób analizy tych tabel będzie przypuszczalnie zbliżony do opisanego wyżej. Zestawienie wniosków z oglądu poszczególnych tabel pozwoli więc odpowiedzieć na pytanie, które z uwzględnionych cech społecznego zróżnicowania bardziej, a które mniej różnicują sposób głosowania w wyborach.

Warto w tym miejscu zastanowić się, do jakich wniosków prowadziłaby analiza tabeli, w której płeć i sposób głosowania zostałyby zamienione miejscami. Dane w tej postaci przedstawiono w tabeli 2.10. Odsetki obliczono w poszczególnych wierszach, co odpowiada horyzontalnemu kierunkowi analizy tabeli.

Odbiorca wyników badania rozpocznie analizę zawartości tabeli od pierwszego wiersza dochodząc do wniosku, że wśród ogółu głosujących jest nieco

więcej kobiet (53%) niż mężczyzn (47%). Następnie przejdzie do analizy odsetków kobiet i mężczyzn wśród głosujących na poszczególne partie. W wypadku PiS-u, PO oraz SLD stwierdzi, że odsetki kobiet głosujących na te partie są zbliżone i wynoszą od 53–55 procent. Konsekwencją jest to, że również odsetki mężczyzn są podobne – od 45 do 47 procent. Łatwo zarazem zauważy, że struktura płci elektoratu tych trzech partii właściwie nie różni się od struktury płci ogółu osób głosujących w wyborach. Przechodząc do wiersza opisującego płeć osób głosujących na Samoobronę stwierdzi, że odsetki kobiet i mężczyzn są odmienne, niż w wypadku poprzednio analizowanych partii. Kobiety stanowiły tylko 39 procent osób głosujących na Samoobronę, co zarazem oznacza, że większość głosów na tę partię oddali mężczyźni. Analiza kolejnego wiersza tabeli pozwoli zauważyć, że analogiczna prawidłowość zachodzi w wypadku PSL. Jeszcze mniejszy odsetek kobiet – niż w przypadku Samoobrony – głosował na kandydatów tej partii. Mężczyzn było ponad dwa razy więcej niż kobiet. W kolejnym wierszu tabeli, odpowiadającym LPR, sytuacja niejako wraca do normy. Wśród głosujących na kandydatów tej partii więcej jest kobiet niż mężczyzn, przy czym odsetek kobiet można uznać za znacząco wyższy niż w wypadku partii w trzech pierwszych wierszach tabeli. Płeć osób głosujących na „pozostałe partie" zapewne nie wywoła już większego zainteresowania,

Tabela 2.10
Odsetki kobiet i mężczyzn wśród głosujących na kandydatów poszczególnych partii w wyborach do Sejmu we wrześniu 2005 roku w badaniu Europejski Sondaż Społeczny 2006

[*w procentach*]

	płeć			
sposób głosowania	kobiety	mężczyźni	ogółem	liczba osób
ogół głosujących	**53**	**47**	**100**	**1062**
PiS	53	47	100	406
PO	54	46	100	281
SLD	55	45	100	86
Samoobrona	39	61	100	84
PSL	32	68	100	28
LPR	63	37	100	21
pozostałe partie	58	42	100	23
odmowa lub nie pamięta	64	36	100	134

Uwzględniono wyłącznie osoby, które zadeklarowały, że głosowały w wyborach. Dane ważone (tabela 2.6). Odsetki zaokrąglono do liczb całkowitych.

ponieważ odsetki są zbliżone do wcześniej oglądanych. Ostatni wiersz tabeli zostanie zapewne uznany za nieistotny, ponieważ *de facto* brak tu informacji, na którą z partii głosowały osoby zaliczone do tej kategorii.

Przypuszczalny wniosek z przeprowadzonej analizy sformułować można następująco: „Elektorat Samoobrony i PSL był specyficzny ze względu na udział kobiet i mężczyzn. Udział mężczyzn był zdecydowanie większy, przy czym stanowili oni 68 procent głosujących na PSL oraz 61 procent głosujących na Samoobronę. W wypadku pozostałych partii struktura elektoratu pod względem płci była podobna. Nieco ponad 50 procent głosujących stanowiły kobiety. Jedynie odsetek kobiet głosujących na LPR okazał się nieco wyższy i wyniósł 62 procent".

Warto zwrócić uwagę, że w przytoczonym rozumowaniu **abstrahuje się od liczby osób** głosujących na poszczególne partie, mimo że liczebności te podane zostały w ostatniej kolumnie tabeli. Wnioski koncentrują się wokół dwóch partii, które w sumie zdobyły niewiele głosów, a więc miały niewielki udział w wybranym Sejmie. Koncentracja na specyfice rozkładu płci osób głosujących na kandydatów tych partii spowodowała, że we wnioskach utracony został obraz zachowań elektoratu jako całości.

Niemniej, do pewnych celów prezentacja danych w tym układzie wydaje się użyteczna. Chociażby z punktu widzenia stratega partii politycznej, na przykład PSL. W centrum zainteresowań stratega leży struktura elektoratu danej partii, nie zaś to, czy płeć różnicuje bardziej, czy mniej sposób głosowania wśród **ogółu** wyborców. Na podstawie analizy tabeli 2.10 strateg PSL wyciągnie wniosek, że komunikacja programu partii w fazie kampanii przedwyborczej trafiła przede wszystkim do mężczyzn, podczas gdy w partiach, które pozyskały najwięcej wyborców, do mężczyzn i do kobiet. Gdyby na PSL głosowało co najmniej tyle kobiet co mężczyzn, to partia ta zyskałaby wyraźnie więcej głosów. Nie wnikając w tym miejscu, czy i na ile komunikacja programu PSL wśród kobiet mogłaby okazać się skuteczna, pozostańmy przy stwierdzeniu, że podczas analizy niektórych problemów zasadna jest zamiana miejscami wierszy i kolumn (transpozycja tablicy), mimo że cecha warunkowana znajdzie się wtedy w boczku, a cecha ją warunkująca w główce tabeli.

Aby uzyskać odsetki podane w tabeli 2.10, nie trzeba było zamieniać cech miejscami, lecz wystarczyło obliczyć odsetki w kolumnach, zamiast w wierszach. Aczkolwiek formalnie obie metody są równoważne, to w fazie prezentacji wyników tabela w odsetkami obliczonymi w wierszach jest prostsza do interpretacji, gdyż sposób czytania jej zawartości jest zgodny z nawykami ukształtowanymi przy czytaniu tekstów. Dlatego lepiej jest procentować tabele wierszami – nawet jeśli wymaga to umieszczenia w wierszach cechy warunkowanej – zamiast prezentować odsetki obliczone w kolumnach.

Aby porównać prezentowane w tabeli wielkości w efektywny sposób, liczba kolumn tabeli nie może być zbyt duża. Przy czym trudno tu o rozstrzygnięcie, jaką liczbę kolumn można uznać za dopuszczalną. Zależy to od konkretnego problemu, a przede wszystkim od charakteru cechy umieszczonej w główce tabeli – między innymi od tego, czy kategorie tej cechy mają naturalny porządek oraz na ile stanowią dla odbiorcy odrębne stany rzeczy. Jeśli prezentacja adresowana jest do osób zainteresowanych polityką, znających niuanse różniące programy poszczególnych partii, to w główce tabeli umieścić można większą liczbę partii. Natomiast adresując prezentację do osób słabiej obeznanych ze sceną polityczną, w zasadzie należałoby ograniczyć się do partii, które w wyborach odegrały największą rolę.

2.2.8 Zasady prezentowania wielkości liczbowych w polach tabeli

Prezentując wyniki badania za pomocą tabeli, pojawia się pytanie o wymagany stopień dokładności podawanych liczb. W tabelach 2.9 i 2.10 odsetki przedstawiono, zaokrąglając je do liczb całkowitych. Generalnie należy kierować się zasadą, aby w tabelach unikać zestawiania ze sobą liczb o większej liczbie cyfr znaczących niż dwie. Porównywanie ze sobą liczb wymaga chwilowego chociaż zapamiętania jednej z nich, co mniej angażuje zasoby percepcyjne gdy liczby mają mniej cyfr znaczących. Adresat tabeli w większym stopniu koncentruje wtedy uwagę na wyciągnięciu wniosków z porównań, niż na samym porównywaniu liczb (Ehrenberg 1981).

Prezentowanie w tabelach liczb z dokładnością do dwóch cyfr znaczących znajduje też swoje uzasadnienie w kryteriach statystycznych. Wiele wyników pochodzi z badań reprezentacyjnych, realizowanych na wybranej próbie, a nie w pełnej populacji. Otrzymane liczebności stanowią więc jedynie estymacje wielkości faktycznych. Precyzja tych estymacji jest na ogół taka, że nawet prezentacja wyniku z dokładnością do dwóch cyfr znaczących wykracza poza tę precyzję.

Weźmy przykładowo odsetek kobiet głosujących na Samoobronę. W badaniu wyniósł on 5,8 (z dokładnością do jednej cyfry po przecinku) i został oszacowany na podstawie odpowiedzi 568 kobiet (dane ważone, tabela 2.9). Uwzględniając te informacje można skonstruować przedział ufności dla szacowanego odsetka. Zakładając losowy dobór próby oraz przyjmując 95-procentowy przedział ufności jego granice dla rozpatrywanego odsetka wynoszą od 3,9 do 7,7 (Ferguson i Takane 2007: 191–194). Mówiąc inaczej, estymowana wartość odsetka kobiet głosujących na Samoobronę w populacji znajduje się z prawdopodobieństwem 0,95 w przedziale, którego granice odbiegają o ±1,9 od wartości otrzymanej w wyniku badania próby, czyli od 5,8.

Podane w tabeli 2.9 odsetki – zaokrąglone do liczb całkowitych – wykraczają więc poza swoją statystyczną precyzję. Wróćmy raz jeszcze do odsetka kobiet głosujących na Samoobronę. Aby mieć w tym przypadku zaufanie do wyniku w postaci 6 procent – czyli odsetka zaokrąglonego do liczby całkowitej – granice przedziału ufności nie powinny odbiegać od szacowanego odsetka o więcej niż o ±0,5 procenta. Aby uzyskać taką precyzję, liczba badanych kobiet powinna wynieść przynajmniej 8425. W badaniach socjologicznych rzadko korzysta się z tak dużych prób. Wielkości typowych projektów badawczych wahają się na ogół od kilkuset do dwóch-trzech tysięcy osób. Prezentacja odsetków zaokrąglonych do liczb całkowitych w zupełności więc wystarcza, a co więcej, biorąc pod uwagę wielkości prób stosowanych w badaniach, na ogół i tak wykracza poza przedziały ufności dla wyników. Podanie odsetków z większą dokładnością wywołać może błędne wrażenie, że dokładność estymacji jest większa, niż faktycznie ma to miejsce.

Zasada prezentacji wielkości liczbowych z dokładnością do dwóch cyfr znaczących koliduje na ogół z wymogiem dotyczącym odtwarzalności danych w tabeli. Interpretacje prezentowane w opracowaniach naukowych powinny dawać ich adresatom możliwość pełnego prześledzenia toku rozumowania autora, co obejmuje również dane, które stanowiły podstawę wnioskowania (Marsh 1988: 138–140). Dlatego w tabelach przedstawiających odsetki umieszcza się z reguły margines liczebności brzegowych (tak jak w tabeli 2.9). Z jednej strony, pozwala to adresatowi opracowania na rozstrzygnięcie, czy ilościową podstawę prezentowanych odsetków można uznać za wystarczającą. Z drugiej zaś strony pozwala odtworzyć liczebności rozkładu łącznego drogą przemnożenia liczebności brzegowych przez odsetki. Odtworzona tablica, poddana analizie inną metodą, stanowić może podstawę wzbogacenia przedstawionego przez autora obrazu badanego zjawiska lub wyrobienia sobie przez odbiorcę opracowania własnego zdania w kwestii interpretacji prezentowanych danych.

Aby jednak wyniki tych obliczeń były zbliżone do liczebności wyjściowych, odsetki musiałyby być podane z większą dokładnością. Odsetki zaokrąglone do liczb całkowitych precyzji takiej nie dają, zwłaszcza że w wielu wypadkach nie sumują się do 100 procent. Na przykład, sytuacja taka ma miejsce w tabeli 2.9 w wypadku mężczyzn, gdyż faktyczna suma odsetków zamieszczonych w polach wnętrza tabeli wynosi 101. Podstawową funkcją tabel jest jednak prezentacja danych, toteż w wypadku istnienia sprzecznych wymogów co do precyzji zamieszczanych wielkości należy w pierwszym rzędzie zadbać o przejrzystość prezentacji. Z tego powodu tabele zawierające dokładne liczebności empiryczne lepiej jest dodatkowo zamieścić w aneksie opracowania, niż zapewnić ich odtwarzalność z tabel użytych do prezentowania wyników.

Omawiany problem stopniowo zresztą traci na znaczeniu, gdyż obecnie tabele coraz częściej sporządza się, wykorzystując dane z badań, których wyniki są publicznie dostępne. W takim wypadku wystarczy właściwie opisać źródło danych, które posłużyły do budowy tabeli, tak aby adresat opracowania mógł we własnym zakresie odtworzyć odpowiednią tablicę liczebności empirycznych.

Jak już wspomniano, na skutek zaokrągleń odsetki w niektórych wierszach lub kolumnach tabeli mogą nie sumować się do 100 procent. Niekiedy wywołuje to niepokój u uważnego odbiorcy opracowania. Rodzić też może przypuszczenie, że autor tabeli przez nieuwagę podał którąś z wielkości błędnie, bądź – gdy suma jest mniejsza od 100 procent – że pewna kategoria bądź kategorie danej cechy zostały pominięte bez stosownego uzasadnienia tego faktu. Aby uniknąć tego rodzaju niejasności dobrze jest stosować dwa zabiegi. Po pierwsze, zawsze zamieszczać w tabeli sumy brzegowe odsetków. Na przykład, w przedostatniej kolumnie tabeli 2.9 podano, że sumy odsetków w wierszach są równe 100. Oznacza to, że żadna z kategorii cechy zamieszczonej w główce tabeli nie została pominięta, gdyż te, które są prezentowane, tworzą całość. Po drugie, w notce do tabeli można umieścić informację o tym, że jej autor świadomy jest faktu, iż pewne wiersze lub kolumny nie sumują się do wielkości brzegowych. Na przykład: „odsetki nie sumują się do 100 ze względu na zaokrąglenia".

Prezentacją w tabeli liczebności empirycznych rządzą nieco inne reguły niż prezentacją odsetków. W pierwszym kroku należy uwzględnić zasadę – omówioną w podrozdziale 2.2.4 – że w wypadku danych ważonych liczebności zaokrąglamy do wielkości całkowitych. W kolejnym kroku należy rozważyć, czy podane wielkości należy dalej zaokrąglić, tak aby zredukować liczbę cyfr znaczących. Wiadomo bowiem, że nadmierna liczba cyfr znaczących utrudnia porównywanie prezentowanych wielkości.

Wielkość typowych prób stosowanych w badaniach socjologicznych powoduje, że liczebności empiryczne prezentowane w tabelach mają na ogół od jednej do trzech cyfr znaczących – tak jak to ma miejsce w tabeli 2.8. W sytuacji takiej nie stosuje się dalszego zaokrąglania liczb, gdyż liczby zaokrąglone – na przykład do dziesiątek – wyglądałyby nienaturalnie wzbudzając niepokój u odbiorcy tabeli. Gdy jednak próba liczy kilkadziesiąt tysięcy badanych jednostek bądź gdy badanie obejmuje pełną populację (na przykład spis powszechny), wtedy liczba cyfr znaczących w prezentowanych wielkościach może wzrosnąć do czterech lub nawet do większej liczby. Czytanie liczb o takiej precyzji nie jest wygodne. Dlatego w takiej sytuacji liczebności empiryczne lepiej jest wyrazić w tysiącach ustalając poziom zaokrągleń tak, aby w wypadku najmniejszych liczb zachować przynajmniej jedną cyfrę znaczącą. Na przykład, liczebność 56 214 można przedstawić jako 56,2 tysiąca, zaś liczebność 832 jako 0,8 tysiąca.

Na zakończenie należy jeszcze wspomnieć o kwestii oznaczania w tabeli pustych pól. Niekiedy zdarza się, że w tabeli sporządzonej na podstawie wyników badania niektóre pola pozostają puste, ponieważ do próby nie trafiła ani jedna osoba o danej kombinacji kategorii obu cech. W sytuacji takiej należy rozpatrzyć, czy dana kombinacja kategorii ma prawo wystąpić w populacji, z której została wylosowana próba. Jeśli tak, to w puste pole należy wpisać liczbę „0". Oznacza to, że dana kombinacja cech występuje w populacji rzadko, toteż 0 jest najbliższą wielkością stanowiącą oszacowanie częstości jej wystąpienia w badanej próbie. Jeśli natomiast dana kombinacja cech jest niemożliwa ze względu na istotę prezentowanego zjawiska, to w polu należy umieścić symbol oznaczający sytuację „nie dotyczy". Symbolem tym jest najczęściej kreska „—" lub symbol przekreślenia „x". Znaczenie użytego symbolu powinno zostać opisane w nocie pod tabelą.

2.2.9 Relacje między tabelą a tekstem

Tabele służące prezentacji danych stanowią na ogół składniki szerszego opracowania: artykułu, książki czy raportu. Muszą więc odpowiednio korespondować z kontekstem, w którym zostały umieszczone. Ma to wpływ na postać wielu elementów składowych tabel, a także decyduje o tym, które dane umieścić w tabelach, a które włączyć do tekstu.

Zamierzając przedstawić dane w tabeli warto zadać najpierw pytanie: czy tabela jest w ogóle potrzebna? Czy tej samej idei lub wniosku nie uda się bardziej klarownie opisać w tekście? Dopiero wtedy, gdy odpowiedź jest negatywna, można posłużyć się tabelą.

Szczególną uwagę warto zwrócić na związki między zawartością tabeli a tekstem, w którym została umieszczona. Należy starać się nie dublować tych samych informacji w tabeli i w tekście (Lang i Secic 2006: 328–331). Wróćmy na chwilę do tabeli 2.9, przyjmując, że została ona opublikowana jako fragment artykułu naukowego. Przypuśćmy, że autor tego artykułu, pod tabelą umieścił następujący opis:

> *Wśród badanych kobiet 38 procent zadeklarowało, że w wyborach parlamentarnych oddało swój głos na kandydata PiS, 27 procent na kandydata PO, 8 procent na kandydata SLD, 6 procent na kandydata Samoobrony, 2 procent na kandydata PSL, 2 procent na kandydata LPR, zaś 2 procent badanych kobiet oddało swój głos na kandydata jednego z pozostałych ugrupowań, które kandydowały do Sejmu. Należy jeszcze dodać, że 15 procent badanych kobiet nie potrafiło powiedzieć, na kandydata jakiej partii oddały swój głos w wyborach. Wśród mężczyzn analogiczne odsetki kształtowały się inaczej. Odsetki mężczyzn głosujących na kandydatów PiS,*

PO czy SLD były zbliżone, jak w wypadku kobiet (odpowiednio 39%, 26% i 8%), natomiast odsetek mężczyzn głosujących na Samoobronę był wyraźnie wyższy, gdyż wynosił 10 procent ...

Nie będę dalej rozwijał tego opisu, gdyż przykładowy fragment w zupełności wystarcza, aby przekonać się o zaletach tabel jako syntetycznego i dogodnego dla odbiorcy sposobu prezentacji danych. Tym bardziej nie ma więc potrzeby opisywania jej zawartości w tekście. Zdublowana prezentacja danych wywoła u czytelnika co najwyżej konfuzję, gdyż zostanie potraktowana jako wskazówka, że coś w tabeli przeoczył lub czegoś nie zrozumiał.

Podejmując decyzje, czy dane umieścić w tekście czy w tabeli, warto kierować się zasadą DRY (*don't repeat yourself*). Jest to jedna z dobrych praktyk pisania oprogramowania komputerowego i sprowadza się do tego, aby powtarzające się fragmenty kodu źródłowego pisać jednokrotnie, zaś w miejscach, w których należy je wykorzystać, umieszczać odwołania. Podstawowa korzyść polega na tym, że ewentualne poprawki czy modyfikacje wystarczy wprowadzić w jednym miejscu. Współczesne języki programowania obiektowego zorganizowane są wokół tej właśnie zasady i zawierają szereg narzędzi, które ułatwiają jej stosowanie, a w niektórych wypadkach wręcz to narzucają.

Przyjmijmy przez analogię, że tabele pełnią funkcje takich właśnie narzędzi. Zawierają informacje, do których odwoływać się można wielokrotnie. Podejście takie ułatwia przygotowanie opracowania czy raportu. Pozostawia bowiem swobodę wyboru fragmentów tekstu, w których odwołamy się do konkretnych danych, a także pozwala odwoływać się do nich więcej niż jeden raz. Umieszczenie danych w tekście ogranicza te możliwości. Tekst nie posiada bowiem naturalnych markerów pozwalających konstruować odwołania. Nie napiszemy przecież „odsetek wyborców podany pięć wierszy wyżej", lecz będziemy musieli daną wielkość powtórzyć. W rezultacie korekty i modyfikacje opracowania wymagały będą szczególnej uwagi, tak aby odpowiadające sobie wielkości pozamieniać w jednakowy sposób. Stosowanie zasady DRY nie tylko poprawia efektywność przygotowania opracowania lecz również ogranicza ryzyko sformułowania niespójnych wniosków.

2.3 Elementy tabeli i ich edycja

Poniżej przedstawiono zasady formatowania tabel, które z jednej strony sprzyjają czytelności komunikatu, z drugiej zaś pozwalają wyeksponować jego wybrane elementy. Ogólny schemat budowy tabeli przedstawia rycina 2.3. Bloki oznaczone kolorem szarym obejmują jej część zasadniczą. Bloki leżące powyżej to część nagłówkowa, zaś bloki leżące poniżej to stopka tabeli. Omó-

wienie zasad formatowania poszczególnych elementów tabeli rozpocznijmy od części nagłówkowej.

Identyfikator tabeli składa się ze słowa „Tabela", po którym następuje jej kolejny numer. W obrębie artykułów naukowych i opracowań o podobnym charakterze stosuje się ciągłą numerację tabel, to znaczy 1, 2, i tak dalej. Nawet gdyby w artykule umieszczona została tylko jedna tabela, to i tak powinna być opatrzona identyfikatorem (w tym wypadku „Tabela 1"), który pozwoli odwoływać się w tekście do jej zawartości. W książkach i innych opracowaniach podzielonych na rozdziały lub posiadających wyodrębnione części – jak na przykład aneks – tabele numeruje się osobno w obrębie każdej z nich, poprzedzając numer tabeli symbolem rozdziału lub części opracowania, na przykład „Tabela 2.1" czy „Tabela A-2".

Kolejnym elementem tabeli jest jej tytuł. Formułując tytuły tabel, warto przestrzegać dwóch zasad (March 1988: 140; Lang i Secic 2006: 334–5).

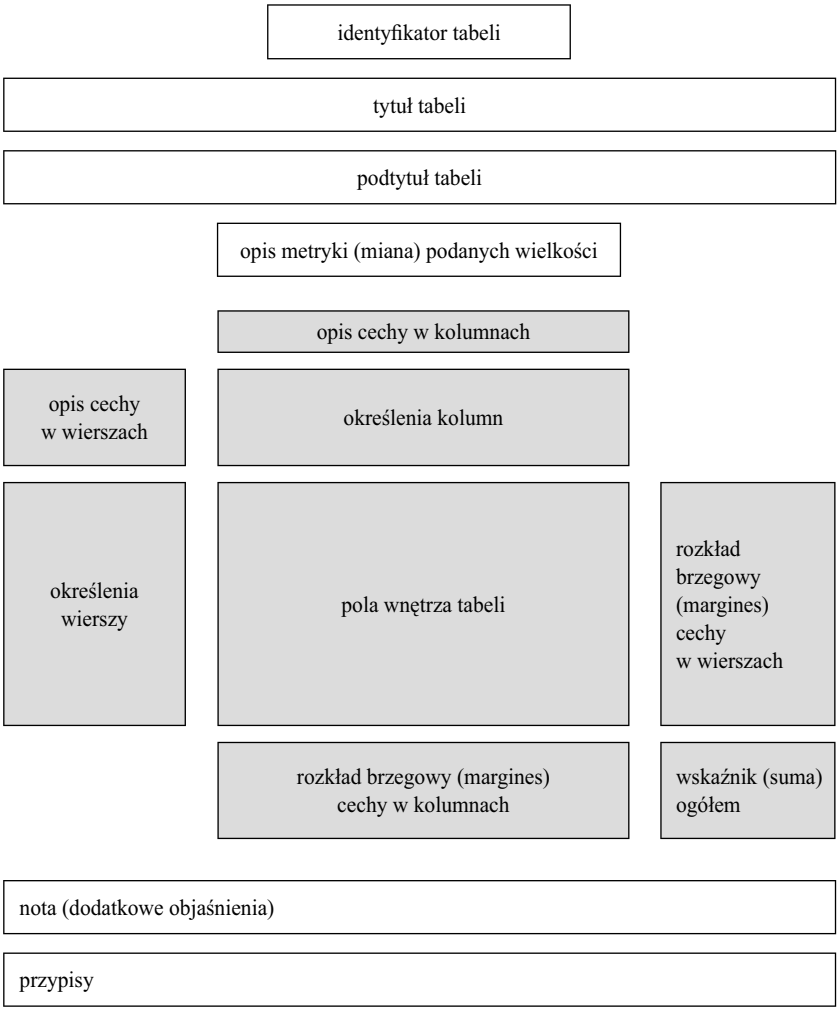
Po pierwsze, tytuł powinien umożliwić właściwe zrozumienie danych w tabeli bez odwoływania się do tekstu opracowania. Nie należy w żadnym wypadku dążyć do tego, aby tytuł był maksymalnie krótki. Prezentacja wyników badania to nie utwór poetycki. Odbiorca opracowania musi być jednoznacznie poinformowany o tym, w jaki sposób zostały zebrane i czego dotyczą dane przedstawione w tabeli. Na przykład, w tytułach wszystkich tabel przedstawionych w tym rozdziale umieszczono informację, że prezentowane dane dotyczą odpowiedzi respondentów w badaniu Europejski Sondaż Społeczny 2006. Bez tej informacji osoba, która rozpoczęła lekturę tej książki od przeglądania tabel mogłaby dojść do błędnego przekonania, że prezentowane dane (na przykład odsetki) dotyczą ogółu głosujących w wyborach do sejmu we wrześniu 2005 roku.

Po drugie, w tytule należy w pierwszym rzędzie opisać zawartość wnętrza tabeli, zaś dopiero w drugiej kolejności cechy umieszczone w wierszach i w kolumnach. Jeśli w komórkach wnętrza tabeli podano liczbę badanych o poszczególnych kombinacjach kategorii obu cech (tak jak w tabeli 2.8), to tytuł tabeli powinien rozpoczynać się od sformułowania „liczba badanych..." czy odpowiedniego równoważnego. Jeśli wnętrze tabeli zawiera odsetki, to tytuł najlepiej rozpocząć od słowa „odsetki..." (tabela 2.9).

Gdy ilość informacji wymaganych do pełnego opisu zawartości tabeli jest na tyle duża, że umieszczenie ich wszystkich w tytule spowoduje, iż jego treść przestanie być przejrzysta, to wtedy należy mniej istotne z tych informacji przenieść do noty zamieszczonej pod tabelą. W opracowaniach naukowych w notach umieszcza się na ogół informacje o źródle prezentowanych danych. Można tak zrobić tylko wtedy, gdy z tytułu jednoznacznie wynika, że prezentowane dane dotyczą wyników badania (na przykład „odsetki **odpowiedzi na**

pytanie o stosunek do socjalizmu wśród mieszkańców miast i wsi"). W notach umieszcza się też informacje o sposobie uzyskania podanych wielkości (zob. tabela 2.10) bądź wskazówki dotyczące sposobu ich interpretowania. Natomiast rodzaj czy miano wskaźnika użytego w tabeli („w procentach", „w tysiącach osób") przyjęto opisywać w osobnym wierszu umieszczanym między tytułem a tabelą – umieszczając niekiedy stosowny opis w nawiasach kwadratowych.

Rycina 2.3 Schematyczna budowa tabeli



Specyficzne funkcje pełni podtytuł tabeli. W zasadzie stosuje się go w sytuacjach, gdy tabela obejmuje fragment szerszego zasobu danych, omawianych w całości w danym opracowaniu – przy czym inne zamieszczone tabele prezentują pozostałe fragmenty danych. Przyjmijmy, że przedmiotem opracowania jest porównanie wyników Europejskiego Sondażu Społecznego dla różnych krajów. Wtedy w tytule tabeli można opisać rodzaj porównywanych danych, zaś w podtytule podać nazwę kraju, na przykład „Polska", „Niemcy", „Holandia". Niekiedy przedmiotem porównania są wyniki różnych badań dotyczące tego samego zjawiska. W takiej sytuacji w podtytule można umieścić nazwę badania, na przykład „Polacy 1990", „Europejski Sondaż Społeczny 2002".

Przejdźmy obecnie do omówienia zasad formatowania samej tabeli. Kolumna umieszczona najbardziej z lewej strony, zwana „boczkiem" tabeli, służy do opisu jednej z dwóch prezentowanych cech. Na ogół najwyższe pole tej kolumny zawiera nazwę cechy (w tabeli 2.9 cechą tą jest „płeć"), natomiast w kolejnych polach tej kolumny podane są określenia kategorii tej cechy – w tym wypadku „kobiety" oraz „mężczyźni". Z kolei najwyższej umieszczony wiersz tabeli, zwany jej „główką", zawiera określenia kategorii drugiej cechy. W tabeli 2.9 są nimi określenia partii, na kandydatów których można było oddać głos, nazwa kategorii pogrupowanych odpowiedzi oraz nazwa kategorii rezydualnej. Nad wierszem umieścić można określenie drugiej cechy – w tym wypadku „sposób głosowania".

Prezentując dane za pomocą tabeli, przestrzega się konwencji, że każdy wiersz i każda kolumna ma swój indywidualny deskryptor (ang. *heading, label*) stanowiący **skrócony** opis zawartości danego wiersza lub kolumny. Przy czym, jeśli w tekście wystąpią odwołania do wierszy lub do kolumn tabeli, to należy posłużyć się deskryptorami w identycznej formie, tak aby odpowiedniość między tekstem a zawartością tabeli była jednoznaczna (March 1988: 140; Lang i Secic 2006: 336). Przykładowo, w główce tabeli 2.9 dla identyfikacji kolumn posłużono się skrótami nazw partii politycznych. W tekście należałoby raczej posłużyć się pełnymi nazwami partii, chociażby z tego względu, że za kilka czy kilkanaście lat czytelnik opracowania może nie wiedzieć, co oznaczał skrót LPR. Niemniej jednak, używając w tekście pełnych nazw partii, warto w nawiasach umieścić ich skróty, tak aby ułatwić czytelnikowi zidentyfikowanie odpowiednich kolumn tabeli.

Blok pozostałych pól tabeli – zwany „wnętrzem" tabeli – składa się z pól, w których umieszcza się wskaźniki odpowiadające złożeniom kategorii obu cech. Prezentacją liczb zamieszczanych w polach wnętrza tabeli również rządzą określone zasady. Przede wszystkim przyjmuje się, że jeśli są to liczby dziesiętne, to wszystkie powinny mieć jednakową liczbę cyfr po przecinku (Lang i Secic 2006: 338–340). Gdy jedynie część pól wnętrza tabeli wypeł-

niają liczby dziesiętne, w pozostałych zaś polach występują liczby całkowite, to wtedy liczby całkowite powinny być uzupełnione o określoną liczbę zer, na przykład „12,0". Liczby podane w polach stanowiących kolumny tabeli powinny zostać sformatowane w taki sposób, aby wyrównany był prawy margines każdej kolumny. Należy unikać środkowania liczb w kolumnach, gdyż utrudnia to odczytanie wielkości o niejednakowej liczbie cyfr i może prowadzić do pomyłek w ich interpretacji.

W większości tabel prezentuje się również wskaźniki zagregowane odpowiadające poszczególnym wierszom i kolumnom. Niekiedy informacje te są redundantne – to znaczy wynikają z zawartości wnętrza tabeli – niemniej ułatwiają analizę tabeli w całości. Rozkłady brzegowe umieszcza się w tabeli jako ostatni wiersz i jako ostatnią kolumnę bądź bezpośrednio obok określeń kategorii każdej z cech. Drugi ze sposobów stosuje się wtedy, gdy rozkład brzegowy stanowi kontekst dla poprawnej interpretacji wnętrza tabeli (Ehrenberg 1981).

Przykładowo, zamieszczony w tabeli 2.10 profil brzegowy kolumn informuje o udziale kobiet i mężczyzn wśród badanych osób. Podczas interpretacji zawartości tabeli rozkład ten ułatwia zauważenie, że w zbiorowości tej jest więcej kobiet, co może pomóc zrozumieć fakt wyższych odsetków kobiet wśród głosujących na poszczególne partie. Rozkład brzegowy dla wierszy składa się z dwóch kolumn umieszczonych z prawej strony wnętrza tabeli. W pierwszej podano sumy odsetków, zaś w drugiej liczbę badanych stanowiących podstawę obliczenia odsetków w każdym z wierszy. Omawiając przypuszczalne wnioski, jakie można wysnuć z oglądu tabeli 2.10, sygnalizowałem problem koncentracji uwagi na udziale kobiet i mężczyzn w elektoratach poszczególnych partii przy jednoczesnym pominięciu samej wielkości elektoratu – czyli znaczenia poszczególnych partii w wybranym sejmie. Przeniesienie kolumny prezentującej liczbę badanych osób na lewą stronę sprzyjałoby uwzględnieniu tych wielkości.

Ostatnim elementem tabeli są przypisy. Stosuje się je w wypadkach, gdy niezbędne są dodatkowe wyjaśnienia poszczególnych elementów tabeli: tytułu, podtytułu, określeń cech, określeń kategorii, czy niektórych wielkości w polach wnętrza tabeli. Na przykład, w tabelach 2.9–2.10 można posłużyć się przypisem w celu opisania pełnej zawartości kategorii „pozostałe partie". Ponieważ część przypisów odnosić się może do wielkości liczbowych, stąd jako symbole przypisów stosuje się nie liczby, lecz kolejne małe litery alfabetu: a, b, c itd.

Aby tabela przekazywała pewne treści w sposób skuteczny, zastosowane rozwiązania graficzne powinny w maksymalnym stopniu uwypuklać jej zawartość merytoryczną. Dlatego dąży się do redukcji liczby i rodzajów stosowanych symboli graficznych, a w szczególności unika się oddzielania od siebie

poszczególnych kolumn i wierszy za pomocą linii. Podział tabeli na wiersze i kolumny powinien wynikać z pozostałych rozwiązań edycyjnych: kroju stosowanej czcionki, wyrównania kolumn, odstępów między wierszami i innych. Już wiele lat temu statystycy doszli do wniosku, że przejrzysta tabela powinna zawierać jak najmniej linii rozdzielających (Croxton i Cowden 1955: 64–65). Ostatnio coraz częściej – zwłaszcza w publikacjach książkowych – spotyka się tabele, w których w ogóle nie stosuje się linii rozdzielających, zaś struktura tabeli wynika z odpowiednich rozwiązań edycyjnych[2].

Umieszczenie w tabeli dwóch lub trzech linii poziomych na ogół jednak nie budzi zastrzeżeń, gdy chodzi o jej czytelność (Lang i Secic 2006: 328). Liniami poziomymi najczęściej oddziela się: główkę tabeli (określenia kolumn) od liczebności wnętrza tabeli, a także liczebności wnętrza tabeli od wielkości brzegowych dla kolumn. W tabelach prezentowanych w tym rozdziale zdecydowałem się dodatkowo użyć linii rozdzielającej określenie cechy umieszczonej w główce od określeń zawartości poszczególnych kolumn. Nie powinno to utrudniać przeglądania zawartości tabel.

Obowiązująca współcześnie estetyka formatowania tabel nie dopuszcza natomiast stosowania linii pionowych! Wynika to z logiki budowy tabel, która zakłada umieszczenie prezentowanych wielkości w kolejnych polach każdego wiersza. Wprowadzenie linii oddzielających od siebie kolumny nie tylko utrudnia te porównania, lecz wręcz sugeruje, że poszczególne wielkości powinny być interpretowane jako odrębne.

Na koniec o stosowaniu w formatowaniu tabel zasad wspólnych dla wszystkich tekstach drukowanych. Warto przede wszystkim pamiętać o dość fundamentalnej zasadzie **unikania podwójnych wyróżnień**. Jeśli dany element tabeli wyodrębniony został za pomocą linii poziomej bądź interlinii, to nie ma potrzeby dodatkowego wyodrębniania go za pomocą innych środków typograficznych, jak na przykład pogrubienie czcionki czy kursywa. Zastosowanie wielu środków jednocześnie wygląda po prostu nieestetycznie.

Omawiana zasada dotyczy nie tylko prezentacji liczb, lecz również innych elementów tabeli, a szczególnie jej tytułu. Tytuł to wizytówka tabeli. Wiadomo, jakie funkcje pełni i w którym miejscu go szukać, przez co trudno go pomylić z jakimkolwiek innym elementem. Mnie osobiście nie podoba się rozwiązanie

[2] Posłużenie się tego rodzaju rozwiązaniem zaproponowałem wydawcy w książce na temat klasyfikacji i skal zawodów, napisanej wspólnie z Henrykiem Domańskim i Kazimierzem M. Słomczyńskim (Domański, Sawiński i Słomczyński 2007). Wspominam o tym z tego powodu, że dwa lata później ukazało się nowe wydanie tej książki, tym razem przygotowane do druku przez edytora amerykańskiego, który w tabelach stosował poziome linie rozdzielające (Domański, Sawiński i Słomczyński 2009). Stwarza to unikalną możliwość porównania obu layoutów i oceny, który z nich jest bardziej przejrzysty.

stosowane w niektórych publikacjach amerykańskich, między innymi w *American Journal of Sociology*. Polega ono na tym, że tytuły tabel składane są KAPITALIKAMI. Tradycja tradycją, lecz we współczesnych tekstach wygląda to dość archaicznie.

2.4 Uwagi końcowe

Zebrane w tym rozdziale wskazówki dotyczące budowy tabel w zasadzie nie wykraczają poza dość podstawową wiedzę na temat zasad tworzenia i formatowania komunikatów mediowych. W badaniach przez wiele lat do spraw tych nie przywiązywano szczególnego znaczenia. Obecnie jednak, w dobie ekspansji informacyjnej, kwestie te stały się ważne. Prezentacja wyników badania, tak jak każdy komunikat, musi utorować sobie drogę wśród tysięcy przekazów. Aby uczynić to w sposób skuteczny, musi uwypuklać elementy, które są najbardziej istotne. Tylko wtedy ma szanse przykuć uwagę audytorium, do którego jest adresowana.

Nadanie tabeli odpowiedniej formy to oczywiście nie wszystko. Aby komunikat wywołał zainteresowanie, musi kryć w sobie pewną ideę czy myśl, dostatecznie atrakcyjną i odkrywczą. W poszukiwaniu twórczych interpretacji użyteczna okazać się może perspektywa analityczna proponowana w dalszych rozdziałach.

ROZDZIAŁ 3

Czy niezależność istnieje?

Któryś z badaczy powiedział kiedyś złośliwie, że niezależność istnieje tylko w matematyce. W prowadzonych badaniach nigdy nie natknął się bowiem na niezależność. Jest w tym poniekąd sporo prawdy, gdyż rzeczywiście trudno wskazać przykład tablicy, w której badane cechy byłyby niezależne. Niemniej jednak, niezależność pełni w metodach analizy tablic dość fundamentalną funkcję. Stanowi punkt odniesienia dla oceny uzyskanych w badaniu rezultatów.

Jak jednak coś, co nie istnieje, może pełnić funkcję ram odniesienia dla badanych zjawisk? Otóż właśnie – ten problem chciałbym w tym rozdziale poddać dyskusji. Albowiem niezależność nie tylko potencjalnie może pełnić taką rolę, lecz faktycznie ją pełni. I to nie tylko w obrębie metod analizy danych – mających w gruncie rzeczy matematyczne podstawy – lecz również w koncepcjach teoretycznych, stanowiących podstawę wyjaśniania badanych zjawisk.

Pierwszy z podrozdziałów ma charakter pomocniczy. Przedstawiam w nim symbolikę stosowaną w opisach tablic i używaną w dalszych fragmentach tekstu, głównie we wzorach. Podrozdział ten nie ma wiele wspólnego z merytorycznym wątkiem książki, toteż Czytelnicy obeznani z tego rodzaju symboliką mogą go pominąć. W razie potrzeby zawsze można do niego wrócić podczas dalszej lektury.

Wątek merytoryczny rozpoczyna się w podrozdziale 3.2. Omawiam w nim ideę podejścia do analizy zjawisk polegającego na porównywaniu ich z określonymi modelami. Modele mogą mieć status teoretyczny bądź formalny. Te ostatnie opisują relacje między elementami zjawiska niezależnie od tego, jakie zjawisko jest przedmiotem badania.

W podrozdziale 3.3 wprowadzam definicję pojęcia niezależności wyjaśniając zarazem sposób utworzenia modelu niezależności dla rozpatrywanej tablicy. Do interpretacji niezależności jako identyczności profili nawiązuję w podrozdziale 3.4. Interpretacja ta jest doskonale znana każdemu badaczowi, który analizuje bądź prezentuje wyniki badań za pomocą tablic.

Podrozdział 3.5 rozpoczyna rozważania dotyczące istoty modelu niezależności. Kwestie te podejmuje się w literaturze rzadko, gdyż przydatność tych zagadnień w analizie danych nie jest oczywista. W podrozdziale 3.5 proponuję rozumienie mechanizmu niezależności jako wypadkowej potencjałów odpowiadających wierszom i kolumnom tablicy. Taki sposób rozumienia niezależności nawiązuje do tabliczki mnożenia. Ważne jest w związku z tym ustalenie roli rozkładów brzegowych w tablicy. Czy można traktować je jako czynnik zewnętrzny, wyznaczający ramy badanego zjawiska, czy też stanowią one jeden z jego elementów. W podrozdziale 3.6 przedstawiam użyteczność kategorii analitycznej, jaką stanowią dyspozycje do zachowań. Ułatwia ona interpretację tablic, których marginesy trudno byłoby traktować jako integralny składnik zjawiska będącego przedmiotem analizy. Podrozdział 3.7 poświęcony jest omówieniu sytuacji, w których marginesy konstruowanych tablic nie odzwierciedlają należycie kształtu badanego zjawiska. Może to być skutkiem przyjęcia określonego schematu badawczego bądź uchybień proceduralnych, które powstają w fazie gromadzenia danych. W podrozdziale 3.8 omawiam relacje, jakie zachodzą między modelem niezależności a losowością. Pokazuję, że losowość nie stanowi warunku koniecznego powstania niezależności. Z drugiej strony zjawiska odbiegające znacznie od modelu niezależności, mogą być po części wynikiem działania czynników losowych. I wreszcie, w podrozdziale 3.9 wskazuję związki, jakie istnieją między stosowaną w naukach społecznych koncepcją równych szans, a zastosowaniami w analizach tablic modelu niezależności.

Ostatnia część poświęcona jest ilościowym kryteriom oceny stopnia, w jakim tablica skonstruowana na podstawie wyników badania odbiega od modelu niezależności. W podrozdziale 3.10 omawiam znany badaczom test niezależności chi-kwadrat. Wymieniam korzyści stosowania tego testu, wskazując również na ograniczenia. W podrozdziale 3.11 nawiązuję natomiast do pomiaru wielkości odstępstw od niezależności w tablicy stanowiącej wynik badania.

Rozdział zamyka krótka dyskusja (3.13), zogniskowana na kwestii funkcji pełnionych przez pojęcie niezależności w wyjaśnianiu zjawisk. Anonsuję też sposoby wykorzystania pojęcia niezależności w metodach analizy tablic prezentowanych w dalszych rozdziałach książki.

3.1 Notacja stosowana w opisie tablic

Dla ułatwienia opisu zależności między elementami tablicy wprowadzimy symbolikę, która odpowiada formalnemu modelowi tablicy. Pełny zestaw proponowanych symboli przedstawiony został w tabeli 3.1.

W modelu formalnym tablicy cechom odpowiadają **zmienne**, zaś kategoriom cech **wartości zmiennych**. Zmienne będziemy oznaczać dużymi literami X oraz Y. Przyjmiemy, że zmienna umieszczona w boczku tablicy oznaczona jest literą X, natomiast zmienna w główce tablicy literą Y. Odpowiada to konwencji stosowanej do oznaczania zmiennych w modelu regresji, w którym zmienną zależną przyjęto oznaczać symbolem Y, zaś zmienną niezależną symbolem X. W rozdziale 2 przedstawiono powody, dla których cechę warunkowaną (skutek) dogodnie jest umieścić w główce tablicy, zaś zmienną ją warunkującą (przyczynę) w boczku. Przyjęte w tym miejscu oznaczenia zmiennych są spójne z tymi ustaleniami.

Tabela 3.1
Symbole stosowane w opisie tablic

		Zmienna Y					
		y_1	y_2	...	y_{k-1}	y_k	ogółem
Zmienna X	x_1	n_{11}	n_{12}	...	$n_{1,k-1}$	n_{1k}	a_1
	x_2	n_{21}	n_{22}	...	$n_{2,k-1}$	n_{2k}	a_2
	...	...	...	...	...	...	...
	x_{w-1}	$n_{w-1,1}$	$n_{w-1,2}$	...	$n_{w-1,k-1}$	$n_{w-1,k}$	a_{w-1}
	x_w	n_{w1}	n_{w2}	...	$n_{w,k-1}$	n_{wk}	a_w
	ogółem	b_1	b_2	...	b_{k-1}	b_k	n

Wartości zmiennych oznaczymy odpowiednio małymi literami x_i oraz y_j. Przyjmiemy, że zmienna X jest reprezentowana w tablicy przez w wartości, zaś zmienna Y przez k wartości. Przyjęcie akurat tych liter dla oznaczenia liczby wartości zmiennych jest konwencją dogodną o tyle, że litera w rozpoczyna słowo „wiersze", zaś litera k słowo „kolumny". Zmienna X posiada zatem w wartości, które dla odróżnienia od siebie oznaczymy za pomocą subskryptów, jako $x_1, x_2, ..., x_{w-1}, x_w$. Przyjmiemy zarazem, że wartość x_1 odpowiada najwyższemu wierszowi tablicy, zaś wartość x_w wierszowi położonemu w tablicy najniżej. Analogicznie oznaczymy poszczególne wartości zmiennej Y jako $y_1, y_2, ..., y_{k-1}, y_k$, przy czym y_1 odpowiada pierwszej kolumnie tablicy – licząc od lewej strony, zaś y_k kolumnie ostatniej.

Małą literą n będziemy oznaczać liczebności pól wnętrza tablicy. Liczebności tych jest tyle, ile wynosi iloczyn liczb wartości obu zmiennych, czyli $w \times k$. Poszczególne liczebności – zwane też liczebnościami rozkładu łącznego obu zmiennych – oznaczać będziemy podwójnym subskryptem, przy czym

pierwszy subskrypt odpowiada oznaczeniu wiersza, zaś drugi oznaczeniu kolumny tablicy. A zatem n_{11} oznacza liczebność leżącą na przecięciu pierwszego wiersza i pierwszej kolumny tablicy, $n_{1,k-1}$ przedostatnią liczebność pierwszego wiersza, zaś n_{wk} oznacza liczebność na przecięciu ostatniego wiersza i ostatniej kolumny. Zgodnie z ustaleniami poczynionymi w podrozdziale 2.2.4 liczebności rozkładu łącznego wyrażają się liczbami niecałkowitymi.

Ostatnie z wprowadzonych oznaczeń dotyczą rozkładów brzegowych, czyli marginesów tablicy. Liczebności rozkładu brzegowego zmiennej X oznaczymy jako a_i. Symbole $a_1, a_2,..., a_{w-1}, a_w$ odpowiadają więc sumom liczebności w kolejnych wierszach tablicy. Analogicznie, symbole $b_1, b_2,..., b_{k-1}, b_k$ odpowiadają sumom kolejnych kolumn[1]. Sumę wszystkich liczebności rozkładu łącznego obu zmiennych oznaczymy małą literą n. W analizach wyników badań reprezentacyjnych odpowiada ona liczbie zbadanych osób. Gdy zaś badanie obejmuje pełną populację (badanie wyczerpujące), to n jest równe wielkości populacji.

Relacje między oznaczeniami liczebności rozkładu łącznego obu zmiennych oraz liczebnościami rozkładów brzegowych wyrazić można za pomocą poniższych wzorów.

$$a_i = \sum_{j=1}^{k} n_{ij} \tag{3.1}$$

dla wszystkich $i = 1,2,..., w$

$$b_j = \sum_{i=1}^{w} n_{ij} \tag{3.2}$$

dla $j = 1,2,..., k$. Jednocześnie zachodzi

$$n = \sum_{i=1}^{w} \sum_{j=1}^{k} n_{ij} = \sum_{i=1}^{w} a_i = \sum_{j=1}^{k} b_j \tag{3.3}$$

Zarówno więc liczebności brzegowe rozkładów obu zmiennych, jak też suma wszystkich jednostek n mogą być obliczone na podstawie liczebności rozkładu łącznego n_{ij}. Nie dostarczają więc niezależnych informacji o rozkładzie obu zmiennych. W dalszych rozważaniach przyjmiemy, że mówiąc o tablicy opisującej związek między dwiema zmiennymi mamy na myśli tablicę li-

[1] Gdy liczebności rozkładu łącznego oznaczone zostają jako n_{ij}, to na ogół liczebności rozkładów brzegowych oznacza się konsekwentnie jako $n_{i.}$ (margines zmiennej X) oraz $n_{.j}$ (margines zmiennej Y), gdzie symbol kropki oznacza zmienną, po której następuje sumowanie. Ponieważ w książce wielokrotnie odwoływać się będę do liczebności brzegowych tablicy, stąd dla ich oznaczenia przyjąłem odrębne symbole literowe a i b. Poprawia to czytelność odwołań do marginesów w tekście i we wzorach.

czebności rozkładu łącznego n_{ij}. Tego typu prostokątne tablice liczb nazywane są też macierzami i oznacza się je dużymi literami: *A, B,..., M, N,...*. Niekiedy mówiąc o własnościach bądź modelach rozkładu łącznego dwóch zmiennych, wygodniej jest posługiwać konwencją zapisu tablic w postaci macierzy, zamiast symboli pól rozkładu łącznego n_{ij}.

Na podstawie tablicy liczebności n_{ij} utworzyć można tablicę proporcji p_{ij}, zwanych też częstościami względnymi. Wielkościami tymi są uzyskane w badaniu liczebności znormalizowane do całkowitej liczby jednostek w tablicy.

$$p_{ij} = \frac{n_{ij}}{n} \tag{3.4}$$

Suma proporcji dla całej tablicy jest w tej sytuacji równa 1.

$$\sum_{i=1}^{w}\sum_{j=1}^{k} p_{ij} = 1 \tag{3.5}$$

Wielkości proporcji wyrazić można za pomocą ułamków dziesiętnych, na przykład 0,10 czy 0,7822. Prezentując wyniki badań częściej przedstawia się je jednak w postaci odsetków, na przykład 10% czy 78,22%, które w całej tablicy sumują się do 100 procent. Obie konwencje zapisu proporcji są równoważne, gdyż 0,10 to ta sama wielkość co 10%. Na stosowany format zapisu warto jednakże zwrócić uwagę, gdyż odsetki w tablicach podaje się na ogół w wielkościach niemianowanych – na przykład 10 zamiast 10%.

Niekiedy wielkości w tablicy przedstawia się w postaci rozkładów warunkowych, zwanych też profilami. Obliczyć je można w dwóch wariantach: w obrębie wierszy tablicy lub w obrębie kolumn. Niech pw_{ij} oznaczają wielkości pól profilu – inaczej mówiąc wielkosci proporcji warunkowych – dla *i*-tego wiersza, zaś pk_{ij} wielkości pól profilu w *j*-tej kolumnie tablicy. Wówczas

$$pw_{ij} = \frac{n_{ij}}{a_i} \tag{3.6}$$

$$pk_{ij} = \frac{n_{ij}}{b_j} \tag{3.7}$$

dla wszystkich $i = 1,2,..., w$ oraz $j = 1,2,..., k$[2]. Wielkości w polach profili obliczonych w wierszach sumują się do jedności w każdym z wierszy z osobna

[2] Prezentując dalsze wzory, pomijać będziemy przedział zmienności indeksów *i* oraz *j*, co nie powinno prowadzić do nieporozumień. Przedział ten zostanie podany tylko w wypadku, gdy będzie specyficzny dla danego wzoru.

$$\sum_{j=1}^{k} pw_{ij} = 1 \tag{3.8}$$

Analogiczna prawidłowość zachodzi dla profili obliczonych w kolumnach tablicy

$$\sum_{i=1}^{w} pk_{ij} = 1 \tag{3.9}$$

Wielkości w polach profili, podobnie jak proporcje obliczone dla całego rozkładu, przedstawić można w postaci odsetków. Forma ta jest na ogół stosowana w prezentacjach danych (zob. podrozdział 2.2.7).

Bazową formą zapisu tablicy jest macierz liczebności n_{ij} obejmująca wszystkie pola rozkładu łącznego obu zmiennych. Analizując niektóre zagadnienia stosuje się alternatywny sposób zapisu tablicy, w którym jako znane przyjmuje się wielkości pól obu rozkładów brzegowych oraz pewną liczbę pól rozkładu łącznego.

Ilustrację takiego sposobu zapisu stanowi tablica o dwóch wierszach i trzech kolumnach, przedstawiona w tabeli 3.2. Aby odtworzyć wszystkie liczebności rozkładu łącznego wystarczy znajomość wielkości pól obu rozkładów brzegowych (a_1, a_2, b_1, b_2, b_3)) oraz dwóch pól rozkładu łącznego. W tabeli 3.2 znane wielkości zaznaczone zostały kolorem szarym. Przykładowymi polami rozkładu łącznego, których wielkość jest znana, są w tym przykładzie n_{11} oraz n_{12} – aczkolwiek mogą nimi być dowolne dwa z sześciu pól, o ile spełnią warunek braku redundancji. Na przykład, gdyby jako znane wielkości wybrać n_{11} i n_{12}, to informacje w nich zawarte byłyby redundantne, ponieważ liczebność n_{21} można otrzymać odejmując od b_1 liczebność n_{11} i analogicznie – liczebność n_{11} otrzymać można poprzez odjęcie n_{21} od b_1. Znając tylko dwie wielkości wzajemnie redundantne nie uda się odtworzyć wielkości wszystkich sześciu pól rozkładu łącznego przedstawionej tablicy.

Tabela 3.2
Ilustracja zapisu tablicy z wykorzystaniem rozkładów brzegowych

	y_1	y_2	y_3	
x_1	n_{11}	n_{12}	n_{13}	a_1
x_2	n_{21}	n_{22}	n_{23}	a_2
	b_1	b_2	b_3	n

Aby odtworzyć liczebności wszystkich pól tablicy rozkładu łącznego o w wierszach i k kolumnach w oparciu o znane rozkłady brzegowe, wystarcza znajomość $(w-1) \times (k-1)$ wielkości w polach rozkładu łącznego, o ile żadna z tych wielkości nie jest redundantna. Liczbę tę nazywa się też **liczbą stopni swobody** tablicy. Pojęcie to wykorzystywane jest w wielu metodach analizy tablic.

3.2 Modele referencyjne w ujęciu konfirmacyjnym i eksploracyjnym

Współcześnie stosowane metody analizy tablic korzystają z różnych kryteriów oceny lub klasyfikowania liczebności w poszczególnych polach tablicy. Podejście bodajże najczęściej stosowane polega na zestawieniu tablicy otrzymanej w wyniku badania – składającej się z $w \times k$ liczebności rozkładu łącznego – z inną tablicą o tych samych rozmiarach, utworzoną według określonych zasad. Tę ostatnią tablicę nazwiemy **modelem referencyjnym**.

W podrozdziale 1.2 wprowadzone zostało rozróżnienie między zasięgiem, czy też rozmiarami zjawiska, a jego mechanizmem. Budowa modelu referencyjnego na ogół odwołuje się do mechanizmu zjawiska. Wiedza na temat tego mechanizmu może mieć status hipotezy. Wtedy wyniki badania służą jej weryfikacji. Poprzez porównanie tablicy otrzymanej w wyniku badania z modelem referencyjnym – skonstruowanym przed realizacją badania na mocy hipotezy dotyczącej mechanizmu zjawiska – badacz rozstrzyga, czy hipoteza może być utrzymana, czy też powinna być odrzucona, gdyż wyniki badania jej nie potwierdzają.

Opisane podejście nazywane jest **konfirmacyjnym**. Wymaga ono apriorycznej wiedzy na temat mechanizmów badanych zjawisk. Współcześnie podejście to stosuje się coraz rzadziej, gdyż ekspansja badań wyprzedza rozwój wiedzy teoretycznej na temat badanych zjawisk. W wypadku wielu badań brakuje teorii, która pozwoliłaby na sformułowanie hipotez. W związku z tym wiedzę gromadzi się w sposób ekstensywny, poprzez kumulację wyników kolejnych badań. Podejście to nazywane jest **eksploracyjnym**. Dominuje ono w naukach społecznych i w ich zastosowaniach.

W podejściu eksploracyjnym również korzysta się z metod analizy tablic opartych na modelach referencyjnych. Jednakże modeli tych nie konstruuje się na bazie wiedzy na temat mechanizmów badanych zjawisk, gdyż tego rodzaju wiedza na ogół nie jest dostępna. W zamian badacze odwołują się do mechanizmów uniwersalnych, które potencjalnie stanowić mogą o istocie każdego zjawiska. Powszechnie wykorzystuje się w tym celu mechanizm zwany **niezależnością** (ang. *independence*).

W wypadku tablic rozumie się to w ten sposób, że układ liczebności w tablicy stanowiącej wynik badania pozwala traktować obie cechy jako niezależne. Nawiązując do omawianego w rozdziale 2 przykładu sposobów głosowania kobiet i mężczyzn powiedzielibyśmy, że sposób głosowania **nie zależy** od płci. Łatwo zauważyć, że model niezależności nie ma statusu hipotezy w takim sensie, jak to ma miejsce w podejściu konfirmacyjnym. Teoria, która głosiłaby jedynie, że coś nie zależy od czegoś, nie byłaby w stanie dostarczyć pogłębionej wiedzy na temat mechanizmów badanego zjawiska. Dlatego w podejściu eksploracyjnym model referencyjny traktuje się na ogół jako pewien **stan nierzeczywisty** (językowo pasuje tu słowo „hipotetyczny", ale kłóci się ono z wcześniejszą wykładnią funkcji hipotezy), który potencjalnie mógłby zaistnieć, aczkolwiek w praktyce nie należy się raczej tego spodziewać. Do tego stanu porównuje się wyniki badania, co stanowi podstawę sformułowania wniosków, w których fragmentach tablicy uzyskane w badaniu liczebności odbiegają od modelu niezależności, zaś w których są z nim zgodne. W oparciu o tego typu ustalenia formułuje się wyjaśnienia mechanizmu badanego zjawiska (por. Agresti 2002: 85).

3.3 Niezależność stochastyczna

Model niezależności przyjmowany jako referencyjny w analizach tablic nazwiemy – jak to się niekiedy czyni[3] – modelem niezależności **stochastycznej**. Słowo „stochastyczny" służy podkreśleniu, że stan określony w modelu ma charakter losowy, że rządzi nim przypadek (Kopaliński 2007). Model niezależności stochastycznej opisuje domniemaną sytuację, w której rozkład łączny dwóch cech nie podlega wpływom żadnych czynników zakłócających przypadkowe łączenie się kategorii obu cech w pary. Mówiąc inaczej, brak jest powodów, dla których pewne kombinacje kategorii obu cech występowałyby częściej, inne zaś rzadziej, niż miałoby to miejsce, gdyby powstawały w sposób całkowicie przypadkowy.

Do dyskusji związków między pojęciami niezależności i losowości wrócimy w podrozdziale 3.8. W tym miejscu przedstawimy zasadę budowy modelu

[3] Termin „stochastyczny" używany jest w literaturze polskiej do odróżnienia omawianego modelu niezależności od modeli niezależności skonstruowanych według innych zasad (Lissowski, Haman i Jasiński 2008: 214–216, 321–333). W literaturze anglojęzycznej z dziedziny nauk społecznych i behawioralnych termin „stochastic independence" stosowany jest w psychologii do opisu zjawiska z dziedziny procesów zapamiętywania, polegającego na odrębności uświadamianych i nieuświadamianych obszarów pamięci (Toth 2000: 248–249). W literaturze poświęconej metodom analizy tablic określenie „stochastic independence" raczej nie występuje.

niezależności stochastycznej dla tablic. Oznaczmy liczebności poszczególnych pól modelu niezależności stochastycznej jako e_{ij}. Ich wielkości są funkcją liczebności w polach rozkładów brzegowych obu zmiennych

$$e_{ij} = \frac{a_i \times b_j}{n} \tag{3.10}$$

Przy czym liczebności pól modelu sumują się do rozkładów brzegowych tablicy, na podstawie której model został skonstruowany

$$\sum_{j=1}^{k} e_{ij} = a_i \tag{3.11}$$

$$\sum_{i=1}^{w} e_{ij} = b_j \tag{3.12}$$

Tym samym liczebności wszystkich pól modelu sumują się do sumy ogółem wszystkich liczebności oryginalnej tablicy

$$\sum_{i=1}^{w}\sum_{j=1}^{k} e_{ij} = n \tag{3.13}$$

Omawiany model niezależności zachowuje więc marginesy rozpatrywanej tablicy. Być może właśnie dzięki temu zyskał popularność jako model referencyjny w metodach analizy tablic, w których przyjmuje się, że rozkłady brzegowe tablicy są ustalone i stanowią kontekst oceny układu liczebności w jej wnętrzu.

Wróćmy do tabeli 2.6 z rozdziału 2, w której przedstawiono sposoby głosowania kobiet i mężczyzn w wyborach do sejmu we wrześniu 2005 roku. Dla tabeli tej skonstruujmy model niezależności korzystając z wzoru (3.10). Pełny zestaw liczebności tego modelu został przedstawiony w tabeli 3.3. Przykładowo, liczebność e_{11} – odpowiadająca liczbie kobiet głosujących na PiS – obliczona została jako

$$\frac{568{,}194... \times 405{,}851...}{1061{,}834...} = 217{,}174...$$

Liczebności w polach modelu niezależności z reguły wyrażają się liczbami niecałkowitymi. Ze względu na ilorazową postać wzoru (3.10) ma to miejsce również wtedy, gdy rozkłady brzegowe obu cech wyrażają się liczbami całkowitymi. W tabeli 3.3 obliczone wielkości podano w postaci zaokrąglonej do jednej cyfry po przecinku.

Tabela 3.3
Liczebności modelu niezależności dla sposobów głosowania kobiet i mężczyzn w wyborach do Sejmu we wrześniu 2005 roku
Europejski Sondaż Społeczny 2006

	sposób głosowania								
płeć	**ogółem**	PiS	PO	SLD	Samo-obrona	PSL	LPR	inne partie	nie pamięta
ogółem	**1061,8**	**405,9**	**280,8**	**85,5**	**84,2**	**27,5**	**20,8**	**23,2**	**133,9**
kobiety	**568,2**	217,2	150,3	45,8	45,1	14,7	11,1	12,4	71,7
mężczyźni	**493,6**	188,7	130,6	39,8	39,2	12,8	9,7	10,8	62,3

Liczebności w polach modelu obliczono według wzoru (3.10) na podstawie rozkładów brzegowych płci i sposobu głosowania (tabela 2.6). Wielkości podano zaokrąglone do jednej cyfry po przecinku.

3.4 Niezależność jako identyczność profili

Z postaci modelu niezależności prezentowanej w tabeli 3.3 niewiele wynika, gdy chodzi o interpretację liczebności w poszczególnych polach tego modelu. Lepiej zaprezentować je w postaci odsetków obliczonych w wierszach lub w kolumnach. Model niezależności stochastycznej jest bowiem równoważny **identyczności profili**. Jeśli dla tablicy E zawierającej wielkości pól modelu niezależności obliczone zostaną profile w wierszach, to we wszystkich wierszach tablicy będą one identyczne a zarazem takie same jak profil brzegowy zmiennej umieszczonej w kolumnach

$$pw_{i_1 j} = pw_{i_2 j} = \frac{b_j}{n} \tag{3.14}$$

dla dowolnych $i_1, i_2 = 1,2,...,w$ oraz dla każdego $j = 1,2,...,k$. Analogicznie, wszystkie profile obliczone dla kolumn będą identyczne a zarazem jednakowe jak profil brzegowy zmiennej umieszczonej w wierszach tablicy

$$pk_{ij_1} = pk_{ij_2} = \frac{a_i}{n} \tag{3.15}$$

dla każdego $i = 1,2,...,w$ oraz dla dowolnych $j_1, j_2 = 1,2,...,k$.

Podana własność modelu niezależności stochastycznej zachodzi również w drugą stronę. Jeśli w wyniku przeprowadzonego badania stwierdzilibyśmy, że wszystkie profile w wierszach lub w kolumnach tablicy są identyczne, to otrzymane w badaniu liczebności tablicy tworzyłyby model niezależności. Ich

wielkości byłyby wtedy ściśle zgodne z wzorem (3.10). Co prawda, sytuacja taka jest w praktyce mało prawdopodobna, lecz dobrze oddaje ideę niezależności w tablicy otrzymanej w wyniku realizacji badania. Z tego zapewne powodu w wielu podręcznikach metod statystycznych w ten właśnie sposób wyjaśnia się pojęcie niezależności (Agresti i Finlay 2008: 223; Lissowski, Haman i Jasiński 2008: 214–216)

W celu zilustrowania takiego rozumienia niezależności rozważmy profile obliczone dla wierszy (tabela 3.4) i kolumn (tabela 3.5) modelu niezależności stochastycznej z tabeli 3.3.

Rozpocznijmy od analizy odsetków podanych w tabeli 3.4. Gdyby w faktycznie zrealizowanym badaniu otrzymano odsetki podanej wielkości, to zapewne poczynione obserwacje byłyby następujące. Wśród kobiet odsetek głosujących na kandydatów PiS jest taki sam, jak wśród mężczyzn (po 38,2%). Ponadto, te same odsetki kobiet i mężczyzn oddały głos na kandydatów PO (po 26,4%). Identyczne są również odsetki kobiet i mężczyzn, którzy głosowali na SLD ... i tak dalej, w wypadku każdego z uwzględnionych w tabeli sposobów głosowania. Podsumowaniem tych obserwacji jest bez wątpienia wniosek, że zgodnie z wynikami badania sposób głosowania kobiet w wyborach **nie różnił się** od sposobu głosowania mężczyzn. Stąd blisko do stwierdzenia, że sposób głosowania **nie zależał** od płci. Identyczność profili przekłada się więc na intuicje związane z pojęciem niezależności w takim sensie, w jakim używamy tego słowa, mówiąc, że jedna cecha nie zależy od innej.

Tabela 3.4
Odsetki osób głosujących na kandydatów poszczególnych partii wśród ogółu badanych oraz wśród kobiet i mężczyzn w modelu niezależności
Europejski Sondaż Społeczny 2006
[*w procentach*]

	sposób głosowania								
płeć	**ogółem**	PiS	PO	SLD	Samo-obrona	PSL	LPR	inne partie	nie pamięta
ogółem	**100,0**	**38,2**	**26,4**	**8,1**	**7,9**	**2,6**	**2,0**	**2,2**	**12,6**
kobiety	**100,0**	38,2	26,4	8,1	7,9	2,6	2,0	2,2	12,6
mężczyźni	**100,0**	38,2	26,4	8,1	7,9	2,6	2,0	2,2	12,6

Obliczono na podstawie liczebności modelu niezależności prezentowanego w tabeli 3.3.

Do analogicznych intuicji prowadzi analiza odsetków kobiet i mężczyzn wśród respondentów, którzy oddali swoje głosy na kandydatów poszczególnych partii (tabela 3.5). Rozumowanie przebiegałoby tu następująco. Wśród

osób, które oddały głos na kandydata PiS, było 53,5% kobiet i 46,5% mężczyzn. Wśród respondentów głosujących na PO było również 53,5% kobiet i 46,5% mężczyzn. Idąc dalej – te same odsetki kobiet i mężczyzn powtarzają się w wypadku każdej partii uwzględnionej w tabeli. Stąd wniosek, że odsetki kobiet i mężczyzn są takie same **niezależnie** od partii, na którą głosowano.

Tabela 3.5
Odsetki kobiet i mężczyzn wśród ogółu głosujących oraz wśród głosujących na kandydatów poszczególnych partii w modelu niezależności
Europejski Sondaż Społeczny 2006
[*w procentach*]

	płeć		
sposób głosowania	kobiety	mężczyźni	ogółem
ogół głosujących	**53,5**	**46,5**	**100,0**
PiS	53,5	46,5	100,0
PO	53,5	46,5	100,0
SLD	53,5	46,5	100,0
Samoobrona	53,5	46,5	100,0
PSL	53,5	46,5	100,0
LPR	53,5	46,5	100,0
pozostałe partie	53,5	46,5	100,0
nie pamięta	53,5	46,5	100,0

Obliczono na podstawie liczebności modelu niezależności prezentowanego w tabeli 3.3.

3.5 Nadrzędna rola marginesów

Aby lepiej zrozumieć istotę modelu niezależności warto raz jeszcze odwołać się do budowy tabliczki mnożenia. W tabliczce mnożenia rezultat w każdym z pól jest iloczynem liczb odpowiadających danemu wierszowi i kolumnie. Układ liczebności w polach modelu niezależności stochastycznej ma z tym wiele wspólnego. Liczebności te – jak wynika z wzoru (3.10) – są proporcjonalne do liczebności brzegowych obu cech. Im większe są liczebności brzegowe w wierszu i kolumnie, tym większa będzie liczebność w polu leżącym na przecięciu danego wiersza i kolumny. We wzorze (3.10) czynnik podany w mianowniku (n) odpowiada łącznej liczbie obiektów w tablicy, przez co jest identyczny dla wszystkich pól modelu. Nie wpływa więc na zasady różnicujące wielkości pól w modelu niezależności.

Analogia z tabliczką mnożenia może być nawet pełna, jeśli wzór (3.10) przekształcimy do następującej postaci:

$$e_{ij} = \frac{a_i}{\sqrt{n}} \times \frac{b_j}{\sqrt{n}} \tag{3.16}$$

Obecnie liczebność e_{ij} jest iloczynem dwóch wielkości. Nazwijmy je **potencjałami** mx_i oraz my_j odpowiadającymi kategoriom umieszczonym w danym wierszu i danej kolumnie tablicy. Wielkości potencjałów wyrażają się więc za pomocą wzorów

$$mx_i = \frac{a_i}{\sqrt{n}} \tag{3.17}$$

oraz

$$my_j = \frac{b_j}{\sqrt{n}} \tag{3.18}$$

zaś wielkość e_{ij} jest ich iloczynem, czyli

$$e_{ij} = mx_i \times my_j \tag{3.19}$$

Analogia modelu niezależności do tabliczki mnożenia jest obecnie pełna.

Tabela 3.6
Potencjały i liczebności modelu niezależności między płcią a sposobem głosowania w wyborach do Sejmu we wrześniu 2005 roku obliczone na podstawie wyników Europejskiego Sondażu Społecznego 2006.

		sposób głosowania oraz potencjały my_j							
		PiS	PO	SLD	Samo-obrona	PSL	LPR	inne partie	nie pamięta
płeć i potencjały mx_i		**12,5**	**8,6**	**2,6**	**2,6**	**0,8**	**0,6**	**0,7**	**4,1**
kobiety	**17,4**	217,2	150,3	45,8	45,1	14,7	11,1	12,4	71,7
mężczyźni	**15,1**	188,7	130,6	39,8	39,2	12,8	9,7	10,8	62,3

Liczebności w polach modelu obliczono według wzoru (3.19) na podstawie potencjałów określonych przez wzory (3.17) i (3.18).

W tabeli 3.6 przedstawiono potencjały dla kobiet i mężczyzn oraz dla poszczególnych sposobów głosowania. Podano również liczebności modelu niezależności obliczone jako iloczyny tych potencjałów. Liczebności są zgodne z przedstawionymi w tabeli 3.3. Ponieważ potencjały są proporcjonalne do liczebności brzegowych, stąd między innymi potencjał dla kobiet (17,4) jest

większy od potencjału dla mężczyzn (15,1). Badanych kobiet było bowiem więcej, toteż ich znaczenie w wyznaczaniu wielkości pól modelu niezależności jest większe. Znajduje to odzwierciedlenie w wielkościach podanych we wnętrzu tablicy. Dla każdego sposobu głosowania, to jest w każdej kolumnie tablicy, liczebność modelu niezależności dla kobiet jest większa od liczebności dla mężczyzn.

Różny jest też potencjał poszczególnych sposobów głosowania. Największe potencjały odpowiadają PiS i PO – to jest partiom, na kandydatów których głosowało najwięcej badanych osób. Fakt ten oznacza, że w modelu niezależności liczebności dla obu tych partii będą proporcjonalnie większe, niż dla pozostałych. Najniższy potencjał odpowiada Lidze Polskich Rodzin (LPR), gdyż najmniej badanych wskazało LPR jako partię, na którą oddali głos. W efekcie liczebności modelu niezależności dla tej partii są również najmniejsze – zarówno wśród kobiet, jak i wśród mężczyzn.

Dane w tabeli 3.6 obrazują mechanizm generujący niezależność. Liczebności w poszczególnych polach modelu niezależności są **wypadkową potencjałów** odpowiadających im kategorii obu cech. W ujęciu tym wielkość potencjałów proporcjonalna jest do liczby badanych jednostek w tych kategoriach.

Idea niezależności jako wypadkowej potencjałów przynależnych wierszom i kolumnom znalazła zastosowanie w wielu metodach analizy tablic. W metodach modelowania log-liniowego, o których nieco szerzej piszę w rozdziale 4, model estymacji liczebności pól tablicy opiera się na iloczynie obu potencjałów, nazywanych efektami wiersza i kolumny. W ramach metody zwanej analizą korespondencji, której poświęcony został rozdział 7, potencjały wiersza i kolumny tablicy nazywane są „masami". Stanowi to nawiązanie do klasycznej mechaniki, gdzie efekt oddziaływania na siebie obiektów zależy od wielkości ich mas.

Jeśli wielkości w polach modelu niezależności uznamy za wypadkową liczebności brzegowych, czyli marginesów tablicy, to tym samym zakładamy **pierwotną**, czy też **nadrzędną rolę marginesów** wobec kształtu modelu. Zastanówmy się przez chwilę, jak można to interpretować. Czy na poziomie badanej rzeczywistości marginesy można traktować jako ustalone i niezależne od kształtu rozpatrywanego zjawiska.

W wypadku niektórych zjawisk interpretacja taka wydaje się uzasadniona. Weźmy jako przykład tablicę, w której skrzyżowano wykształcenie ojców absolwentów gimnazjów z rodzajem szkoły ponadgimnazjalnej, w której absolwenci ci kontynuują naukę. Interesuje nas w tym wypadku jak silnie wykształcenie ojca określa wybór szkoły ponadgimnazjalnej przez dziecko. Przyjmijmy, że przedmiotem badania są wszyscy absolwenci z danego rocznika. Przyjmijmy też, że interesuje nas podział szkół ponadgimnazjalnych na trzy kategorie: licea ogólnokształcące, technika oraz szkoły zasadnicze zawodowe.

Liczba miejsc dostępnych w wyodrębnionych kategoriach szkół jest określona przez aktualny stan systemu edukacyjnego. Określony jest również rozkład wykształcenia ojców. Zjawisko polegające na wyborze szkoły w zależności od wykształcenia ojca można więc wyobrazić sobie jako **wypełnianie pustej tablicy**, która **ma określone oba marginesy**. Przy czym nie ma możliwości, aby wielkości tych marginesów zostały przekroczone. Jeśli zdarzyłoby się, że dzieci, których ojcowie mają wykształcenie wyższe, zajęłyby wszystkie miejsca w liceach ogólnokształcących – chociażby ze względu na lepsze wyniki testów kompetencyjnych – to dzieci ojców o pozostałych poziomach wykształcenia musiałyby z konieczności wybrać inne szkoły. Marginesy wyznaczają więc pewne ramy, które należy uwzględnić tworząc model niezależności, czy też dowolny model badanego związku. Traktowanie marginesów jako nadrzędnych czy pierwotnych wobec kształtu konstruowanego modelu znajduje tu dobre osadzenie w istocie zjawiska, które badamy.

3.6 Marginesy jako dyspozycje do zachowań

Wiele zjawisk przebiega jednak według innego schematu, niż opisany wyżej, przez co nie można ich traktować jako wypełnianie pustej tablicy o ustalonych z góry marginesach. Wyjaśnijmy to na przykładzie głosowania w wyborach. W tym wypadku chodzi jednak nie o omawiane wcześniej odpowiedzi udzielane przez respondentów na temat swojego głosowania w momencie, gdy wybory już dawno się zakończyły, lecz faktyczny mechanizm oddawania głosów przez wyborców. Przyjmijmy dla zachowania analogii, że interesuje nas związek sposobu głosowania z płcią. Uwagę skupimy więc na tym, na kogo głosują mężczyźni i kobiety.

Omawiane zjawisko przedstawić można w następujący sposób. O 6.00 rano otwarte zostają drzwi lokali wyborczych. W tym momencie pola tablicy są jeszcze niewypełnione. Nie znamy ani liczebności wnętrza tablicy, ani też jej marginesów. Do jednego z lokali wchodzi kobieta i oddaje swój głos na kandydata PiS. Do odpowiedniego pola tablicy wpisujemy więc jedynkę. Następnie do lokalu wchodzi mężczyzna i oddaje głos na kandydata Samoobrony. Kolejną jedynkę wpisujemy w kratce odpowiadającej mężczyznom głosującym na Samoobronę. Procedurę kontynuujemy aż do zamknięcia drzwi lokali wyborczych. Dopiero w tym momencie określone zostały liczebności w polach tablicy, odpowiadające liczbie mężczyzn i kobiet głosujących na kandydatów poszczególnych partii. Na ich podstawie obliczyć można margines sposobu głosowania w wyborach (czyli rzeczywiste wyniki wyborów), jak też ustalić liczbę kobiet i mężczyzn, którzy zdecydowali się w wyborach wziąć udział.

Marginesy tak otrzymanej tablicy są wypadkową liczebności w polach jej wnętrza. Twierdzenie, że mają one pierwotny czy zewnętrzny charakter wobec badanego zjawiska, nie znajduje tu dobrej operacyjnej interpretacji. Sumujemy bowiem zdarzenia, które stanowią wyraz indywidualnych decyzji jednostek. Przypomnijmy sobie atmosferę przed każdymi wyborami. Zażarte dyskusje w domach i wśród znajomych, komentowanie wystąpień kandydatów, napięcia związane z wynikami sondaży przedwyborczych. Przynajmniej znaczna część wyborców trafia do poszczególnych pól tablicy mając przekonanie, że ich wybór jest świadomy i dobrze uzasadniony. Zresztą twierdzenie, że margines tablicy – czyli wyniki głosowania – ma charakter pierwotny wobec aktu głosowania byłoby ewidentnie sprzeczne z samą ideą wyborów, których celem jest właśnie ustalenie owego marginesu drogą niczym nie skrępowanych decyzji wyborców.

Można jednak do problemu podejść w inny sposób. Zamiast przyjmować końcowy wynik wyborów jako swoiste ramy, w których osadzone są indywidualne decyzje wyborców, można założyć, że wśród osób idących do urn wyborczych istnieją pewne ukryte dyspozycje co do oddania głosu na kandydatów poszczególnych partii. Są one pochodną przekonań, sympatii politycznych, stopnia akceptacji argumentów komunikowanych w fazie kampanii, bądź też innych czynników, które potencjalnie mogą wpływać na sposób głosowania. Owe ukryte rozkłady skłonności decydują o nierównych szansach głosowania na poszczególne partie. A dowodów, że takie ukryte dyspozycje istnieją, dostarczają sondaże przedwyborcze, które niekiedy z dużą trafnością są w stanie przewidzieć faktyczne rezultaty wyborów. Model oparty na założeniu, że niczym nieskrępowane decyzje wyborców składają się w sumie na wynik do przewidzenia, wydaje się wart przynajmniej rozważenia.

Dyspozycje działają przy tym inaczej niż ograniczenia związane z ustaloną liczbą miejsc w poszczególnych rodzajach szkół. Nie należy ich traktować tak jak zachowania, lecz jako prawdopodobieństwa tych zachowań. Przy takiej interpretacji nie jest ważne, czy głos na daną partię odda Iksiński, czy Ygrekowski. Nie jest też ważne, czy będzie to kobieta, czy mężczyzna. Rozkład dyspozycji jest bowiem atrybutem kolektywnym, przynależnym ogółowi wyborców. I może być niezależnie badany – jak dowodzą tego sondaże przedwyborcze. Interpretacja marginesów w kategoriach **dyspozycji** do określonych zachowań wykracza więc poza zjawisko opisane w tablicy, lecz stwarza dogodny punkt odniesienia dla wyjaśniania zachowań składających się rozpatrywane zjawisko.

Innym przykładem, który ilustruje użyteczność pojęcia dyspozycji do zachowań, jest związek między wiekiem mężczyzny i kobiety w momencie ślubu. Przypuśćmy, że rozważamy małżeństwa zawarte w okresie danego roku.

Urzędy stanu cywilnego działają inaczej niż Ministerstwo Edukacji. Nie wyznaczają w swoich planach, że w ciągu roku udzielą ślubu 10 tysiącom mężczyzn w wieku 25 lat, a tylko stu mężczyznom w wieku 59 lat, w związku z czym sto pierwszy 59-letni mężczyzna ślubu nie dostanie. Podane wielkości są jednak w stanie dość dokładnie przewidzieć. Można zrobić to chociażby na podstawie danych z poprzedniego roku.

Skąd jednak bierze się omawiana regularność? Jej podstawą są właśnie dyspozycje do zawarcia małżeństwa w określonym wieku. Mają one bardzo mocne podstawy, nie tylko demograficzne, społeczne, czy kulturowe, ale przede wszystkim biologiczne – związane z imperatywem utrzymania ciągłości gatunku. Dyspozycje te tworzą ramy dla rynku małżeńskiego. A ponieważ nieco różnią się dla mężczyzn i kobiet, stąd strategie poszukiwania partnera są niejednakowe dla obu płci. Byłoby sporą rozrzutnością, gdyby rozkładu owych dyspozycji, czyli marginesów tablicy, nie uwzględniać wyjaśniając mechanizmy zawierania małżeństw.

3.7 Czy marginesy tablicy poprawnie odzwierciedlają kształt zjawiska

Jeżeli przyjmiemy, że marginesy tablicy stanowią podstawę konstruowania modeli opisujących układ liczebności w jej wnętrzu, to z miejsca pojawia się pytanie o poprawność oszacowania tych marginesów. Sprawa nie jest jednak prosta, gdyż w wielu wypadkach kształtu marginesów nie można ustalić inaczej, niż przeprowadzając stosowne badanie. Wyjaśnijmy to na przykładzie fikcyjnego badania, którego przedmiotem są wypadki drogowe. Przyjmijmy, że interesującym nas zjawiskiem jest związek między faktem zapięcia pasów bezpieczeństwa a tym, czy wypadek okazał się śmiertelny[4]. Do opisu badanego zjawiska skonstruujmy tablicę, która w boczku zawiera informację o tym, czy w momencie wypadku kierowca miał zapięty pas bezpieczeństwa (tak lub nie), zaś w główce skutek wypadku (śmiertelny bądź nie). Niech badanie polega na rejestracji przez okres roku wszystkich wypadków, do których dochodzi w obrębie dużego skrzyżowania, gdzie niebezpieczne kolizje zdarzają się szczególnie często. Procedura badawcza polega więc na systematycznym uzupełnianiu liczebności w odpowiednich polach tablicy, w zależności od skutków wypadku

[4] Przykład zaczerpnięty został z podręcznika Agrestiego (2002: 40–41), aczkolwiek w oryginale posłużono się nim w innym celu (dla uzasadnienia wyboru funkcji generującej liczebności w polach tablicy w zależności od schematu realizacji badania). Przedstawione w tabeli 3.7 liczebności oraz ich interpretacje pochodzą całkowicie ode mnie.

oraz tego, czy kierowca miał zapięty pas. Liczebność w każdym z pól generuje się niezależne od liczebności w pozostałych polach, zaś suma wszystkich czterech pól tablicy odpowiada liczbie wypadków w czasie objętym obserwacją. Przyjmijmy, że w ciągu roku zarejestrowano 1000 wypadków, zaś poszczególne pola osiągnęły wielkości podane w tabeli 3.7 (schemat 1). Dla podkreślenia, że pola te mają charakter pierwotny wobec pozostałych liczebności modelu – w tabeli 3.7 oznaczono je szarym kolorem.

Pomimo tego, że marginesy obu cech są wyłącznie odbiciem liczebności wnętrza tablicy – traktowanie ich jako zewnętrznych ram zjawiska dostarcza pewnych korzyści. Dostarcza przede wszystkim interpretacji dla modelu niezależności, który stanowi dogodny punkt odniesienia analizy badanego zjawiska. Liczebności modelu niezależności podane zostały w prawym panelu tabeli 3.7. Wynika z nich, że gdyby śmiertelny skutek wypadku nie zależał od faktu zapięcia pasa, to wśród kierowców, którzy pasa nie zapięli, należałoby spodziewać się tylko 8 wypadków śmiertelnych. Faktycznie było ich 24. Model pozwala więc sformułować wniosek na temat mechanizmu badanego zjawiska. Wniosek ten brzmiałby następująco: „nie zapięcie pasów trzykrotnie zwiększa ryzyko śmiertelnych obrażeń w stosunku do sytuacji, w której zapięcie pasów nie miało żadnego znaczenia dla śmiertelnych skutków wypadku". Na podstawie uzyskanych wyników można wypowiadać się również na temat rozmiarów zjawiska. Wiemy bowiem, ile wypadków zdarza się na badanym skrzyżowaniu w ciągu roku oraz jaka część z nich kończy się śmiertelnym skutkiem. Jesteśmy w stanie wręcz prognozować, o ile ofiar byłoby mniej, gdyby wszyscy kierowcy zapinali pasy. Wszystko to dzięki temu, że marginesy – będące sumą czterech rodzajów zdarzeń składających się na badane zjawisko – poprawnie odzwierciedlają jego kształt.

Rozważmy obecnie nieco zmodyfikowany schemat realizacji tego badania. Badacz doszedł do wniosku, że czekanie rok to zbyt długo, aby dowiedzieć się, czy fakt nie zapięcia pasów zwiększa ryzyko śmiertelnych skutków wypadku. Dlatego zrezygnował z obserwacji i postanowił wylosować z kartotek policyjnych 1000 opisów wypadków z poprzednich lat, tak aby na ich podstawie odtworzyć liczebności tablicy. Okazało się jednak, że opisy wypadków ze skutkiem śmiertelnym i bez takich skutków archiwizowane są odrębnie, zaś połączenie obu rejestrów jest niewykonalne. Badacz stanął więc przed decyzją, ile rekordów wylosować osobno z każdego rejestru. Nie mając lepszego pomysłu, wylosował ich po 500. Analizując rekordy wylosowane z rejestru wypadków bez skutków śmiertelnych, ustalił, że 408 kierowców miało zapięty pas, zaś 92 nie miało zapiętego pasa w momencie wypadku. Liczebności te wpisał do pierwszej kolumny tabeli (tabela 3.7, schemat [2]). Analogiczne obliczenia dla rekordów wylosowanych z rejestru wypadków śmiertelnych pozwoliły mu

ustalić, że 200 kierowców miało zapięty pas, zaś 300 nie. Po wpisaniu dwóch ostatnich wielkości do wnętrza tabeli obliczył margines cechy – czy kierowca miał zapięty pas.

Z utworzonej w ten sposób tablicy wynika dużo mniej, niż poprzednio. Przede wszystkim, żadnych wniosków nie dostarcza analiza liczebności w wierszach. Na przykład wniosek, że skutkiem śmiertelnym kończy się 76 procent wypadków, w których kierowca nie miał zapiętego pasa (tabela 3.7,

Tabela 3.7
Liczebności obserwowane oraz liczebności modelu niezależności dla dwóch schematów badania związku między zapięciem pasów bezpieczeństwa a uniknięciem śmiertelnych urazów w wypadkach drogowych. Dane fikcyjne.

Schemat 1
Rejestracja wszystkich wypadków

[1] liczebności

	skutki śmiertelne		
pas zapięty	nie	tak	ogółem
nie	176	24	200
tak	784	16	800
ogółem	960	40	1000

[2] odsetki w kolumnach (w procentach)

	skutki śmiertelne		
pas zapięty	nie	tak	ogółem
nie	18	60	20
tak	82	40	80
ogółem	100	100	100

[3] odsetki w wierszach (w procentach)

	skutki śmiertelne		
pas zapięty	nie	tak	ogółem
nie	88	12	100
tak	98	2	100
ogółem	96	4	100

[4] liczebności modelu niezależności

	skutki śmiertelne		
pas zapięty	nie	tak	ogółem
nie	192	8	200
tak	768	32	800
ogółem	960	40	1000

Schemat 2
Losowanie z osobnych kartotek

[1] liczebności

	skutki śmiertelne		
pas zapięty	nie	tak	ogółem
nie	92	300	392
tak	408	200	608
ogółem	500	500	1000

[2] odsetki w kolumnach (w procentach)

	skutki śmiertelne		
pas zapięty	nie	tak	ogółem
nie	18	60	39
tak	82	40	61
ogółem	100	100	100

[3] odsetki w wierszach (w procentach)

	skutki śmiertelne		
pas zapięty	nie	tak	ogółem
nie	23	77	100
tak	67	33	100
ogółem	50	50	100

[4] liczebności modelu niezależności

	skutki śmiertelne		
pas zapięty	nie	tak	ogółem
nie	196	196	392
tak	304	304	608
ogółem	500	500	1000

Kolorem szarym oznaczono liczebności, których wielkości wynikają z przyjętego schematu realizacji badania.

część [3]), jest nieprawdziwy. Znając zasadę budowy tablicy, można jedynie odczytać, że w 60 procentach wypadków kończących się śmiertelnymi obrażeniami kierowca nie miał zapiętych pasów, zaś w wypadkach bez skutku śmiertelnego odsetek ten wynosił tylko 18 procent (tabela 3.7, część [2]). Pozwala to sformułować wniosek na temat mechanizmu zjawiska (zapięcie pasów zmniejsza ryzyko śmiertelnych obrażeń), chociaż wniosek ten nie jest już tak oczywisty jak poprzednio i trzeba się nieco nagłowić, aby do niego dojść. Zupełnie nic nie można natomiast powiedzieć na temat rozmiarów zjawiska. Nie znamy bowiem ani odsetka wypadków kończących się skutkiem śmiertelnym, ani też odsetka kierowców niezapinających pasów.

Właściwie można powiedzieć, że utworzenie tabeli nie było tu potrzebne. Wystarczyło porównać ze sobą odsetki kierowców niezapinających pasów w kategoriach wypadków kończących się skutkiem śmiertelnym i pozostałych. To wszystko, co można było zrobić przy zastosowanym schemacie realizacji badania. Tablica okazała się bezużytecznym narzędziem z tego powodu, że jej marginesy nie odzwierciedlają kształtu badanego zjawiska. Z tego samego powodu nie ma sensu tworzyć modelu niezależności. Spójrzmy, co mówi model dla drugiego schematu badania, przedstawiony w części [4] tabeli 3.7. Liczebności w jego wierszach powielają wyłącznie decyzję badacza dotyczącą jednakowych proporcji, w jakich dobrał opisy wypadków z obu rejestrów. Nie dają natomiast żadnego wglądu w hipotetyczną sytuację, w której skutki śmiertelne wypadków byłyby niezależne od faktu zapięcia pasów.

Niezależność jako model referencyjny sprawdza się więc wyłącznie w tablicach, których marginesy poprawnie odzwierciedlają kształt badanego zjawiska. Modelu tego nie uda się jednak skonstruować w taki sposób, aby jego marginesy były inne, niż rozpatrywanej tablicy, co wynika z wzorów (3.10)–(3.12). Sprowadza to problem adekwatności marginesów modelu niezależności do problemu poprawności marginesów samej tablicy, a te z kolei są wyłącznie sumami liczebności w jej wnętrzu. Czyli, jeśli kształt zjawiska w tablicy przedstawiony jest w sposób wypaczony, to wnioski z analizy jej zawartości będą również wypaczone. Odnoszą się bowiem do tego, co przedstawione zostało w tablicy, nie zaś do rzeczywistego kształtu zjawiska.

Doszliśmy tym samym do dość oczywistej konkluzji, że fałszywe dane prowadzą do fałszywych wniosków. Cóż więc jest w tym szczególnego, gdy rozpatrujemy model niezależności? Otóż jest pewien punkt, na który warto zwrócić uwagę. Ilustruje go właśnie rozpatrywany przykład badania wypadków drogowych.

W drugim ze schematów badacz losował określoną liczbę przypadków z dwóch kartotek, w których gromadzono informacje o wypadkach kończących się skutkiem śmiertelnym oraz bez takich skutków. Z obu kartotek badacz

wylosował jednakową liczbę rekordów co równoważne było nieprawdziwemu założeniu, że oba rodzaje wypadków zdarzają się równie często. Zwróćmy jednak uwagę, że w obrębie każdej z kartotek zjawisko było opisane poprawnie, to znaczy odsetki kierowców, którzy mieli zapięte pasy bezpieczeństwa są takie same jak otrzymane w badaniu polegającym na rocznej obserwacji faktycznie zdarzających się wypadków (tabela 3.7, część [2]). Aby więc utworzyć tablicę poprawnie opisującą zarówno mechanizm jak i rozmiary badanego zjawiska wystarczyło ustalić, jaki odsetek wypadków ma skutki śmiertelne. Być może informują o tym wielkości obu kartotek, o ile archiwizowane są w nich wszystkie wypadki. Być może informację tę uzyskać można w inny sposób. Gdyby była ona znana, to wtedy postąpić można na jeden z dwóch sposobów. Albo wylosować z każdej kartoteki liczbę rekordów proporcjonalną do odsetka danego rodzaju wypadków. Bądź też dokonać ważenia wylosowanych rekordów, tak aby margines obu rodzajów wypadków odzwierciedlał swoim kształtem rzeczywiste odsetki ich występowania.

Ważenie jest metodą powszechnie obecnie stosowaną. Badacze pytają o wiele cech, których rozkłady są znane z innych źródeł. Należą do nich cechy demograficzne. Jeśli, przykładowo, wśród zbadanych osób jest 65 procent kobiet[5], zaś z danych GUS wynika, że kobiety stanowią 52 procent badanej zbiorowości, to ważenie stanowi rutynową metodę uzyskania odsetków płci zgodnych z danymi dla populacji. W ten sposób korygować można elementy procedury badania, które prowadzić mogą do niepoprawnych szacunków udziałów różnych segmentów badanej populacji wśród zbadanych osób. Prowadzi to zarazem do skorygowania marginesów wielu tablic.

Ważenie niewiele jednak pomoże, gdy profile w ramach ważonych kategorii oszacowane zostały nieprawidłowo. Powróćmy do omawianego przykładu wypadków drogowych. Przyjmijmy, że w kartotece opisującej wypadki bez skutków śmiertelnych archiwizowane są wszystkie wypadki, w których kierowca nie miał zapiętych pasów, zaś gdy kierowca miał zapięte pasy, to odnotowywane są tylko te wypadki, w których w samochodzie byli oprócz niego pasażerowie. Realizując badanie według drugiego ze schematów nie będziemy w stanie poprawnie odtworzyć obu wielkości w polach kolumny odpowiadającej wypadkom bez skutków śmiertelnych. Schemat ten staje się wtedy bezużyteczny.

Z podobnym zagrożeniem mamy do czynienia w rzeczywistych badaniach. Szereg czynników spowodować może, że nieprawidłowo oszacowane zostaną

[5] Feminizacja badanych prób jest obecnie jednym z najpoważniejszych problemów realizacji badań. Szczególnie ujawnia się w badaniach telefonicznych, w których ankieter dysponuje ograniczonymi środkami uniknięcia odmowy ze strony wylosowanej osoby.

profile odpowiedzi w podgrupach badanych. Jedynym ratunkiem jest przestrzeganie standardów jakości realizacji badań (Sztabiński, Sawiński i Sztabiński 2005). Nie eliminuje to rzecz jasna wszystkich potencjalnych zagrożeń, lecz zmniejsza szanse ich wystąpienia dzięki uwzględnieniu wiedzy i doświadczeń płynących z wcześniej zrealizowanych badań.

3.8 Niezależność a losowość

Niezależność utożsamia się niekiedy z losowym, czy przypadkowym powstawaniem konfiguracji dwóch cech. Tymczasem związki między oboma pojęciami są złożone. Warto im przyjrzeć się bliżej, aby uniknąć zbyt powierzchownych interpretacji mechanizmów kształtowania się zjawisk. Należy przy tym mieć na uwadze, że o ile pojęcie niezależności cech w tablicy można precyzyjnie zdefiniować, co uczyniliśmy w podrozdziale 3.3, to pojęcie losowości definicji takiej nie posiada. Odpowiada ono wyłącznie pewnym intuicjom dotyczącym mechanizmów łączenia w pary kategorii dwóch cech. Najczęściej ma się tu na myśli przypadkowość czy brak powodów, dla których kategorie łączą się w pary w taki, a nie w inny sposób. Można też przywołać model urny, z której na chybił trafił wyciągamy kule. Intuicyjnego podłoża pojęcia losowości dowodzi fakt, że nie da się go zaimplementować w urządzeniach imitujących ludzkie myślenie. Tak zwane generatory liczb losowych w komputerach w rzeczywistości odczytują dokładny czas komputerowego zegara i przekształcają go do postaci liczb losowych (czy raczej pseudolosowej, jak też się niekiedy je nazywa). Komputer jest urządzeniem zdeterminowanym, przez co pojęcie losowości będzie mu zawsze obce.

Dla określenia relacji między niezależnością a losowością kwestią kluczową jest to, że tablica obejmuje cechy o wartościach skategoryzowanych. Poszczególne pola tablicy odpowiadają przez to nie pojedynczym jednostkom, lecz kategoriom liczącym wiele jednostek. Losowość może więc zarówno mieć miejsce w obrębie poszczególnych pól, jak też określać zasady przydziału jednostek do różnych pól tablicy.

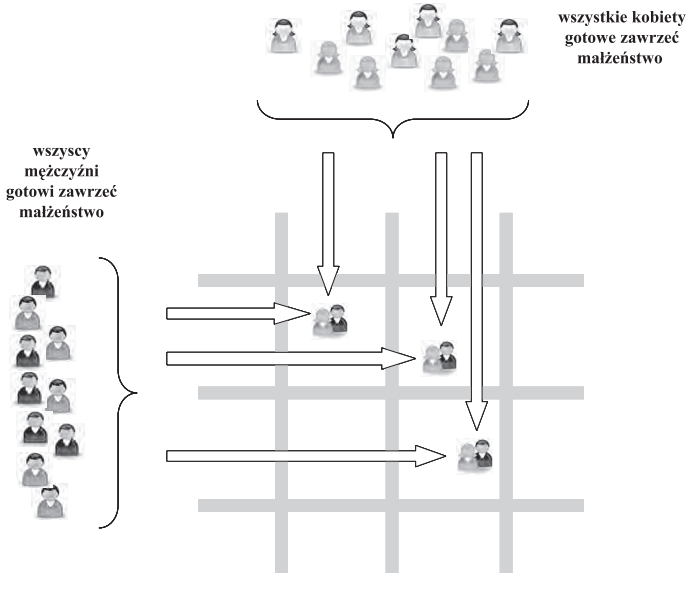
Wyjaśnijmy to na przykładzie tablicy, która opisuje związek wieku mężczyzn i kobiet zawierających małżeństwa. Tablice takie GUS publikuje corocznie i obejmują one małżeństwa zawarte w roku poprzednim. Tego rodzaju tablicę poddamy szczegółowej analizie w rozdziale 4, toteż aby przekonać się, jak ona w rzeczywistości wygląda, można spojrzeć do tabeli 4.5. W tym miejscu wystarczy wiedzieć, że wiek, zarówno mężczyzn, jak i kobiet, jest w tych tablicach grupowany w przedziały: do 19 lat, 20–24 lata, 25–29 lat, i tak dalej aż do ostatniego przedziału – 60 lub więcej lat. Każde z pól tablicy obejmuje

więc wiele małżeństw. Przykładowo, w 2006 roku w polu odpowiadającym małżeństwom zawartym między mężczyznami a kobietami w wieku 25–29 lat było 47,3 tysiąca małżeństw.

Wyobraźmy sobie dość absurdalną sytuację, że państwo wydało dekret, który określa, że małżeństwa zawierane będą drogą losowania. Wszystkie osoby, które pragną zawrzeć małżeństwo w danym roku, zgłaszają swój akces do urzędu stanu cywilnego. Po zebraniu puli wszystkich chętnych urząd z puli tej losuje pary osób różnej płci, które zostaną małżonkami. Utworzone w ten sposób małżeństwa trafiają do poszczególnych pól tablicy w sposób przedstawiony symbolicznie na rycinie 3.1. Każde małżeństwo klasyfikowane jest do określonego pola tablicy w zależności od wieku mężczyzny i wieku kobiety, którzy wylosowani zostali jako para małżeńska. Po rozlosowaniu wszystkich par układ liczebności we wnętrzu tablicy utworzy model niezależności stochastycznej.

Rozważmy teraz inny schemat losowania małżeństw. Zanim wydano dekret, dokonano analizy wieku mężczyzn i kobiet zawierających małżeństwa, uwzględniając dane z wcześniejszych lat. Na tej podstawie oszacowano profile wieku żon osobno dla każdej kategorii wieku mężów. Chodziło o to, aby

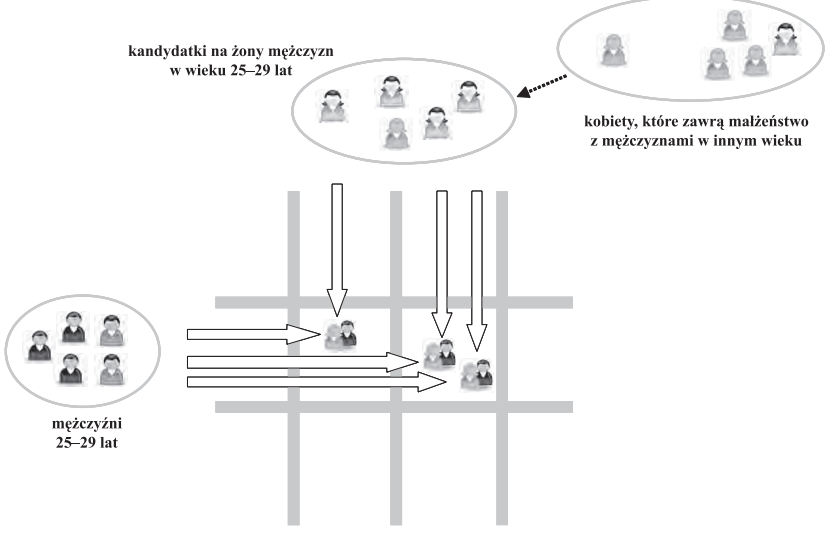
Rycina 3.1
Sposób tworzenia się tablicy przy losowaniu małżeństw spośród wszystkich osób, które zgłosiły gotowość zawarcia małżeństwa



procedura losowania odtwarzała naturalne prawa doboru małżeńskiego pod względem wieku. W przeciwnym wypadku krajowi groziłby spadek przyrostu naturalnego. Procedura losowania ustawiona została więc w taki sposób, aby w odtwarzać liczebności pól tablicy, które znano z lat poprzednich. Można to przedstawić następująco. Weźmy wszystkich mężczyzn w wieku 25–29 lat. Na podstawie wcześniejszych danych ustalono ile kobiet w jakim wieku zawierało małżeństwa z mężczyznami z tej grupy wiekowej. Odpowiednią liczbę kobiet w odpowiednim wieku odlosowano więc z puli wszystkich kobiet, które zgłosiły chęć zawarcia małżeństwa. Kolejny etap procedury przebiegał analogicznie jak poprzednio. Małżeństwa kojarzono poprzez wylosowanie mężczyzny z grupy 25–29 lat oraz kobiety z wcześniej wyłonionej puli. Procedurę tę obrazuje rycina 3.2. Przedstawia ona przykład wypełnienia jednego wiersza tablicy – odpowiadającego mężczyznom w wieku 25–29 lat. W analogiczny sposób wypełniane są wiersze tablicy odpowiadające pozostałym kategoriom wieku mężczyzn.

Druga z zastosowanych procedur losowania odtwarza rzeczywiste liczebności w polach tablicy wieku mężczyzn i kobiet zawierających małżeństwo. Czyli prowadzi do uzyskania analogicznych liczebności, jak dokonujący się

Rycina 3.2
Sposób tworzenia się tablicy przy losowaniu małżeństw spośród wszystkich mężczyzn w wieku 25–29 lat oraz odlosowanej puli kobiet o ustalonym profilu wieku



w rzeczywistości wybór partnera według indywidualnych preferencji. Tablica stanowi zagregowaną formę danych, przez co nie pozwala odtworzyć mechanizmów zjawiska na poziomie **indywidualnym**. A więc, nawet silna zależność między cechami w tablicy otrzymanej w badaniu może mieć podłoże w losowym powstawaniu kombinacji obu cech! Wszystko zależy od tego, w jakiej fazie zjawiska i w jaki sposób działają mechanizmy losowe.

W pierwszym przykładzie mechanizm losowania prowadził do uzyskania stochastycznej niezależności. Nie będę w tym miejscu dowodził, że mechanizm losowości w takiej postaci, w jakiej został przedstawiony, zawsze prowadzić będzie do niezależności stochastycznej. Nie to jest bowiem warte zapamiętania. Bardziej istotne jest, że nigdy nie wolno wyciągać wniosków w drugą stronę. Jeśli w wyniku badania stwierdzimy, że cechy w tablicy są stochastycznie niezależne, to nie oznacza to wcale, że kombinacje obu cech powstają w losowy sposób. Analizowany wcześniej związek płci ze sposobem głosowania w wyborach w niewielkim stopniu odbiegał od niezależności. Nie można więc wykluczyć, że gdybyśmy analizie poddali rzeczywisty sposób głosowania mężczyzn i kobiet, to otrzymalibyśmy tablicę o liczebnościach bardzo bliskich modelowi niezależności stochastycznej. A przecież nie oznacza to, że wyborcy wchodząc do lokalu wyborczego rzucają kostką. Wręcz przeciwnie, większość z nich podejmuje decyzje w świadomy i przemyślany sposób. Niezależność zaobserwować można więc zarówno wtedy, gdy kombinacje obu cech tworzą się losowo, jak też wtedy, gdy tworzą się świadomie.

Fakt, że obserwowane na poziomie społecznym zjawiska mogą mieć różnorodne podłoże, gdy chodzi o indywidualne dążenia, znany jest od dawna. W XIX wieku belgijski badacz i statystyk Adolphe Quételet pisał (Szacki 1981: 289)

> *Cóż bardziej zależnego od wolnej woli jednostek ludzkich jak małżeństwa, przestępstwa, zbrodnie czy samobójstwa? A przecież dane statystyczne wskazują, iż wszystkie te zjawiska odznaczają się daleko posuniętą regularnością.*

Gdy analizujemy zjawisko za pomocą tablicy interesują nas wyłącznie owe regularności. Dopiero po ich ustaleniu, gdy wyjaśniamy badane zjawisko, możemy odwoływać się do motywów ludzkich zachowań, czy też innych mechanizmów, które powodują, że badani trafiają do tych, a nie do innych pól tablicy. Jednych z tych mechanizmów może być losowość. Nie oznacza to jednak, że musi ona wystąpić. A jeśli już wystąpiła, to nie musi przełożyć się na niezależność.

3.9 Niezależność a równość szans

Socjologowie często utożsamiają model niezależności z modelem równych szans. Przyznam się szczerze, że sam nie potrafię się tego wyzbyć. Po napisaniu książki zmuszony byłem drobiazgowo przejrzeć jej zawartość i dokonać korekt we wszystkich fragmentach, w których zamiast napisać „model niezależności" bezwiednie użyłem sformułowania „model równych szans".

Utożsamianie obu pojęć wynika zapewne stąd, że dziedzina badań nad strukturą społeczną jest być może jedyną, w której modelu niezależności używa się *explicite*. W innych dziedzinach nauk społecznych modelu tego niekiedy w ogóle się nie zauważa, gdyż jest szczelnie obudowany innymi pojęciami. Wielu badaczy nie jest przez to świadomych, że metody analizy danych, którymi się posługują, są w rzeczywistości nadbudowane na modelu niezależności. W wypadku badań nad strukturą społeczną, równością, czy sprawiedliwością, od modelu niezależności uciec się nie da. Stanowi on bowiem operacjonalizację pojęcia **równych szans**, które z kolei stanowią warunek *sine qua non* istnienia **merytokratycznego** społeczeństwa (White 2008: 75–95). Społeczeństwa, w którym każdy może dojść do najwyższych pozycji, co zależy wyłącznie od uzdolnień i pracowitości. We wpływowej na rozwój socjologicznego myślenia teorii funkcjonalnej równość szans stanowiła warunek efektywności systemu społecznego (Domański 2007: 49–51). Z tego powodu wszystkie późniejsze ujęcia musiały odnieść się do równości szans (Sawiński 2009). Dzięki temu model równych szans jest znany każdemu socjologowi.

Istota stosowania modelu równych szans w analizach społeczeństw jest analogiczna, jak stosowania modelu niezależności w analizach tablic. Jeden i drugi stanowi wyłącznie punkt odniesienia wobec badanych zjawisk, natomiast nie oczekuje się, aby opisywał te zjawiska. Analogie te trafnie przedstawił australijski badacz F. Lancaster Jones (1985: 839)

> [...] *model niezależności jest modelem utopijnego społeczeństwa, społeczeństwa o całkowitej równości szans. Ze względu na ten fundamentalny powód nie stanowi problemu fakt, że model statystycznej niezależności „... nie jest dopasowany do danych ..."* [Hauser 1978: 924; przyp. *Z.S.*]*, ponieważ tego wymogu nigdy mu nie stawiano. Traktowano go raczej jako normatywne kryterium pozwalające określić, jak blisko dane społeczeństwo doszło do tego konkretnego ideału, o który demokratyczne społeczeństwa walczyły od czasu rewolucji francuskiej* [...] *Z tego powodu statystyczny model niezależności jest oczywistym sposobem operacjonalizacji makrosocjologicznego pojęcia „otwartości społecznej" i podstawowym celem zastosowania tego modelu jest dostarczenie miary pozwalającej wyrazić, jak blisko dane społeczeństwo doszło (bądź odeszło) od tego normatywnego kryterium* [...]

W końcowym fragmencie przytoczonej wypowiedzi Jones nawiązuje do możliwości określenia ilościowego kryterium, które pozwoliłoby porównać badane zjawisko z modelem niezależności. Zagadnieniom tym poświęcona zostanie dalsza część tego rozdziału.

3.10 Statystyczny test niezależności

Dla oceny wielkości różnic między opisanym w tablicy zjawiskiem a modelem niezależności stosuje się dwie metody, oparte na dwóch odmiennych koncepcjach. Pierwsza polega na weryfikacji, czy obserwowane różnice mogą być sprowadzone do **losowych fluktuacji** powstających na skutek tego, że badaniu podlega jedynie próba wylosowana z populacji stanowiącej przedmiot badania. Druga koncepcja polega natomiast na ilościowym określeniu **stopnia**, w jakim obserwowany w tabeli związek między cechami odbiega od modelu niezależności. W podrozdziale tym omówimy wyłącznie pierwszą z koncepcji. Drugą zaś przedstawimy w kolejnym podrozdziale (3.11).

W większości badań dane pochodzą z próby dobranej z badanej populacji. Uzyskany w próbie obraz zjawiska może różnić się w większym lub mniejszym stopniu od jego rzeczywistego kształtu. Im próba jest bardziej liczna, tym rozbieżności te powinny być mniejsze. Wyniki uzyskane w badaniu reprezentacyjnym należy więc oceniać w kontekście wielkości próby. Do tego celu służy test χ^2 (chi-kwadrat), który obraz zjawiska zestawia z modelem niezależności.

$$\chi^2 = \sum_{i=1}^{w} \sum_{j=1}^{k} \frac{(n_{ij} - e_{ij})^2}{e_{ij}} \tag{3.20}$$

Z podanego wzoru odczytać można na czym polega porównanie zjawiska z modelem. W każdym polu tablicy obliczana jest różnica między liczebnością otrzymaną w badaniu a liczebnością modelu niezależności. Następnie każda z tych różnic podnoszona jest do kwadratu i dzielona przez liczbę jednostek, których należałoby się spodziewać w danym polu tablicy, gdyby cechy były niezależne. Ostatnią wymienioną operację interpretować można jako normalizację obserwowanych różnic. Jeśli różnica o danej wielkości odpowiada polu o niewielkiej liczebności, to jej względne znaczenie jest większe niż w wypadku pola, w którym należy oczekiwać dużej liczebności. Owe znormalizowane wielkości różnic sumuje się po wszystkich polach, co daje miarę odstępstw dla całej tablicy, znaną jako współczynnik chi-kwadrat.

Z wzoru (3.21) nie wynika wprost, że suma odstępstw „ma prawo" być większa, gdy próba jest bardziej liczna. Wzór przedstawić jednak można w innej postaci (Mirkin 2001: 112; Lissowski i inni 2008: 331) jako

$$\chi^2 = n * \sum_{i=1}^{w} \sum_{j=1}^{k} \frac{(p_{ij}^{(N)} - p_{ij}^{(E)})^2}{p_{ij}^{(E)}} \tag{3.21}$$

Wielkości w polach uzyskanej w badaniu tablicy N oraz liczebności modelu niezależności E są zastąpione proporcjami, czyli liczebnościami poszczególnych pól znormalizowanymi do liczebności próby. Wyrażenie objęte znakiem sumowania wyraża więc wielkość odstępstw niejako w „czystej" postaci, niezależnej od wielkości próby. Znajdujący się przed znakiem sumowania mnożnik n dowodzi, że przy danej wielkości „czystych" odstępstw wielkość współczynnika χ^2 jest wprost proporcjonalna do liczebności próby.

Podstawą konstrukcji testu statystycznego jest twierdzenie (Lissowski i inni 2008: 642–652)

> *jeśli cechy w populacji są niezależne stochastycznie zaś wyniki w tablicy pochodzą z próby losowej dobranej z tej populacji, to statystyka* χ^2 *o wartościach wyrażających się wzorem (3.21) ma rozkład chi-kwadrat z liczbą stopni swobody równą* $(w - 1) * (k - 1)$, *gdzie w i k oznaczają liczbę wierszy i kolumn tablicy*

W twierdzeniu tym oraz we wcześniejszym fragmencie tekstu termin „chi-kwadrat" pojawia się w trzech różnych znaczeniach, co dla uniknięcia nieporozumień warto wyjaśnić.

Pierwsze z nich to „współczynnik chi-kwadrat". Wynikiem zastosowania wzoru (3.21) jest obliczenie wartości liczbowej. Wszystkie wielkości w tym wzorze oparte są bowiem na liczebnościach pól w tablicy uzyskanej w badaniu. Wielkość chi-kwadrat określona przez wzór (3.21) nazywana jest **współczynnikiem** w znaczeniu konkretnej wartości liczbowej.

Znaczenie to ulega zmianie, gdy mówimy o niezależności cech w populacji oraz o obrazie tej niezależności uzyskanym w próbie. Mamy wtedy do czynienia z różnymi potencjalnymi kształtami tego obrazu, co odpowiada różnym możliwościom wylosowania próby o danej liczebności. Wielkość chi-kwadrat nie jest już więc konkretną liczbą, lecz wartością zmiennej losowej. Tego rodzaju zmienne, których dziedziną jest przestrzeń możliwych prób, nazywane są **statystykami**. W tym sensie użyto określenia „statystyka chi-kwadrat" w podanym twierdzeniu.

Trzecie wystąpienie omawianego terminu to „rozkład chi-kwadrat". Jest to rozkład pewnej zmiennej **teoretycznej**. Co więcej, nie jest to pojedyncza zmienna, lecz pewna klasa zmiennych różniących się wartością parametru nazywanego liczbą stopni swobody. Podane wyżej twierdzenie określa, którą zmienną z tej klasy należy wybrać do opisu rozkładu statystyki chi-kwadrat.

Być może tłumaczę sprawy dość elementarne, lecz bez ich rozumienia nie sposób wyjaśnić zasad stosowania testu chi-kwadrat do interpretacji związków w tablicach. Otóż w teście tym hipoteza zerowa (H_0) głosi, że próba pochodzi z populacji, w której cechy są niezależne, zaś hipoteza alternatywna (H_1) określa, że tak nie jest. Przyjmijmy poziom istotności α równy 0,05 (chwilowo nie wnikając w istotę tej decyzji), po czym korzystając z wzoru (3.21), obliczmy wartość statystyki χ^2 dla przedstawionej w rozdziale 2 tablicy 2.8, w której skrzyżowano płeć ze sposobem głosowania w wyborach parlamentarnych. Wartość ta wynosi 18,84. Rozpatrywana tablica ma 2 wiersze i 8 kolumn, więc liczba stopni swobody jest równa 7. W tablicach rozkładu zmiennej χ^2 o 7 stopniach swobody odnajdujemy wartość krytyczną dla poziomu istotności 0,05 (Zieliński 1972: 114–115). Wynosi ona 14,07. Ponieważ obliczona wartość statystyki wykracza poza wartość krytyczną, to hipotezę zerową – o niezależności sposobu głosowania od płci w populacji – należy odrzucić.

Omawiany sposób analizy tablic pozwala podzielić je w gruncie rzeczy na dwie kategorie. Do pierwszej należą tablice, w wypadku których nie można odrzucić hipotezy o niezależności cech w populacji. Do drugiej pozostałe tablice, w wypadku których hipotezę tę należy odrzucić. Test chi-kwadrat nie jest więc zbyt subtelną metodą i nadaje się przede wszystkim do wstępnego przeselekcjonowania rozpatrywanych tablic i wybrania tych, którym warto przyjrzeć się bliżej. W praktyce nawet i do tego celu okazuje się jednak mało przydatny.

Pierwszy problem wiąże się z wyborem poziomu istotności. Zgodnie z regułami wnioskowania statystycznego wielkość tę należy ustalić jeszcze przed obliczeniem statystyki testu. Nie wszyscy badacze rozumieją jednak do końca zasady ustalania poziomu istotności. Pół biedy, gdy badacz rutynowo przyjmie poziom 0,05 – najczęściej chyba wybierany w naukach społecznych. Gorzej, gdy uczyni to dopiero po obliczeniu wartości statystyki chi-kwadrat. Oprogramowanie do analizy tablic stwarza ku temu silną pokusę podając na ogół tak zwaną wartość graniczną poziomu istotności. Dla omawianej tabeli wynosi ona 0,009 co oznacza, że w wypadku przyjęcia poziomu istotności poniżej tej wielkości – na przykład 0,001 – nie będzie podstaw do odrzucenia hipotezy zerowej. I wtedy zaczynają się dywagacje w stylu: „właściwie to sposób głosowania **nie zależy tak bardzo** do płci, gdyż przyjmując niższy poziom istotności, nie byłoby podstaw do odrzucenia hipotezy o niezależności w populacji". Inaczej mówiąc, metodzie pozwalającej wyłącznie na podjęcie decyzji na tak lub na nie, przypisuje się własności gradacyjne, których nie posiada.

Drugi problem stosowania testu chi-kwadrat wiąże się z faktem, że wartość obliczanej statystyki jest wprost proporcjonalna do liczebności próby (wzór 3.22). Przypuśćmy, że Polska zaliczona została do grupy krajów, które Euro-

pejski Sondaż Społeczny realizują na próbie 800 osób, czyli około dwa razy mniejszej od krajów o liczniejszych populacjach[6]. Gdyby przy tej wielkości próby struktura związku płci ze sposobem głosowania – rozumiana jako rozkłady procentowe w wierszach i kolumnach – okazała się identyczne jak to miało miejsce w faktycznie zrealizowanej próbie liczącej 1621 osób, to wartość statystyki χ^2 obliczona dla omawianej tabeli byłaby o połowę mniejsza od wyliczonej uprzednio, czyli wynosiłaby 9,42. Wielkość ta nie wykracza poza wartość krytyczną zmiennej chi-kwadrat o 7 stopniach swobody, toteż nie można byłoby odrzucić hipotezy o niezależności płci i sposobu głosowania w populacji, z której wylosowano próbę.

W wypadku badań realizowanych na mniejszych próbach test chi-kwadrat jest więc bardziej selektywny, to znaczy w fazie wstępnego przeglądania tablic mniejszy ich odsetek klasyfikuje się do kategorii zawierających wyniki warte dalszej uwagi. Jest to zrozumiałe, gdyż mniejsza próba nakazuje ostrożność w interpretacji wyników a zwłaszcza tych, które zaznaczają się niezbyt wyraźnie.

Co ciekawe, problemu nie rozwiązuje duża próba. Przyjmijmy, że badanie zrealizowane zostało na próbie liczącej 10 tysięcy osób. Praktyka dowodzi, że w wypadku większości tablic skonstruowanych na podstawie wyników takiego badania hipotezę o niezależności cech w populacji należy odrzucić. Mówiąc w slangu badawczym, „wszystkie wyniki stają się istotne". Tym samym test chi-kwadrat traci swoją użyteczność jako narzędzie selekcjonowania wyników.

3.11 Określenie siły związku

Drugie podejście do oceny rozbieżności między tablicą uzyskaną w badaniu a modelem niezależności polega na ilościowym określeniu wielkości tych rozbieżności. Tablicy przypisana zostaje pewna wartość liczbowa nazywana współczynnikiem siły bądź natężenia związku między obiema cechami.

Aby dobrze zrozumieć istotę omawianego podejścia warto odwołać się do jego historycznych korzeni. Można wyodrębnić dwa nurty myślenia na ten te-

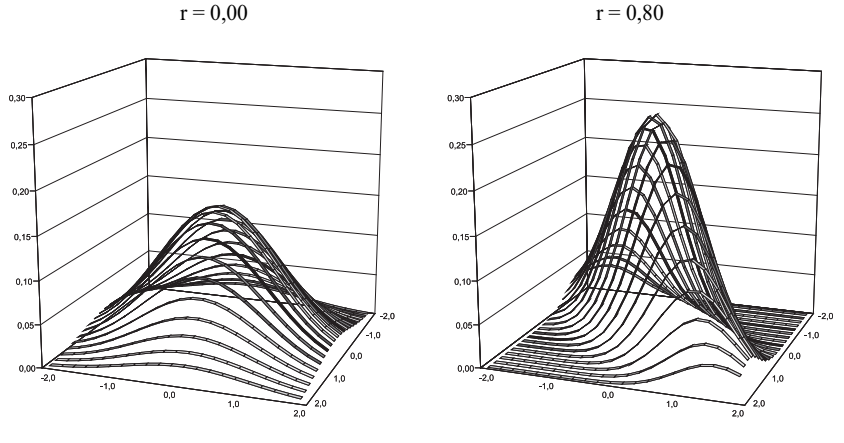
[6] Wielkość tak zwanej „efektywnej próby" wymagana od większości krajów w Europejskim Sondażu Społecznym wynosi 1500 osób. Jednakże kraje, w których wielkość populacji nie przekracza 2 milionów (np. Słowenia), realizują badanie na próbie 800 osób. Europejski Sondaż Społeczny stanowi więc przykład projektu, w którym wielkość próby jest ustalana niezależnie od przedmiotu badania, wspólnego w tym wypadku dla wszystkich krajów. Tego rodzaju okoliczności również warto brać pod uwagę przy wyborze narzędzi analizy danych. Może się bowiem okazać, że wartości testów statystycznych w większym stopniu odzwierciedlają schemat badania niż zróżnicowania będące jego przedmiotem.

mat. Pierwszy związany był z grupą badaczy, działających na przełomie XIX i XX wieku, skupionych wokół Karla Pearsona. Uważali oni, że cechy podlegające badaniom mają z natury charakter ciągły, tak jak waga czy wzrost, przez co ich rozkład łączny powinien być zbliżony do dwuwymiarowego rozkładu normalnego. Rozkłady należące do tej klasy charakteryzują się tym, że mają ustalony kształt, różniąc się między sobą wyłącznie jednym parametrem, zwanym **współczynnikiem korelacji** (rycina 3.3). Większym wartościom współczynnika korelacji odpowiadają rozkłady bardziej skupione wokół przekątnej, co na wykresach nadaje im bardziej „szpiczasty" wygląd.

Logiczną konsekwencją tego stanowiska było przekonanie, że dla trafnego opisania związku między dwiema cechami wystarcza znajomość jednego parametru – współczynnika korelacji. Przekonania tego nie osłabiał fakt, że większość danych z badań dostępna była w postaci tablic. Tablice traktowano bowiem jako konsekwencję kategoryzowania cech w fazie pomiaru. Na przykład, gdy w badaniu pytano nie o dokładny wiek, lecz proszono o zaklasyfikowanie wieku do jednego z przedziałów. Założenie o dwu-normalnym kształcie rozkładów cech przetrwało bardzo długo, zwłaszcza w ekonometrii. Dopiero w ostatnich latach pojawiły się próby wykazania, że nawet w rozkładach „klasycznych" cech ciągłych – jak wzrost czy waga – pojawiają się lokalne anomalie naruszające ich dwu-normalną logikę, przez co ograniczenie się do współczynnika korelacji pomija istotne aspekty badanego związku (Levine 2005).

Drugi sposób opisu tablic za pomocą pojedynczego parametru ma jeszcze dłuższą genezę, sięgającą połowy XIX wieku. W konstruowanych w tamtych

Rycina 3.3
Wykresy gęstości dwuwymiarowych rozkładów normalnych różniących się współczynnikiem korelacji



czasach tablicach krzyżowano ze sobą najczęściej dwie cechy dychotomiczne. Tablice o rozmiarach 2 na 2 charakteryzują się tym, że pojedynczy parametr całkowicie wystarcza dla opisu ich zawartości, gdyż mają tylko jeden stopień swobody. Względna prostota analizowanych tablic doprowadziła do upowszechnienia się stanowiska, że wskaźniki należy konstruować nie jako narzędzia uniwersalne, lecz pod kątem rozwiązywania konkretnych problemów badawczych. Swoje propozycje w tym zakresie stworzyli nie tylko prekursorzy współczesnych metod analizy wyników badań, jak Karl Pearson (1857–1936) czy George Udny Yule (1871–1951), lecz również klasycy podejścia socjologicznego, jak Ferdinand Tönnies (1855–1936), czy ekonomicznego, jak Corrado Gini (1894–1965). W artykule wydanym w 1959 roku Goodman i Kruskal podjęli trud usystematyzowania podejść do pomiaru siły związku w tablicach. W sumie omawiają kilkadziesiąt różnych koncepcji, zaznaczając, że nie są w stanie objąć wszystkich pomysłów w tym zakresie.

Czytelników poszukujących inspiracji do tworzenia własnych wskaźników odsyłam do cytowanego artykułu Goodmana i Kruskala (1959). W tym miejscu zajmę się wyłącznie jednym z wątków, który przewijał się przez całą historię prac nad stworzeniem użytecznej miary siły związku w tablicach. Chodzi o próby stworzenia takiej miary, której wartości miałyby interpretację operacyjną. Co to na przykład znaczy, że wartość miary wynosi 0,5.

Spójną wykładnię takiego stanowiska jako pierwsi sformułowali Goodman i Kruskal w 1954 roku. Zaproponowali oni, aby siłę związku w tablicy oceniać w kategoriach redukcji błędu przewidywania jednej cechy na podstawie drugiej, którą to redukcję umożliwia wykorzystanie informacji o łącznym rozkładzie cech w tablicy.

Prezentację koncepcji zilustrujemy dość nietypowym przykładem, pochodzącym z badania o charakterze metodologicznym. Zostało ono zrealizowane w ramach Europejskiego Sondażu Społecznego 2008 i polegało na zadaniu tym samym respondentom tych samych pytań w dwóch wersjach. Jedną z wersji zadawano na początku wywiadu, drugą zaś na końcu, tak aby zminimalizować możliwość zapamiętania wcześniej udzielonej odpowiedzi. Celem badania było oszacowanie trafności pomiaru poprzez uzyskanie dwóch niezależnych wskaźników tych samych zjawisk. Z zestawu zadanych w ten sposób pytań wybrałem pytanie dotyczące czasu oglądania telewizji. W obu wersjach pytanie miało identyczne brzmienie, różniło się natomiast zestawem odpowiedzi (tabela 3.8). Ze względu na charakter danych i sposób ich gromadzenia należy spodziewać się wysokiej zgodności odpowiedzi w wypadku obu wersji pytania, czyli silnego związku między cechami przedstawionymi w tablicy. Zobaczmy, jak intuicje dotyczące siły związku przekładają się na wielkości wskaźników zaproponowanych przez Goodmana i Kruskala.

Przypuśćmy, że chcielibyśmy przewidzieć czy też „odgadnąć" – jak to określają Goodman i Kruskal (1954: 741) – odpowiedź udzieloną przez respondenta na pytanie zadane w drugiej wersji (umieszczonej w kolumnach tabeli 3.8). Najrozsądniej jest w tym wypadku podać odpowiedź „trochę czasu", gdyż pojawiała się ona najczęściej. Podając tę odpowiedź trafialibyśmy średnio w 33 procentach dokonywanych przewidywań, zaś w pozostałych 67 procentach przewidywalibyśmy błędnie.

Gdyśmy jednak znali odpowiedź na pytanie zadane w pierwszej wersji, to wtedy skuteczność przewidywania poprawiłaby się. Na przykład, gdybyśmy wiedzieli, że respondent odpowiedział „w ogóle" na pytanie zadane w pierwszej wersji, to w 86 przypadkach na 100 zgadlibyśmy, że odpowiedzią na pytanie zadane w drugiej wersji jest również „w ogóle". Jak łatwo zauważyć, optymalna strategia polega w tym wypadku na wyborze odpowiedzi, której odsetek w danym wierszu tablicy jest najwyższy. Pola te w tabeli 3.8 zaznaczone zostały szarym kolorem.

Goodman i Kruskal zaproponowali miarę dla opisanej strategii przewidywania. W ogólnym przypadku jej postać wyraża się wzorem (1954: 741)

$$\lambda = \frac{\left(\begin{array}{c}\text{odsetek pomyłek}\\ \text{bez uwzględnienia}\\ \text{cechy w wierszach}\end{array}\right) - \left(\begin{array}{c}\text{odsetek pomyłek}\\ \text{przy znanej kategorii}\\ \text{cechy w wierszach}\end{array}\right)}{\left(\begin{array}{c}\text{odsetek pomyłek}\\ \text{bez uwzględnienia}\\ \text{cechy w wierszach}\end{array}\right)} \qquad (3.22)$$

Odsetek pomyłek bez uwzględnienia cechy w wierszach jest równy $1 - \max(p_{\bullet j})$, czyli od 100 procent odejmowany jest odsetek odpowiadający kategorii zmiennej kolumnowej, która występuje najczęściej. Z kolei jako odsetek pomyłek przy znajomości kategorii cechy w wierszach przyjmuje się średnią analogicznych wyrażeń dla poszczególnych wierszy, obliczoną z uwzględnieniem jako wag wielkości brzegowych cechy umieszczonej w wierszach, gdyż poszczególne odpowiedzi na pierwszą wersję pytania występują z różnymi częstościami. Po podstawieniu tej wielkości do wzoru (3.22) oraz po dokonaniu prostych przekształceń arytmetycznych wzór na wartość współczynnika λ_{YX} przyjmuje postać:

$$\lambda_{YX} = \frac{\sum_{i=1}^{w} p_{i\bullet} * \max(p_{j|X=x_i}) - \max(p_{\bullet j})}{1 - \max(p_{\bullet j})} \qquad (3.23)$$

Wyjaśniając zasadę obliczania współczynnika λ_{YX}, odwoływaliśmy się do odsetków udzielanych odpowiedzi. Jedną z zalet omawianego współczynnika jest to, że do wyjaśnienia jego istoty równie dobrze posłużyć się można tablicą liczebności. Jest to dogodne dla osób, których wyobraźnia łatwiej operuje na zbiorze badanych respondentów niż na proporcjach. Wzór na wartość współczynnika λ_{YX} oparty na liczebnościach przedstawia się w podany niżej sposób

$$\lambda_{YX} = \frac{\sum_{i=1}^{w} \max_j (n_{ij}) - \max(b_j)}{n - \max(b_j)} \tag{3.24}$$

zaś poszczególne jego elementy interpretować można następująco.

Przypuśćmy, że odpowiedź na drugą wersję pytania przewidujemy tyle razy, ile jest w sumie osób w tabeli. Pierwszy składnik wyrażenia w liczniku $\sum_{i=1}^{w} \max_j (n_{ij})$ odpowiada liczbie trafnych przewidywań w sytuacji znajomości odpowiedzi na pytanie zadane w pierwszej wersji. Rozważmy pierwszy wiersz tabeli 3.8. W wierszu tym znajdziemy się 19 razy, gdyż tylu respondentów odpowiedziało „w ogóle" na pytanie zadane w pierwszej wersji. Optymalnym przewidywaniem jest w tym wypadku odpowiedź „w ogóle" na pytanie zadane w wersji drugiej. Z rozpatrywanego wiersza tabeli odczytujemy, że w 17 przypadkach nasze przewidywanie będzie trafne. Liczbę 17 wpisujemy więc do ostatniej kolumny tabeli, po czym rozumowanie powtarzamy dla kolejnych wierszy. Po zsumowaniu liczb wypisanych w ostatniej kolumnie otrzymujemy sumę poprawnie odgadniętych odpowiedzi dla całej tablicy. Wynosi ona 268. Odejmujemy od niej liczbę odpowiedzi na pytanie w wersji drugiej, które odgadlibyśmy bez znajomości odpowiedzi udzielonej na pytanie w pierwszej wersji. Liczba ta wynosi 176, co odpowiada odpowiedzi najczęściej udzielanej w drugiej wersji pytania. Różnica tych wielkości, czyli 268 – 176 = 92, jest wartością licznika we wzorze (3.24). W wypadku tylu badanych osób poprawiło się przewidywanie odpowiedzi udzielonej na pytanie w drugiej wersji w sytuacji, gdy znaliśmy odpowiedź na pytanie zadane w wersji pierwszej.

Mianownik wzoru (3.24) określa z kolei, w wypadku ilu osób przewidywanie okazałoby się nie trafione, gdybyśmy nie znali odpowiedzi udzielonej na pytanie zadane w pierwszej wersji. Miałoby to miejsce w 355 (= 531 – 176) przypadkach. Iloraz obliczonych wielkości, czyli 92/355, stanowi wartość współczynnika λ_{YX}, która w omawianym przykładzie jest w zaokrągleniu równa 0,26.

Niektórzy czytelnicy mogą czuć się zawiedzeni. Dlaczego wartość okazała się tak niska, skoro przedstawiony w tabeli 3.8 związek wydaje się silny? Ci

Tabela 3.8
Liczebności oraz odsetki odpowiedzi na drugą wersję pytania o przeciętny czas oglądania telewizji w kategoriach odpowiedzi na pierwszą wersję tego pytania

		wersja 2											
wersja 1		w ogóle	bardzo mało czasu	mało czasu	trochę czasu	dość dużo czasu	dużo czasu	bardzo dużo czasu	trudno powiedzieć	ogółem	odsetek osób [%]	maksymalna liczba w wierszu
w ogóle	n	17	1	1	1					**19**	*4*	17
	%	*86*	*5*	*4*	*5*					*100*		
poniżej ½ godziny	n		26	6						**32**	*6*	26
	%		*81*	*19*						*100*		
½ godziny do 1 godziny	n		27	48	14	2				**90**	*17*	48
	%		*30*	*53*	*15*	*2*				*100*		
ponad 1 godzinę do 1 ½ godziny	n		8	27	29	3	2			**69**	*13*	29
	%		*12*	*39*	*42*	*4*	*3*			*100*		
ponad 1 ½ godziny do 2 godzin	n		10	29	54	9	2			**103**	*19*	54
	%		*9*	*28*	*52*	*9*	*2*			*100*		
ponad 2 godziny do 2 ½ godziny	n			8	25	6				**39**	*7*	25
	%			*19*	*65*	*16*				*100*		
ponad 2 ½ godziny do 3 godzin	n			9	36	17	5	1		**68**	*13*	36
	%			*13*	*54*	*25*	*7*	*1*		*100*		
ponad 3 godziny	n			3	17	32	32	24		**108**	*20*	32
	%			*3*	*16*	*29*	*30*	*23*		*100*		
trudno powiedzieć	n								1	**1**	*0*	1
	%								*100*	*100*		
ogółem	**n**	**17**	**72**	**131**	**176**	**69**	**41**	**25**	**1**	**531**	*100*	268
	%	*3*	*14*	*25*	*33*	*13*	*8*	*5*	*0*	*100*		
maksymalna liczba w kolumnie		17	27	48	54	32	32	24	1	234		

Europejski Sondaż Społeczny 2008 (dane ważone). Obie wersje różniły się jedynie zestawem odpowiedzi, zaś zadawane pytanie brzmiało identycznie: „Ile w sumie czasu spędza P. na oglądaniu telewizji w przeciętny dzień powszedni? Odpowiadając proszę posłużyć się następującą kartą". Odsetki oraz liczebności przedstawiono w postaci zaokrąglonej do liczb całkowitych. Kolorem szarym zaznaczono pola zawierające największe odsetki w wierszach.

sami respondenci odpowiadali na to samo pytanie podczas tego samego wywiadu. Intuicja podpowiada, że zgodność odpowiedzi na obie wersje pytania powinna wyrażać się wyższą wartością współczynnika.

Ciekawe, jakie powody kierowały Goodmanem i Kruskalem, gdy ponad 50 lat temu zaproponowali, aby wartościom wskaźnika siły związku w tablicy nadać interpretację operacyjną. Być może jednym z powodów była świadomość, że badacze wykazują skłonność do szacowania siły związku między cechami jako wyższej, niż ma to faktycznie miejsce. Co jest poniekąd zrozumiałe, biorąc pod uwagę emocjonalne zaangażowanie badacza w projekt i oraz oczekiwanie, że badane prawidłowości zarysują się wyraziście.

Tymczasem wartość współczynnika siły związku, gdy ma on operacyjną interpretację, bezlitośnie obnaża niedoskonałości pomiaru w badaniach. Rozważmy bowiem osoby, które na pytanie zadane w pierwszej wersji odpowiedziały, że telewizję oglądają ponad 3 godziny dziennie. Wśród osób tych 23 procent odpowiedziało na pytanie zadane w wersji drugiej, że zajmuje im to „bardzo dużo czasu", 30 procent, że „dużo czasu", 29 procent, że „dość dużo czasu", a 16 procent, że tylko „trochę czasu". Nie mówiąc już o 3 procentach respondentów, którzy stwierdzili, że ponad 3 godziny oglądania telewizji to „mało czasu". Spójrzmy na problem z tej strony i odpowiedzmy sobie sami na pytanie – czy rzeczywiście mamy tu do czynienia z wysoką zgodnością odpowiedzi udzielanych na pytanie zadane w obu wersjach?

Przedstawiony przykład uzmysławia jeszcze jeden aspekt tej sprawy. W wypadku związku między odpowiedziami udzielonymi na pytanie zadane w dwóch wersjach narzuca się niejako, że odpowiedzi te należałoby dla każdego badanego „zestawić" ze sobą w celu sprawdzenia, czy są jednakowe, czy też nie. Jest to naturalny sposób wyobrażania sobie zgodności dwóch wielkości. W rozważanym przykładzie tego zrobić się jednak nie da. Kafeterie odpowiedzi są bowiem różne (sama liczba odpowiedzi jest niejednakowa), stąd też nie można wskazać ścisłej odpowiedniości między jednym zestawem odpowiedzi a drugim. Jedynym narzędziem przedstawienia tego związku jest tablica krzyżująca odpowiedzi uzyskane w obu wersjach pytania.

Interpretacja siły związku jako poziomu redukcji błędów przewidywania nie zachwyca może, gdy chodzi o jej naturalność i przejrzystość, lecz jest to jedyna **uniwersalna** interpretacja operacyjna, jaką dotychczas zaproponowano. Jej uniwersalność należy rozumieć w ten sposób, że gdy w tablicy zamiast cech jakościowych umieścimy cechy ilościowe – na przykład wyrażone na skalach porządkowych lub interwałowych – lub też skrzyżujemy ze sobą cechę ilościową i jakościową, to również jesteśmy w stanie wskazać współczynnik pomiaru siły związku, który posiada interpretację w języku redukcji błędów przewidywania. Propozycje tego rodzaju mierników znaleźć można nie tylko

u Goodmana i Kruskala (1954), lecz również u licznych kontynuatorów tego podejścia. Ujęcie to jest z pewnością nieobce badaczom, którzy swoje szlify zdobywali w Instytucie Socjologii Uniwersytetu Warszawskiego. Kurs statystyki od ponad 30 lat oparty jest na tym podejściu, zaś odpowiadający programowi kursu podręcznik zalicza się do najlepiej przemyślanych, aktualnych i kompletnych prezentacji tych zagadnień (Lissowski i inni 2008).

Wróćmy jednak do tablicy opisującej związek odpowiedzi uzyskanych na pytanie zadane w dwóch różnych wersjach. Dotychczas próbowaliśmy przewidywać odpowiedź udzieloną w drugiej wersji na podstawie odpowiedzi na pytanie w wersji pierwszej. Problem można oczywiście odwrócić i zapytać o to, jaka byłaby skuteczność przewidywania odpowiedzi na pierwsze pytanie, gdy znamy odpowiedź na drugie. Wartość współczynnika oblicza się wtedy analogicznie, szukając największych liczebności w każdej z kolumn tabeli, a następnie zestawiając ich sumę z największą liczebnością brzegową rozkładu odpowiedzi na pierwsze pytanie, która odpowiada szansom „zgadnięcia" odpowiedzi w pierwszym z zadanych pytań bez znajomości odpowiedzi udzielonej w drugim. Oznaczmy ten współczynnik jako λ_{XY}. Wzór na wartość λ_{XY} analogiczny jest do wzoru (3.24)

$$\lambda_{XY} = \frac{\sum_{j=1}^{k} \max_i(n_{ij}) - \max(a_i)}{n - \max(a_i)} \tag{3.25}$$

Wartość współczynnika obliczona dla danych z tabeli 3.8 wynosi w zaokrągleniu 0,30 – a więc jest inna niż poprzednio. Niejednakowość obu wartości wynika z braku symetrii tablicy. W obu wersjach pytania zastosowano odmienne zestawy odpowiedzi, przez co byłoby raczej dziełem przypadku, gdyby skuteczność przewidywania w obie strony okazała się taka sama.

Ze względu na niesymetryczność proponowanego współczynnika Goodman i Kruskal zdefiniowali także jego wariant symetryczny (1954: 742–743). Wzór, oparty na liczebnościach, przedstawia się następująco

$$\lambda = \frac{\sum_{i=1}^{w} \max_j(n_{ij}) - \max(b_j) + \sum_{j=1}^{k} \max_i(n_{ij}) - \max(a_i)}{n - \max(b_j) + n - \max(a_i)} \tag{3.26}$$

Łatwo zauważyć, że licznik zawiera sumę liczników obu wariantów niesymetrycznych, określonych przez wzory (3.24) i (3.25), zaś mianownik sumę ich mianowników. Interpretacja wskaźnika w kategoriach przewidywania jest więc następująca. Wyobraźmy sobie, że przewidujemy **tę samą liczbę razy**

kategorię cechy X na podstawie znajomości kategorii cechy Y co odwrotnie – kategorię cechy Y na podstawie znajomości kategorii cechy X. Wtedy wyrażenie (3.26) określa poziom redukcji błędów przewidywania wobec sytuacji, w której podczas przewidywania nie bralibyśmy pod uwagę kategorii drugiej z cech.

Ramka 3.1
Alan Agresti: spojrzenie środowiska akademickiego na metody analizy zjawisk

Alan Agresti urodził się w 1947 roku. Stopień B.A. w dziedzinie matematyki uzyskał na uniwersytecie Rochester w roku 1968. Był to okres szczególnego nasilenia walk w Wietnamie. Tysiące młodych Amerykanów z niepokojem oczekiwało na transmitowane przez telewizję kolejne sesje „loterii poborowej" (ang. *draft lottery*), której wyniki pokrzyżować mogły życiowe plany. W takiej atmosferze Alan Agresti zdecydował się robić doktorat na Uniwersytecie Wisconsin, gdyż zaoferowano mu asystenturę chroniącą wtedy od poboru. Jego opiekunem stał się Stephen Stigler, ekspert w dziedzinie historii statystyki. W ten sposób zaczął się romans Alana Agrestiego ze statystyką, który trwa po dzień dzisiejszy.

Alan Agresti jest autorem najbardziej znanych i szeroko używanych podręczników na temat metod analizy danych kategorialnych. Jego fundamentalnym dziełem jest prawie 800-stronicowe *Categorical Data Analysis* (2002). Praca ta doczekała się również skróconego wydania, które lepiej sprawdza się na uniwersyteckich kursach statystyki (*An Intoduction to Categorical Data Analysis*, Agresti 2007). Wspólnie z Barbarą Finlay jest też autorem wielokrotnie wznawianego podręcznika metod statystycznych dla studentów nauk społecznych (*Statistical Methods for the Social Sciences*, Agresti i Finlay 2008).

Karierę zawodową Alan Agresti związał z Uniwersytetem Floryda, gdzie jest profesorem statystyki. Praca na uniwersytecie jest specyficznym rodzajem kariery. W większym stopniu wymaga porządkowania wiedzy i metod niż znajdowania dla nich kolejnych zastosowań. Alan Agresti ma co prawda w swoim dorobku szereg artykułów – zgodnie z zasadą „publish or perish" – lecz swój wysiłek koncentruje na pisaniu podręczników. Jak sam twierdzi, jest to wdzięczniejsze zajęcie, gdyż każdy kolejny fragment daje poczucie pójścia do przodu, gdyż coś zostało zrobione. Aczkolwiek dostrzega też negatywy: „You can be watching a movie at night, and your mind wanders to that section you wrote today that really could be improved" (Dietz 2004).

W swoich książkach Agresti kładzie duży nacisk na historyczne korzenie współcześnie stosowanych metod. Studenci uważają – jak twierdzi – że statystyka to metody istniejące setki lat, z którymi uczelnia zapoznaje w niezmienionej postaci kolejne roczniki absolwentów. Perspektywa historyczna pozwala pokazać, że metody statystyczne jak cała wiedza podlegają ewolucji. Obecnie prowadzi ona ku metodom, które znajdą zastosowanie w analizach dużych zasobów danych (ang. *data mining*). Tradycyjna statystyka matematyczna stopniowo zaś będzie tracić na znaczeniu.

Za swój wkład w rozwój metod analizy danych Alan Agresti uhonorowany został tytułem Statystyka Roku 2003 przez American Statistical Association. Cztery lata wcześniej otrzymał doktorat *honoris causa* na brytyjskim uniwersytecie De Montfort w Leicester. Związki Agrestiego z Anglią, a szczególnie z Londynem, są zresztą bliskie. W Londynie przebywał na stażu naukowym (*sabbatical year*) w latach osiemdziesiątych. Miasto to traktuje jako swój drugi dom. Alan Agresti lubi zresztą podróżować. Długa jest lista miast i krajów, w których wygłaszał wykłady lub prowadził kursy na temat metod analizy danych kategorialnych.

Symetryczna wersja omawianego współczynnika wydaje się mniej przydatna w praktyce od wersji niesymetrycznej. Na ogół bowiem przyjmuje się określony kierunek zależności między cechami w tablicy. Na przykład, badacza częściej będzie interesować zależność sposobu głosowania od płci, niż zależność w drugą stronę (o ile tej ostatniej nie uzna za pozbawioną empirycznego sensu). Dlatego interpretacja siły związku w tablicy jako stopnia redukcji błędu przewidywania jest bardziej naturalna, gdy przewidywania dokonuje się w jedną stronę. Niemniej jednak, można wskazać sytuacje, gdy potraktowanie związku jako wzajemnego jest uzasadnione, co dotyczy też rozpatrywanego związku między odpowiedziami na dwie wersje pytania. Argumenty w tej kwestii przedstawię w podrozdziale 4.6, natomiast w tym miejscu dla porządku podam wartość symetrycznego współczynnika λ dla omawianego przykładu. W zaokrągleniu wyniosła ona 0,28.

Wartość symetrycznego współczynnika leży zawsze pomiędzy wartościami współczynników niesymetrycznych

$$\min(\lambda_{YX}, \lambda_{XY}) \leq \lambda \leq \max(\lambda_{YX}, \lambda_{XY}) \tag{3.27}$$

Przewidując jednocześnie w obie strony nie uda się więc zrobić tego dokładniej, niż w tę stronę, w którą przewidywanie jest lepsze. Symetryczną wartość wskaźnika można też rozumieć jako wypadkową obu jego wariantów niesymetrycznych.

Współczynniki lambda mają własności, których na ogół wymaga się od mierników siły związku w tablicy.

(1) Wartości każdego z tych współczynników zawierają się w przedziale od 0 do 1 (są znormalizowane do tego przedziału).
(2) Dowolna permutacja wierszy tablicy bądź kolumn nie zmienia wartości żadnego z tych współczynników. Można je zatem stosować do charakteryzowania tablic krzyżujących ze sobą cechy jakościowe, w wypadku których porządek wierszy i kolumn nie ma znaczenia.
(3) Współczynnik λ_{Yx} przybiera maksymalną wartość równą 1 tylko wtedy, gdy w każdym wierszu tablicy tylko w jednym z pól jest niezerowa liczebność. Odpowiada to sytuacji, gdy znając wartość zmiennej w wierszach jesteśmy w stanie bezbłędnie przewidzieć wartość zmiennej umieszczonej w kolumnach. Analogiczna prawidłowość, lecz w drugą stronę, zachodzi dla współczynnika λ_{XY}. Natomiast symetryczny współczynnik λ przybiera wartość 1 wtedy i tylko wtedy, gdy jednocześnie mają miejsce obie prawidłowości. W sytuacji tej tablica musi mieć jednakową liczbę wierszy i kolumn, zaś na przecięciu każdego wiersza i kolumny występuje tylko jedna niezerowa liczebność.
(4) Jeżeli zmienne są stochastycznie niezależne, to każdy ze współczynników λ osiąga swoją najniższą możliwą wartość równą 0.

Należy zaznaczyć, że ostatnia z własności nie jest spełniona w drugą stronę. Jeśli dla danej tablicy dowolny z współczynników λ przyjmie wartość 0, to z faktu tego nie wynika, że cechy w tablicy są stochastycznie niezależne. Można wykazać to na prostym przykładzie.

W tabeli 3.9 przedstawiono fikcyjny związek dwóch cech dychotomicznych. Przypuśćmy, że przewidujemy kategorie cechy Y na podstawie kategorii cechy X. Gdy cecha X przybiera wartość x_1, to jest zupełnie obojętne, którą z wartości cechy Y będziemy przewidywać. Obie występują bowiem z jednakowymi częstościami. Obojętnie więc jak będziemy postępować, to średnio rzecz biorąc powinniśmy poprawnie odgadnąć 100 razy na 200. Gdy natomiast cecha X przybiera wartość x_2, to każdorazowo należy przewidywać, że cecha Y przybierze wartość y_1. Przewidywanie będzie za każdym razem spełnione, gdyż w drugim wierszu tabeli cecha Y przybiera jedynie tę wartość. W sumie więc, przy znajomości wartości cechy X przewidywanie będzie trafne w 200 na 300 przypadków. Ogląd rozkładu brzegowego cechy Y pozwala stwierdzić, że nie znając wartości cechy X również jesteśmy w stanie trafnie przewidzieć wartość tej cechy 200 na 300 razy, przewidując za każdym razem wartość y_1. Podstawienie obu wielkości do wzoru (3.24) zeruje jego licznik. Stąd współczynnik λ_{YX} dla omawianej tablicy jest równy 0. Analogiczne rozumowanie przeprowadzić można w drugą stronę, dla współczynnika λ_{XY}. Z wzoru (3.27) wynika, że symetryczna wersja współczynnika λ również jest równa 0. Powstaje w związku z tym pytanie, czy współczynniki λ uznać można za adekwatne miary siły związku w tablicy, gdy związek ten rozumiany jest (przynajmniej w tym rozdziale) jako odstępstwo od modelu niezależności stochastycznej?

Tabela 3.9
Liczebności rozkładu łącznego dwóch cech dychotomicznych, dla którego niesymetryczne wersje λ_{YX} oraz λ_{XY}, a także symetryczna wersja współczynnika λ Goodmana-Kruskala przybierają wartość 0, pomimo że związek nie jest niezależnością stochastyczną

Dane fikcyjne

	cecha Y		
cecha X	y_1	y_2	ogółem
x_1	100	100	200
x_2	100	0	100
ogółem	200	100	300

Na pytanie to nie ma dobrej odpowiedzi. Współczynnik λ Goodmana-Kruskala wybrałem do prezentacji ze względu na prostotę przyjętej reguły prze-

widywania, jako „zgadywania" najczęściej udzielanej odpowiedzi. Sposób obliczania wartości tego współczynnika pozwala dość łatwo prześledzić, jak układy liczebności w wierszach czy w kolumnach tablicy przekładają się na liczbę skutecznych przewidywań. W ramach koncepcji redukcji błędu przewidywania zaproponowano wiele innych współczynników, które mają między innymi tę własność, że przyjmują wartość 0 wtedy i tylko wtedy, gdy związek cech w tablicy spełnia ściśle warunki niezależności stochastycznej. Przykład takiego współczynnika, oznaczonego jako τ (tau), znaleźć chociażby można w cytowanym artykule Goodmana i Kruskala (1954: 759–760). Do nowszych propozycji należy współczynnik omawiany przez Lissowskiego i in. (2008: 321–329). W tym ostatnim wypadku reguła przewidywania oparta jest na pojęciach entropii oraz ilości informacji niezbędnej do identyfikacji wartości zmiennej. Dla wielu badaczy ten sposób spojrzenia na rozkłady warunkowe w tablicy wydawać się może bardziej atrakcyjny czy użyteczny. Nie zmienia jednak faktu, że reguła przewidywania nie ma tak prostej interpretacji, jak proponowana w wypadku współczynnika λ reguła „zgadywania" wartości najczęściej występującej. W sumie – coś za coś. Wybierając współczynnik λ Goodmana-Kruskala dostajemy przejrzystą i bezpośrednią interpretację. Narażeni jesteśmy jednak na niebezpieczeństwo, że wartość bliska zeru wcale nie musi wcale oznaczać, że kształt związku przypomina niezależność stochastyczną. Wybierając zaś współczynnik, który chroni przed takim niebezpieczeństwem, musimy liczyć się z bardziej złożoną interpretacją uzyskanej wartości.

3.12 Dyskusja

Pojęcie niezależności pełni specyficzną funkcję w wyjaśnianiu zjawisk. Stanowi punkt odniesienia wielu metod, za pomocą których analizuje się wyniki badań przedstawione w tablicach. Ułatwić też może zrozumienie istoty i mechanizmów badanych zjawisk.

Warto jednak zaznaczyć, że część badaczy nie odwołuje się w budowanych wyjaśnieniach do pojęcia niezależności. Jak twierdzi cytowany w podrozdziale 3.9 Robert M. Hauser, z modelu niezależności nie warto korzystać opisując zjawiska, gdyż „nie jest dopasowany do danych" (1978: 924). Gdyby zwolennikom tego podejścia zadać pytanie sformułowane w tytule rozdziału, to ich odpowiedź zapewne byłaby negatywna. I to bynajmniej nie dlatego, że stan niezależności nie występuje w świecie badanych zjawisk. Lecz raczej dlatego, że nie jest on właściwym przedmiotem wyjaśnienia. Zdaniem zwolenników tej opcji, badacza obowiązuje oszczędność w konstruowaniu wyjaśnień. Im wyjaśnienie jest prostsze, tym jest lepsze. Szkoda więc wysiłku na rozbijanie

zjawisk na dwa komponenty: na niezależność oraz na to, co niezależnością nie jest. Komplikuje to budowane wyjaśnienia.

W rozdziale tym prezentuję odmienne stanowisko. Większość rozważań poświęciłem określeniu wpływu, jaki na kształt związku w tablicy mają jej marginesy. Starałem się też uzasadnić, że w wielu wypadkach marginesy kształtowane są przez czynniki zewnętrzne, stanowiąc ramy rozpatrywanych zjawisk. Same zjawiska można zaś wyobrazić sobie jako mechanizmy wypełniania pustych pól tablic o z góry ustalonych marginesach.

Gdy badacz zdecyduje się traktować marginesy tablicy jako czynnik zewnętrzny, to wyodrębnienie modelu niezależności bywa pomocne. Marginesy wyznaczają bowiem postać tego modelu, dzięki czemu stanowić on może punkt odniesienia dla oceny badanego związku. Część związku odpowiadająca modelowi niezależności wnosi na ogół sporo do obrazu badanego zjawiska, gdy obraz ten rozumiemy jako układ liczebności w polach tablicy. Jednakże część ta odzwierciedla wielkości kategorii obu cech, kształtowane niekiedy poza obszarem badanego zjawiska. W związku z tym wartość eksplanacyjną przypisuje się przede wszystkim temu komponentowi badanego zjawiska, który odbiega od niezależności.

W podrozdziałach 3.10 i 3.11 przedstawiłem narzędzia analizy tablic, które służą ilościowemu oszacowaniu wielkości owego komponentu. Pozwalają one odpowiedzieć na dwa pytania. Czy badany związek w ogóle różni się od modelu niezależności oraz w jakim stopniu. W następnym rozdziale zaproponuję metody pozwalające zidentyfikować konkretne fragmenty tablicy, które odbiegają od modelu niezależności. Pozwali to wyodrębnić najbardziej znaczące aspekty badanego zjawiska, dla których warto podjąć próbę sformułowania wyjaśnień.

Pojęcie niezależności towarzyszyło nam również będzie w kolejnych rozdziałach. W rozdziale 5 przedstawię podejście, które odwołuje się do przedstawionego w podrozdziale 3.4 sposobu rozumienia niezależności jako zgodności profili w wierszach i w kolumnach. W rozdziale 6 zaproponuję model stanowiący integrację obu podejść – to jest analizy odstępstw pewnych fragmentów tablicy od modelu niezależności oraz analizy zróżnicowania profili. Pozwoli to pokazać, że różne podejścia prowadzą w gruncie rzeczy do tego samego sposobu rozumienia przedstawionego w tablicy zjawiska. Rozdział 7 poświęcony zostanie omówieniu metody zwanej analizą korespondencji, która pozwala na wizualizację związku cech w tablicy. Dostarczy ona dodatkowych interpretacji pojęcia niezależności, odwołujących się do intuicji geometrycznych.

ROZDZIAŁ 4

W poszukiwaniu modelu zjawiska

Przez model tablicy rozumieć będziemy regułę, która opisuje układ liczebności w jej wnętrzu. W poprzednim rozdziale omówiliśmy jedną z ważniejszych reguł tego rodzaju, nazywaną niezależnością. Każda reguła – o ile jest poprawnie zidentyfikowana – stanowi odzwierciedlenie praw rządzących badanym zjawiskiem, jest podstawą sformułowania jego przybliżonego obrazu, zwanego też modelem.

W praktyce każdy wgląd badacza w liczebności tablicy zawiera w sobie elementy identyfikacji jej modelu. Ma to miejsce zarówno w wypadku stosowania prostych technik, jak porównywanie ze sobą profili wierszy bądź kolumn, jak też przy posługiwaniu się bardziej złożonymi metodami. Pomimo że tworzenie modelu tablicy jest dość rudymentarną operacją, występującą w codziennej praktyce badawczej, jej metodologia nie doczekała się dotychczas systematycznego wykładu. Celem rozdziału jest przynajmniej częściowe zapełnienie tej luki.

Podrozdział 4.1 rozpocznę od określenia, co należy rozumieć przez model tablicy. Na przykładzie przedstawię związki liczebności wnętrza tablicy z modelem badanego zjawiska, a następnie nawiążę do dwóch pojęć użytecznych przy porównywaniu wielkości w tablicy: pojęć stosunku i różnicy (4.2). W podrozdziale 4.3 omówię korzyści i ograniczenia posługiwania się różnicami porównywanych liczebności, zaś w podrozdziale 4.4 przedstawię dość często stosowany w porównaniach wskaźnik, jakim jest indeks. Następnie przejdę do omówienia propozycji w tym zakresie sformułowanej przez Adolphe Quételeta w pierwszej połowie XIX wieku (4.5). W ostatnim czasie nastąpił swoisty powrót do idei zawartych w tej koncepcji, gdyż pozwalają one w zrównoważony sposób uwzględnić różne aspekty badanego zjawiska. Wykład propozycji Quételeta zawieszę w podrozdziale 4.6, aby omówić jeden z aspektów badanych zjawisk, który nazwę wzajemnością oddziaływań. Aspekt ten wyznacza istotę propozycji, którą przedstawię w podrozdziale 4.7. Stanowi ona modyfikację idei Quételeta, stosowaną między innymi w ramach podejścia zwanego analizą korespondencji.

W podrozdziale 4.8 omówię krok po kroku sposób zastosowania proponowanych wskaźników do identyfikacji modelu związku cech w tablicy. Przykładowe dane dotyczyć będą zależności między wiekiem kobiety i mężczyzny w momencie zawierania małżeństwa. Przykład wybrałem z tego względu, że odtworzenie modelu zjawiska jest w tym wypadku sprawą dość złożoną i rodzić może szereg wątpliwości. W kolejnym podrozdziale (4.9) przedstawię sposób weryfikacji zgodności znalezionego modelu z liczebnościami pól tablicy otrzymanymi w badaniu. Wykorzystam do tego celu rozwiązania zaproponowane w ramach metod modelowania log-liniowego. W ostatnim podrozdziale (4.10) omówię podejście zogniskowane na identyfikacji jedynie najbardziej znaczących pól w tablicy. Podejście to należy do chętnie stosowanych, zwłaszcza w badaniach marketingowych.

4.1 Czym jest model związku w tablicy

Przez **model związku** (ang. *pattern of association*) będziemy rozumieć podział pól wnętrza tablicy na pewną liczbę kategorii. Przy czym pola należące do danej kategorii podziału charakteryzuje pewna wspólna własność, czy też – mówiąc inaczej – cechują się one pewną specyfiką wobec innych pól czy podzbiorów pól w tablicy. Podejście przedstawione w tym rozdziale sprowadza się do znalezienia owego modelu i odtworzenia za jego pomocą liczebności w polach tablicy. Model stanowić może podstawę wyjaśnienia zarówno mechanizmu zjawiska, jak i jego rozmiarów.

Tabela 4.1
Odsetki głosujących na kandydatów poszczególnych partii w wyborach do Sejmu we wrześniu 2005 roku wśród kobiet i mężczyzn. Europejski Sondaż Społeczny 2006
[*w procentach*]

	sposób głosowania									
płeć	PiS	PO	SLD	LPR	pozostałe partie	Samoobrona	PSL	odmowa lub nie pamięta	ogółem	liczba osób
kobiety	38	27	8	2	2	6	2	15	100	568
mężczyźni	39	26	8	2	2	10	4	10	100	494
ogółem	38	26	8	2	2	8	3	13	100	1062

Dane prezentowane uprzednio w tabeli 2.9.

Identyfikacji modelu rozpatrywanego związku dokonaliśmy już *de facto* w podrozdziale 2.2.7, interpretując sposób głosowania mężczyzn i kobiet (tabela 2.9). W tabeli 4.1 przytoczone zostały jeszcze raz te same dane w celu ilustracji zasady podziału pól tablicy na kategorie. Pola odpowiadające głosowaniu na PiS, PO, SLD, LPR lub na pozostałe partie zgrupowane zostały w lewej części tabeli 4.1. Tworzą blok pól posiadających tę wspólną własność, że odsetki kobiet i mężczyzn głosujących na poszczególne partię są do siebie podobne. Aby blok ten graficznie wyodrębnić, każde z należących do niego pól zaznaczone zostało ramką. Pola drugiego bloku zaznaczone zostały kolorem jasno szarym. Pola te odpowiadają partiom, na które mężczyźni głosowali wyraźnie częściej niż kobiety. Trzeci blok obejmuje pola oznaczone kolorem ciemno szarym. Obejmują one odmowy ujawnienia ankieterowi, na które z ugrupowań badana osoba oddała głos. Postawy te częściej występowały u kobiet.

Tabela 4.1 ma dość prostą strukturę, toteż podziału pól na kategorie – czyli utworzenia modelu związku – dokonać można, porównując dla obu płci profile sposobów głosowania. Gdy struktura tablicy jest bardziej złożona lub gdy z góry niewiele wiadomo o badanym zjawisku, to wtedy bardziej efektywne okazuje się zestawienie pól tablicy z liczebnościami modelu referencyjnego. Powszechnie korzysta się w tym celu z omówionego w rozdziale 3 modelu niezależności. Aczkolwiek przy badaniu niektórych problemów lepiej posłużyć się innymi modelami (Sawiński 1984; Krauze i Słomczyński 1985).

4.2 Narzędzia porównywania liczebności: stosunek i różnica

Aby zobrazować rozbieżności między liczebnościami w tablicy skonstruowanej na podstawie wyników badania a liczebnościami modelu niezależności korzysta się z dwóch elementarnych pojęć: stosunku i różnicy. Stosunki oblicza się, dzieląc liczebności empiryczne w polach tablicy przez właściwe dla tych pól liczebności modelu niezależności

$$s_{ij} = \frac{n_{ij}}{e_{ij}} \tag{4.1}$$

Im bardziej wielkość stosunku przewyższa 1, tym większa jest uzyskana w badaniu liczebność n_{ij} w stosunku do liczebności e_{ij} w modelu niezależności. Na przykład, gdy wartość stosunku (4.1) wynosi 2, to liczebność uzyskana w badaniu jest 2 razy od niej większa. Z kolei wartości mniejsze od 1 świadczą o tym, że uzyskana w badaniu liczebność jest mniejsza od wielkości danego pola w modelu niezależności. Wartości bliskie 1 interpretuje się natomiast jako

świadectwo faktu, że liczebności empiryczne w niewielkim stopniu odbiegają od wielkości, które powinny znaleźć się w polach tablicy zgodnie z modelem niezależności.

Drugi sposób pomiaru wielkości rozbieżności pomiędzy liczebnościami w tablicy a liczebnościami modelu niezależności polega na obliczeniu różnic porównywanych liczebności

$$d_{ij} = n_{ij} - e_{ij} \tag{4.2}$$

Licząc różnicę z zasady od liczebności uzyskanych w badaniu odejmuje się liczebności modelu niezależności, gdyż ten sposób interpretowania znaku różnicy jest bardziej naturalny. Dodatnie wartości różnic świadczą wtedy, że liczebności uzyskane w badaniu odbiegają *in plus* od liczebności, których należałoby się spodziewać, gdyby wyniki badania spełniały założenia modelu. Różnice ujemne świadczą natomiast, że w danym polu tablicy jest za mało jednostek w stosunku do sytuacji niezależności.

4.3 Korzyści i ograniczenia różnic

Obliczenie różnic pozwala na wyrażenie ocenianych rozbieżności w liczbie badanych jednostek. Jest to korzyść nie do przecenienia, gdyż pozwala budować interpretacje oparte na wielkościach istniejących na poziomie empirycznym.

Tabela 4.2
Różnice między liczbą kobiet i mężczyzn głosujących na kandydatów poszczególnych partii a liczebnościami modelu niezależności
Europejski Sondaż Społeczny 2006

	sposób głosowania								
płeć	PiS	PO	SLD	Samo-obrona	PSL	LPR	pozo-stałe partie	odmowa lub nie pamięta	ogółem
kobiety	-2	3	1	-12	-6	2	1	14	0
mężczyźni	2	-3	-1	12	6	-2	-1	-14	0
ogółem	0	0	0	0	0	0	0	0	0

Różnice wielkości zamieszczonych w tabelach 2.8 i 3.3.

W tabeli 4.2 przedstawiono różnice między liczebnościami pól tablicy opisującej związek sposobu głosowania z płcią a liczebnościami modelu niezależności. Sumy różnic w poszczególnych wierszach, kolumnach, a także

w całej tablicy są równe 0. Wynika to z faktu, że model niezależności ma te same rozkłady brzegowe co tablica liczebności uzyskanych w badaniu (wzory 3.11–3.13). Aby ocenić stopień, w jakim cała tablica odbiega od niezależności, wykorzystać można sumę różnic dodatnich. Jest ona równa sumie wartości bezwzględnych różnic we wszystkich polach tablicy podzielonej przez 2, gdyż suma różnic dodatnich jest zawsze równa co do wielkości sumie różnic ujemnych. Otrzymaną w ten sposób wielkość miary rozbieżności nazwiemy **minimalną liczbą przemieszczeń** (Sawiński 1984: 39–59) i oznaczymy jako *mlp*

$$mlp = \frac{1}{2}\sum_{i=1}^{w}\sum_{j=1}^{k}\left|n_{ij} - e_{ij}\right| \tag{4.3}$$

Dla różnic przedstawionych w tabeli 4.2 wielkość ta wynosi 41 osób.

Minimalna liczba przemieszczeń posiada interpretację operacyjną. W omawianym przykładzie wielkość ta określa, ile co najmniej osób musiałoby zmienić swoją odpowiedź, aby zamiast tablicy faktycznie otrzymanej uzyskać liczebności modelu niezależności. Rozważmy to na przykładzie konkretnego pola tablicy. Według wyników badania 153 kobiety głosowały na PO (tabela 2.8). Gdyby wyniki spełniały założenia modelu niezależności, to powinno być ich 150 (tabela 3.3). Różnica tych dwóch wielkości, równa +3, określa więc nadmiar kobiet w stosunku do modelu niezależności. Gdyby kobiety stanowiące nadmiar udzieliły w badaniu odpowiedzi należącej do jednej z kategorii, w których występuje niedomiar kobiet (odpowiedzi: PiS, Samoobrona lub PSL; tabela 4.2), to ów nadmiar zostałby zlikwidowany. Liczba kobiet głosujących na PO byłaby wtedy równa wielkości oczekiwanej przy założeniu, że sposób głosowania jest niezależny od płci[1].

W opisany sposób zlikwidować można każdy z nadmiarów (dodatnich różnic d_{ij}), „przenosząc" badanych do tych pól tablicy, w których występuje niedomiar (ujemne różnice d_{ij}). W sumie, aby tablicę liczebności uzyskanych w badaniu przekształcić w tablicę niezależności, należałoby „przenieść" 41 osób. Jest to niedużo biorąc pod uwagę fakt, że tablica została skonstruowana na podstawie odpowiedzi 1062 osób. Płynie stąd wniosek, że związek płci ze sposobem głosowania odbiega niewiele od niezależności.

[1] Liczebności w modelu niezależności mają na ogół postać liczb niecałkowitych. Wprowadza to komplikację do operacyjnej interpretacji liczby przemieszczeń, gdyż „przenoszenie" ułamkowych części jednostek (na przykład osób) nie ma odpowiednika na poziomie empirycznym. Niedokładność z tym związana nie prowadzi jednak do trudności interpretacyjnych, zwłaszcza gdy rozbieżności między tablicą empiryczną a modelem niezależności są wyraźne. Dlatego zasadnie jest przyjąć, że w większości zastosowań skutki związane z niecałkowitą postacią liczb w modelu niezależności mogą być pominięte bez szkody dla uzyskanych wniosków.

Odtwarzając model związku na podstawie różnic, uwagę ogranicza się do największych rozbieżności. Dwie największe różnice (+14 i –14) wystąpiły w wypadku osób, które nie odpowiedziały ankieterowi, na które z ugrupowań oddały głos (tabela 4.2). Nie powinno budzić wątpliwości, że odpowiadające im pola tablicy warto wyodrębnić w postaci osobnego bloku. Wątpliwości nie powinno też być w wypadku dwóch pól obejmujących kobiety i mężczyzn głosujących na Samoobronę. Ponieważ znaki różnic są w tym wypadku odwrotne niż w wyodrębnionym już bloku „odmów"; osoby głosujące na Samoobronę należy wyodrębnić w postaci osobnej kategorii.

Na marginesie warto zauważyć, że decyzję o odrębności obu bloków pól podjęlibyśmy nawet w wypadku, gdyby główka i boczek tablicy zostały zakryte i nie byłoby wiadomo, jakie cechy skrzyżowano w tablicy. Przeciwstawne znaki różnic stanowią bowiem dość sugestywne kryterium, przez co często są stanowią podstawę wyodrębnienia bloków podobnych pól. W fazie interpretacji zidentyfikowanego w ten sposób modelu należy wziąć pod uwagę, że tablica stanowi całość. Niedobór osób w niektórych polach wiąże się z ich nadmiarem w innych. W rozpatrywanym przykładzie nie można wykluczyć, że niedomiar kobiet wśród głosujących na Samoobronę wiąże się z ich nadmiarem w kategorii odmów. Oznaczałoby to, że część kobiet nie ujawniła ankieterom faktu głosowania na Samoobronę.

Wróćmy jednak do wyodrębnianych bloków. Kwestia sklasyfikowania pól obejmujących osoby, które głosowały na PSL, nie jest tak oczywista, jak w wypadku poprzednich kategorii. Z jednej strony znaki różnic są identyczne, jak w wypadku bloku obejmującego osoby głosujące na Samoobronę, co skłaniałoby do wyodrębnienia obu kategorii w jednym bloku. Z drugiej zaś strony wielkości różnic są niewielkie (+/– 6 osób), co skłaniałoby do zaliczenia ich do bloku obejmującego pola niewiele odbiegające od modelu niezależności. Gdy powstają tego rodzaju wątpliwości, to można spróbować utworzyć więcej niż jeden model, a następnie sprawdzić, który z nich okaże się bardziej przydatny do wyjaśnienia badanego zjawiska.

Wyodrębniając model na podstawie różnic warto mieć na uwadze fakt, że nie uwzględniają one wielkości brzegowych obu cech, czyli potencjału poszczególnych kategorii. Różnice ±2 osoby uzyskano dla PiS-u, na który głosowało 38 procent badanych, jak też dla LPR-u wymienianego jedynie przez 2 procent respondentów. Wydaje się uzasadnione, aby wielkości stwierdzonych różnic relatywizować do wielkości kategorii obu cech. W przeciwnym wypadku pominięty zostanie ten aspekt związku, który we wcześniejszych rozdziałach nazwaliśmy rozmiarami czy zasięgiem badanego zjawiska.

Ramka 4.1
Adolphe Quételet (1796–1874). Prekursor współczesnych badań ilościowych

Adolphe Jacques Quételet urodził się w 1796 roku w Belgii. W 1819 roku uzyskał na uniwersytecie w Gent tytuł doktora matematyki. W późniejszych latach działał w obrębie wielu dyscyplin naukowych. Jako astronom był inicjatorem powstania i pierwszym dyrektorem Królewskiego Obserwatorium Astronomicznego w Brukseli. Najbardziej znaczący ślad pozostawił jednak Quételet w trzech innych dyscyplinach, które obecnie zaliczylibyśmy do statystyki, socjologii i kryminologii.

Nas najbardziej interesuje wkład Quételeta w dziedzinie badań. Quételet dużo czasu poświęcał szukaniu prawidłowości w danych antropometrycznych, a także w danych statystycznych, które w owych czasach zaczynano systematycznie gromadzić. Rezultatem tych zainteresowań była koncepcja „przeciętnego człowieka" (*l'homme moyen*), którą Quételet sformułował w 1831 roku, rozbudowując ją w następnych latach (Stigler 1999: 51–65). Przeciętny człowiek stanowił kategorię pojęciową, która służyła Quételetowi do wyjaśniania społecznych uwarunkowań zjawisk zaobserwowanych w danych statystycznych. Quételet stwierdził między innymi, że wskaźniki przestępczości są niejednakowe wśród osób w różnym wieku, o różnym wykształceniu, czy mieszkających w różnych regionach. Jego zdaniem dowodziło to, że poszczególne kategorie społeczne cechuje niejednakowa skłonność do popełniania przestępstw. Średni poziom owej skłonności, który można wyliczyć z danych, to właśnie ów przeciętny człowiek. Inny jest wzór przeciętnego człowieka wśród mężczyzn i kobiet, wśród osób w różnym wieku, czy o różnym wykształceniu. Na wzory te nakładają się cechy indywidualne, które powodują, że poszczególne osoby odbiegają od wzorca właściwego dla swojej grupy. Wzory przeciętnego człowieka są przy tym immanentną charakterystyką danego społeczeństwa. Stanowią element porządku społecznego i społecznej integracji (Szacki 1981: 287–293; Stigler 1986: 161–182).

Wzór przeciętnego człowieka zarazem jest taką charakterystyką społeczeństwa, którą można badać. Warunkiem uzyskania wiarygodnej wiedzy jest jednak możliwość abstrahowania od indywidualnych różnic między ludźmi. Aby cel ten zrealizować Quételet postulował prowadzenie badań w odpowiednio dużych zbiorowościach: „[...] im większa jest liczba obserwowanych jednostek, tym bardziej swoiste cechy indywidualne [...] zacierają się i ustępują miejsca serii faktów ogólnych, dzięki którym społeczeństwo istnieje i przedłuża swoje istnienie" (cytuję za Szacki 1981: 290). Twierdził także, iż „precyzja wyników wzrasta wraz z wzrostem pierwiastka z liczby dokonanych obserwacji. Gdy wyniki pochodzą z niewielkiej liczby obserwacji, ich odchylenia od przeciętnej będą większe" (cytuję za: Stigler 1986, s. 180). Quételeta można więc uznać za prekursora współczesnych badań ilościowych. Przedstawił on nie tylko istotę tego rodzaju badań, lecz również zaproponował narzędzie analizy ich wyników – w postaci konceptu człowieka przeciętnego.

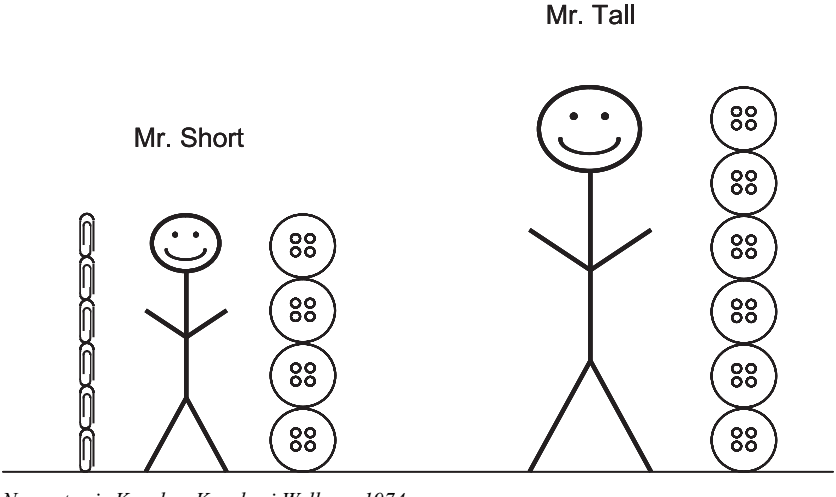
Z innych dokonań Quételeta, które znalazły zastosowanie w badaniach, warto wymienić miarę diagnostyczną dotyczącą wagi osób w populacji, znaną jako BMI (Body Mass Index). Propozycja ta wywodzi się z antropometrycznych zainteresowań Quételeta i zaczęła być stosowana gdzieś pomiędzy 1830 a 1850 rokiem. Zapewne ponadprzeciętny autorytet naukowy Quételeta przyczynił się do wzrostu zainteresowania tym wskaźnikiem w owych czasach. Dalsza kariera tej w sumie prostej miary wydaje się jednak zaskakująca. Współcześnie wskaźnik ten stosowany jest nie tylko w badaniach antropometrycznych czy medycznych, lecz również w badaniach sondażowych.

Pośrednio Quételetowi zawdzięczamy również nazwanie naszej dziedziny socjologią. Quételet określał przedmiot swojego zainteresowania jako *physique sociale*. Tak się jednak złożyło, że tego samego określenia używał niezależnie August Comte (1798–1857). Comte nie cenił koncepcji Quételeta, toteż aby zaznaczyć odrębność własnej teorii, zdecydował się zmienić jej określenie. I stąd wziął się termin *socjologia* (Szacki 1981: 270).

4.4 Użyteczność indeksów

Aby wyjaśnić, dlaczego stosunek dwóch wielkości jest bardziej uniwersalnym narzędziem porównywania dwóch wielkości niż różnica, można odwołać się do koncepcji stadiów rozwoju poznawczego człowieka (Piaget 1972). Rozumowanie w kategoriach proporcjonalności (ang. *proportional reasoning*) kształtuje się w najpóźniejszym stadium rozwoju, to jest dopiero w okresie wchodzenia w wiek dorosłości (Noelting 1980). Najbardziej znany test umiejętności proporcjonalnego rozumowania opracował Robert Karplus i zatytułował go „Mr. Short–Mr. Tall Problem" (Karplus, Karplus i Wollman 1974). Na rysunku przedstawione zostały figury niskiego oraz wysokiego człowieczka (rycina 4.1). Wzrost niskiego człowieczka określony został jako 6 spinaczy biurowych lub 4 duże guziki. Wysoki człowieczek mierzy zaś 6 dużych guzików. Polecenie brzmi: oblicz, ile spinaczy wynosi wzrost wysokiego człowieczka i wyjaśnij dlaczego.

Rycina 4.1
Rysunek w teście „Mr Short–Mr Tall Problem"



Na postawie Karplus, Karplus i Wollman 1974.

Liczne zastosowania testu pozwoliły stwierdzić, że u dzieci a także u osób dorosłych, które nie opanowały w pełni myślenia symbolicznego, dominuje addytywne podejście do rozwiązania przedstawionego zadania. Na przykład: „różnica wzrostu wysokiego i niskiego człowieczka wynosi 2 guziki, toteż ich wzrost będzie się różnił o 2 spinacze. Wzrost wysokiego człowieczka wyno-

si więc 8 spinaczy". Test pozwala też wykazać, że u osób posługujących się rozumowaniem proporcjonalnym jego zastosowanie przybiera różną postać. Na przykład: „w wypadku niskiego człowieczka jeden guzik jest równy półtora spinacza. Mnożąc wzrost wysokiego człowieczka przez półtora uzyska się 9 spinaczy". Bądź inaczej: „mierząc w guzikach wysoki człowieczek jest półtora raza wyższy od niskiego. Wzrost niskiego w spinaczach należy więc przemnożyć przez półtora, co daje 9 spinaczy". Bądź jeszcze inaczej: „2 guziki to 3 spinacze. Ponieważ wysoki człowieczek jest o 2 guziki wyższy, jest też wyższy o 3 spinacze. Jego wzrost wynosi więc 9 spinaczy". Ostatni z przykładów świadczy, że niekiedy myślenie proporcjonalne można skutecznie łączyć z addytywnym (Karplus, Karplus i Wollman 1974).

Badania prowadzone na gruncie psychologii rozwojowej dowodzą więc, że rozumowanie proporcjonalne w stosunku do podejścia addytywnego jest bardziej elastyczną a zarazem skuteczną strategią przy rozwiązywaniu problemów opartych na porównywaniu dwóch wielkości. Jest to chyba powodem, dla którego miara oparta na stosunku jest narzędziem najczęściej przez badaczy stosowanym. Chodzi o współczynnik zwany popularnie **indeksem**[2], będący ilorazem porównywanych wielkości przemnożonym przez 100. W wypadku porównywania liczebności w polach tablicy z liczebnościami modelu niezależności, wartość indeksu dla danego pola wyraża się wzorem

$$i_{ij} = 100 \frac{n_{ij}}{e_{ij}} \tag{4.4}$$

W tabeli 4.3 przedstawiono wartości indeksów dla pól tablicy opisującej związek między płcią a sposobem głosowania. Największe rozbieżności między liczebnościami tablicy a modelem niezależności ujawniają się w wypadku dwóch pól tablicy odpowiadających kobietom i mężczyznom głosującym na PSL. Dla mężczyzn wartość indeksu wyniosła 147 co oznacza, że liczebność pola aż o 47 procent przewyższa liczebność oczekiwaną przy założeniu, że sposób głosowania jest niezależny od płci. Dla kobiet wartość indeksu wyniosła 59, czyli respondentki głosujące na PSL stanowią 59 procent liczebności modelu niezależności.

[2] Pisząc o tym, że indeks jest najczęściej stosowanym narzędziem porównywania ze sobą dwóch wielkości, opieram się na obserwacji sposobów analizowania danych przez badaczy. Wniosek znajduje też potwierdzenie w rozwiązaniach stosowanych w programach komputerowych wspomagających analizę danych w badaniach marketingowych. Trudno bowiem wskazać oprogramowanie zorientowane na tę klasę badań, w którym procedura liczenia indeksów nie byłaby dostępna na równie podstawowym poziomie, jak odsetki. Pewna niechęć do tego wskaźnika cechuje natomiast autorów oprogramowania adresowanego do przedstawicieli badań akademickich. Procedury liczenia indeksów nie znajdzie się na przykład w pakiecie SPSS.

Uzyskany wynik warto zestawić z wnioskami otrzymanymi na podstawie analizy różnic (tabela 4.2), które ujawniły specyfikę grupy głosujących na PSL dopiero w trzeciej kolejności. W wypadku różnic zarysowała się ona wyraźnie słabiej zarówno od specyfiki osób, które nie odpowiedziały, na kogo głosowały, jak też osób, które głosowały na Samoobronę. Tym samym analiza indeksów prowadzi do znalezienia odmiennego modelu związku, niż analiza różnic. Nie ulega wątpliwości, że posługując się wartościami indeksów, jako osobny blok należy wyodrębnić pola tablicy obejmujące respondentów głosujących na Samoobronę lub na PSL. Natomiast kwestia utworzenia dalszych bloków nie jest już tak oczywista. Jeśli zdecydowalibyśmy się utworzyć blok z osób, które nie udzieliły odpowiedzi na kogo głosowały, to w zasadzie należałoby do nich dołączyć osoby, które głosowały na LPR. Wielkości indeksów w obu tych kategoriach są bowiem podobne.

Tabela 4.3
Indeksy dla kobiet i mężczyzn głosujących na kandydatów poszczególnych partii

Europejski Sondaż Społeczny 2006

	sposób głosowania							
płeć	PiS	PO	SLD	Samo-obrona	PSL	LPR	pozostałe partie	odmowa lub nie pamięta
kobiety	99	102	102	73	59	117	109	119
mężczyźni	101	98	97	131	147	80	90	78

Obliczono według wzoru (4.4) na podstawie danych prezentowanych w tabelach 2.8 i 3.3.

Gdy dwa podejścia prowadzą do odmiennych konkluzji, to warto ustalić, co jest tego przyczyną. Języczek u wagi stanowi kategoria osób głosujących na LPR. Wielkości różnic w tej kategorii są niewielkie i wynoszą ±2 osoby. Ponieważ jednak w wypadku kobiet głosujących na LPR liczebność modelu niezależności wynosi zaledwie 11 osób (tabela 3.3), stąd owe dwie stanowią nadwyżkę aż 17-procentową. Podobna co do wielkości nadwyżka 19 procent występuje wśród kobiet, które nie odpowiedziały na pytanie na kogo głosowały. W tym jednak wypadku liczebność modelu niezależności wynosi 72 osoby, stąd też 19-procentowa nadwyżka odpowiada różnicy 14 kobiet. Widać stąd, że indeksy premiują kategorie o niewielkich liczebnościach brzegowych, które na ogół mają mniejsze znaczenie dla obrazu zjawiska.

Zarówno więc różnice, jak i indeksy posiadają swoje ograniczenia, przez co identyfikacja modelu związku na bazie któregokolwiek z tych wskaźników nie zawsze prowadzić będzie do użytecznych wniosków.

4.5 Wskaźniki Quételeta

Wybór wskaźnika trafnie charakteryzującego specyfikę poszczególnych pól w tablicy nie jest problemem nowym. Rozwiązania stosowane współcześnie wywodzą się bowiem z propozycji, jaką w pierwszej połowie XIX wieku sformułował belgijski badacz i statystyk Adolphe Quételet (1796–1874). O wkładzie Quételeta w rozwój nauki – nie tylko zresztą w dziedzinie statystyki, lecz również socjologii – piszę osobno w ramce 4.1. W tym miejscu ograniczę się do przedstawienia zaproponowanych przez Quételeta wskaźników przeznaczonych do analizy pól w tablicach.

Nie potrafię powiedzieć, czy proponując swoje wskaźniki, Quételet dostrzegał zalety i ograniczenia różnic i indeksów. Nie wykazałem bowiem aż tyle determinacji, aby dotrzeć do tekstów źródłowych (Quételet 1832, cytuję za Mirkinem 2001; Quételet 1849, cytuję za Goodmanem i Kruskalem 1959). W każdym razie istota zaproponowanych przez niego wskaźników polega na złożeniu pojęć różnicy i stosunku. Pozwala to uwypuklić korzyści obu rozwiązań a zarazem zredukować ich ograniczenia. Budowa wskaźników Quételeta jest zresztą tak oczywista, że w historii statystyki niejednokrotnie proponowano podobne wskaźniki, przy czym należy sądzić, że w większości wypadków bez znajomości oryginalnej propozycji. Goodman i Kruskal (1959: 133) na temat jednego ze wskaźników Quételeta piszą wręcz, że

this ratio probably has been used since nearly the beginning of arithmetic

co zresztą nie umniejsza rangi dokonań Quételeta, lecz wręcz przeciwnie. Gdy te same idee przewijają się w coraz to nowych propozycjach od prawie 200 lat to oznacza, że kryje się w nich ponadczasowa mądrość. Zauważył to już sto lat temu George Udny Yule (1912: 586), który jeden ze swoich współczynników oznaczył literą Q dla uhonorowania wkładu Quételeta w badania struktury tablic.

Zaproponowany przez Quételeta wskaźnik – którym się tu zajmiemy – jest różnicą proporcji warunkowej i proporcji brzegowej (Mirkin 2001: 114)

$$w_{ij} = \frac{n_{ij}}{a_i} - \frac{b_j}{n} \tag{4.5}$$

Zanim jednak przejdziemy do interpretacji wskaźników Quételeta dla pól konkretnej tablicy konieczne jest uzgodnienie sposobu prezentowania ich wielkości numerycznych. Wyrażenie we wzorze 4.5 jest różnicą proporcji. Przypomnijmy, że suma proporcji dla wszystkich pól tablicy wynosi 1. W efekcie, wartość wyrażenia (4.5) jest ułamkiem dziesiętnym o niewielkiej wartości. Na

przykład, dla pola odpowiadającego mężczyznom głosującym na PSL wartość wskaźnika Quételeta wynosi 0,0121 (tabela 4.4). Posługiwanie się tak małymi wielkościami, aczkolwiek matematycznie uzasadnione, jest w praktyce niewygodne. Dlatego we wszelkich prezentacjach wskaźników Quételeta będziemy podawać ich wartości przemnożone przez 100. Jest to konwencja analogiczna do powszechnie przyjętej zasady mnożenia przez 100 wartości indeksów omawianych w podrozdziale 4.4.

Co więcej, przemnożone przez 100 wskaźniki Quételeta mają bardziej bezpośrednią interpretację niż oryginalne wielkości. Zastąpmy bowiem we wzorze 4.5 odjemną n_{ij} / a_i przez odsetek, jaki stanowi wielkość w danym polu tablicy do liczebności *i*-tego wiersza, zaś odjemnik b_j / n przez odsetek wielkości brzegowej w danej kolumnie w stosunku do wszystkich badanych. Wtedy wartość współczynnika Quételeta interpretować można jako różnicę odsetków. Dla mężczyzn głosujących na PSL wygląda to następująco. Odsetek mężczyzn głosujących na PSL wśród ogółu badanych mężczyzn wynosi 3,80 (w procentach, z dokładnością do dwóch cyfr po przecinku), odsetek głosujących na PSL wśród ogółu badanych jest równy 2,59, stąd różnica wynosi 1,21. Przemnożone przez 100 wartości wskaźników Quételeta mają więc interpretację jako różnice odsetków w porównywanych polach.

Jak można interpretować obliczoną wartość? Po pierwsze, świadczy ona, że w polu, dla którego obliczono wartość wskaźnika Quételeta, występuje nadwyżka mężczyzn. Obliczona wartość ta okazała się bowiem dodatnia. Po drugie, wartość wskaźnika określa wielkość tej nadwyżki. Obejmuje ona 1,21 procenta badanych mężczyzn. Czy jest to znacząca nadwyżka można przekonać się obliczając wartości wskaźników Quételeta dla pozostałych pól tablicy (część [4] tabeli 4.4). Największy niedobór mężczyzn występuje w polu odpowiadającym odmowie udzielenia odpowiedzi na pytanie (-2,74), zaś największe „nadwyżki" występują wśród mężczyzn głosujących na Samoobronę (2,45) i właśnie PSL (1,21). Wskaźniki Quételeta sumują się do zera w każdym z wierszy tablicy. Ułatwia to zauważenie, że niedobory osób w niektórych polach związane są z ich nadmiarem w innych[3].

[3] Wielkości tych nie należy interpretować jako przepływów między polami tablicy. Przepływy dotyczyć mogą wyłącznie osób, nigdy zaś proporcji. Wyjaśniam to bardziej szczegółowo w podrozdziale 5.2.

Tabela 4.4
Wielkości wybranych parametrów pól tablicy opisującej sposób głosowania kobiet i mężczyzn w wyborach we wrześniu 2005 roku
Europejski Sondaż Społeczny 2006

płeć	sposób głosowania: PiS	PO	SLD	Samo-obrona	PSL	LPR	pozo-stałe partie	odmowa lub nie pamięta	ogółem
[1] Liczebności n_{ij} uzyskane w badaniu									
kobiety	214,71	153,18	46,86	32,98	8,75	13,05	13,45	85,22	568,19
mężczyźni	191,14	127,65	38,65	51,23	18,77	7,76	9,71	48,72	493,64
ogółem	405,85	280,83	85,51	84,21	27,52	20,81	23,15	133,94	1061,83
[2] Liczebności e_{ij} modelu niezależności									
kobiety	217,17	150,28	45,76	45,06	14,73	11,14	12,39	71,67	568,19
mężczyźni	188,68	130,56	39,75	39,15	12,79	9,68	10,76	62,27	493,64
ogółem	405,85	280,83	85,51	84,21	27,52	20,81	23,15	133,94	1061,83
[3] Różnice d_{ij} wielkości [1] i [2]									
kobiety	-2,46	2,91	1,10	-12,08	-5,98	1,91	1,06	13,55	0,00
mężczyźni	2,46	-2,91	-1,10	12,08	5,98	-1,91	-1,06	-13,55	0,00
ogółem	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
[4] Wskaźniki w_{ij} Quételeta dla kobiet i mężczyzn (przemnożone przez 100)									
kobiety	-0,43	0,51	0,19	-2,13	-1,05	0,34	0,19	2,38	0,00
mężczyźni	0,50	-0,59	-0,22	2,45	1,21	-0,39	-0,21	-2,74	0,00
[5] Wskaźniki w_{ij} Quételeta dla sposobów głosowania (przemnożone przez 100)									
kobiety	-0,61	1,04	1,29	-14,35	-21,73	9,20	4,56	10,11	
mężczyźni	0,61	-1,04	-1,29	14,35	21,73	-9,20	-4,56	-10,11	
ogółem	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	
[6] wskaźniki q_{ij} (przemnożone przez 100)									
kobiety	-0,51	0,73	0,50	-5,52	-4,78	1,76	0,92	4,91	
mężczyźni	0,55	-0,78	-0,54	5,93	5,13	-1,89	-0,99	-5,27	

Tabela 4.4 (kontynuacja)

płeć	sposób głosowania: PiS	PO	SLD	Samo-obrona	PSL	LPR	pozo-stałe partie	odmowa lub nie pamięta	ogółem
[7] kwadraty wskaźników q_{ij} przemnożone przez liczbę badanych osób n									
kobiety	0,03	0,06	0,03	3,24	2,43	0,33	0,09	2,56	8,76
mężczyźni	0,03	0,06	0,03	3,73	2,79	0,38	0,10	2,95	10,08
ogółem	0,06	0,12	0,06	6,97	5,22	0,71	0,19	5,51	18,84
[8] udziały kwadratów wskaźników q_{ij} w wartości współczynnika χ^2 (w procentach) (dekompozycja χ^2 między pola tablicy)									
kobiety	0,15	0,30	0,14	17,20	12,89	1,75	0,48	13,59	46,49
mężczyźni	0,17	0,34	0,16	19,80	14,83	2,01	0,55	15,64	53,51
ogółem	0,32	0,64	0,30	37,00	27,72	3,76	1,03	29,23	100,00

Porównanie wskaźników dla kobiet i mężczyzn pozwala zauważyć tę samą prawidłowość, którą zidentyfikowaliśmy, analizując różnice w podrozdziale 4.3. Mianowicie, niedobór kobiet głosujących na Samoobronę i PSL może wynikać z faktu, że część kobiet nie ujawniła swojego sposobu głosowania. Porównanie wskaźników Quételeta z wielkościami różnic – zamieszczonymi dla wygody w części [3] tabeli 4.4 – wykazuje zresztą daleko idącą zbieżność obu miar. We wszystkich polach zgadza się znak porównywanych wielkości (plus lub minus). Ponadto wielkości te są mniej więcej do siebie proporcjonalne. Widoczna prawidłowość uzyskana została nieprzypadkowo. Wzór na omawiany wskaźnik Quételeta można bowiem również przedstawić w następującej postaci

$$w_{ij} = \frac{d_{ij}}{a_i} = \frac{n_{ij} - e_{ij}}{a_i} \tag{4.6}$$

Jego wartość jest więc równa omawianej wcześniej różnicy zrelatywizowanej do liczebności brzegowej cechy w wierszach tablicy. Równoważność obu wzorów świadczy, że wielkości w tablicy interpretować można na wiele różnych sposobów. Analogicznie, jak w wypadku omawianego wcześniej testu „Mr. Short–Mr. Tall". Badani tym testem, którym nieobce było myślenie w języku proporcjonalności, byli w stanie dojść do poprawnego rozwiązania na jeden z kilku sposobów. Wskaźnik Quételeta stwarza podobne możliwości. Części badaczy łatwiej będzie pojąć relację pól tablicy do niezależności po-

przez porównanie profili w wierszach z profilem brzegowym, czyli z profilem sposobu głosowania obliczonym dla wszystkich badanych (wzór 4.5). Innym być może łatwiej jest porównać różnice w poszczególnych polach z marginesem wierszy (wzór 4.6). Jak mówi przysłowie: wszystkie drogi prowadzą do Rzymu. Ostatecznie chodzi o to, aby znaleźć model związku w tablicy.

Wartości omawianego wskaźnika Quételeta obliczaliśmy dotychczas w taki sposób, że poszczególne pola relatywizowane były do rozkładu cechy umieszczonej w kolumnach, czyli do liczby badanych głosujących w różny sposób w wyborach. Równie dobrze obliczyć go można w drugą stronę, porównując profile płci wśród głosujących na poszczególne partie z udziałami mężczyzn i kobiet w badaniu, bądź zestawiając różnice w polach z liczbą głosujących w różny sposób. Odpowiednie wzory analogiczne są do wyrażeń (4.5) i (4.6). Oznaczmy wskaźnik Quételeta obliczany w drugą stronę jako w'_{ij}. Wówczas

$$w'_{ij} = \frac{n_{ij}}{b_j} - \frac{a_i}{n} \tag{4.7}$$

lub

$$w'_{ij} = \frac{d_{ij}}{b_j} = \frac{n_{ij} - e_{ij}}{b_j} \tag{4.8}$$

Warto poświęcić chwilę na analizę otrzymanych tą drogą wartości wskaźników Quételeta. Zostały one zamieszczone w części [5] tablicy 4.4. Pierwsze wrażenie podpowiada, że wartości te są większe od obliczonych poprzednio. Na przykład wartość wskaźnika dla mężczyzn głosujących na Samoobronę wynosi obecnie 14,35, gdy poprzednio wynosiła 2,45. Wynika to stąd, że obecnie różnice d_{ij} w polach tablicy odnosimy nie do liczby badanych kobiet lub mężczyzn, lecz do liczby osób głosujących na kandydatów poszczególnych partii. Te ostatnie wielkości w większości wypadków są mniejsze. Wśród badanych było 494 mężczyzn (w zaokrągleniu), zaś na PSL głosowało 28 respondentów. Jeśli tę samą nadwyżkę mężczyzn, wynoszącą 6 osób, odniesiemy do łącznej liczby osób głosujących na PSL, to wskaźnik Quételeta przybierze wyższą wartość niż w wypadku relatywizacji tej wielkości do ogółu badanych mężczyzn. Warto zwrócić uwagę na fakt, że liczba respondentów głosujących na PiS była wielkością podobnego rzędu co liczba badanych kobiet bądź mężczyzn. Stąd wskaźniki Quételeta w pierwszej kolumnie tablicy przybierają zbliżone wielkości niezależnie od tego, w którą stronę są liczone.

Omawiana dualność wskaźników Quételeta odpowiada dwóm sposobom interpretacji tablic, opartym na analizie profili w wierszach bądź w kolumnach. W wypadku rozważanej tablicy interpretacje te zostały omówione w rozdziale 2. Wskaźniki Quételeta są spójne z tymi wnioskami, gdyż jeden ze spo-

sobów ich liczenia polega właśnie na zestawianiu ze sobą profili (wzory 4.5 i 4.7). Sześciu badanych mężczyzn stanowiących nadwyżkę wśród głosujących na PSL ma **większe znaczenie** dla stratega tej partii niż dla badacza, którego interesują różnice w sposobach głosowania mężczyzn i kobiet w kontekście pełnych wyników wyborów. Doniosłość propozycji Quételeta polega między innymi na tym, że uwzględnił potrzebę **różnej interpretacji tego samego pola tablicy** w zależności od kontekstu, w jakim rozpatruje się badane zjawisko.

4.6 Zasada wzajemności oddziaływań

Istnieje wcale nie mała grupa zjawisk opartych na interakcji bądź wzajemnym oddziaływaniu dwóch cech. Weźmy jako przykład zgodność wieku narzeczonych w momencie ślubu. Związek taki dogodnie jest traktować jako wzajemny, gdyż dobór małżeński jest wynikiem dążeń i starań obu stron. Uproszczeniem byłoby twierdzić, że w naszej kulturze mężczyzna wybiera sobie żonę bądź że to kobieta decyduje, kto zostanie jej mężem. Gdy dobór małżeński przedstawimy w tablicy, to dla jej czytelności jest sprawą obojętną, czy mężczyzn, czy kobiety umieścimy w boczku bądź w główce.

Jako związki wzajemne traktować też należy zjawiska, w wypadku których kwestie przyczynowości czy następstwa w czasie nie mają znaczenia dla ustalenia ich istoty. W rozdziale 3 przedstawiliśmy przykład związku między odpowiedziami ma pytanie zadane w dwóch wersjach tym samym respondentom (tabela 3.8). Pozostaje faktem, że jedno z nich zostało zadane na początku, a drugie pod koniec wywiadu. Rozwiązanie takie przyjęto jednak wyłącznie ze względów technicznych. Trudno bowiem randomizować kolejność pytań w badaniu sondażowym realizowanym za pomocą papierowego kwestionariusza. Badacze zrobili przy tym, co mogli, aby ograniczyć wpływ kolejności zadawania pytań na uzyskane odpowiedzi. Wszystko po to, aby móc traktować związek między odpowiedziami na obie wersje pytania jako **wzajemny**, gdyż w przeciwnym wypadku cały eksperyment nie miałby sensu.

Komponent „wzajemności" związku cech w tablicy wyodrębnić również można wtedy, gdy opisywane zjawisko jest zależnością przyczynową. Jako przykład rozważmy tak zwaną tablicę międzypokoleniowej ruchliwości zawodowej, której wiersze obejmują kategorie zawodowe jednego z rodziców (najczęściej ojca), zaś w kolumnach wyszczególniona jest kategoria zawodowa osoby badanej. Przy czym pozycja zawodowa rodziców odnoszona jest na ogół do tej fazy życia osoby badanej, gdy rozpoczynała ona ponadobowiązkową naukę. Cel takiej prezentacji danych jest przejrzysty. Określić, w jakim stopniu pozycja rodziców wpływa na osiągnięcia dzieci. Związek jest z założenia

jednokierunkowy, gdyż stanowi zależność przyczynową. Pozycja osiągnięta przez osobę badaną nie ma żadnego wpływu na pozycję zajmowaną przez jej rodziców wiele lat wcześniej.

Spójrzmy jednak na tę tablicę od strony osób badanych. Rozważmy osoby, które zaliczono do najwyżej usytuowanej kategorii społecznej, obejmującej dyrektorów firm, urzędników piastujących najwyższe stanowiska, przedstawicieli wolnych zawodów czy wysokiej klasy specjalistów. Wiadomo, że kategoria taka zaznacza w przestrzeni społecznej swoją odrębność, której elementami są style życia czy utrzymywane kontakty i nawiązywane znajomości. Znaczna część więzi społecznych budowana jest na bazie środowiska pochodzenia. Obejmuje to rodziców, rodzeństwo, kolegów i koleżanki ze szkoły czy ze wspólnych wyjazdów. Immanentną charakterystyką każdej grupy społecznej staje się przez to jej struktura pochodzenia. Ważne jest nie tylko to, kim byli moi rodzice, lecz również to, z jakich środowisk wywodzą się osoby, wśród których przebywam. Na ile wśród moich znajomych występują osoby o odmiennych stylach myślenia w związku z tym, że wychowywały się w odmiennych środowiskach. Czy też tworzymy enklawę osób znających się „od zawsze" i dorastających w tym samym systemie wartości.

Na podstawie powyższych przykładów zaproponować można zasadę, która okaże się użyteczna w analizach tablic. Nazwijmy ją zasadą **niepomijania wzajemności oddziaływań**. Ograniczenie się do jednokierunkowej, czy przyczynowej natury zjawiska przedstawionego w tablicy może zubożyć jego obraz. W wielu sytuacjach prowadzi bowiem do pominięcia aspektów kluczowych dla rozumienia zjawiska, bądź ułatwiających prezentację jego istoty.

Proponowana zasada nie jest czymś odkrywczym. Wielu badaczy rutynowo oblicza rozkłady procentowe równocześnie dla wierszy i dla kolumn tablicy, po czym dopiero na tej podstawie stara się sformułować wnioski. O zasadzie tej warto jednak pamiętać chociażby z tego względu, że większość współcześnie proponowanych metod analitycznych – w tym również metod analizy tablic – zakłada jednokierunkowość badanych zjawisk. Jedną ze zmiennych wyodrębnia się jako tak zwaną *response variable*, zaś pozostałe traktuje jako *explanatory variables*. Konstruowane tą drogą modele statystyczne cechuje logika, przejrzystość, a niekiedy spora doza formalnej elegancji. Ich budowa przypomina zaś kształt, strukturę, czy przebieg niektórych zjawisk – gdyż to rzeczywistość stanowi źródło pomysłów dla formułowanych propozycji. Wielu badaczy proponowane modele zauroczyły na tyle, że narzuciły im sposób rozumienia rzeczywistości, którą badają (Firebaugh 2008: 207–209). Zamiast dopasować model do danych postępują odwrotnie – przypisując badanym zjawiskom atrybuty, których wymaga model.

4.7 Zastosowanie wskaźników Quételeta do analizy wzajemnych oddziaływań

W podrozdziale 4.5 pokazaliśmy użyteczność wskaźników Quételeta do odtworzenia modelu związku w sytuacjach, gdy zjawisko rozpatrywane jest jako asymetryczne. To znaczy, że cecha umieszczona w wierszach warunkuje cechę w kolumnach albo odwrotnie. Zaproponowane przez Quételeta wskaźniki można adaptować również do analizy związków wzajemnych. Najprostszym rozwiązaniem jest posłużenie się średnią obu niesymetrycznych wskaźników[4]. Pamiętajmy jednak, że natura tablicy jest multiplikatywna. Idea tabliczki mnożenia trafnie odzwierciedla to, co dzieje się w jej wnętrzu. Z tego powodu obliczając średnią dwóch współczynników Quételeta, bardziej uzasadnione jest posłużenie się formułą średniej geometrycznej, mimo że przy wyliczaniu wypadkowej dwóch wielkości na ogół korzysta się ze średniej arytmetycznej. Oznaczmy średnią geometryczną współczynników w_{ij} i w'_{ij}, czyli pierwiastek z iloczynu obu wielkości, jako q_{ij}. Korzystając z wzorów (4.6) i (4.8), otrzymamy

$$q_{ij} = \sqrt{\frac{d_{ij}}{a_i} * \frac{d_{ij}}{b_j}} = \frac{d_{ij}}{\sqrt{a_i * b_j}} \tag{4.9}$$

czyli różnica d_{ij} w rozpatrywanym polu tabeli zostaje zrelatywizowana jednocześnie do **obu liczebności brzegowych**: wiersza a_i i kolumny b_j. Ponieważ średnia geometryczna jako pierwiastek jest zawsze dodatnia, dogodnie jest przyjąć znak obliczanego wskaźnika q_{ij} zgodnie ze znakiem wskaźników składowych w_{ij} lub w'_{ij} (znaki obu wskaźników są zawsze identyczne, co wynika z wzorów 4.6 i 4.8). Oznaczenie wskaźnika literą q wzięło się z chęci zaznaczenia jego bezpośredniego związku z propozycjami Quételeta[5].

[4] Quételet przewidział potrzebę analizy związków wzajemnych proponując wskaźnik symetryczny w postaci d_{ij} / e_{ij}. Mimo że wskaźnik ten ma szereg własności przydatnych w analizie tablic (Mirkin 2001), świadomie zdecydowałem się posłużyć w tej pracy wskaźnikiem zdefiniowanym jako średnia geometryczna omawianych wcześniej wskaźników niesymetrycznych. Zadecydowała o tym bardziej naturalna interpretacja proponowanego wskaźnika w wypadku sumowania jego wielkości dla pól tablicy, którą to własność wykorzystuje się w wielu metodach analizy tablic, w tym w przedstawionej w rozdziale 7 analizie korespondencji.

[5] Wiele lat temu posłużyłem się podobnie zdefiniowanym wskaźnikiem, nazywając go *dissimilarity index* i również oznaczając literą *q* (Sawiński i Domański 1989: 22). Wtedy jednak nie byłem świadomy analogii między zastosowanym wskaźnikiem a propozycjami Quételeta. Wiedzę na ten temat zawdzięczam artykułowi Mirkina (2001). Również proponowaną wtedy nazwę *dissimilarity index* należy uznać za nieadekwatną, gdyż badaczom kojarzy się raczej z wskaźnikiem różnic profili, który omówię w podrozdziale 5.2. Myślę, że Czytelnicy są w stanie zrozumieć, iż niekiedy przychodzi wycofać się z pewnych wcześniejszych propozycji.

Nazwijmy proponowany współczynnik średnim wskaźnikiem Quételeta. Jego wartości dla związku płci ze sposobem głosowania przedstawiono w części [6] tabeli 4.4. Zgodnie z przyjętą konwencją, dla celów prezentacji zostały one przemnożone przez 100. Wielkości te leżą pomiędzy niesymetrycznymi wskaźnikami Quételeta obliczonymi dla płci (część [4] tabeli 4.4) oraz dla sposobów głosowania (część [5] tej tabeli). W sześciu polach tablicy wielkości q_{ij} niewiele odbiegają od ±5. Na przykładzie tej grupy pól najłatwiej wyjaśnić, na czym polega wypośrodkowanie odchyleń od niezależności rozpatrywanych osobno z punktu widzenia cechy umieszczonej w wierszach oraz cechy umieszczonej w kolumnach tablicy.

Rozpatrzmy pole, które obejmuje mężczyzn głosujących na PSL. Nadwyżka mężczyzn w tym polu, w stosunku do modelu niezależności, wynosi 5,98 (część [3] tabeli 4.4). Gdy rozważymy wszystkich badanych mężczyzn (493,64), to nadwyżkę tę uznać należy za niewielką, gdyż wynosi ona 1,21 procenta liczby badanych mężczyzn. Gdy jednak tę samą różnicę zrelatywizować do liczby osób, które zadeklarowały głosowanie na PSL (27,52), to jej względne znaczenie wzrasta. Z tego powodu wskaźnik Quételeta osiąga wartość 21,73 (część [5] tabeli 4.4). Obliczona wartość q_{ij}, czyli średnia geometryczna wskaźników obu Quételeta, wynosi natomiast 5,13 (część [6] tabeli 4.4). Można ją interpretować jako wypadkową obu sytuacji.

Podobną do omawianej wartość wskaźnika q_{ij}, równą 5,93, otrzymano dla mężczyzn głosujących na Samoobronę. Nadwyżka badanych mężczyzn w tym polu (12,08) okazała się około dwukrotnie większa niż nadwyżka mężczyzn głosujących na PSL, stąd przy spojrzeniu na zjawisko od strony badanych mężczyzn wartość wskaźnika Quételeta jest również dwa razy wyższa i wynosi 2,45. Głosujących na kandydatów Samoobrony było jednak w badaniu ponadtrzykrotnie więcej (84,21) niż głosujących na PSL (27,52). Nadwyżka mężczyzn, mimo że dwukrotnie większa niż w wypadku PSL, ma przez to mniejszą wagę, gdy oceniamy jej wielkość z punktu widzenia głosujących na Samoobronę. Wartość wskaźnika Quételeta wyniosła tu 14,35. Wypadkową tych wielkości jest przytoczona wyżej wartość 5,93. Podobna co do wielkości, jak w wypadku mężczyzn głosujących na PSL, natomiast kryjąca w sobie odmienny mechanizm.

Wskaźniki q_{ij} nie muszą więc trafnie charakteryzować ani jednego, ani drugiego aspektu związku przedstawionego w tablicy. Jeśli badaczowi zależy wyłącznie na jednym z aspektów, to w celu odtworzenia modelu związku powinien raczej wybrać niesymetryczną formę wskaźnika Quételeta i wyciągać wnioski zgodnie z zasadami zaproponowanymi w podrozdziale 4.5. Gdy natomiast pragnie uwzględnić oba aspekty badanego związku, czyli potraktować związek jako wzajemny – jak to ma miejsce w wypadku wieku narzeczonych

w momencie zawarcia ślubu – to posłużenie się symetryczną wersją wskaźnika wydaje się lepiej uzasadnione[6].

Dla identyfikacji wzorca związku za pomocą średnich wskaźników Quételeta użyteczne okazuje się podniesienie ich wartości do kwadratu i przemnożenie przez łączną liczbę badanych osób. Wielkości w tej postaci przedstawione zostały w części [7] tabeli 4.4. Sumują się one do wartości omawianego w 3.8 współczynnika χ^2 (chi-kwadrat)

$$\chi^2 = \sum_{i=1}^{w}\sum_{j=1}^{k} n * q_{ij}^2 = n * \sum_{i=1}^{w}\sum_{j=1}^{k} q_{ij}^2 \tag{4.10}$$

Warto to sprawdzić i przekonać się, że suma ogółem wskaźników podanych w części [7] tabeli 4.4 jest równa wartości statystyki testu niezależności chi-kwadrat obliczonej w podrozdziale 3.8.

Omawiana własność ma kapitalne znaczenie dla analizy tablic z dwóch powodów. Po pierwsze, wiąże powszechnie stosowany test niezależności chi-kwadrat z podejściem polegającym na odtworzeniu modelu tablicy. Pozwala tym samym ustalić, które z pól tablicy odpowiadają za wysoką wartość statystyki χ^2 w sytuacji, gdy hipotezę o niezależności cech w populacji należy odrzucić. Po drugie, ułatwiają rozstrzygnięcie, w których fragmentach tablicy szukać pól charakteryzujących się specyfiką wobec modelu niezależności. Kwadraty q_{ij} przemnożone przez liczbę badanych osób sumują się bowiem do χ^2 nie tylko w obrębie rozkładu łącznego obu cech w tablicy, lecz również jako sumy brzegowe **osobno** w wierszach i w kolumnach.

Aby zilustrować ostatnią z podanych własności, obliczmy udziały procentowe wielkości $n * q_{ij}^2$ w obrębie wnętrza tablicy, a także w obrębie marginesów. W części [8] tabeli 4.4 przedstawione one zostały w procentach. Sumy brzegowe dla mężczyzn i kobiet nie pomogą w identyfikacji modelu związku, gdyż obie wielkości są zrównoważone. Natomiast ogląd sum brzegowych obliczonych dla sposobów głosowania dostarcza sporo informacji. Trzy kategorie: głosowanie na Samoobronę, na PSL oraz odmowa odpowiedzi obejmują w sumie prawie 94 procent odstępstw obserwowanego związku od modelu niezależności. Tak znaczny odsetek przekona chyba każde-

[6] W pakiecie SPSS dostępny jest podobnie zdefiniowany współczynnik nazwany **resztą standaryzowaną** (Górniak i Wachnicki 2008: 136). Jego wartości dla poszczególnych pól tablicy wyrażają się wzorem $d_{ij} / \sqrt{e_{ij}}$, czyli są to wartości proponowanego tu średniego wskaźnika Quételeta przemnożone przez $\sqrt{n}$, które jest stałą dla wszystkich pól. Resztami standaryzowanymi można więc posługiwać się w sposób analogiczny do wskaźników Quételeta, zaś otrzymane wnioski – ze względu na liniową zależność między wskaźnikami – powinny okazać się identyczne. Przypomnę też, że sposobom wykonania stosownych obliczeń poświęcony jest aneks B.

go badacza, że pól specyficznych warto szukać przede wszystkim w owych trzech kolumnach. Dodatkowe uwzględnienie znaku wskaźników q_{ij} pozwala wyodrębnić dwa bloki specyficznych pól. Pierwszy obejmuje głosujących na Samoobronę i PSL (nadwyżki mężczyzn, niedobór kobiet), drugi zaś osoby, które nie odpowiedziały, na kogo głosowały (nadwyżka kobiet, niedobór mężczyzn).

Ramka 4.2
Problem trafności oceny ilościowych aspektów zjawisk za pomocą funkcji liniowych i kwadratowych

Proponowana w podrozdziale 4.7 metoda dekompozycji współczynnika χ^2 oparta jest na kwadratowej funkcji odchyleń różnic w poszczególnych polach. Funkcja kwadratowa uwypukla znaczenie dużych odchyleń, natomiast obniża znaczenie niewielkich. Nasuwa się pytanie, czy taki sposób opisu ilościowych aspektów zjawisk jest w ogóle zasadny? Stosowane wskaźniki tracą bowiem interpretację w języku opisu badanej rzeczywistości, gdyż pojęcia „więcej" lub „mniej" w realnym świecie wyrażają się w metryce liniowej.

Dominacja metod analitycznych opartych na funkcji kwadratowej ma swoje uwarunkowania historyczne. Funkcje kwadratowe w stosunku do funkcji liniowych mają szereg własności, które pozwalają na badanie ich przebiegu (na przykład wyznaczenie maksimum) metodami analizy matematycznej. Fakt ten spowodował, że przez wiele lat modele powiązań między dwiema lub większą liczbą zmiennych (na przykład model regresji, model analizy czynnikowej) konstruowano, korzystając z klasy funkcji kwadratowych.

Omawiana tradycja jest silnie zakorzeniona, toteż na co dzień wielu badaczy przechodzi do porządku dziennego nad tym, że rozwiązania oparte na funkcjach kwadratowych mogą w rzeczywistości być nieadekwatne do analizy niektórych zjawisk. Jednym z niewielu zagadnień z tego zakresu, które wywołują refleksję, jest kwestia zasadności stosowania średniej jako miary wielkości dochodów. Ponieważ średnia minimalizuje kwadraty odchyleń elementów składowych (Lissowski i in. 2008: 165-167), stąd czuła jest na wielkości skrajne. Niewielka liczba badanych o wysokich dochodach jest w stanie „przeciągnąć" średnią w swoją stronę, podczas gdy większość członków zbiorowości osiąga dochody niewielkie. Dlatego wielu badaczy decyduje się posługiwać w tym wypadku medianą, która minimalizuje liniową sumę odchyleń. Miary oparte na funkcjach liniowych nie mają jednak tak bogato opracowanego warsztatu. Większość badaczy potrafi dokonać estymacji przedziału ufności dla średniej. Natomiast rzadko kto potrafi wskazać analogiczny schemat wnioskowania dla mediany.

Możliwości obliczeniowe komputerów spowodowały, że obecnie tworzyć można metody analityczne bazujące na innych – niż funkcja kwadratowa – kryteriach oceny dopasowania modelu do danych (Wolfram 2002). Pójście w tym kierunku odwróciłoby jednak do góry nogami dotychczasową analitykę, co między innymi zmniejszyłoby możliwości zestawiania ze sobą wyników badań, zwłaszcza tych wcześniej opracowywanych. Nie mówiąc o tym, że badacze musieliby zmienić swoje przyzwyczajenia. Z wymienionych powodów przedstawione w tej książce propozycje mieszczą się w obszarze tradycyjnych metod analitycznych, opartych na paradygmacie funkcji kwadratowej. Wybór tej perspektywy bez wątpienia zubaża możliwości przełożenia stosowanych wskaźników na język opisywanych zjawisk. Należy to jednak traktować jako wybór mniejszego zła. Jedyny środek zaradczy jaki można zaproponować, to wyraźne zwrócenie uwagi na te z elementów proponowanych metod, które prowadzić mogą do interpretacji niezgodnych z przyjętymi celami.

Uzyskany tą drogą model związku pokrywa się więc z modelem stanowiącym punkt wyjścia rozważań tego rozdziału (tabela 4.1), zidentyfikowanym na podstawie prostej analizy odsetków kobiet i mężczyzn głosujących na kandydatów poszczególnych partii. Nie zmniejsza to jednak użyteczności wskaźników Quételeta z dwóch powodów.

Po pierwsze, pozostawiają one badaczowi swobodę wyboru perspektywy, z której pragnie analizować badany związek. Posłużenie się jednym z wariantów niesymetrycznej wersji współczynników Quételeta odpowiada spojrzeniu na związek z jednej strony. W omawianym przykładzie jest to spojrzenie bądź z punktu widzenia różnic w sposobach głosowania mężczyzn i kobiet, bądź kompozycji elektoratów poszczególnych partii ze względu na płeć głosujących. Gdy badacz pragnie połączyć obie perspektywy, to wtedy skorzystać może z symetrycznej wersji wskaźników Quételeta. Należy podkreślić, że ostatnia z wymienionych perspektyw, w odróżnieniu od dwóch pozostałych, nie ma swojego odpowiednika w operacji obliczania odsetków.

Po drugie, w wypadku tablic o większych rozmiarach, czy też o bardziej złożonej strukturze, analiza odsetków nie musi prowadzić do uzyskania tak klarownego modelu, jak miało to miejsce w przykładzie głosowania mężczyzn i kobiet. Możliwość skorzystania z przejrzyście zdefiniowanego kryterium specyfiki poszczególnych pól może okazać się jedyną drogą dojścia do wartościowych konkluzji.

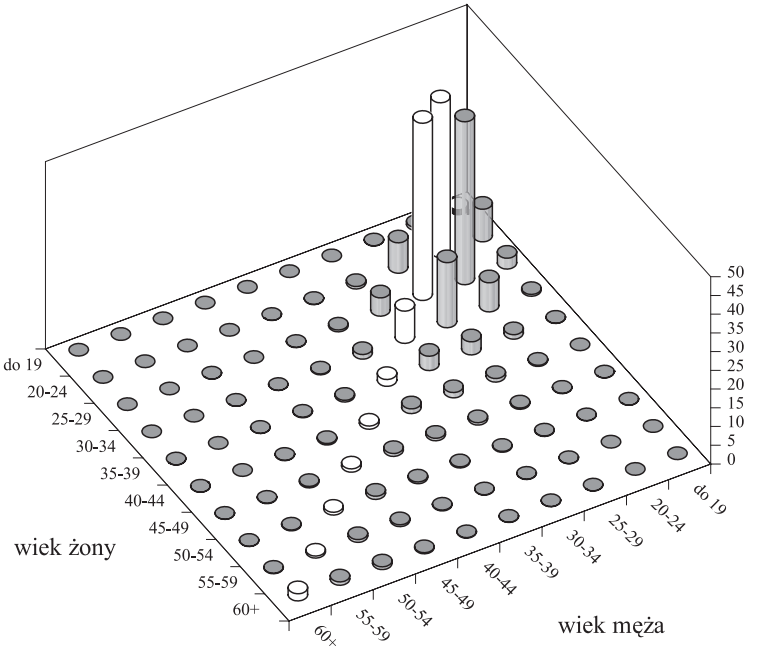
4.8 Przykład budowy modelu: homogamia małżeńska ze względu na wiek

Aby przekonać się o zaletach i ograniczeniach wskaźników Quételeta do identyfikacji modelu związku rozważmy przykład tablicy, w której zestawione zostały wiek mężczyzny i wiek kobiety w momencie zawierania małżeństwa. Przykład wybrałem z tego względu, że wbrew pozorom jest trudny do interpretacji. Przypuszczam też, że nie wszyscy Czytelnicy zaakceptują wnioski, które ostatecznie zaproponuję. Tym bardziej zachęcam do prześledzenia proponowanego toku rozumowania.

Główny Urząd Statystyczny co roku publikuje dane na temat małżeństw zawartych w roku poprzednim, w tym informacje o wieku mężczyzn i kobiet zawierających małżeństwo w podziale na 5-letnie kohorty. W części [1] tabeli 4.5 przedstawiony został wiek mężczyzn i kobiet, którzy zawarli małżeństwo w 2006 roku (GUS 2007). Dla ułatwienia przeglądania danych pola na przekątnej tablicy wyodrębnione zostały za pomocą ramek. Odpowiadają one sytuacjom, gdy wiek męża i żony jest zgodny z dokładnością do przyjętych

przez GUS przedziałów wieku. Tego rodzaju zgodność miała miejsce w wypadku prawie połowy małżeństw zawartych w 2006 roku (46,5%). Biorąc to pod uwagę, można sądzić, że istotę badanego związku wyznacza homogamia małżonków pod względem wieku. Wniosek taki potwierdza graficzny obraz rozkładu liczebności (rycina 4.1). Na rycinie tej wysokości walców proporcjonalne są do liczby małżeństw w poszczególnych polach tablicy. Walce odpowiadające polom leżącym na przekątnej oznaczono kolorem białym.

Rycina 4.1
Obraz rozkładu liczby małżeństw zawartych w 2006 roku sklasyfikowanych ze względu na wiek żony i męża



Aby ocenić zasadność wniosku, że istotę związku wyznacza homogamia ze względu na wiek, przyjrzyjmy się dokładniej poszczególnym wielkościom w polach tablicy. Analizę rozpocznijmy od profili w wierszach i w kolumnach, prezentowanych w częściach [2] i [3] tabeli 4.5. Obliczone odsetki nie do końca potwierdzają wniosek o pełnej homogamii małżeństw pod względem wieku. Rozważmy profile wieku żon w kategoriach wieku mężów. Dla ułatwienia ich analizy maksymalny odsetek w każdym wierszu zaznaczony został szarym ko-

lorem. W kategorii mężczyzn do 19 lat największy odsetek, bo aż 60, wybiera partnerkę z tej samej kategorii wieku. Podobnie jest wśród mężczyzn w dwóch kolejnych grupach wiekowych: 20–24 i 25–29 lat. Natomiast wśród mężczyzn starszych tak rozumiana zasada homogamii przestaje obowiązywać. Mężczyźni w wieku 35–39 lat wybierają partnerki młodsze (30–34 lata) ponad dwa razy częściej niż kobiety należące do tej samej grupy wiekowej. W wypadku mężczyzn w wieku 40–44 lata różnica jest jeszcze większa, bo jako kandydatki na żonę wybiera się najczęściej kobiety średnio o 10 lat młodsze. W starszych grupach wiekowych omawiane rozbieżności ulegają zmniejszeniu. Homogamia wraca zaś w najstarszej rozpatrywanej grupie wiekowej – mężczyzn powyżej 60. roku życia. Prawie połowa z nich wybiera żonę należącą do tej samej grupy wiekowej.

W wypadku kobiet obraz jest bardziej spójny (część [3] tabeli 4.5). Poza kobietami należącymi do dwóch najmłodszych kohort pozostałe zawierają małżeństwo najczęściej z mężczyzną należącym do tej samej grupy wiekowej.

Omawiane profile przekładają się na wartości niesymetrycznych wskaźników Quételeta, które obliczone zostały osobno w wierszach i w kolumnach rozpatrywanej tablicy (części [4] i [5] tabeli 4.5). Do wartości tych jeszcze wrócimy, natomiast analizę rozpoczniemy od wartości średnich obu wskaźników (część [6] tabeli 4.5). Niektóre z pól tej tablicy zostały wyodrębnione za pomocą różnych odcieni tła. Pola, w których wskaźnik przybiera wartość wyższą niż 40, oznaczono kolorem czarnym. Kolorem ciemno szarym oznaczono pola zawierające wskaźniki o wielkościach od 10 do 39. Jasno szare pola zawierają pozostałe wielkości dodatnie, zaś pola, w których wskaźnik przybiera wartość ujemną, pozostawiono nie zaznaczone.

Wprowadzone oznaczenia pól ułatwiają ocenę, w których obszarach tablicy średnie wskaźniki Quételeta przybierają najwyższe wartości. Na pierwszy rzut oka zastanawia – czy wręcz zaskakuje – dlaczego otrzymany obraz związku jest aż tak rozbieżny z obrazem uzyskanym dla liczebności. Szczególnie jest to widoczne, gdy średnie wskaźniki Quételeta zobrazujemy graficznie w taki sam sposób, jak na rycinie 4.1 zobrazowane zostały liczebności tablicy. Tego rodzaju prezentację wskaźników Quételeta przedstawia rycina 4.2.

Tabela 4.5
Wskaźniki dla pól tablicy przedstawiającej wiek męża oraz wiek żony dla małżeństw zawartych w 2006 roku

[1] liczba małżeństw (w tysiącach)

	wiek żony										
wiek męża	do 19	20–24	25–29	30–34	35–39	40–44	45–49	50–54	55–59	60 lub więcej	ogółem
do 19	1,3	0,8	0,1	0,0	0,0	0,0	0,0	0,0	0,0	0,0	2,1
20–24	8,7	41,3	8,1	0,7	0,1	0,0	0,0	0,0	0,0	0,0	58,9
25–29	2,9	43,5	47,3	4,9	0,5	0,1	0,0	0,0	0,0	0,0	99,2
30–34	0,5	7,8	17,4	8,6	1,2	0,3	0,1	0,0	0,0	0,0	35,8
35–39	0,1	1,2	3,5	3,8	1,8	0,5	0,2	0,1	0,0	0,0	11,1
40–44	0,0	0,3	1,0	1,4	1,3	0,8	0,4	0,1	0,0	0,0	5,3
45–49	0,0	0,1	0,4	0,6	0,7	0,8	0,9	0,4	0,1	0,0	4,0
50–54	0,0	0,0	0,2	0,3	0,3	0,5	0,9	0,8	0,2	0,1	3,2
55–59	0,0	0,0	0,1	0,1	0,1	0,2	0,5	0,7	0,5	0,1	2,4
60 lub więcej	0,0	0,0	0,0	0,0	0,1	0,1	0,3	0,7	0,9	1,9	4,1
ogółem	13,5	95,1	77,9	20,5	6,1	3,2	3,2	2,8	1,8	2,2	226,3

[2] profile wieku żony w kategoriach wieku męża (w procentach)

	wiek żony										
wiek męża	do 19	20–24	25–29	30–34	35–39	40–44	45–49	50–54	55–59	60 lub więcej	ogółem
do 19	60,2	36,2	3,2	0,4	0,0	0,0	0,0	0,0	0,0	0,0	100,0
20–24	14,8	70,2	13,8	1,1	0,1	0,0	0,0	0,0	0,0	0,0	100,0
25–29	2,9	43,9	47,7	4,9	0,5	0,1	0,0	0,0	0,0	0,0	100,0
30–34	1,3	21,7	48,5	24,1	3,4	0,7	0,1	0,0	0,0	0,0	100,0
35–39	0,7	10,8	31,3	34,5	16,4	4,2	1,6	0,4	0,1	0,1	100,0
40–44	0,4	5,7	18,2	26,5	23,4	15,7	7,5	2,1	0,5	0,0	100,0
45–49	0,3	2,9	9,3	15,9	16,7	19,8	21,9	9,7	2,9	0,5	100,0
50–54	0,2	1,3	5,0	8,0	10,3	14,3	27,3	24,3	7,0	2,3	100,0
55–59	0,1	0,6	2,2	3,9	4,8	8,9	21,0	30,6	22,0	5,9	100,0
60 lub więcej	0,1	0,4	0,4	1,2	1,3	3,3	7,7	17,1	21,8	46,7	100,0
ogółem	6,0	42,0	34,4	9,0	2,7	1,4	1,4	1,2	0,8	1,0	100,0

[3] profile wieku męża w kategoriach wieku żony (w procentach)

	wiek żony										
wiek męża	do 19	20–24	25–29	30–34	35–39	40–44	45–49	50–54	55–59	60 lub więcej	ogółem
do 19	9,5	0,8	0,1	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,9
20–24	64,6	43,5	10,4	3,2	1,2	0,3	0,1	0,0	0,1	0,0	26,0
25–29	21,5	45,8	60,7	23,9	8,5	2,3	0,5	0,1	0,2	0,0	43,9
30–34	3,6	8,2	22,3	42,1	20,3	7,8	1,7	0,5	0,2	0,1	15,8
35–39	0,6	1,3	4,5	18,8	30,2	14,4	5,4	1,8	0,4	0,3	4,9
40–44	0,1	0,3	1,2	6,9	20,6	26,0	12,4	4,0	1,6	0,0	2,4
45–49	0,1	0,1	0,5	3,1	10,9	24,5	27,3	13,9	6,3	1,0	1,8
50–54	0,0	0,0	0,2	1,2	5,4	14,0	26,9	27,6	12,1	3,4	1,4
55–59	0,0	0,0	0,1	0,5	1,9	6,6	15,8	26,5	29,2	6,5	1,1
60 lub więcej	0,0	0,0	0,0	0,2	0,9	4,2	9,9	25,5	49,9	88,6	1,8
ogółem	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0

Tabela 4.5 (kontynuacja)

[4] wskaźniki Quételeta w_{ij} w kategoriach wieku męża[a]

	wiek żony										
wiek męża	do 19	20–24	25–29	30–34	35–39	40–44	45–49	50–54	55–59	60 lub więcej	ogółem
do 19	54,2	-5,9	-31,2	-8,7	-2,6	-1,4	-1,4	-1,2	-0,8	-1,0	0,0
20–24	8,8	28,1	-20,7	-7,9	-2,6	-1,4	-1,4	-1,2	-0,8	-1,0	0,0
25–29	-3,0	1,8	13,2	-4,1	-2,2	-1,4	-1,4	-1,2	-0,8	-1,0	0,0
30–34	-4,6	-20,3	14,1	15,0	0,8	-0,7	-1,3	-1,2	-0,8	-1,0	0,0
35–39	-5,3	-31,3	-3,1	25,4	13,8	2,7	0,1	-0,8	-0,7	-0,9	0,0
40–44	-5,6	-36,3	-16,3	17,5	20,7	14,3	6,1	0,9	-0,3	-0,9	0,0
45–49	-5,7	-39,2	-25,1	6,9	14,0	18,4	20,5	8,5	2,1	-0,4	0,0
50–54	-5,8	-40,7	-29,5	-1,0	7,6	12,9	25,9	23,1	6,1	1,4	0,0
55–59	-5,9	-41,4	-32,2	-5,2	2,2	7,4	19,6	29,3	21,2	5,0	0,0
60 lub więcej	-5,9	-41,6	-34,1	-7,9	-1,4	1,8	6,3	15,9	21,0	45,8	0,0

[5] wskaźniki Quételeta w'_{ij} w kategoriach wieku żony[a]

	wiek żony									
wiek męża	do 19	20–24	25–29	30–34	35–39	40–44	45–49	50–54	55–59	60 lub więcej
do 19	8,6	-0,1	-0,9	-0,9	-0,9	-0,9	-0,9	-0,9	-0,9	-0,9
20–24	38,6	17,4	-15,6	-22,8	-24,8	-25,8	-25,9	-26,0	-26,0	-26,0
25–29	-22,4	1,9	16,9	-19,9	-35,4	-41,6	-43,4	-43,7	-43,7	-43,9
30–34	-12,2	-7,6	6,5	26,3	4,5	-8,0	-14,2	-15,3	-15,7	-15,7
35–39	-4,4	-3,7	-0,4	13,8	25,3	9,5	0,5	-3,1	-4,5	-4,7
40–44	-2,2	-2,0	-1,1	4,6	18,3	23,6	10,1	1,7	-0,8	-2,3
45–49	-1,7	-1,6	-1,3	1,3	9,2	22,7	25,5	12,2	4,6	-0,8
50–54	-1,4	-1,4	-1,2	-0,2	4,0	12,6	25,5	26,2	10,7	2,0
55–59	-1,1	-1,1	-1,0	-0,6	0,9	5,6	14,7	25,4	28,2	5,5
60 lub więcej	-1,8	-1,8	-1,8	-1,6	-0,9	2,4	8,1	23,6	48,0	86,8
ogółem	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0

[6] średnie wskaźniki Quételeta q_{ij} [a]

	wiek żony									
wiek męża	do 19	20–24	25–29	30–34	35–39	40–44	45–49	50–54	55–59	60 lub więcej
do 19	21,5	-0,9	-5,2	-2,8	-1,6	-1,2	-1,2	-1,1	-0,9	-1,0
20–24	18,4	22,1	-18,0	-13,4	-8,0	-6,0	-6,0	-5,7	-4,6	-5,0
25–29	-8,3	1,9	14,9	-9,1	-8,7	-7,5	-7,8	-7,3	-5,9	-6,5
30–34	-7,5	-12,4	9,5	19,9	1,8	-2,4	-4,2	-4,3	-3,5	-3,9
35–39	-4,8	-10,7	-1,2	18,8	18,6	5,1	0,3	-1,6	-1,8	-2,1
40–44	-3,5	-8,6	-4,3	8,9	19,5	18,3	7,8	1,2	-0,4	-1,5
45–49	-3,1	-8,0	-5,7	3,0	11,3	20,4	22,9	10,2	3,1	-0,6
50–54	-2,8	-7,4	-5,9	-0,4	5,5	12,8	25,7	24,6	8,1	1,7
55–59	-2,5	-6,6	-5,7	-1,8	1,4	6,4	17,0	27,3	24,4	5,2
60 lub więcej	-3,3	-8,7	-7,9	-3,5	-1,1	2,1	7,1	19,4	31,8	63,0

Tabela 4.5 (kontynuacja)

[7] różnice $d_{ij} = n_{ij} - e_{ij}$ (w tysiącach osób)

	wiek żony										
wiek męża	do 19	20–24	25–29	30–34	35–39	40–44	45–49	50–54	55–59	60 lub więcej	
do 19	1,2	-0,1	-0,7	-0,2	-0,1	-0,0	-0,0	-0,0	-0,0	-0,0	0,0
20–24	5,2	16,6	-12,2	-4,7	-1,5	-0,8	-0,8	-0,7	-0,5	-0,6	0,0
25–29	-3,0	1,8	13,1	-4,1	-2,1	-1,3	-1,4	-1,2	-0,8	-1,0	0,0
30–34	-1,7	-7,3	5,0	5,4	0,3	-0,3	-0,5	-0,4	-0,3	-0,3	0,0
35–39	-0,6	-3,5	-0,3	2,8	1,5	0,3	0,0	-0,1	-0,1	-0,1	0,0
40–44	-0,3	-1,9	-0,9	0,9	1,1	0,8	0,3	0,0	-0,0	-0,1	0,0
45–49	-0,2	-1,6	-1,0	0,3	0,6	0,7	0,8	0,3	0,1	-0,0	0,0
50–54	-0,2	-1,3	-0,9	-0,0	0,2	0,4	0,8	0,7	0,2	0,0	0,0
55–59	-0,1	-1,0	-0,8	-0,1	0,1	0,2	0,5	0,7	0,5	0,1	0,0
60 lub więcej	-0,2	-1,7	-1,4	-0,3	-0,1	0,1	0,3	0,7	0,9	1,9	0,0
ogółem	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0

[8] indeksy $i_{ij} = 100 * n_{ij} / e_{ij}$

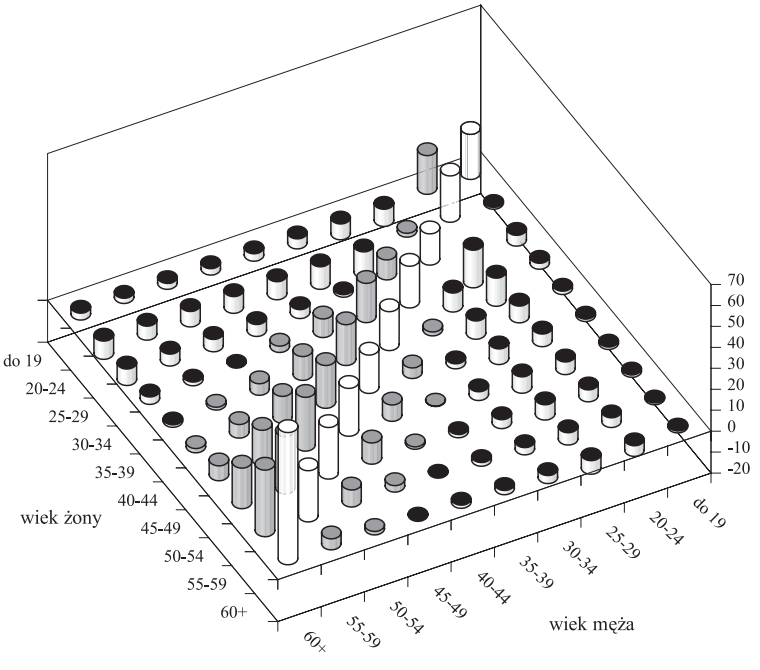
	wiek żony									
wiek męża	do 19	20–24	25–29	30–34	35–39	40–44	45–49	50–54	55–59	60 lub więcej
do 19	1010	86	9	4	2	0	0	0	0	0
20–24	248	167	40	12	5	1	0	0	0	0
25–29	49	104	138	55	19	5	1	0	0	0
30–34	23	52	141	266	128	49	10	3	1	1
35–39	11	26	91	381	613	292	110	36	9	6
40–44	6	14	53	293	873	1099	527	170	68	2
45–49	5	7	27	176	622	1390	1549	791	360	55
50–54	3	3	14	89	384	1003	1924	1975	867	242
55–59	1	1	6	43	181	621	1482	2482	2740	614
60 lub więcej	1	1	1	13	50	228	542	1392	2724	4842

[9] udziały kwadratów średnich wskaźników Quételeta uq^2_{ij} (w procentach)

	wiek żony										
wiek męża	do 19	20–24	25–29	30–34	35–39	40–44	45–49	50–54	55–59	60 lub więcej	ogółem
do 19	3,0	0,0	0,2	0,0	0,0	0,0	0,0	0,0	0,0	0,0	3,2
20–24	2,2	3,1	2,1	1,2	0,4	0,2	0,2	0,2	0,1	0,2	9,9
25–29	0,4	0,0	1,4	0,5	0,5	0,4	0,4	0,3	0,2	0,3	4,5
30–34	0,4	1,0	0,6	2,5	0,0	0,0	0,1	0,1	0,1	0,1	4,9
35–39	0,1	0,7	0,0	2,3	2,2	0,2	0,0	0,0	0,0	0,0	5,6
40–44	0,1	0,5	0,1	0,5	2,4	2,1	0,4	0,0	0,0	0,0	6,2
45–49	0,1	0,4	0,2	0,1	0,8	2,7	3,3	0,7	0,1	0,0	8,3
50–54	0,0	0,4	0,2	0,0	0,2	1,0	4,2	3,9	0,4	0,0	10,4
55–59	0,0	0,3	0,2	0,0	0,0	0,3	1,8	4,8	3,8	0,2	11,4
60 lub więcej	0,1	0,5	0,4	0,1	0,0	0,0	0,3	2,4	6,5	25,4	35,6
ogółem	6,4	6,9	5,4	7,2	6,6	7,0	10,9	12,4	11,2	26,2	100,0

[a] *wielkości przemnożone przez 100.*

Rycina 4.2
Graficzna prezentacja średnich wskaźników Quételeta dla kombinacji wieku męża i żony wśród małżeństw zawartych w 2006 roku



Spróbujmy znaleźć powody, dla których obrazy uzyskane za pomocą graficznej projekcji liczebności oraz projekcji wskaźników Quételeta są aż tak rozbieżne. Rozpocznijmy od pola odpowiadającego małżeństwom, w których kobieta i mężczyzna mają ponad 60 lat. W 2006 roku małżeństw takich było 1,9 tysiąca. Kobiet, które w tym wieku zdecydowały się na zawarcie związku małżeńskiego było 2,2 tysiąca. Stanowiły one 1,0 procent wszystkich kobiet zawierających małżeństwo w 2006 roku. Gdyby wiek męża i wiek żony stanowiły cechy niezależne, to jedynie 1 procent mężczyzn w wieku ponad 60 lat zawierałby związek małżeński z kobietą w tym wieku. Faktycznie odsetek ten wynosi 46,7 procenta. Różnica odsetków, równa w zaokrągleniu 45,8, stanowi więc nadwyżkę mężczyzn w rozpatrywanym wieku, którzy zawarli związek małżeński z kobietą mającą 60 lub więcej lat. Jest to zarazem wartość niesymetrycznego wskaźnika Quételeta w_{ij} (tabela 4.5, część [4]). Gdy małżeństwa w omawianym polu rozpatrzymy z punktu widzenia kobiet, to otrzymamy jeszcze wyższą wartość tego wskaźnika. Mężczyźni w wieku ponad 60 lat stanowili 1,8 procenta

wszystkich mężczyzn, którzy zdecydowali się na zawarcie małżeństwa w 2006 roku, natomiast wśród mężów kobiet z tej grupy wiekowej było ich aż 88,6 procenta. Różnica obu odsetków daje wartość wskaźnika w'_{ij} równą 86,8 (tabela 4.5, część [5]). Czyli, prawie 87 procent kobiet w wieku ponad 60 lat, które zdecydowały się na małżeństwo w 2006 roku, wybrała partnera w tym samym wieku w stosunku do możliwości stwarzanych przez rynek małżeński. Współczynnik będący złożeniem obu niesymetrycznych wskaźników Quételeta, czyli 63,0, odzwierciedla wysoką skłonność do zawierania małżeństw z osobą w wieku ponad 60 lat zarówno wśród mężczyzn, jak i kobiet z tej grupy wiekowej.

Dla porównania rozpatrzmy mężczyzn w wieku 25–29 lat. Wśród nich 47,7 procenta wybrało sobie żonę w tej samej grupie wiekowej. To więcej od odsetka mężczyzn w wieku ponad 60 lat, którzy zawarli małżeństwo z kobietą w tym samym wieku. W wypadku mężczyzn w wieku 25–29 lat przytoczony odsetek nie wydaje się jednak szczególnie duży, jeśli uwzględnić fakt, że aż 34,4 procenta wszystkich kobiet, które zdecydowały się wyjść za mąż w 2006 roku, jest w grupie wiekowej 25–29 lat. Nawet gdyby w urzędzie stanu cywilnego małżonkę przydzielano losowo, spośród wszystkich kobiet, które zgłosiły chęć zawarcia małżeństwa, to i tak mężczyzna w wieku 25–29 lat miałby aż 34,4 procent szans na to, że otrzyma partnerkę w tym wieku. Skłonność mężczyzn w wieku 25–29 lat do wyboru żony z tej samej grupy wiekowej należy więc ocenić jako umiarkowaną. Wskaźnik Quételeta wynosi w tym wypadku 13,2, zaś wskaźnik średni dla grupy 25–29 lat, uwzględniający skłonność kobiet do wyboru męża w ramach tej grupy wiekowej, wynosi 14,9. W analogiczny sposób wyjaśnić można, dlaczego średnie wskaźniki Quételeta w polach odpowiadających pozostałym kategoriom osób młodszych są niższe od wskaźników w najwyższych kategoriach wieku.

Mam poczucie, że części osób nie zadowala przedstawione wyjaśnienie. W naturze ludzkiej leży bowiem przekonanie, że zjawiska należy interpretować uwzględniając ich rzeczywiste rozmiary, czyli w tym wypadku uwzględniając liczbę zawieranych małżeństw. Perspektywie tej odpowiadają różnice między faktyczną liczbą małżeństw a modelem niezależności przedstawione w części [7] tabeli 4.5. Kolorem ciemno szarym oznaczono pola, w których różnice te przekraczają tysiąc małżeństw. Największe różnice skupiają się w obszarze tabeli odpowiadającemu małżeństwom zawieranym przez ludzi młodych. Małżeństw, gdy mąż i żona są w wieku 20–24 lata, jest o 16,6 tysiąca więcej, niż gdyby urząd stanu cywilnego łączył ludzi w pary w sposób losowy. Analogiczna nadwyżka dla małżeństw między osobami należącymi do najstarszej kohorty wynosi zaledwie 1,9 tysiąca.

W tym miejscu nie zamierzam nikogo przekonywać, że na zjawiska warto patrzeć nie tylko z perspektywy liczby osób, które im podlegają, lecz również

ilościowych relacji między różnymi kategoriami osób. Każdy ma bowiem prawo uznać, że dwóch respondentów określonej płci stanowiących nadwyżkę wśród głosujących na PiS wnosi do rozumienia wpływu płci na wyniki wyborów tyle samo, co nadwyżka również dwóch osób, lecz głosujących na LPR (podrozdział 4.3). Umówmy się więc następująco. Jeśli Czytelnik ma poczucie, że analiza różnic wystarcza do uzyskania adekwatnego obrazu badanego zjawiska, to niech dalszą część tego podrozdziału potraktuje jako źródło argumentów przeciwko zestawianiu ze sobą relacji między wielkościami w tablicy. Jeśli natomiast uważa, że stosunki między wielkościami również warto wziąć pod uwagę, to... niech wtedy potraktuje dalsze rozważania jako **ostrzeżenie** przed nadmiernym zaufaniem do tego rodzaju wskaźników.

W części [8] tabeli 4.5 przedstawione zostały wartości indeksów, czyli miary wyrażającej relacje liczebności obserwowanych do liczebności modelu niezależności. Wartość tego współczynnika dla pola tablicy odpowiadającego małżeństwom między osobami w wieku powyżej 60 lat wyniosła aż 4842! Świadczy to, że małżeństwa takie zawierane są ponad 48 razy częściej, niż gdyby współmałżonka losowano spośród wszystkich chętnych. Jest to niewątpliwie bardzo silna homogamia. Na ile jednak silniejsza od homogamii małżeństw między osobami w wieku 25–29, dla której to kategorii indeks wynosi zaledwie 141. Czy można wskazać sposób porównania indeksów dla obu pól, który prowadziłby do wartościowych konkluzji dotyczących siły homogamii w obu kategoriach małżeństw?

Pomijając nawet to, czy indeksy dla pól tablicy można zasadnie ze sobą porównywać, pozostaje niewątpliwie faktem, że ich zastosowanie prowadzi do zasadniczo odmiennego obrazu analizowanego związku niż zastosowanie różnic. Układy ciemnoszarych plam w tablicach [7] i [8] rozmijają się ze sobą. Różnice powielają układ liczebności, które skupione są w lewym górnym narożniku tablicy, gdyż większość małżeństw zawierają ludzie młodzi. Z kolei najwyższe wartości indeksów skupione są w prawym dolnym narożniku, czyli w obszarze małżeństw zawieranych przez ludzi w starszym wieku. Być może w obszarze tym jest wyższa homogamia, lecz z drugiej strony małżeństw tych jest stosunkowo niewiele. Na który z obrazów związku należy się więc zdecydować? Czy kierować się rozmiarami zjawiska czy też jego intensywnością?

Wyjście stanowić może posłużenie się wskaźnikami Quételeta. W podrozdziale 4.7 starałem się wykazać, że stanowią one niejako wypośrodkowanie między spojrzeniem na zjawiska z perspektywy indeksów i różnic, gdyż zostały zdefiniowane jako różnica zrelatywizowana do marginesów obu cech. Rozkład ciemnoszarych pól w części [6] tabeli 4.5 stanowi pewien kompromis między obrazami otrzymanymi za pomocą różnic i indeksów. Niemniej jednak warto zauważyć, że obraz ten ciąży raczej ku małżeństwom osób starszych.

Dowodzi tego chociażby to, że małżeństwom między osobami w wieku ponad 60 lat odpowiada największa wartość średniego wskaźnika Quételeta, równa 63,0. Jest to dużo więcej niż w dowolnym z pozostałych pól tablicy. Z drugiej strony, rozpatrywane pole zawiera zaledwie 2 procent wszystkich małżeństw. Dla porównania, dla małżeństw między osobami w wieku 20–24 lata średni wskaźnik Quételeta wynosi 22,1, zaś małżeństwa te stanowią 21 procent wszystkich rozpatrywanych. Nie jest więc łatwo podjąć decyzji w sytuacji, gdy relacje między wielkościami liczbowymi mającymi pomóc znaleźć model związku kłócą się z intuicjami dotyczącymi jego rozmiarów.

Warto ponadto uwzględnić fakt, że w większości współcześnie stosowanych metod analitycznych używa się funkcji kwadratowych do oceny wagi obserwowanych zróżnicowań (zob. ramka 4.2). Dzieje się tak również w wypadku metod analizy tablic. Na funkcjach kwadratowych oparta jest konstrukcja testu chi-kwadrat, formuła na współczynnik korelacji, modelowanie log-liniowe, czy wiele innych metod – co obejmuje także metody przedstawione w dalszych rozdziałach tej książki. Istota działania kwadratowych funkcji oceny zróżnicowań polega na tym, że nadają one znaczenie przede wszystkim największym ze zróżnicowań, praktycznie ignorując zróżnicowania niewielkie.

Ilustrację niebezpieczeństw z tym związanych stanowi analiza kwadratów średnich wskaźników Quételeta. W podrozdziale 4.7 podałem, że po przemnożeniu tych kwadratów przez liczbę badanych jednostek sumują się one do współczynnika chi-kwadrat, co pozwala zdekomponować wartość tego współczynnika między poszczególne pola tablicy. Dekompozycja taka przedstawiona została w części [9] tabeli 4.5. Wynika z niej, że ponad jedna czwarta związku wieku męża i żony sprowadza się do pojedynczego pola tablicy, które odpowiada małżeństwom zawieranym między osobami ponad 60-letnimi. Z marginesu kategorii wieku mężów odczytać ponadto można, że specyfika wieku żon mężczyzn 60-letnich lub starszych absorbuje aż 35,6 procenta całości zróżnicowań obserwowanych w tablicy. Ma to miejsce pomimo tego, że kategoria ta obejmuje zaledwie 1,8 procenta wszystkich mężczyzn biorących ślub w 2006 roku. Konsekwencje tej sytuacji – upraszczając nieco – wyrazić można następująco. Małżeńskie preferencje mężczyzn w wieku ponad 60 lat, stanowiących 1,8 procenta badanej zbiorowości, decydują w ponad jednej trzeciej o wartości współczynnika chi-kwadrat, o wielkości współczynnika korelacji, czy też o wynikach uzyskanych za pomocą jakiejkolwiek innej metody opartej na kwadratowej funkcji oceny rozbieżności między badanym związkiem a sytuacją niezależności.

Jeśli badacz uzna, że otrzymane tą drogą wyniki deformują obraz badanego związku, to powinien podjąć jakieś środki zaradcze. Przed stosowaniem metod opartych na funkcjach kwadratowych uciec się raczej nie uda. Pozostaje więc

zmodyfikowanie samego przedmiotu analizy, czyli tablicy. Kategoria „60 lub więcej lat" różni się od pozostałych tym, że obejmuje większy zakres roczników demograficznych. Nie można więc wykluczyć, że obserwowana w ramach tej kategorii wysoka homogamia jest pozorna. Gdyby bowiem rozdzielić ją na kategorie 5-letnie, to być może w ramach każdej z nich homogamia nie byłaby wyższa, niż wśród małżeństw zawieranych przez osoby poniżej 60. roku życia. Jednakże GUS udostępnia dane na temat wieku małżonków jedynie w rozpatrywanej postaci. Podziału kategorii „60 lub więcej lat" na 5-letnie podkategorie nie uda się więc dokonać. Jedyne co pozostaje, to **pominięcie** tej kategorii w analizach, poprzez usunięcie odpowiadającego jej wiersza i kolumny z tablicy. Odpowiada to **rezygnacji** z analizy pewnych aspektów zjawiska, którym to zjawiskiem jest zgodność wieku mężczyzny i kobiety zawierających małżeństwo. Mówiąc inaczej, wzory zawierania małżeństw przez osoby w wieku ponad 60 lat są na tyle specyficzne, że zaburzają obraz zjawiska w pozostałych kategoriach. Wymagają przez to odrębnej analizy.

Po podjęciu powyższej decyzji spróbujmy na podstawie tak okrojonej tablicy zidentyfikować model badanego zjawiska. Proponuję skorzystać w tym celu ze średnich wskaźników Quételeta, aczkolwiek warto mieć przed oczami również tablicę różnic. Ułatwia ona uwzględnienie ilościowych aspektów zjawiska. Nie bez znaczenia jest też tablica udziałów kwadratów średnich wskaźników Quételeta. Zdaje ona sprawę z tego, które z elementów badanego związku najbardziej będą ważyły na wynikach uzyskanych w fazie ewentualnych dalszych analiz wyodrębnionego modelu – na przykład podczas weryfikacji stopnia jego spójności z układem liczebności w tabeli. Wszystkie trzy tablice przedstawione zostały w tabeli 4.6. Tak jak poprzednio, różnymi odcieniami szarości zaznaczono pola o największych wartościach wskaźników.

Wstępny ogląd konfiguracji zaszarzonych pól prowadzi do wniosku, że po usunięciu kategorii osób w wieku ponad 60 lat obraz związku zasadniczo się nie zmienił. Nadal największe wartości wskaźników rozkładają się wzdłuż przekątnej tablicy. Wśród osób w starszym wieku homogamia wydaje się silniejsza, lecz zarazem bardziej rozmyta, co objawia się tym, że pola w kolorze szarym rozlewają się na większy obszar. Wśród młodych małżeństw zgodność wieku jest bardziej skoncentrowana wokół przekątnej. Potwierdza to przedstawiona w części [3] tabeli 4.6 dekompozycja wskaźnika chi-kwadrat między poszczególne pola. Sumy brzegowe są wyższe dla małżeństw w starszym wieku, patrząc na zjawisko zarówno ze strony męża, jak i żony. Największa suma brzegowa – równa 24,7 procent – odpowiada mężczyznom w wieku 55–59 lat. Specyfika tej kategorii absorbuje aż jedną czwartą badanego związku – lecz nad faktem tym wypada przejść do porządku dziennego. Przedział wiekowy tej kategorii wyodrębniony został analogicznie tak jak pozostałe przedziały

Tabela 4.6
Wskaźniki dla pól tablicy przedstawiającej wiek męża oraz wiek żony dla małżeństw zawartych w 2006 roku (bez osób w wieku 60 lub więcej lat)

[1] średnie wskaźniki Quételeta q_{ij} [a]

	wiek żony								
wiek męża	do 19	20–24	25–29	30–34	35–39	40–44	45–49	50–54	55–59
do 19	21,5	-1,0	-5,3	-2,8	-1,6	-1,2	-1,1	-0,9	-0,6
20–24	18,2	21,5	-18,6	-13,7	-8,1	-6,0	-5,8	-5,0	-3,3
25–29	-8,6	1,0	14,2	-9,4	-8,9	-7,5	-7,5	-6,4	-4,2
30–34	-7,7	-13,0	9,1	19,7	1,8	-2,3	-4,1	-3,7	-2,5
35–39	-4,9	-11,0	-1,4	18,7	18,7	5,3	0,5	-1,1	-1,2
40–44	-3,6	-8,8	-4,4	8,9	19,5	18,8	8,4	1,9	0,3
45–49	-3,1	-8,2	-5,8	3,0	11,4	21,0	24,3	12,2	5,2
50–54	-2,8	-7,5	-6,0	-0,4	5,6	13,3	27,5	29,2	12,4
55–59	-2,5	-6,5	-5,6	-1,7	1,5	6,9	18,6	33,0	36,2

objaśnienia: ponad 20 | od 10 do 20 | od 0 do 10

[2] różnice $d_{ij} = n_{ij} - e_{ij}$ (w tysiącach osób)

	wiek żony									
wiek męża	do 19	20–24	25–29	30–34	35–39	40–44	45–49	50–54	55–59	**ogółem**
do 19	1,2	-0,1	-0,7	-0,2	-0,1	-0,0	-0,0	-0,0	-0,0	**0,0**
20–24	5,1	16,1	-12,6	-4,8	-1,5	-0,8	-0,8	-0,6	-0,2	**0,0**
25–29	-3,1	1,0	12,5	-4,2	-2,2	-1,3	-1,3	-0,9	-0,4	**0,0**
30–34	-1,7	-7,6	4,8	5,3	0,3	-0,2	-0,4	-0,3	-0,1	**0,0**
35–39	-0,6	-3,6	-0,4	2,8	1,5	0,3	0,0	-0,1	-0,0	**0,0**
40–44	-0,3	-2,0	-0,9	0,9	1,1	0,8	0,3	0,1	0,0	**0,0**
45–49	-0,2	-1,6	-1,0	0,3	0,6	0,7	0,8	0,4	0,1	**0,0**
50–54	-0,2	-1,3	-0,9	-0,0	0,2	0,4	0,8	0,7	0,2	**0,0**
55–59	-0,1	-1,0	-0,7	-0,1	0,1	0,2	0,5	0,7	0,5	**0,0**
ogółem	**0,0**	**0,0**	**0,0**	**0,0**	**0,0**	**0,0**	**0,0**	**0,0**	**0,0**	**0,0**

objaśnienia: ponad 10 | od 1 do 10 | od 0 do 1

[3] udziały kwadratów średnich wskaźników Quételeta q_{ij}^2 (w procentach)

	wiek żony									
wiek męża	do 19	20–24	25–29	30–34	35–39	40–44	45–49	50–54	55–59	**ogółem**
do 19	4,0	0,0	0,2	0,1	0,0	0,0	0,0	0,0	0,0	**4,3**
20–24	2,8	4,0	3,0	1,6	0,6	0,3	0,3	0,2	0,1	**12,9**
25–29	0,6	0,0	1,7	0,8	0,7	0,5	0,5	0,4	0,2	**5,3**
30–34	0,5	1,4	0,7	3,3	0,0	0,0	0,1	0,1	0,1	**6,4**
35–39	0,2	1,0	0,0	3,0	3,0	0,2	0,0	0,0	0,0	**7,5**
40–44	0,1	0,7	0,2	0,7	3,3	3,0	0,6	0,0	0,0	**8,6**
45–49	0,1	0,6	0,3	0,1	1,1	3,8	5,1	1,3	0,2	**12,5**
50–54	0,1	0,5	0,3	0,0	0,3	1,5	6,5	7,3	1,3	**17,8**
55–59	0,1	0,4	0,3	0,0	0,0	0,4	3,0	9,4	11,3	**24,7**
ogółem	**8,5**	**8,5**	**6,7**	**9,5**	**9,0**	**9,8**	**16,1**	**18,7**	**13,1**	**100,0**

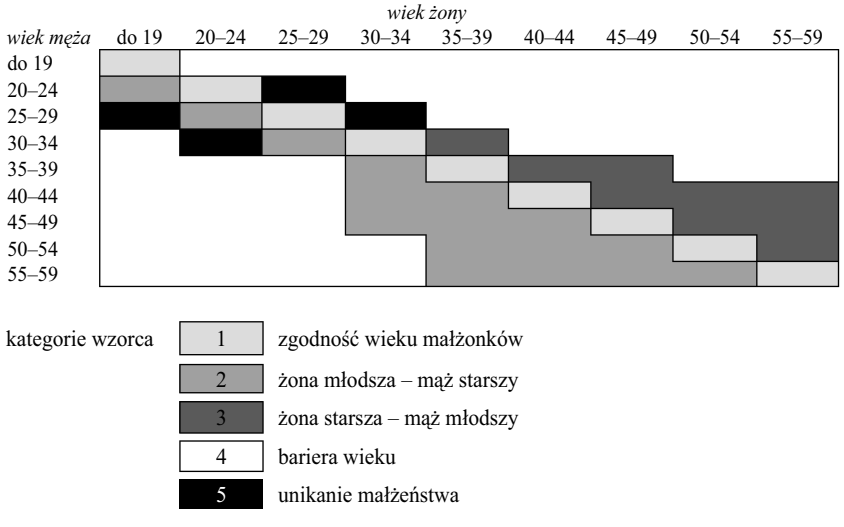
objaśnienia: ponad 10 | od 3 do 10 | od 1 do 3

[a] *wielkości przemnożone przez 100.*

w tablicy, stąd też nie ma podstaw do ewentualnego usunięcia tej kategorii, tak jak postąpiliśmy z osobami w wieku ponad 60 lat.

Jak już zdecydowaliśmy, do znalezienia modelu związku posłużymy się średnimi wskaźnikami Quételeta (część [1] tabeli 4.6). Z niewielkimi wyjątkami najwyższe wartości tych wskaźników leżą na przekątnej tablicy. Przyjmijmy więc, że pola przekątnej stanowić będą pierwszą wyodrębnioną kategorię modelu. Jej interpretacja jest oczywista. Kategoria ta obejmuje małżeństwa, w których wiek mężczyzny i kobiety jest zgodny. Na rycinie 4.3 wyodrębniona kategoria oznaczona została polami w kolorze jasnoszarym.

Rycina 4.3
Model związku wieku męża i żony wśród małżeństw zawartych w 2006 roku



Sprawdźmy teraz, co dzieje się pod przekątną. Pola w tym obszarze odpowiadają sytuacji, gdy żona jest młodsza od męża. Warto zwrócić uwagę na to, że każde pole przekątnej (poza kategorią „do 19 lat") posiada pole o dodatniej wartości wskaźnika Quételeta po swojej lewej stronie. Są to małżeństwa, w których wiek żony zaklasyfikowany jest o jedną kohortę niżej od wieku męża. Wystąpienia nadwyżek w tych polach można było się spodziewać uwzględniając znany fakt, że mężczyzna zawierający małżeństwo jest częściej starszy od kobiety niż odwrotnie. W wypadku małżeństw zawartych w 2006 roku przeciętna różnica wieku mężczyzny i kobiety wynosiła nieco ponad 2 lata. Zakres wieku kobiety decydującej się na małżeństwo z mężczyzną od niej starszym jest przy tym niejednakowy w różnych fragmentach tabeli. W ostat-

nim wierszu, obejmującym mężczyzn w wieku 55–59 lat, na lewo od pola na przekątnej występują 4 pola o dodatnich wartościach wskaźników Quételeta. W wierszach odpowiadających mężczyznom w wieku 50–54 lata i 45–49 lat pól takich jest już 3, w kolejnym wierszu 2, zaś poczynając od kategorii 35–39 lat i rozpatrując kolejne kategorie coraz młodszych mężczyzn stwierdzamy tylko jedno pole tego rodzaju. Obserwacje te pozwalają wyodrębnić drugą kategorię modelu rozpatrywanego zjawiska, obejmującą model małżeństw, w których żona jest młodsza, zaś mąż starszy. Kategoria obejmie wyłącznie te pola, w których można mówić o istnieniu **skłonności** do tego typu wyborów. Odpowiada to dodatnim różnicom między obserwowanymi liczebnościami a modelem niezależności lub – co jest równoważne – dodatnim wartościom wskaźników Quételeta. Pola zaliczone do wyodrębnionej kategorii zostały na rycinie 4.3 zaznaczone kolorem nieco ciemniejszym niż pola odpowiadające homogamii małżeńskiej.

W analogiczny sposób wyodrębnić można trzecią kategorię modelu, obejmującą modele małżeństw, w których żona jest starsza od męża. Pola składające się na tę kategorię zakreślono na rycinie 4.3 kolorem ciemnoszarym. Kategoria ta obejmuje mniejszą liczbę pól niż poprzednia, gdyż małżeństwa tego typu zdarzają się rzadziej od małżeństw, w których mąż jest starszy od żony. Niemniej jednak, w niektórych polach ponad przekątną tablicy obserwujemy nadwyżki małżeństw w stosunku do modelu niezależności. Dowodzi to skłonności do zawierania również tego typu małżeństw a zarazem uzasadnia wyodrębnienie ich w postaci osobnej kategorii modelu.

Wyodrębnienie powyższej kategorii wyczerpało pola tablicy, w których wskaźniki Quételeta miały dodatnie wartości. Analiza pól zawierających wartości ujemne nie prowadzi na ogół do tak wartościowych wniosków, toteż można je wszystkie potraktować jako jedną wspólną kategorię. Na rycinie 4.3 pola te pozostawiono nie zaznaczone. Ujemne wartości wskaźników Quételeta świadczą, że każdemu z tych pól odpowiadają czynniki niesprzyjające zawieraniu małżeństw między osobami w danym wieku. Nie wnikając w ich podłoże, nazwijmy tę kategorię wzorca „barierą wieku".

Przed zaliczeniem wszystkich pól o ujemnych wskaźnikach do wspólnej kategorii warto jednakże sprawdzić, czy wśród nich nie ma pól wyróżniających się szczególnie niskimi wartościami ujemnymi. Wystąpienie tego rodzaju pól może bowiem świadczyć o istnieniu dodatkowych barier, które nie sprzyjają zawieraniu małżeństw między mężczyznami i kobietami o pewnych konfiguracjach wieku. Uważne prześledzenie wskaźników Quételeta w tabeli 4.6 pozwala zauważyć, że w jednym z pól rozpatrywany wskaźnik przybiera szczególnie niską wartość. Pole to odpowiada małżeństwom, w których mężczyzna ma 20–24 lata, zaś kobieta 25–29 lat. W starszych grupach wiekowych

w polach leżących bezpośrednio nad przekątną wskaźniki mają na ogół dodatnie wartości. Odpowiadają one bowiem małżeństwom, w których żona należy do grupy wiekowej o jeden poziom starszej niż mąż. W rozpatrywanym polu mamy natomiast do czynienia z czymś odwrotnym. Małżeństw tego rodzaju jest mniej, niż należałoby oczekiwać zgodnie z modelem niezależności. Z tabeli zawierającej różnice wynika, że niedobór wynosi aż 12,6 tysiąca takich małżeństw. Jest to druga co do wielkości różnica w tabeli, gdy uwzględnimy zarówno różnice dodatnie, jak i ujemne. Działa tu więc specyficzny mechanizm, ograniczający liczbę małżeństw zawieranych między mężczyznami w wieku 20–24 a kobietami w wieku 25–29 lat.

Warto sprawdzić, czy nie jest to symptomem prawidłowości, która ujawnia się w większej liczbie pól tablicy. Prześledzenie kolejnych wierszy prowadzi do wniosku, że jeśli podobna prawidłowość występuje, to tylko w wypadku mężczyzn w wieku 25–29 lat. Wskaźnik Quételeta dla ich małżeństw z kobietami z kolejnej kohorty (30–34 lata) jest bowiem niski (-9,4), zaś wielkość niedoboru małżeństw całkiem spora (4,2 tysiąca).

Analogicznej analizy wymagają kolumny tablicy, gdyż rozpatrywany związek jest wzajemny. Kobiety w wieku do 19 lat są najbardziej skłonne zawrzeć małżeństwo z mężczyzną w tym samym wieku, lecz nie odrzucają kandydatów nieco starszych (20–24 lata). Bariera pojawia się natomiast w wypadku kolejnej kategorii mężczyzn, w wieku 25–29 lat, o czym świadczy niska wartość wskaźnika Quételeta (-8,6) oraz niedobór 3,1 tysiąca małżeństw. Podobne zjawisko ma miejsce w wypadku kobiet w wieku 20–24 lata. Dopuszczają one małżeństwa z kandydatami z sąsiedniej kategorii wiekowej (25–29 lat), natomiast niechętnie wychodzą za mąż za mężczyzn w wieku 30–34 lata. Wśród kobiet z kolejnych kohort wiekowych tak wyraźne bariery nie zarysowują się.

Poczynione obserwacje skłaniają do wyodrębnienia jeszcze jednej kategorii modelu rozpatrywanego związku. Obejmuje ona cztery pola, w których małżeństw jest wyraźnie mniej w porównaniu do sąsiednich pól tablicy, w których obserwuje się skłonność do zawierania małżeństw. Wyodrębnioną kategorię nazwijmy „unikaniem małżeństwa". Na rycinie 4.3 odpowiadające jej pola zaznaczone zostały kolorem czarnym.

Ustalenie modelu związku stanowi na ogół etap wstępny, poprzedzający podjęcie próby wyjaśnienia badanego zjawiska. Warto mieć to na uwadze, gdyż potrzeby związane z wyjaśnianiem zjawiska mają wpływ na liczbę i stopień szczegółowości wyodrębnionych kategorii. Niekiedy wystarcza, gdy model obejmuje jedynie najważniejsze aspekty badanego związku. W innych sytuacjach niezbędna jest bardziej szczegółowa analiza relacji między poszczególnymi polami tablicy. Jeśli dane pochodzą z badania reprezentacyjnego, to decydując się na określony stopień szczegółowości, należy uwzględnić wielkość próby.

Wyodrębnianie kategorii na podstawie niewielkich liczebności jest ryzykowne, gdyż mogą one nie mieć swoich odpowiedników w badanej populacji, a także mogą nie replikować się w dalszych badaniach tego samego zjawiska.

4.9 Ustalenie stopnia dopasowania modelu do danych

Po znalezieniu modelu badanego zjawiska badacz na ogół stawia sobie pytanie, czy należycie odzwierciedla on przedstawione w tablicy dane. Mówiąc inaczej, na ile precyzyjnie **odtworzyć** można otrzymane w badaniu liczebności wnętrza tablicy posługując się modelem związku. Narzędzi pozwalających na rozstrzygnięcie tej kwestii dostarczają metody modelowania log-liniowego. Przy czym samej metody modelowania log-liniowego, jako techniki analizy tablic, nie będę w tym miejscu prezentować, gdyż wykraczałoby to poza ramy książki (zob. Bishop, Fienberg i Holland 1975; Domański i Przybysz 2007; Treiman 2009). Ograniczę się wyłącznie do omówienia tych elementów metody, które przydatne są do oceny modelu związku.

Przyjętą na gruncie modelowania log-liniowego zasadę odtwarzania liczebności w polach tablicy przedstawia następujący wzór

$$\hat{F}_{ij} = \alpha_i \beta_j \gamma_{ij} \eta \tag{4.11}$$

w którym estymowane liczebności pól tablicy $\hat{F}_{ij}$ są funkcją czterech czynników:

- α_i efektu wiersza;
- β_j efektu kolumny;
- γ_{ij} interakcji, czyli wartości specyficznej dla każdego pola tablicy, która pozwala dopasować estymowaną liczebność do faktycznej względem sytuacji, w której działałyby jedynie efekty wiersza i kolumny;
- η efektu głównego, czyli jednakowej dla wszystkich pól stałej, która pozwala dopasować estymacje do faktycznych liczebności[7]

Jako funkcję dopasowania testowanego modelu do danych przyjąć można wartość statystyki χ^2 obliczoną według wzoru

$$\chi^2 = \sum_{i=1}^{w} \sum_{j=1}^{k} \frac{(n_{ij} - \hat{F}_{ij})^2}{\hat{F}_{ij}} \tag{4.12}$$

Jest to ta sama statystyka, której używa się, testując hipotezę o niezależności cech w populacji (podrozdział 3.10). Różnica sprowadza się do tego, że

[7] Potrzeba dołączenia tego czynnika bierze się stąd, że na wartości pozostałych czynników nakłada się pewne warunki normalizujące ich wielkości. Na przykład, iloczyn wszystkich czynników α_i przyjmuje się jako 1.

w teście niezależności otrzymane w wyniku badania liczebności n_{ij} porównuje się z liczebnościami e_{ij} estymowanymi na mocy modelu niezależności. W modelowaniu log-liniowym liczebności n_{ij} zestawiane są natomiast z liczebnościami $\hat{F}_{ij}$ estymowanymi na mocy utworzonego modelu związku.

Jako punkt odniesienia dla budowy modelu związku przyjmuje się niezależność cech w tablicy. W modelu tym interakcje γ_{ij} dla wszystkich pól są równe 1. Estymowane liczebności są wtedy wprost proporcjonalne do efektów α_i dla wierszy i efektów β_j dla kolumn

$$\hat{F}_{ij} = \alpha_i \, \beta_j \, \eta \tag{4.13}$$

Testowany model buduje się, dodając do powyższego modelu parametry interakcji γ_{ij} w określonych polach tablicy. Dodanie parametru interakcji oznacza, że w danym polu dopuszcza się odchylenie obserwowanej liczebności od niezależności. Gdyby parametry interakcji dodać do wszystkich pól tablicy, to wtedy model byłby dokładnie dopasowany do otrzymanych w badaniu liczebności n_{ij}. Model taki nazywany jest **nasyconym** i nie ma on żadnej użyteczności poznawczej. Cała sztuka polega bowiem na tym, aby związek opisać za pomocą jak najmniejszej liczby parametrów interakcji, dobierając je zarazem w taki sposób, aby stopień dopasowania liczebności modelu do liczebności obserwowanych okazał się zadowalający. Im wyjaśnienie zjawiska wymaga mniejszej liczby czynników, tym uznawane jest za bardziej wartościowe.

Technikę modelowania log-liniowego można stosować do oceny stopnia, w jakim znaleziony model pozwala wyjaśnić, czy odtworzyć liczebności pól tablicy uzyskane w badaniu. Niekiedy mówi się też o **dopasowaniu** modelu do otrzymanych w badaniu liczebności. W przykładzie płci i sposobu głosowania model związku sprowadzał się do wyodrębnienia trzech kategorii badanych osób. Przypomnijmy, że pierwsza obejmowała respondentów głosujących na Samoobronę lub PSL, co częściej cechowało mężczyzn niż kobiety. Druga kategoria obejmowała osoby, które nie określiły na kogo oddały głos. Postawy takie charakterystyczne były dla kobiet. Ostatnia kategoria objęła pozostałe sposoby głosowania. Jak stwierdziliśmy uprzednio, liczebności uzyskane w badaniu nie odbiegały w wypadku pól tej kategorii od modelu niezależności.

Aby ocenić metodą modelowania log-liniowego stopień dopasowania do danych omawianego modelu, wystarczy wprowadzić dwa parametry interakcyjne. Pierwszy, oznaczmy go γ_1, byłby wspólny dla dwóch pól tablicy obejmujących mężczyzn głosujących na Samoobronę lub PSL. Jak łatwo zauważyć, wartość tego parametru musi być większa od 1, gdyż służy on zwiększeniu liczebności modelu w tych polach ponad to, co wynikałoby z niezależności (zob. wzór 4.11). Drugi z kolei parametr testowanego modelu, oznaczony jako γ_2, byłby specyficzny dla kobiet, które nie odpowiedziały na kogo oddały głos.

Ramka 4.3
Leo A. Goodman: sylwetka twórcy.

Leo A. Goodman jest profesorem statystyki i socjologii na Uniwersytecie Kalifornijskim w Berkeley. Uznawany jest za jednego z czołowych twórców metod analizy danych kategorialnych współcześnie stosowanych w badaniach. Dorobek Goodmana jest imponujący i obejmuje ponad 150 artykułów, z których większość ukazała się w czasopismach z najwyższej półki. Goodman publikuje nie tylko w tytułach o profilu statystycznym, lecz również socjologicznym (między innymi: *American Journal of Sociology, American Sociological Review, Contemporary Sociology, Social Science Quarterly, Research in Social Stratification and Mobility, Sociological Methodology*). Dzięki temu jego propozycje znane są badaczom w naukach społecznych.

Leo A. Goodman urodził się w 1928 roku w Nowym Jorku. W 1950 roku, a więc już w wieku 22 lat, uzyskał w Princeton stopień doktora w dziedzinie statystyki matematycznej. W latach pięćdziesiątych rozpoczął współpracę z amerykańskim matematykiem i statystykiem Williamem Henry Kruskalem (1919–2005). Artykuły będące owocem tej współpracy porządkowały wcześniejszą wiedzę na temat metod analizy tablic dwuzmiennowych oraz zawierały szereg innowacyjnych propozycji (Goodman i Kruskal 1954, 1959, 1963, 1972). Należy do nich model interpretacji siły zależności między cechami w kategoriach skuteczności przewidywania, który przedstawiam w podrozdziale 3.11.

W latach 1950–1986 Goodman związany był z Uniwersytetem w Chicago, zajmując się głównie metodami analizy tablic dwu- i wielozmiennowych za pomocą modeli log-liniowych i logitowych. W tej dziedzinie jego wkład do nauki uważa się za fundamentalny (Agresti 2002: 627). Goodman kładł szczególny nacisk na elastyczność metod należących do tej grupy a także na szerokie możliwości ich stosowania do rozwiązywania zagadnień pojawiających się w praktyce badawczej. Jego ważniejsze artykuły z tego okresu zostały zebrane i opublikowane w postaci osobnych książek (Goodman 1978, 1984). Za sprawą Goodmana uniwersytet w Chicago stał się jednym z wiodących na świecie ośrodków rozwoju metod analizy danych kategorialnych. Do jego doktorantów bądź współpracowników należeli między innymi Shelby Haberman, Clifford Clogg, Zvi Gilula czy wymieniony już wcześniej William Kruskal. Do wkładu samego Goodmana, jak też wymienionych osób, odwołuję się w wielu miejscach książki.

W 1996 roku Goodman sformułował propozycję, która była swoistym ukoronowaniem prawie 50 lat dociekań na temat prawidłowości rządzących zależnościami między cechami kategorialnymi. Propozycję tę stanowił uniwersalny model analityczny. Objął on nie tylko metody, w rozwój których wkład Goodmana należy uznać jako istotny, lecz również niektóre z metod wywodzących się z odmiennych podejść czy stylów myślenia. Do tych ostatnich należy analiza korespondencji (omawiana w rozdziale 7 tej książki), która ma rodowód francuski i na dobrą sprawę zaczęto ją stosować dopiero na przełomie lat osiemdziesiątych i dziewięćdziesiątych. Zaproponowany przez Goodmana model uświadomił badaczom, że większość współczesnych podejść do analizy tablic ma wspólne matematyczne podstawy. Użyteczność poszczególnych metod sprowadza się zaś do umiejętności ich twórczego wykorzystania w rozwiązywaniu zagadnień praktycznych.

Leo A. Goodman jest członkiem Amerykańskiej Akademii Nauk, a także towarzystw naukowych w dziedzinie statystyki i socjologii, w których przez lata pełnił różne funkcje. Między innymi w latach 1969–1992 był członkiem sekcji metodologicznej American Sociological Association, zaś w latach 1974–1975 jej przewodniczył.

W 2003 roku Leo A. Goodman otrzymał doktorat *honoris causa* Uniwersytetu Michigan. Fakt ten jest o tyle znaczący, że będący częścią tego uniwersytetu Institute for Social Research w Ann Arbor od wielu lat stanowi jeden z wiodących na świecie ośrodków w dziedzinie rozwoju metodologii badań społecznych.

Zauważmy, że w obu wypadkach ma potrzeby definiowania osobnych parametrów dla pól odpowiadających drugiej płci. Liczebności modelu dla mężczyzn i kobiet sumują się bowiem do liczebności brzegowej. Na przykład, niedobór mężczyzn odmawiających odpowiedzi na pytanie o sposób głosowania wynika z nadwyżki kobiet w tej kategorii.

Trzecia kategoria, która obejmuje osoby o pozostałych sposobach głosowania, nie wymaga wprowadzenia osobnego parametru interakcji. W wypadku tych osób zakładamy bowiem, że ich zachowania są zgodne z modelem niezależności. Liczebności przewidywane w polach tablicy opisuje więc równanie (4.13).

Przyjęty sposób parametryzacji wzorca przedstawiony został w części [1] tabeli 4.7. Znając wartości tych parametrów obliczyć można estymowaną liczebność w każdym z pól tablicy. Na przykład, liczebność w polu odpowiadającym mężczyznom głosującym na Samoobronę zgodnie z modelem związku powinna być równa iloczynowi $\alpha_4\beta_2\gamma_1\eta$. Opracowane na gruncie modelowania log-liniowego procedury analityczne pozwalają oszacować wartości parametrów w sposób minimalizujący kryterium (4.12). Odpowiada to estymacji parametrów modelu w taki sposób, aby przewidywane liczebności $\hat{F}_{ij}$ były jak najlepiej dopasowane do faktycznych liczebności n_{ij} uzyskanych w badaniu.

W części [2] tabeli 4.7 podano wielkości parametrów modelu oszacowane za pomocą programu LEM[8] (Vermunt 1997). Wartości parametrów brzegowych α i β odzwierciedlają liczbę osób w kategoriach obu cech. Dla uświadomienia sobie sposobu ich działania można raz jeszcze odwołać się do tabliczki mnożenia, gdyż porównanie takie okazuje się użyteczne przy wyjaśnianiu istoty każdego modelu multiplikatywnego, a do takich należy model log-liniowy. Wartość parametru α dla PiS jest większa niż dla PO, gdyż więcej respondentów głosowało na kandydatów pierwszej z wymienionych partii. Należy więc oczekiwać, że liczba osób głosujących na PiS będzie proporcjonalnie większa od liczby osób głosujących na PO zarówno wśród kobiet, jak i wśród mężczyzn. Potwierdzają to estymacje przedstawione w części [3] tabeli 4.7. Parametry γ, mimo że nie odpowiadają wyborowi wiersza i kolumny w tabliczce mnożenia, interpretować można w analogiczny sposób. Parametr γ_1 mówi, że mężczyzn głosujących na Samoobronę lub PSL należy spodziewać się około 2 razy więcej (1,97), niż wynikałoby z częstości głosowania na Samoobronę lub PSL wśród wszystkich badanych oraz z udziału mężczyzn w próbie.

[8] Program ten jest bezpłatnie dostępny w sieci. Opis sposobu wykorzystania programu LEM do estymacji parametrów modeli log-liniowych wraz z licznymi przykładami zastosowań znaleźć można w pracy Domańskiego i Przybysza (2007). Procedury estymacji parametrów modeli log-liniowych dostępne są też w większości pakietów statystycznych.

Parametr γ_2 informuje natomiast, że kobiet, które nie odpowiedziały na kogo oddały głos, powinno być o prawie połowę więcej, niż należałoby oczekiwać w sytuacji, gdyby reakcja na zadane w badaniu pytanie zależna była jedynie od efektu wiersza i kolumny.

Różnice między liczebnościami uzyskanymi w badaniu oraz liczebnościami estymowanymi za pomocą modelu związku przedstawiono w części [4] tabeli 4.7. W wypadku dwóch pól tablicy obejmujących osoby, które nie odpowiedziały na pytanie na kogo głosowały, model dokładnie dopasował esty-

Tabela 4.7
Parametry modelu log-liniowego wzorca tablicy przedstawiającej sposób głosowania kobiet i mężczyzn w wyborach we wrześniu 2005 roku
Europejski Sondaż Społeczny 2006

	sposób głosowania								
płeć	PiS	PO	SLD	Samo-obrona	PSL	LPR	pozostałe partie	odmowa lub nie pamięta	ogółem
[1] Sposób parametryzacji tablicy[a]									
kobiety								γ_2	α_1
mężczyźni				γ_1	γ_1				α_2
ogółem	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	η
[2] Estymowane wartości parametrów									
kobiety								1,49	1,08
mężczyźni				1,97	1,97				0,92
ogółem	5,83	4,03	1,23	0,84	0,27	0,30	0,33	1,52	34,69
[3] Estymowane liczebności $\hat{F}_{ij}$ modelu									
kobiety	219,42	151,83	46,23	31,45	10,28	11,25	12,52	85,22	568,20
mężczyźni	186,44	129,01	39,28	52,76	17,24	9,56	10,64	48,72	493,64
ogółem	405,85	280,83	85,51	84,21	27,52	20,81	23,15	133,94	1061,83
[4] Różnice liczebności obserwowanych n_{ij} i estymowanych $\hat{F}_{ij}$									
kobiety	-4,71	1,36	0,63	1,53	-1,53	1,80	0,93	0,00	0,00
mężczyźni	4,71	-1,35	-0,63	-1,53	1,53	-1,80	-0,93	0,00	0,00
ogółem	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00

Liczebności uzyskane w badaniu, liczebności modelu niezależności stochastycznej oraz różnice tych liczebności przedstawiono w tabeli 4.4. Wszystkie wielkości podano w zaokrągleniu do dwóch cyfr po przecinku.

[a] *Liczba faktycznie estymowanych parametrów jest mniejsza o dwa od liczby wyszczególnionych w tablicy. Jeden z parametrów* α *oraz jeden z parametrów* β *mogą być bowiem obliczone w oparciu o wartości pozostałych.*

mowane liczebności do uzyskanych w badaniu. Dzieje się tak zawsze, gdy wprowadzony parametr interakcyjny odpowiada pojedynczemu polu tabeli. W wypadku kategorii osób głosujących na Samoobronę lub PSL dopasowanie nie jest pełne co wynika z faktu, że pojedynczy parametr odpowiada więcej niż jednemu polu (założono ten sam efekt w wypadku głosowania na Samoobronę i PSL). Przy czym zestawienie różnic w tych polach z różnicami między obserwowanymi liczebnościami a modelem niezależności (część [3] tabeli 4.4) dowodzi, że uwzględnienie testowanego modelu związku znacząco poprawiło przewidywanie liczebności uzyskanych w badaniu.

Wartość statystyki χ^2 obliczonej według wzoru (4.13) dla testowanego modelu wyniosła 1,52. Wartość ta jest nieistotna statystycznie dla poziomów istotności <0,90 (przy 5 stopniach swobody). Wyniki badania nie upoważniają więc do odrzucenia hipotezy zerowej, w myśl której w badanej populacji kształt zjawiska jest zgodny z testowanym modelem. Proponowany model trafnie więc opisuje specyfikę głosowania kobiet i mężczyzn.

Do oceny stopnia odtwarzalności wyników badania przez znaleziony model korzysta się też z dwóch innych wskaźników. Pierwszy z nich określa poziom redukcji wielkości statystyki χ^2 w stosunku do modelu niezależności. Wielkość chi-kwadrat dla modelu niezależności obliczyliśmy w podrozdziale 3.8 testując hipotezę, że płeć i sposób głosowania są w populacji niezależne. Wyniosła ona 18,84. Dla testowanego modelu wartość χ^2 wynosi 1,52, czyli zmniejszyła się aż o 92 procent w stosunku do pierwotnej wielkości. Odsetek ten interpretować można jako stopień, w jakim za pomocą modelu związku wyjaśnić można obserwowane w tablicy odstępstwa od niezależności.

Drugi wskaźnik dotyczy odsetka badanych osób, które można poprawnie zaklasyfikować do pól rozpatrywanej tablicy na podstawie znajomości modelu związku. W części [4] tabeli 4.7 podano dla każdego z pól liczbę osób, których nie udało się poprawnie zaklasyfikować. Suma różnic dodatnich wynosi w zaokrągleniu 12 osób, co stanowi 1,2 procenta wszystkich badanych[9]. Oznacza to, że znając podaną w tabeli 4.7 parametryzację modelu, do właściwych pól tablicy zaklasyfikować można aż 98,8 procenta wszystkich badanych. To wystarczająco dużo aby uznać, że znaleziony model należycie opisuje badane zjawisko.

Modelowanie log-liniowe zalicza się do metod wnioskowania statystycznego, gdyż zasadniczym celem metody jest rozstrzyganie hipotez o zgodności modelu z populacją. Atrakcyjność koncepcji tworzenia modeli według określonych założeń a także klarowna interpretacja stosowanych wskaźników po-

[9] Wielkość ta odpowiada *minimalnej liczbie przemieszczeń* (podrozdział 4.3) zrelatywizowanej do liczby badanych osób.

wodują jednak, że modelowanie log-liniowe bywa również stosowane w badaniach obejmujących całą populację. Użyteczność metody do oceny modelu odtworzonego na podstawie wyników badania wyczerpującego przedstawić można na przykładzie wieku mężczyzn i kobiet zawierających małżeństwa. W podrozdziale 4.8 zaproponowaliśmy, aby model tego związku sprowadzić do pięciu kategorii (rycina 4.3). W tabeli 4.8 podana została parametryzacja tego modelu, oszacowane wielkości parametrów modelu, a także różnice między rzeczywistymi liczebnościami a liczebnościami estymowanymi na podstawie modelu.

Wartość χ^2 dla modelu odpowiadającego proponowanemu modelowi wynosi 16 489, podczas gdy analogicznie obliczona wartość – lecz przy przyjęciu jako punktu odniesienia modelu niezależności – jest równa 258 400. Stopień redukcji współczynnika χ^2 wynosi więc 93,6 procenta. Zarazem proponowany model pozwala sklasyfikować poprawnie do poszczególnych pól tablicy 94,6 procenta małżeństw. Jest to sporo, lecz wyraźnie mniej niż w wypadku uprzednio analizowanego związku płci ze sposobem głosowania. Co więcej, dla rozstrzygnięcia, czy proponowany model jest dostatecznie dobrze dopasowany do danych, nie można posłużyć się w tym wypadku testem statystycznym, gdyż model obejmuje wszystkie małżeństwa zawarte w 2006 roku. Na jakiej podstawie podjąć w takim razie decyzję, czy model odzwierciedla liczebności empiryczne w akceptowalnym stopniu?

Tabela 4.8
Parametry modelu log-liniowego modelu tablicy przedstawiającej wiek męża oraz wiek żony dla małżeństw zawartych w 2006 roku (bez osób w wieku 60 lub więcej lat)

[1] sposób parametryzacji tablicy [a]

	wiek żony									
wiek męża	do 19	20–24	25–29	30–34	35–39	40–44	45–49	50–54	55–59	ogółem
do 19	γ_1									α_1
20–24	γ_2	γ_1	γ_5							α_2
25–29	γ_5	γ_2	γ_1	γ_5						α_3
30–34		γ_5	γ_2	γ_1	γ_3					α_4
35–39				γ_2	γ_1	γ_3	γ_3			α_5
40–44				γ_2	γ_2	γ_1	γ_3	γ_3	γ_3	α_6
45–49				γ_2	γ_2	γ_2	γ_1	γ_3	γ_3	α_7
50–54					γ_2	γ_2	γ_2	γ_1	γ_3	α_8
55–59					γ_2	γ_2	γ_2	γ_2	γ_1	α_9
ogółem	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	η

kategorie wzorca: γ_1 zgodność wieku; γ_2 żona młodsza mąż starszy; γ_3 żona starsza mąż młodszy; γ_5 unikanie małżeństwa

Tabela 4.8 (kontynuacja)

[2] Estymowane wartości parametrów

	wiek żony									
wiek męża	do 19	20–24	25–29	30–34	35–39	40–44	45–49	50–54	55–59	ogółem
do 19	19,67									0,39
20–24	14,49	19,67	4,97							2,79
25–29	4,97	14,49	19,67	4,97						3,91
30–34		4,97	14,49	19,67	5,94					2,05
35–39				14,49	19,67	5,94	5,94			1,69
40–44				14,49	14,49	19,67	5,94	5,94	5,94	0,71
45–49				14,49	14,49	14,49	19,67	5,94	5,94	0,49
50–54					14,49	14,49	14,49	19,67	5,94	0,52
55–59					14,49	14,49	14,49	14,49	19,67	0,37
ogółem	1,44	5,93	4,91	1,52	0,57	0,46	0,49	0,49	0,25	127,24

kategorie wzorca: zgodność wieku; żona młodsza mąż starszy; żona starsza mąż młodszy; unikanie małżeństwa

[3] Różnice liczebności obserwowanych n_{ij} i estymowanych $\hat{F}_{ij}$ (w tysiącach)

	wiek żony									
wiek męża	do 19	20–24	25–29	30–34	35–39	40–44	45–49	50–54	55–59	ogółem
do 19	-0,12	0,48	-0,17	-0,07	-0,03	-0,02	-0,02	-0,02	-0,01	0,00
20–24	1,27	-0,13	-0,55	0,12	-0,13	-0,15	-0,17	-0,17	-0,09	0,00
25–29	0,67	0,75	-0,71	1,15	0,23	-0,15	-0,23	-0,24	-0,12	0,00
30–34	0,11	0,08	-1,23	0,82	0,35	0,13	-0,08	-0,11	-0,06	0,00
35–39	-0,23	-0,08	2,44	-0,88	-0,58	-0,12	-0,46	-0,05	-0,05	0,00
40–44	-0,11	-0,23	0,52	-0,58	0,50	0,02	0,13	-0,15	-0,11	0,00
45–49	-0,08	-0,26	0,06	-0,74	0,15	0,38	0,27	0,21	0,02	0,00
50–54	-0,09	-0,35	-0,17	0,15	-0,22	0,02	0,39	0,14	0,12	0,00
55–59	-0,07	-0,27	-0,18	0,02	-0,28	-0,10	0,17	0,40	0,30	0,00
ogółem	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00

kategorie wzorca: zgodność wieku; żona młodsza mąż starszy; żona starsza mąż młodszy; unikanie małżeństwa

Liczebności pól oryginalnej tabeli, liczebności modelu niezależności stochastycznej oraz różnice tych liczebności przedstawiono w tabeli 4.5. Wszystkie wielkości podano w zaokrągleniu do dwóch cyfr po przecinku.

[a] Liczba faktycznie estymowanych parametrów jest mniejsza o dwa od liczby wyszczególnionych w tablicy. Jeden z parametrów α *oraz jeden z parametrów* β *mogą być bowiem obliczone w oparciu o wartości pozostałych.*

Proponuję, aby kierować się przede wszystkim przydatnością modelu do wyjaśnienia badanego zjawiska. Jeśli bowiem zachodziłaby taka potrzeba, to stopień dopasowania modelu do danych zawsze można zwiększyć. Rozważmy wielkości niewyjaśnionych różnic przedstawione w części [3] tabeli 4.8. Największe niedoszacowanie rzeczywistej liczebności występuje w polu tabeli

odpowiadającemu małżeństwom zawieranym przez mężczyzn w wieku 35–39 lat z kobietami w wieku 25–29 lat. Pole to pierwotnie zaliczyliśmy do kategorii „bariera wieku", gdyż wskaźnik Quételeta dla tego pola miał wartość ujemną, równą -1,4 (część [1] tabeli 4.6). Wartość tę należy uznać jako stosunkowo niewielką, gdy porównamy ją z wartościami w pozostałych polach należących do tej kategorii. W rezultacie model – w którym przyjmuje się jednakowy niedobór małżeństw dla wszystkich pól zaliczonych do kategorii „bariera wieku" – nadmiernie obniża estymowaną liczebność w rozważanym polu. Gdyby pole to opisać za pomocą dodatkowego, specyficznego parametru interakcji, to wtedy obserwowaną w tym polu liczebność można byłoby wyjaśnić dokładnie, a jednocześnie poprawić dopasowanie w pozostałych polach kategorii „bariera wieku". Pozwoliłoby to zwiększyć poziom redukcji współczynnika χ^2 do 95,9 procenta, zaś odsetek małżeństw poprawnie sklasyfikowanych do 96,2 procenta. Jeśli to nadal nie zadowoliłoby badacza, to do modelu wprowadzić można dalsze parametry. Na przykład, dzieląc małżeństwa w kategorii „żona młodsza–mąż starszy" na dwie grupy: na takie, w których żona należy do kohorty wieku o jeden niższej niż mąż, oraz na takie, w których różnica wieku męża i żony jest większa.

Należy jednak mieć na uwadze, że celu nigdy nie stanowi osiągnięcie maksymalnego możliwego dopasowania modelu do danych. Celem jest wyjaśnienie badanego zjawiska. Dodawanie kolejnych parametrów interakcyjnych ma sens wyłącznie wtedy, gdy można wskazać ich realne odpowiedniki w postaci mechanizmów dobierania się małżonków. Wprowadzenie osobnego parametru dla pola obejmującego mężczyzn w wieku 35–39 lat i kobiety w wieku 25–29 lat uzasadnione jest tylko wtedy, gdy potrafimy uzasadnić, dlaczego bariery zawierania małżeństw są w tym wypadku słabsze. Czy chodzi o mężczyzn, którzy po osiągnięciu pewnego stopnia stabilności zawodowej zdecydowali się założyć rodzinę, zaś żony poszukują wśród kobiet, które mają przed sobą perspektywę urodzenia i wychowania przynajmniej dwójki dzieci. Bądź rozpatrując ten sam mechanizm z punktu widzenia kobiet. Czy chodzi tu o kobiety, które kończąc studia i pragnąc założyć rodzinę, poszukują partnerów już zawodowo ustabilizowanych? To tylko jedno z możliwych wyjaśnień. Można sformułować także inne. Czy istotna dla osłabienia omawianych barier jest kategoria mężczyzn mających już za sobą nieudane małżeństwo i pragnących założyć rodzinę na nowo? Tak mogą brzmieć przykładowe hipotezy, które należy mieć na uwadze. Jeśli hipotezy te nie potwierdzą się w konfrontacji z wiedzą na temat badanego zjawiska, to wyodrębnienie danego pola jako specyficznego, czy też grupy pól jako osobnej kategorii, nie ma większego sensu. Cóż bowiem za korzyść z tego, że model związku jest dobrze dopasowany do danych, jeśli nie przekłada się na wyjaśnienie badanego zjawiska.

4.10 Identyfikacja pól o największej specyfice

Niekiedy badaczowi nie zależy na identyfikacji pełnego modelu związku, lecz jedynie na znalezieniu pól tablicy charakteryzujących się największą specyfiką. Przy czym przez specyfikę danego pola rozumie się stopień, w jakim obserwowana liczebność odbiega od liczebności, której należałoby się spodziewać, gdyby cechy były niezależne. Problem sprowadzić więc można do znalezienia pól o największych bezwzględnych wartościach wybranego współczynnika: indeksu, różnicy czy wskaźnika Quételeta. Prezentowane w tym miejscu podejście różni się od omawianych wcześniej propozycji tym, że zawiera dodatkową regułą rozstrzygania, czy dane pole należy uznać za specyficzne, czy też nie. Spośród wszystkich pól tablicy selekcjonowane są w ten sposób pola, które najwięcej wnoszą do wyjaśnienia zjawiska.

Reguła rozstrzygania oparta jest na kryterium statystycznym. Dla każdego pola tablicy obliczana jest wielkość

$$sr_{ij} = \frac{n_{ij} - e_{ij}}{\sqrt{e_{ij} \times (1 - a_i/n)(1 - b_j/n)}} \tag{4.14}$$

którą nazwiemy resztą skorygowaną[10]. Jeśli dane pochodzą z badania zrealizowanego na próbie losowej dobranej z populacji, w której obie rozpatrywane cechy są niezależne, to dla każdego pola tablicy statystyka odpowiadająca reszcie skorygowanej ma rozkład normalny o średniej 0 i odchyleniu 1 (Haberman 1978: 17-23; Garcia-Pérez i Núñez-Antón 2003). Pozwala to skonstruować test statystyczny, którego hipoteza zerowa głosi, że liczebność obserwowana w danym polu jest w populacji równa liczebności wynikającej z modelu niezależności. Jeśli, przykładowo, poziom istotności przyjąć jako 0,05, to otrzymanie reszty skorygowanej większej od 1,96 lub mniejszej od -1,96 wymagałoby odrzucenia hipotezy zerowej. Według zasady tej działa reguła selekcji pól o największej specyfice. Dla każdego z pól konstruuje się podany test a następnie wybiera pola, których liczebności „istotnie" – w górę lub w dół – różnią się od modelu niezależności.

Omawiana metoda znajduje zastosowanie przede wszystkim w sytuacjach, gdy analizie trzeba poddać wiele tablic, koncentrując uwagę wyłącznie na najbardziej znaczących aspektach rozpatrywanych związków. Dlatego rozpowszechniona jest w badaniach marketingowych, których cel ograniczony nie-

[10] W literaturze anglojęzycznej wielkość tę nazywa się *adjusted residual* (Haberman 1978: 17–23) bądź *standardized residual* (Agresti 2007: 38–39). W polskojęzycznej wersji pakietu SPSS przyjęty został termin *reszta skorygowana*, którym posługiwać się będę w tej książce (Górniak i Wachnicki 2008: 137–138).

kiedy bywa do znalezienia *insightu* bądź wskazówek ułatwiających określenie strategii działań rynkowych. Odtworzenie pełnego modelu każdej uzyskanej w badaniu tablicy mija się z celem, gdyż chodzi wyłącznie o uzyskanie syntetycznego wglądu w badany problem – na co na ogół jest zresztą niewiele czasu. Dlatego też oprogramowanie stosowane w badaniach marketingowych ma zaimplementowaną z reguły procedurę domyślnego obliczania skorygowanych reszt dla pól przetwarzanych tablic. Wielkości te porównuje się następnie z dystrybuantą rozkładu normalnego i na tej podstawie określa, czy liczebność w polu tablicy odbiega w sposób statystycznie istotny od niezależności, a jeśli tak, to na jakim poziomie istotności. Różnice istotne statystycznie oznaczane są w tablicach raportów w ustalony sposób, najczęściej za pomocą symbolu gwiazdki lub ciągu dwóch-trzech gwiazdek, w zależności od poziomu istotności. Omawiany system oceny specyfiki pól w tablicach stosuje się niezależnie od tego, czy tablice wypisywane są w formie liczebności, odsetków, czy innych wskaźników[11].

Tabela 4.9
Skorygowane reszty sr_{ij} *wraz z krytycznymi poziomami istotności dla sposobów głosowania kobiet i mężczyzn w wyborach do Sejmu we wrześniu 2005 roku*
Europejski Sondaż Społeczny 2006

[1] skorygowane reszty sr_{ij}

płeć	PiS	PO	SLD	Samo-obrona	PSL	LPR	pozostałe partie	odmowa lub nie pamięta
kobiety	-0,312	0,406	0,250	-2,752	-2,315	0,850	0,445	2,510
mężczyźni	0,312	-0,406	-0,250	2,752	2,315	-0,850	-0,445	-2,510

[2] graniczny poziom istotności dla testu dwustronnego

0,755	0,685	0,802	0,006	0,020	0,395	0,656	0,012

Obliczono na podstawie danych prezentowanych w tabeli 2.9.

[11] W latach dziewięćdziesiątych miałem okazję implementować procedurę obliczania skorygowanych reszt w oprogramowaniu do tworzenia raportów tabelarycznych w firmie GfK-Polonia, a także w pakiecie Soft Data Explorer – używanym obecnie przez wiele firm badania rynku (m.in. Millward Brown SMG/KRC, TNS OBOP) oraz przez ich klientów. Ostatnia wymieniona grupa użytkowników jest szczególnie liczna, gdyż obejmuje działy marketingu i działy badań w największych firmach produkujących dobra czy oferujących usługi, a także w mediach oraz w agencjach obsługujących biznes. Nie można więc wykluczyć, że najpowszechniej stosowaną w Polsce metodą analizy tablic jest właśnie metoda skorygowanych reszt.

W tabeli 4.9 przedstawione zostały wielkości skorygowanych reszt dla pól tablicy opisującej sposób głosowania mężczyzn i kobiet. Ze względu na fakt, że płeć jest cechą dychotomiczną, reszty obliczone dla mężczyzn są co do wielkości bezwzględnej równe resztom obliczonym dla kobiet. Cechy o liczbie kategorii większej niż 2 własności tej nie mają. W części [2] tabeli 4.9 podano wartości granicznego poziomu istotności dla obliczonych reszt. Wartości upoważniające do odrzucenia hipotezy o niezależności sposobu głosowania od płci w populacji otrzymano jedynie dla 6 spośród 16 pól tablicy (przyjmując poziom istotności 0,05). Są to pary pól obejmujących kobiety i mężczyzn głosujących na Samoobronę, głosujących na PSL, a także para pól odpowiadających kategorii osób, które nie odpowiedziały, na kogo oddały głos. Kryterium istotności skorygowanych reszt koncentruje więc uwagę badacza jedynie na owych sześciu polach tablicy. W wypadku pozostałych nie można bowiem wykluczyć, że różnice między liczebnościami obserwowanymi a wielkościami modelu niezależności są wyłącznie konsekwencją losowych wahań związanych z faktem realizacji badania na próbie, nie zaś w pełnej populacji.

Analiza skorygowanych reszt w wypadku rozpatrywanego związku pozwoliła odtworzyć jego pełny model. Związek sposobu głosowania z płcią w niewielkim jednak stopniu odbiega od niezależności, toteż akurat wszystkie pola, które różnią się od niej w sposób statystycznie istotny, tworzą zarazem pełny model związku. Na ogół jednak tak się nie dzieje. Bezpieczniej jest zatem przyjąć, że metoda skorygowanych reszt pozwala wskazać jedynie najbardziej wyraźne odstępstwa od niezależności, natomiast nie dostarcza informacji wystarczających do odtworzenia pełnego modelu badanego związku.

Warto też mieć na uwadze dwa dalsze ograniczenia metody skorygowanych reszt. Po pierwsze, gdy badanie zrealizowane zostało na próbie o dużej liczebności, to metoda wykazać może istotność różnic w większości pól rozpatrywanej tablicy. Nie dostarcza tym samym korzyści polegającej na identyfikacji niewielkiej liczby pól najbardziej specyficznych. Po drugie, gdy badanie ma charakter wyczerpujący, czyli obejmuje całą badaną populację, to metoda traci swoją rację bytu. Co prawda niekiedy badacze próbują stosować ją także wtedy, co wymaga dość karkołomnych założeń w stylu „gdyby była to próba dobrana z nieograniczonej populacji ...” Na ogół jednak badanie wyczerpujące obejmuje dużą liczbą jednostek co powoduje, że metoda przestaje być użyteczna ze względu na powód podany wyżej jako pierwszy. Gdyby obliczyć skorygowane reszty dla pól tablicy przedstawiającej zgodność wieku współmałżonków, to okazałyby się istotne na poziomie 0,05 we wszystkich 81 polach tablicy.

4.11 Dyskusja

Podejście proponowane w tym rozdziale ma charakter eksploracyjny. Jest ukierunkowane na odtworzenie modelu badanego związku bez żadnych wstępnych założeń, wyłącznie na podstawie uzyskanego w badaniu układu liczebności w polach tablicy.

Istotę podejścia można również wyjaśnić, porównując je z dwoma innymi. Pierwsze polega na przyjęciu pewnego modelu zjawiska na mocy założenia, a następnie jego weryfikacji w oparciu o wyniki badania. Podejście to występuje w badaniach, których celem jest weryfikacja określonej teoretycznej koncepcji badanego zjawiska. Narzędziem weryfikacji jest na ogół modelowanie log-liniowe, chociaż stosuje się również modele regresji logitowej bądź inne modele uwzględniające specyfikę procedury badawczej (Agresti 2002). Przykładów omawianego podejścia można wskazać wiele, gdyż od ponad 30 lat większość badań empirycznych nad strukturą społeczną stanowi jego egzemplifikacje. Przykładów dostarcza też książka, która ukazała się jako pierwszy tom tej serii wydawniczej (Domański i Przybysz 2007). Wyjaśnia ona istotę podejścia oraz prezentuje jego możliwości.

Drugie z podejść, do którego chciałbym w tym miejscu nawiązać, stanowi przedmiot rozważań zamieszczonych w dalszych rozdziałach tej książki. Cechuje się ono zalgorytmizowaniem procedur tworzenia modeli związków. Poszukiwanie specyficznych pól czy też podział pól na kategorie przestaje być zadaniem badacza, lecz staje się elementem metody. Po stronie badacza pozostaje wyłącznie interpretacja proponowanego modelu.

Metody identyfikacji wzorca związku przedstawione w tym rozdziale nie zastępują żadnego z tych podejść, lecz stanowią ich uzupełnienie. W wypadku pierwszego podejścia jest bowiem tak, że weryfikacja hipotez dotyczących modelu związku wymaga wstępnego przyjrzenia się wynikom, tak aby móc trafnie je zinterpretować na gruncie koncepcji teoretycznych. Podobnie rzecz się ma w wypadku korzystania z metod automatycznej identyfikacji modeli związków. Aby ocenić, czy otrzymany model jest trafny, bądź też móc go poprawnie zinterpretować, zasadne jest dokonanie wglądu w tablicę uzyskaną w badaniu. Proponowane w tym rozdziale metody identyfikacji modeli badanych zjawisk wspomóc mogą realizację wymienionych celów.

ROZDZIAŁ 5

Dystanse między profilami

W rozdziale proponuję podejście różniące się koncepcyjnie od dotychczas omawianych metod analizy tablic. Opis struktury pól tablicy zastąpiony zostanie analizą układów dystansów między kategoriami cech. Podejście to stanowi uogólnienie dość elementarnej metody analizy tablic, jaką jest porównywanie ze sobą profili, czyli rozkładów procentowych w wierszach bądź w kolumnach.

Rozpocznę od wyjaśnienia sposobu, w jaki porównania profili przekładają się na dystanse między kategoriami cech uwzględnionych w tablicy (5.1). Nie wymaga to stosowania żadnych wskaźników ilościowych. Same intuicje wystarczają, aby dość precyzyjnie oszacować wielkości tych dystansów.

W podrozdziale 5.2 wprowadzę wskaźnik, pozwalający liczbowo wyrazić dystanse między profilami w wierszach lub w kolumnach. Jedną z korzyści stanowi możliwość przypisania wierszom i kolumnom tablicy wartości skalowych, odzwierciedlających owe dystanse. W 5.3 przedstawię, jak na podstawie tego wskaźnika odtworzyć można jednowymiarowy układ dystansów między kategoriami każdej z cech posługując się metodą skalowania wielowymiarowego.

W podrozdziałach 5.4 i 5.5 zaproponuję i omówię inną metodę analityczną, prowadzącą do przypisania wierszom i kolumnom wartości liczbowych. Nazwę ją metodą dopasowania średnich. Metoda jest mało znana badaczom zjawisk społecznych, gdyż stosuje się ją głównie w botanice. Niemniej jednak dostarcza zadziwiająco trafnych intuicji dotyczących mechanizmów zjawisk społecznych, co pokażę na dwóch przykładach.

W podrozdziale 5.6 dokonam porównania obu metod. Pokażę, że w wypadku tablic, w których jedna z cech ma tylko dwie kategorie, metoda porównywania dystansów oraz metoda dopasowania średnich prowadzą do identycznych rezultatów. Przedstawię też przykład tablicy o większych rozmiarach w celu wyjaśnienia, dlaczego nie można w pełni zobrazować jej struktury za pomocą dystansów między wierszami bądź kolumnami.

5.1 Wnioskowanie na podstawie podobieństwa profili

Najczęściej stosowaną metodą analizy zawartości tablicy jest porównywanie ze sobą profili w wierszach lub w kolumnach. Z metody tej korzystaliśmy w rozdziale 2. Przypomnę, że jeśli profile zbliżone są do siebie, to związek między cechami należy uznać jako słaby. Jak bowiem pokazałem w rozdziale 3, profile w modelu niezależności są identyczne – zarówno w wierszach, jak i w kolumnach tablicy. Jeśli natomiast zestawiane profile różnią się od siebie, to fakt ten interpretuje się jako przejaw istnienia związku między cechami. Omawiana metoda opiera się na intuicyjnej ocenie, co to znaczy, że profile są do siebie podobne, bądź że różnią się od siebie.

W rozdziale tym sformułujemy kryteria, które pozwalają określić wielkość różnic między profilami w sposób ilościowy. Rozważania zilustrujemy danymi dotyczącymi związku między wykształceniem rodziców a wyborem szkoły przez dziecko po ukończeniu gimnazjum. W 2006 roku Polska uczestniczyła w ogólnoświatowym projekcie badawczym *Programme for International Students Assessment* (PISA). Badanie to objęło reprezentatywną próbę uczniów pierwszych klas szkół ponadgimnazjalnych. Pozwoliło to między innymi ustalić, jaką szkołę wybrało dziecko po ukończeniu gimnazjum. Dla naszych celów szkoły te podzielimy na trzy kategorie: licea ogólnokształcące, technika – przy czym kategoria ta obejmie również licea profilowane – oraz szkoły zasadnicze zawodowe. Rodzice badanych uczniów wypełniali kwestionariusz, w którym znalazły się między innymi pytania o ich wykształcenie. Dane te posłużyły do skonstruowania tablicy, w której wykształcenie ojca – pogrupowane w 4 kategorie – skrzyżowano z rodzajem szkoły, w którym uczyło się dziecko. Część [1] tabeli 5.1 przedstawia otrzymane liczebności, zaś w częściach [2] i [3] podano profile w kolumnach i w wierszach.

W pierwszej kolejności ustalmy, w których spośród rozpatrywanych rodzajów szkół uczniowie różnią się najbardziej pod względem wykształcenia swoich ojców. W liceach ogólnokształcących ojcowie 23 procent uczniów mają wykształcenie wyższe. Odsetek ten jest wyraźnie niższy w dwóch pozostałych rodzajach szkół. W technikach 3 procent uczniów ma ojców z wykształceniem wyższym, zaś w szkołach zasadniczych zawodowych zaledwie 1 procent. Największa różnica występuje pod tym względem między szkołami zasadniczymi a liceami ogólnokształcącymi.

Informacje o wykształceniu ojca pochodzą z kwestionariuszy wypełnianych przez rodziców. Uwzględniono wyłącznie uczniów, dla których podano zarówno wykształcenie ojca, jak i matki. O wykształcenie obojga rodziców pytano niezależnie od tego, czy uczeń mieszkał z obojgiem rodziców czy też nie.

Tabela 5.1
Liczebności oraz odsetki uczniów w poszczególnych rodzajach szkół ponadgimnazjalnych ogółem oraz w podziale ze względu na wykształcenie ojca

Badanie PISA 2006

[1] liczebności

wykształcenie ojca	rodzaj szkoły ponadgimnazjalnej: liceum ogólnokształcące	technikum	zasadnicza zawodowa	ogółem
wyższe	399	45	10	454
średnie	627	454	92	1173
zasadnicze zawodowe	625	1036	465	2126
podstawowe	95	204	162	461
ogółem	1746	1739	729	4214

[2] odsetki uczniów o różnym wykształceniu ojca w poszczególnych rodzajach szkół

wykształcenie ojca	rodzaj szkoły ponadgimnazjalnej: liceum ogólnokształcące	technikum	zasadnicza zawodowa	ogółem
wyższe	23	3	1	11
średnie	36	26	13	28
zasadnicze zawodowe	36	60	64	50
podstawowe	5	12	22	11
ogółem	100	100	100	100

[3] odsetki uczniów poszczególnych rodzajów szkół w kategoriach wykształcenia ojca

wykształcenie ojca	rodzaj szkoły ponadgimnazjalnej: liceum ogólnokształcące	technikum	zasadnicza zawodowa	ogółem
wyższe	88	10	2	100
średnie	53	39	8	100
zasadnicze zawodowe	29	49	22	100
podstawowe	21	44	35	100
ogółem	41	41	17	100

Idąc dalej, rozpatrzmy odsetki uczniów, których ojcowie mają wykształcenie średnie. W liceach ogólnokształcących jest ich 36 procent, w technikach 26 procent zaś w szkołach zasadniczych zawodowych 13 procent. Również w wypadku tego poziomu wykształcenia najbardziej różnią się od siebie licea ogólnokształcące oraz szkoły zasadnicze zawodowe, zaś technika lokują się między nimi. Stwierdzona prawidłowość powtarza się w wypadku odsetków uczniów, których ojcowie mają wykształcenie zasadnicze zawodowe.

Pozostała jeszcze kategoria ojców o wykształceniu podstawowym. Najniższy odsetek uczniów, których ojcowie mają wykształcenie podstawowe, odnotowano w liceach ogólnokształcących, zaś najwyższy w szkołach zasadniczych zawodowych. Odsetek dla techników przyjmuje zaś wartość pośrednią.

Poczynione obserwacje prowadzą do konkluzji, że pod względem wykształcenia ojców najbardziej różnią się uczniowie liceów ogólnokształcących i szkół zasadniczych zawodowych. Prawidłowość ta wystąpiła w wypadku wszystkich czterech wyodrębnionych poziomów wykształcenia. Jeśli przyjąć, że oba rodzaje szkół wyznaczają skrajne wartości pewnej osi, to technika lokują się na tej osi pomiędzy nimi. Pod względem rozpatrywanego kryterium są przy tym bardziej podobne do szkół zasadniczych niż do liceów ogólnokształcących.

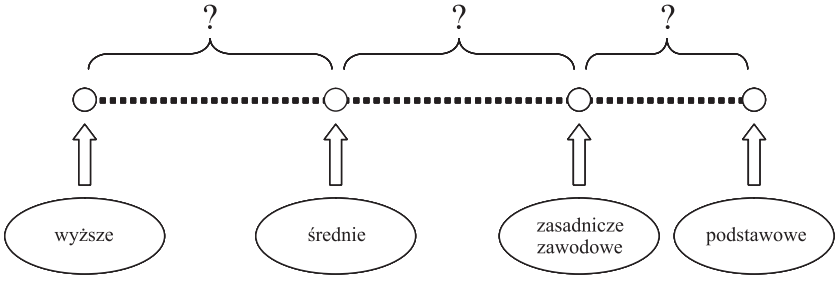
Analogiczne rozumowanie można przeprowadzić, analizując wybór szkoły przez dzieci o różnym wykształceniu ojców (tabela 5.1, część [3]). Na uwagę zasługują w tym wypadku dzieci ojców mających wykształcenie wyższe. W zdecydowanej większości uczą się one w liceach ogólnokształcących (88 procent), zaś tylko 2 procent wybrało szkoły zasadnicze zawodowe. Kategorią przeciwstawną stanowią dzieci, których ojcowie mają wykształcenie podstawowe. Tylko 21 procent tej grupy uczy się w liceach ogólnokształcących, zaś w szkołach zasadniczych zawodowych aż 35 procent. Ojcowie mający wykształcenie wyższe oraz ojcowie mający wykształcenie podstawowe stanowią więc dwie przeciwstawne kategorie ze względu na wybór szkoły ponadgimnazjalnej, w której kształci się dziecko.

Dalsza analiza pozwala zauważyć, że ojcowie o dwóch pozostałych wyodrębnionych poziomach wykształcenia – to jest średnim oraz zasadniczym zawodowym – lokują swoje dzieci gdzieś pomiędzy owymi dwiema skrajnymi kategoriami. Przy czym uzasadniony wydaje się wniosek, że dzieci ojców o wykształceniu zasadniczym wybierają szkoły w sposób zbliżony do dzieci ojców mających wykształcenie podstawowe. W obu grupach podobne są bowiem odsetki dzieci uczęszczających do liceów ogólnokształcących. Również odsetki dzieci uczęszczających do szkół zasadniczych zawodowych są w obu grupach wyraźnie wyższe niż w dwóch pozostałych. Z kolei profil szkół, do których uczęszczają dzieci ojców o wykształceniu średnim, zbliżony jest naj-

bardziej do wyborów drogi szkolnej przez dzieci mające ojców o wykształceniu wyższym. Na przykład, w obu kategoriach niski jest odsetek dzieci uczących się w szkołach zasadniczych zawodowych. Podobieństw takich można wskazać więcej.

Ogół dokonanych ustaleń można sobie wyobrazić zatem jako pewną oś, na której umiejscowione są punkty odpowiadające kategoriom ojców o różnym wykształceniu. Skrajne pozycje na osi zajmują ojcowie o wykształceniu podstawowym i wyższym. Pomiędzy nimi umiejscowione są punkty odpowiadające kategoriom ojców o wykształceniu zasadniczym zawodowym oraz średnim bądź pomaturalnym. Przy czym pierwszy z tych punktów leży bliżej tego skraju osi, który zajmują ojcowie o wykształceniu podstawowym, zaś drugi leży bliżej skrajnego punktu odpowiadającego ojcom o wykształceniu wyższym. Wynik ten zobrazowany został graficznie na rycinie 5.1.

Rycina 5.1
Obraz hipotetycznej osi porządkującej kategorie wykształcenia ojców ze względu na rodzaje szkół ponadgimnazjalnych, w których uczą się ich dzieci



Przeprowadzone rozumowanie pozwoliło pokazać, że analiza podobieństwa profili w wierszach i w kolumnach przekłada się na intuicje dotyczące dystansów między kategoriami uwzględnionych w tablicy cech. Rozumowanie to należy uznać za dość rudymentarne, toteż trudno rozstrzygnąć, kto i kiedy po raz pierwszy doszedł do takich wniosków. W każdym razie potrzeba przełożenia profili w tablicy na układ dystansów stała się punktem wyjścia dla zaproponowania metod analitycznych pozwalających wyrazić te dystanse w sposób ilościowy. Zarówno dalsza część tego rozdziału, jak też w dużym stopniu dwa kolejne rozdziały, poświęcone zostały omówieniu najbardziej znaczących propozycji w tym zakresie.

5.2 Wskaźnik różnic między profilami i jego interpretacja

Punkt wyjścia dla tego rodzaju metod stanowi ilościowe określenie wielkości różnic między profilami w wierszach lub w kolumnach tablicy. W rozumowaniu przedstawionym w poprzednim podrozdziale podobieństwa i różnice oceniane były intuicyjnie za pomocą relacji: większe–mniejsze. Pozwoliło to ustalić porządek kategorii, lecz nie stanowiło dostatecznej podstawy dla precyzyjnego wyrażenia wielkości dystansów między nimi.

W celu dokonania ilościowego porównania profili w tablicy skorzystać można z wielkości, którą nazwiemy **wskaźnikiem różnic między profilami** (ang. *dissimilarity index*)[1]. Wskaźnik ten odpowiada intuicyjnemu rozumieniu zróżnicowania profili. Należy zaznaczyć, że odzwierciedla on nie podobieństwa, a różnice. Oba pojęcia są jednak komplementarne, toteż stosowanie wskaźników różnic jest równie efektywne dla budowania interpretacji opartych na porównywaniu profili, jak posługiwanie się wskaźnikami podobieństwa.

Wartość wskaźnika różnic między profilami określają wzory (5.1), gdy porównywane są profile w wierszach tablicy oraz (5.2), kiedy przedmiot porównań stanowią profile w jej kolumnach.

$$\Delta_{i_1 i_2} = \tfrac{1}{2} \sum_{j=1}^{k} \left| pw_{i_1 j} - pw_{i_2 j} \right| \tag{5.1}$$

dla dowolnych $i_1, i_2 = 1,..., w$; oraz

$$\Delta_{j_1 j_2} = \tfrac{1}{2} \sum_{i=1}^{w} \left| pk_{ij_1} - pk_{ij_2} \right| \tag{5.2}$$

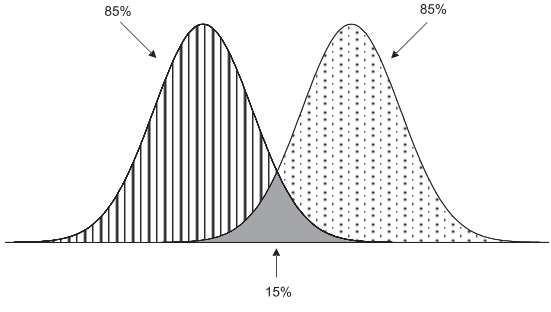
dla dowolnych $j_1, j_2 = 1,..., k$. Przy czym występujące w powyższych wzorach wyrażenia pw_{ij} oraz pk_{ij} oznaczają wielkości w polach profili obliczonych dla wierszy i kolumn tablicy według wzorów (3.6) i (3.7).

Z postaci wzorów (5.1) i (5.2) wynika, że jako wielkość miary różnic między porównywanymi profilami przyjmuje się sumę różnic wielkości w odpowiadających sobie polach tych profili. Ponieważ jednak różnica może być dodatnia bądź ujemna, pod uwagę zawsze jest brana jej wielkość dodatnia, czemu

[1] Propozycję wskaźnika wiąże się na ogół z artykułem Duncan i Duncan (1955), w którym omówiono jego własności. Jednakże wskaźnik ten ma dość oczywistą budowę, toteż spotyka się go również wcześniej (np. Jahn, Schmid i Schrag 1947). Dyskusję praktycznych aspektów stosowania wskaźnika można znaleźć w artykułach Cortese i in. (1976), Tauber i Tauber (1976), Cohen i in. (1976). Sakoda (1981) sformułował propozycję uogólnienia wskaźnika na większą liczbę porównywanych kategorii.

odpowiada w podanych wzorach symbol wielkości bezwzględnej w postaci dwóch pionowych kresek. Obliczona wielkość sumy jest następnie dzielona przez 2, co normalizuje wartość wskaźnika do wielkości 1 – gdy profile podane są w postaci proporcji – lub do 100, gdy podane są w procentach.

Rycina 5.2
Obraz zachodzenia na siebie dwóch profili



Na rycinie 5.2 graficznie zobrazowano zasadę liczenia wartości wskaźnika różnic. Każdy z przedstawionych profili obejmuje 100 procent zróżnicowania w swojej kategorii. Odpowiada to jednakowym wielkościom pola pod obiema krzywymi. Jednakże tylko 15 procent owego pola jest wspólne dla obu profili. Na rycinie 5.2 zostało ono wypełnione kolorem szarym. Pozostała część każdego profilu nie ma swojego odpowiednika w drugim z nich. Obliczając wartość wskaźnika różnic od wielkości pola każdego z profili, czyli od 100 procent, należy odjąć część wspólną, czyli 15 procent. W sumie otrzymamy 170 procent, co po podzieleniu przez 2 daje 85 procent. Tyle wynosi wartość wskaźnika różnic między profilami w przykładzie prezentowanym na rycinie 5.2. Gdyby oba profile zachodziłyby na siebie bardziej, to wartość omawianego wskaźnika byłaby niższa. Od pola każdego z profili należałoby bowiem odjąć większą niż na rysunku część wspólną. Gdyby profile pokryły się całkowicie, to część wspólna wynosiłaby 100 procent, a więc wartość wskaźnika wyniosłaby 0.

Niekiedy wartość wskaźnika różnic interpretuje się w języku przesunięć czy przemieszczeń, które miałyby doprowadzić do sytuacji, w której profile byłyby identyczne (Duncan i Duncan 1955: 211; Sakoda 1981: 245). Należy jednak mieć na uwadze fakt, że nie „przenosi" się w tym wypadku odsetków, lecz z pola do pola przesunąć można wyłącznie jednostki, które stanowią podstawę obliczenia owych odsetków. A ponieważ każdemu z profili odpowiadają na ogół różne liczebności brzegowe, stąd tego rodzaju przesunięcia mogą zmienić konfigurację innych wierszy czy kolumn tablicy.

Aby lepiej to wyjaśnić, wróćmy do prezentowanych w części [3] tabeli 5.1 profili szkół, w których uczą się dzieci ojców zaliczonych do różnych kategorii wykształcenia. W wypadku ojców z wykształceniem wyższym 88 procent ich synów i córek uczy się w liceach ogólnokształcących. Przyjmijmy, że chcielibyśmy osiągnąć ten sam odsetek wśród dzieci ojców, którzy mają wykształcenie zasadnicze zawodowe. Obecnie w polu tabeli odpowiadającym tej sytuacji jest 625 dzieci, co odpowiada odsetkowi 29 procent. Aby uzyskać 88 procent, w polu tym powinno być 1871 dzieci. Brakujące 1246 należałoby dobrać spośród tych, którzy uczą się w technikach bądź szkołach zasadniczych zawodowych, a których ojcowie mają wykształcenie zasadnicze. Potencjalnie jest to do wykonania, gdyż suma wielkości w obu polach przekracza 1246. Można byłoby więc podjąć działania, aby wśród dzieci, których ojcowie mają wykształcenie zasadnicze zawodowe, odsetek uczących się w liceach ogólnokształcących był identyczny jak wśród dzieci, których ojcowie ukończyli wyższe uczelnie. Jednakże wymagałoby to jednoczesnego zwiększenia liczby miejsc w liceach ogólnokształcących dla tych dzieci o owe 1246.

Jeżeli z natury rozpatrywanego problemu wynika, że dopuszcza się zmianę marginesów tablicy, to wtedy wielkości wskaźnika różnic między profilami można interpretować w kategoriach przemieszczania jednostek między polami tablicy. Gdybyśmy mogli odpowiednio zwiększyć liczbę miejsc w liceach ogólnokształcących, to wtedy można byłoby podjąć starania, aby 88 procent młodzieży, której ojcowie mają zasadnicze wykształcenie, skłonić do nauki w liceum. Jeśli takiej możliwości nie ma, to wyrównanie profili w dwóch kategoriach na ogół pociąga za sobą konieczność dokonania odpowiednich przesunięć również w innych kategoriach. Gdyby skłonić do rezygnacji z nauki w liceum większość młodzieży, której ojcowie mają wykształcenie średnie lub podstawowe, to wtedy zwolniłaby się odpowiednia liczba miejsc dla dzieci, których ojcowie ukończyli szkoły zasadnicze.

5.3 Przekształcenie różnic między profilami w układ dystansów

Zdefiniowany w poprzednim podrozdziale wskaźnik zastosować można do oceny wielkości dystansów między kategoriami, dla których obliczone zostały profile. W tabeli 5.2 zamieszczone zostały wartości wskaźnika różnic między profilami wybieranych szkół wśród uczniów o różnym wykształceniu ojca. Aby wielkości wskaźników traktować jako miarę dystansu konieczne jest jednak, aby spełniały pewien warunek, który nazwiemy wymogiem **addytywności**. Jak ustaliliśmy poprzednio, kategorie wykształcenia ojca ze względu na rodzaj wybieranej przez dziecko szkoły uporządkować można w hierarchię: wyż-

sze–średnie–zasadnicze–podstawowe. Wymóg addytywności spełniony byłby wtedy, gdyby dystans pomiędzy dwiema dowolnymi kategoriami byłby równy sumie dystansów pośrednich – o ile są między nimi kategorie pośrednie.

Tabela 5.2
Wartości wskaźnika różnic między profilami rodzaju wybranej szkoły w kategoriach uczniów o różnym wykształceniu ojca. Badanie PISA 2006
[*w procentach*]

	wykształcenie ojca			
wykształcenie ojca	wyższe	średnie	zasadnicze zawodowe	podstawowe
wyższe	0,0	34,4	58,5	67,3
średnie		0,0	24,1	32,9
zasadnicze zawodowe			0,0	13,3
podstawowe				0,0

Obliczono według wzoru (5.1) na podstawie profili podanych w części [2] tabeli 5.1. Pominięto prezentację wartości pod przekątną tabeli, gdyż są one równoważne wartościom nad przekątną. Kolorem szarym oznaczono pola odpowiadające sąsiadującym ze sobą kategoriom wykształcenia.

Sprawdźmy więc, czy podane wskaźniki różnic spełniają wymóg addytywności. Wartość wskaźnika między dwoma skrajnymi poziomami wykształcenia – wyższym i podstawowym – wynosi 67,3. Między nimi są dwie kategorie pośrednie: wykształcenie średnie i zasadnicze. Dystans między wykształceniem wyższym a podstawowym powinien być więc równy sumie trzech dystansów: między wyższym a średnim, między średnim a zasadniczym oraz między zasadniczym a podstawowym. Zsumujmy te trzy dystanse. W tabeli 5.2 odpowiadające im pola oznaczone zostały kolorem szarym.

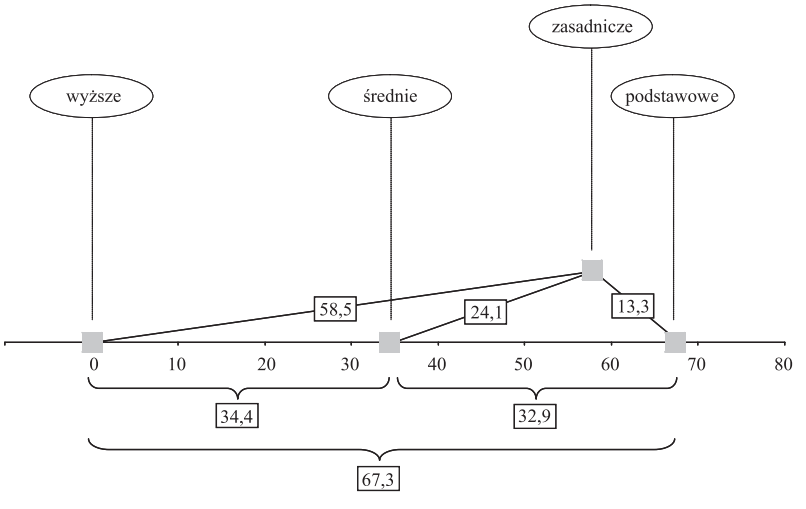
Otrzymana wartość, która wynosi 71,8, różni się od „bezpośredniego" dystansu między kategorią wykształcenia wyższego i podstawowego. Oznacza to, że prezentowane wskaźniki nie spełniają wymogu addytywności.

Brak spełnienia wymogu addytywności oznacza, że bez dokonania ingerencji w dane, wszystkich wskaźników nie można sprowadzić w spójny sposób do pojedynczego wymiaru, który można byłoby traktować jako wymiar dystansów. Ilustruje to obraz układu wskaźników przedstawiony na rycinie 5.3. Kategorie wykształcenia wyższego, średniego oraz podstawowego spełniają wymóg addytywności, przez co można je umiejscowić na jednej osi. Wymogu tego nie spełnia natomiast kategoria wykształcenia zasadniczego zawodowego. Dlatego musieliśmy ją narysować ponad osią, gdyż inaczej

nie dałyby się przedstawić graficznie wielkości wskaźników różnic profili między tą kategorią, a kategoriami wykształcenia wyższego, średniego i podstawowego.

Przekształcenie wskaźników różnic w dystanse między kategoriami wymaga więc dodatkowego kroku, polegającego na „wpasowaniu" wielkości wskaźników w jeden wymiar. Rycina 5.3 sugeruje, aby po prostu zrzutować je na oś poziomą. Sposobem tym można posłużyć się w tym wypadku, lecz nie jest on dogodnym rozwiązaniem, gdy cecha ma więcej kategorii. Na ogół konfiguracja wskaźników nie daje się wtedy przedstawić na płaszczyźnie, lecz wymaga przestrzeni trzy- lub więcej wymiarowej, a w konsekwencji sposób wykonania rzutowania przestaje być sprawą oczywistą.

Rycina 5.3
Układ dystansów miedzy wartościami wskaźników różnic profili dla kategorii wykształcenia ojca



Dlatego lepiej posłużyć się jedną z technik specjalnie przygotowanych do rozwiązywania tego typu problemów. Proponuję, aby do sprowadzenia konfiguracji wskaźników w jeden wymiar skorzystać z technik skalowania wielowymiarowego (ang. *multidimensional scaling*). Ich przeznaczeniem jest przekształcanie tablic wskaźników podobieństwa lub różnic w jedno- lub więcej wymiarowy układ dystansów między obiektami. Technik skalowania wielowymiarowego nie będziemy w tym miejscu omawiać, gdyż poświęcone im zostało wiele odrębnych pozycji (Kruskal i Wish 1978; Koseła i Utzig 1980).

Są one zaimplementowane w większości komputerowych pakietów analitycznych, a także dostępne w sieci. Skorzystanie z jednej z nich nie powinno więc rodzić trudności[2].

Tabela 5.3
Wartości wskaźnika różnic profili wykształcenia ojców wśród uczniów różnych rodzajów szkół ponadgimnazjalnych. Badanie PISA 2006
[*w procentach*]

	rodzaj szkoły ponadgimnazjalnej		
rodzaj szkoły ponadgimnazjalnej	liceum ogólnokształcące	technikum	zasadnicza zawodowa
liceum ogólnokształcące	0,0	30,1	44,8
technikum		0,0	14,7
zasadnicza zawodowa			0,0

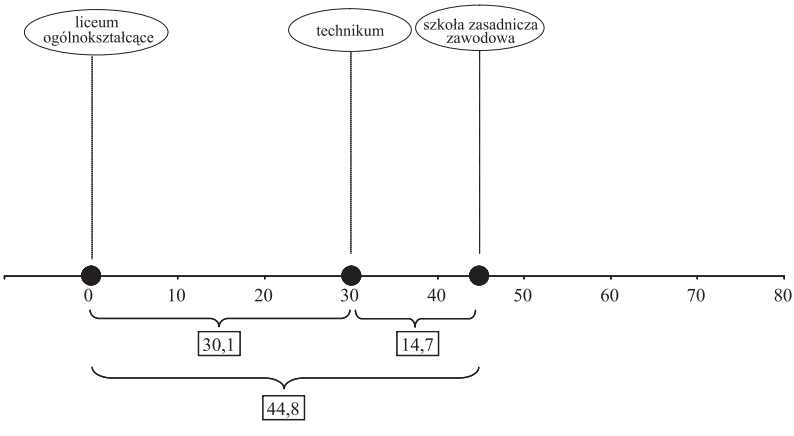
Porównywane profile podano w części [*2*] *tabeli 5.1. Kolorem szarym zaznaczono pola odpowiadające kategoriom sąsiadującym ze sobą.*

Zanim tego dokonamy, poświęćmy kilka słów wskaźnikom różnic między profilami wykształcenia ojców wśród uczniów poszczególnych rodzajów szkół (tabela 5.3). Analiza ich konfiguracji nie prowadzi do takich problemów, jak uprzednio. Spełniają one bowiem warunek addytywności. Zobrazowane to zostało na rycinie 5.4, z której wynika, że dystans między liceami ogólnokształcącymi a szkołami zasadniczymi zawodowymi jest dokładnie równy sumie dystansów między liceami a technikami oraz między technikami a szkołami zasadniczymi. Należy jednak zastrzec, że spełnienie warunku addytywności zastosowanych wskaźników zależy od konfiguracji liczebności w tablicy. Wyniki badania ułożyły się akurat w ten sposób, że wskaźniki różnic profili są addytywne. Gdyby ułożyły się inaczej (na przykład, gdyby posłużyć się inną

[2] Należy mieć na uwadze fakt, że poszczególne techniki skalowania wielowymiarowego dają nieco różniące się rozwiązania. Ponadto, uruchamiając daną procedurę, należy w odpowiedni sposób ustawić dostępne opcje – gdyż od tego również zależą uzyskane wyniki. W szczególności warto sprawdzić, czy podane wskaźniki zostaną zinterpretowane jako „dissimilarities", gdyż większość procedur domyślnie działa na wskaźnikach podobieństwa. Jedna z dostępnych opcji procedury skalowania wielowymiarowego zawsze dotyczy maksymalnej liczby wymiarów rozwiązania. W tym miejscu zalecałbym, aby liczbę tę ustawiać jako 1 – gdyż celem jest przekształcenie tablicy wskaźników różnic w **pojedynczą** oś, na której określony zostanie porządek kategorii i wielkości dystansów między nimi. Do projekcji dystansów między kategoriami w więcej niż jednym wymiarze lepiej jest posłużyć się metodami, które zostaną omówione w rozdziałach 6 i 7.

edycją badania PISA), to warunek addytywności mógłby nie zostać spełniony. Z góry nie wiadomo jak skonfigurują się liczebności w tablicy, więc nie należy liczyć na to, że spełniony będzie. Nie jest też regułą, że w wypadku cechy o trzech kategoriach warunek addytywności jest zawsze spełniony, zaś gdy kategorii jest więcej niż trzy, to spełniony być nie musi.

Rycina 5.4
Układ dystansów między profilami wykształcenia ojców uczniów różnych rodzajów szkół ponadgimnazjalnych



W tabeli 5.4 przedstawione zostały wartości otrzymane metodą skalowania wielowymiarowego dla kategorii obu cech, to jest wykształcenia ojca i rodzaju szkoły. W pierwszej kolumnie podano wartości surowe, co w tym wypadku oznacza wartości uzyskane poprzez poddanie skalowaniu wielowymiarowemu wskaźników przedstawionych w tabelach 5.2 i 5.3. Wynikiem zastosowanej procedury skalowania są zestawy współrzędnych, których średnie wynoszą 0. Część współrzędnych jest przez to dodatnia, część ujemna – co utrudnia porównanie dystansów między kategoriami w wymiarze utworzonym przez procedurę.

Otrzymane wartości odpowiadają skali, którą nazywa się interwałową (zob. podrozdział 1.6). Oznacza to między innymi, że układ dystansów określony jest z dokładnością do przekształcenia liniowego. Należy to rozumieć w ten sposób, że współrzędne punktów odpowiadających kategoriom cechy przemnożyć można przez dowolną stałą, a także dodać do nich dowolną stałą. Nie zmieni to wzajemnych relacji między punktami. Własność tę wykorzystać można do

Tabela 5.4
Surowe i znormalizowane wartości skalowe dla rodzajów szkół ponadgimnazjalnych oraz dla kategorii wykształcenia ojca uzyskane metodą skalowania wskaźników różnic rozkładów warunkowych

cechy i kategorie	wartości skalowe	
	surowe	znormalizowane
rodzaj szkoły ponadgimnazjalnej		
liceum ogólnokształcące	0,773	100
technikum	-0,159	33
szkoła zasadnicza zawodowa	-0,614	0
wykształcenie ojca		
wyższe	0,942	100
średnie	0,124	49
zasadnicze zawodowe	-0,411	15
podstawowe	-0,654	0

Obliczono na podstawie wskaźników różnic profili (tabele 5.2 oraz 5.3). Wartości surowe uzyskano za pomocą procedury PROXSCAL pakietu SPSS PC, przyjmując opcję wyboru metody jako „Torgerson" oraz maksymalną liczbę wymiarów jako 1. Otrzymane wartości podano w rubryce „surowe". W ostatniej kolumnie podano te same wartości znormalizowane do przedziału <0, 100> w postaci zaokrąglonej do liczb całkowitych.

normalizacji otrzymanego rozwiązania. Proponuję normalizację sprowadzającą rozpiętość dystansów do przedziału od 0 do 100. Wybór takiego przedziału sprawdza się w praktyce. Zestawiając i porównując ze sobą liczby w zakresie od 0 do 100 łatwiej jest ocenić wielkości różnic między nimi a także wzajemne relacje tych różnic. Pozwala to porównywać dystanse między kategoriami uzyskane dla cech skrzyżowanych w tej samej tablicy, a także dystanse otrzymane dla cech w różnych tablicach. Normalizacja zestawu współrzędnych do przedziału <0, 100> wymaga wybrania spośród nich wielkości największej $x_{\max}$ oraz najmniejszej $x_{\min}$, a następnie skorzystania z wzoru (5.3).

$$zx = 100 * \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{5.3}$$

Wyniki skalowania w znormalizowanej postaci zamieszone zostały w ostatniej kolumnie tabeli 5.4. Potwierdzają one wcześniejszy wniosek, że dystans między technikami a liceami ogólnokształcącymi jest większy niż między szkołami zasadniczymi a technikami. Wartości odpowiadające kategoriom

wykształcenia ojców również potwierdzają wcześniejsze intuicje. Dystans między ojcami o wykształceniu zasadniczym zawodowym a podstawowym jest stosunkowo niewielki, co odzwierciedla fakt sporego podobieństwa dróg szkolnych dzieci ojców zaliczonych do obu kategorii. Największy dystans, równy 51 umownych jednostek, zarysował się pomiędzy ojcami o wykształceniu wyższym a średnim. Odzwierciedla to fakt, że wybory szkół przez dzieci ojców z wyższym wykształceniem odbiegają znacznie od wszystkich pozostałych rozważanych kategorii. Przypomnijmy, że aż 88 procent dzieci z rodzin, w których ojciec ma wykształcenie wyższe, po ukończeniu gimnazjum kontynuuje naukę w liceum ogólnokształcącym. Wyraźnie inne są wybory szkolne dzieci w rodzinach, w których ojciec nie ma wykształcenia wyższego.

5.4 Wartości skalowe jako wynik dopasowania średnich

W podrozdziale tym przedstawię metodę wyznaczania dystansów między kategoriami cech w tablicy opartą na odmiennym podejściu. Metoda ta jest mało znana badaczom w naukach społecznych zapewne z tego powodu, że znalazła zastosowanie w dość odległej dziedzinie, określanej w języku angielskim jako *ecology* – co nie ma dobrego polskiego odpowiednika. Przyjmijmy więc, że chodzi o botanikę. Klasyfikowanie środowisk ze względu na warunki występowania i wzrostu różnych gatunków roślin jest zadaniem o tyle złożonym, że w grę wchodzi duża liczba wzajemnie zależnych od siebie czynników, jak rodzaj gleby, wilgotność, średnia temperatura, nasłonecznienie oraz wiele innych. Z obserwacji wynika, że w pewnych środowiskach niektóre gatunki roślin rozwijają się lepiej, inne gorzej. Zaś te gatunki, które rozwijają się gorzej, w innych środowiskach mogą rozwijać się lepiej. Tworzy to rodzaj swoistej interakcji między środowiskiem a gatunkami roślin, którym środowisko stwarza sprzyjające warunki dla rozwoju. Środowiska można więc klasyfikować ze względu na gatunki roślin, zaś gatunki roślin ze względu na środowiska.

Wprowadzenie omawianej metody do badań przypisuje się amerykańskiemu botanikowi Robertowi H. Whittakerowi (1920–1980), aczkolwiek sama metoda była znana i sporadycznie stosowana wiele lat wcześniej[3]. Whittaker (1967) wyodrębnił dwa sposoby analizy związku między środowiskiem a występującymi w nim gatunkami roślin. Pierwszy sposób, który nazwał bezpo-

[3] Jako pierwsze zastosowanie metody uważa się utworzenie przez Richardsona i Kudera (1933) narzędzia oceny pracowników dla firmy Procter and Gamble. Związki metody dopasowania średnich z innymi metodami analizy danych omawiają Tenenhaus i Young (1985: 94–107). Twierdzą przy tym, że metoda stanowiła pierwowzór zaproponowanej w latach sześćdziesiątych we Francji *analizy korespondencji* (rozdział 7).

średnim, polega na wskazaniu elementu środowiska, który można uznać za najważniejszy (np. wilgotność gleby), a następnie uporządkowaniu rozpatrywanych gatunków według średnich wielkości wyodrębnionego elementu. Z naszego punktu widzenia ciekawszy jest jednak drugi ze sposobów, który Whittaker nazywa pośrednim.

Sposób pośredni opiera się na założeniu, że znaczących elementów środowiska jest wiele. A ponieważ są one ze sobą skorelowane, więc wyodrębnienie ich specyficznych wpływów drogą analityczną (na przykład za pomocą regresji) nie prowadzi do konkluzywnych rezultatów. Dlatego lepiej traktować je jako wspólny, hipotetyczny czynnik, stanowiący wypadkową elementów środowiska o największym znaczeniu. W pierwszym kroku proponowanej przez Whittakera metody rozpatrywane gatunki roślin porządkuje się, przypisując im średnie któregokolwiek ze składników owej wypadkowej (na przykład średnie wilgotności gleby). A następnie – i tu tkwi sedno omawianej metody – oblicza się średnie dla środowisk na podstawie średnich przypisanych gatunkom. Te nowe średnie nie muszą dokładnie pokrywać się ze średnimi obliczonymi na podstawie pierwotnie uwzględnionego składnika, gdyż biorą pod uwagę wszystkie gatunki roślin występujące w poszczególnych środowiskach. Stanowią więc **lepsze przybliżenie** hipotetycznej wypadkowej wszystkich elementów środowiska sprzyjających wzrostowi poszczególnych gatunków, niż stanowił je ów pierwotnie wybrany składnik. Procedurę można kontynuować, obliczając dla gatunków nowe średnie na podstawie średnich przypisanych poszczególnym środowiskom, po czym znów środowiskom średnie na podstawie nowych średnich przypisanych gatunkom. Po każdym kolejnym kroku tej procedury – nazwijmy go iteracją – uzyskuje się bardziej precyzyjne oszacowania zarówno średnich przypisanych środowiskom ze względu na występujące gatunki, jak też gatunkom ze względu na środowiska, które sprzyjają ich rozwojowi. Po którymś kroku następne iteracje nie przynoszą już widocznej poprawy precyzji oszacowań, gdyż wyniki stabilizują się. Opisaną metodę Whittaker nazwał *reciprocal averaging*. Na użytek polski nazwiemy ją **metodą dopasowania średnich**.

Omawiana metoda jest we współczesnej botanice mocno ugruntowana (Ter Braak 1987), zaś w podręcznikach poświęca się jej równie wiele uwagi co innym metodom analizy danych (McGarigal, Cushman i Stafford 2000: 67–68). Może więc warto zainteresować metodą dopasowania średnich również badaczy w naukach społecznych. Zwłaszcza że jej istota odpowiada dość podstawowym intuicjom dotyczącym wzajemnych relacji między cechami w tablicy. Pomaga między innymi zrozumieć sposób, w jaki kształt związku przekłada się na dystanse między kategoriami obu cech.

Istotę algorytmu wzajemnego dopasowania średnich przedstawmy na przykładzie związku wykształcenia ojca z wyborem szkoły ponadgimnazjalnej. Po-

zwoli to porównać dopasowane średnie z wartościami skalowymi uzyskanymi w poprzednim podrozdziale na podstawie analizy zróżnicowania profili w tablicy. Przypuśćmy, że skądinąd wiemy, że ojcowie dzieci uczęszczających do liceów ogólnokształcących mają stosunkowo wysokie wykształcenie, a w każdym razie bardziej korzystne niż dzieci wybierające po gimnazjum pozostałe rodzaje szkół. Umieśćmy więc licea ogólnokształcące najwyżej w rankingu

Tabela 5.5
Surowe i znormalizowane wartości skalowe dla rodzajów szkół ponadgimnazjalnych oraz dla kategorii wykształcenia ojców uzyskane w kolejnych iteracjach procedury dopasowania średnich

		rodzaj szkoły ponadgimnazjalnej			
nr iteracji	*typ wielkości*	liceum ogólnokształcące	technikum	zasadnicza zawodowa	*korelacja kanoniczna*[a]
0	znormalizowane	100,000	0,000	50,000	
2	surowe	37,939	14,979	8,847	
	znormalizowane	100,000	21,081	0,000	0,405209
4	surowe	47,602	26,021	18,372	
	znormalizowane	100,000	26,167	0,000	0,412330
6	surowe	48,310	26,829	19,070	
	znormalizowane	100,000	26,538	0,000	0,412372
8	surowe	48,363	26,890	19,112	
	znormalizowane	100,000	26,565	0,000	0,412372
10	surowe	48,367	26,895	19,126	
	znormalizowane	100,000	26,568	0,000	0,412372

		wykształcenie ojca				
nr iteracji	*typ wielkości*	wyższe	średnie	zasadnicze zawodowe	podstawowe	*korelacja-kanoniczna*[a]*i*
1	surowe	88,987	57,374	40,334	38,178	
	znormalizowane	100,000	37,782	4,244	0,000	0,341639
3	surowe	89,975	61,612	39,670	29,936	
	znormalizowane	100,000	52,759	16,214	0,000	0,411817
5	surowe	90,479	63,580	42,149	32,187	
	znormalizowane	100,000	53,856	17,090	0,000	0,412369
7	surowe	90,516	63,724	42,330	32,351	
	znormalizowane	100,000	53,938	17,156	0,000	0,412372
9	surowe	90,519	63,735	42,343	32,363	
	znormalizowane	100,000	53,944	17,161	0,000	0,412372

[a] *Wskaźnik ten omówiony zostanie w rozdziale 6.*

szkół, skonstruowanym ze względu na poziom wykształcenia ojców uczniów, przypisując temu rodzajowi szkół rangę 100. Przyjmijmy dalej, że nie mamy pewności, czy szkoły zasadnicze zawodowe grupują dzieci o bardziej korzystnym wykształceniu ojców, czy też bardziej korzystne wykształcenie ojców cechuje dzieci uczęszczające do techników. Na mieście mówi się, że w szkołach zasadniczych oferujących możliwość zdobycia atrakcyjnych zawodów liczba kandydatów na jedno miejsce przewyższa nawet liczbę kandydatów w renomowanych liceach. Przypiszmy więc w sposób arbitralny 50 punktów szkołom zasadniczym zawodowym, zaś technikom 0 punktów, przyznając im tym samym najniższe miejsce w rankingu. Wartości te przedstawiono w tabeli 5.5 w wierszu oznaczonym jako iteracja 0.

Rozważmy obecnie, w jakich szkołach lokują **przeciętnie** swoje dzieci ojcowie o różnych poziomach wykształcenia. Jesteśmy co prawda w stanie określić to dokładnie, gdyż znamy rozkłady warunkowe wyboru rodzaju szkoły w kategoriach wykształcenia ojca (tabela 5.1, część [3]), lecz przyjmijmy, że chcielibyśmy ocenić, jak wygląda to **średnio**. Na przykład, czy ojcowie o wykształceniu wyższym lokują swoje dzieci przeciętnie w lepszych szkołach niż ojcowie o wykształceniu zasadniczym zawodowym. Aby odpowiedzieć na tak sformułowane pytanie obliczmy dla każdej kategorii ojców średnie wartości rankingowe szkół, do których uczęszczają ich dzieci. Posłużymy się w tym celu wzorem (5.4), który pozwala obliczyć średnią dla każdego wiersza tablicy w oparciu o wartości skalowe $y_1, y_2,..., y_k$ przypisane jej kolumnom

$$x_i = \frac{1}{a_i} \sum_{j=1}^{k} y_j n_{ij} \tag{5.4}$$

W dotychczasowym rankingu szkoły mają wartości skalowe przypisane im arbitralnie: 100 – licea ogólnokształcące, 50 – szkoły zasadnicze zawodowe, zaś 0 – technika. Dla ojców o wykształceniu wyższym średnia wartości rankingowych szkół, do których uczęszczają ich dzieci, wyniesie więc (korzystamy z liczebności podanych w tabeli 5.1 [1])

$$1/454 * (100 * 399 + 0 * 45 + 50 * 10) = 88{,}987$$

Średnie obliczone dla wszystkich kategorii wykształcenia ojców przedstawione zostały w wierszu tabeli 5.5 oznaczonym jako iteracja 1 – wielkości surowe. Znormalizujmy te wartości do przedziału <0, 100>, a następnie policzmy średnie w drugą stronę. Chodzi o określenie pozycji rankingowych szkół na podstawie wykształcenia ojców. Średnie obliczymy, korzystając z wzoru

$$y_j = \frac{1}{b_j} \sum_{i=1}^{w} x_i n_{ij} \tag{5.5}$$

Na przykład, dla liceów ogólnokształcących średnia wartości skalowych przypisanych kategoriom wykształcenia ojców po pierwszej iteracji wynosi

$$1/1746 * (100 * 399 + 37{,}782 * 627 + 4{,}244 * 625 + 0 * 95) = 37{,}939$$

Wartości te podano w wierszu oznaczonym jako iteracja 2 w postaci otrzymanej bezpośrednio z obliczeń oraz w postaci znormalizowanej.

I stała się rzecz zaskakująca. Pomimo, że szkołom zasadniczym zawodowym przydzieliliśmy arbitralnie dość korzystne miejsce w rankingu, to po uwzględnieniu **faktycznego** profilu wykształcenia ojców szkoły zasadnicze zawodowe spadły na dno hierarchii, zaś technika zajęły pozycję od nich wyższą. Przy czym wyraźnie większy dystans zaznaczył się między liceami ogólnokształcącymi a technikami niż między technikami a szkołami zasadniczymi zawodowymi. Przypomina to układ dystansów uzyskanych poprzez porównywanie ze sobą profili (tabela 5.4).

Spróbujmy ustalić, skąd wzięło się to podobieństwo. W poprzednim kroku – oznaczonym jako iteracja 1 – obliczyliśmy średnie dla kategorii wykształcenia ojców. Po ich znormalizowaniu najwyższa wartość równą 100 otrzymali ojcowie z wykształceniem wyższym. Aż 88 procent ich dzieci uczęszcza bowiem do liceów ogólnokształcących (tabela 5.1, część [3]), stąd też odpowiadająca im wartość średnia jest w największym stopniu wyznaczona przez wartość rankingową przypisaną liceom ogólnokształcącym. Dla porównania, ojcom mającym wykształcenie podstawowe odpowiada najniższa średnia, gdyż 44 procent ich dzieci uczęszcza do techników, które w pierwotnym rankingu miały przypisaną najniższą pozycję, równą 0. Średnia niewiele się od niej różniąca odpowiada ojcom o wykształceniu zasadniczym. Ich dzieci wybierają bowiem szkoły w podobny sposób – może nieco częściej technika (49 procent), lecz z drugiej strony częściej też licea ogólnokształcące. Reasumując, obliczone średnie dla kategorii wykształcenia ojców ułożyły się po pierwszej iteracji w kolejności: wyższe–średnie–zasadnicze–podstawowe. Jeśli na tej podstawie odświeży się w kolejnej iteracji wartości rankingowe dla szkół, to wynikiem musi być kolejność: licea ogólnokształcące–technika–szkoły zasadnicze zawodowe. Wynika to z kształtu profili wykształcenia ojców w kategoriach uczniów poszczególnych rodzajów szkół (część [2] tabeli 5.1).

Procedurę dopasowywania średnich można kontynuować dalej. Rezultaty dla kolejnych iteracji przedstawione zostały w tabeli 5.5. Wynika z nich, że

zarówno średnie dla kategorii wykształcenia ojców, jak też dla rodzajów szkół, stopniowo stabilizują się. Po ostatniej prezentowanej iteracji różnice w stosunku do iteracji poprzedniej nie przekraczają jednego miejsca po przecinku. Zakończmy więc na tym proces dopasowania średnich i przyjmijmy wyniki otrzymane po ostatniej wykonanej iteracji jako wartości skalowe odpowiadające wierszom i kolumnom tablicy.

Tabela 5.6
Znormalizowane wartości skalowe dla rodzajów szkół ponadgimnazjalnych oraz dla kategorii wykształcenia ojców otrzymane metodą skalowania wskaźników różnic rozkładów warunkowych oraz metodą dopasowania średnich

cechy i kategorie	skalowanie wskaźników różnic	wzajemne dopasowywanie średnich
rodzaj szkoły ponadgimnazjalnej		
liceum ogólnokształcące	100	100
technikum	33	27
szkoła zasadnicza zawodowa	0	0
wykształcenie ojca		
wyższe	100	100
średnie	49	54
zasadnicze zawodowe	15	17
podstawowe	0	0

Wartości skalowe podano w postaci zaokrąglonej do liczb całkowitych. Wartości otrzymane metodą skalowania wskaźników różnic pochodzą z tabeli 5.4, zaś wartości uzyskane metodą wzajemnego dopasowania średnich z tabeli 5.5.

Wartości uzyskane metodą dopasowania średnich zestawione zostały w tabeli 5.6 z analogicznymi wartościami otrzymanymi w poprzednim podrozdziale metodą skalowania wielowymiarowego. Okazuje się, że metoda dopasowania średnich odtwarza podobny kształt układu dystansów, jak metoda oparta na bezpośrednim porównywaniu ze sobą profili. Wniosek wydaje się uzasadniony, pomimo że między oboma rozwiązaniami występują niewielkie różnice. Czy obu podejść nie można więc sprowadzić do wspólnego mianownika? Do problemu tego wrócimy w podrozdziale 5.6, poprzedzając to jeszcze jednym przykładem zastosowania metody dopasowania średnich. Pozwoli to lepiej zrozumieć związek, jaki zachodzi między dopasowanymi średnimi a kształtem rozkładu cech w tablicy.

5.5 Kształt związku a porządek wierszy i kolumn: homogamia małżeńska ze względu na wiek

Rozważmy związek między wiekiem mężczyzny i kobiety w momencie zawarcia ślubu. W rozdziale 4 przedstawiłem dane GUS na ten temat, dotyczące małżeństw zawartych w 2006 roku. Pozwoliły one zidentyfikować wzorzec tego związku poprzez analizę wskaźników Queteleta dla poszczególnych pól tablicy. Nic nie mówiliśmy wtedy na temat możliwości przedstawienia istoty tego związku za pomocą wartości skalowych przypisanych wierszom i kolumnom. Spróbujmy do tego samego przykładu podejść od tej strony. W tym celu rozpatrzmy fikcyjne zdarzenie.

Urzędy stanu cywilnego mają obowiązek raportować do urzędu statystycznego dane o wieku osób zawierających małżeństwa. Nie wiadomo już który z urzędników wpadł kiedyś na pomysł, aby ze względu na wymogi ochrony danych osobowych informacje te szyfrować. Wprowadzono w związku z tym klucz kodowy, w którym przedziały wieku oznaczono literami, przypisując je do poszczególnych przedziałów w sposób przypadkowy. Przy czym osobne oznaczenia zastosowano dla przedziałów wieku kobiet, osobne zaś dla wieku mężczyzn (tabela 5.7).

Tabela 5.7
Oznaczenia przedziałów wieku mężczyzn i kobiet zawierających małżeństwo

Przykład fikcyjny

wiek mężczyzny	symbol kodowy	wiek kobiety	symbol kodowy
do 19 lat	X	do 19 lat	M
20–24 lata	C	20–24 lata	S
25–29 lat	P	25–29 lat	W
30–34 lata	G	30–34 lata	B
35–39 lat	N	35–39 lat	L
40–44 lata	F	40–44 lata	A
45–49 lat	Y	45–49 lat	H
50–54 lata	U	50–54 lata	T
55–59 lat	R	55–59 lat	Z
ponad 60 lat	K	ponad 60 lat	E

Wszystkie urzędy stanu cywilnego otrzymały ten sam klucz kodowy i za jego pomocą szyfrowały dane przesyłane do urzędu statystycznego. Gdy w urzędzie statystycznym zebrano dane dla całego roku, ważny urzędnik we-

zwał swojego pracownika i polecił mu sporządzić w trybie pilnym tablicę wieku mężczyzn i kobiet zawierających małżeństwo. Za dwa dni ukazać miał się nowy rocznik statystyczny, a konieczność uwzględnienia w nim tej tablicy wcześniej przeoczono. Pracownik zabrał się rączo do pracy, uruchomił w komputerze odpowiednią procedurę i otrzymał tablicę w postaci przedstawionej w części [1] tabeli 5.8.

Jak widać, to, co otrzymał, nie nadawało się do publikacji. Komputer uporządkował bowiem wiersze i kolumny tablicy według symboli literowych przypisanych przedziałom wieku. Pracownik znał zakresy wieku w tych przedziałach, gdyż takie same stosowano w poprzednich latach. Nie mógł natomiast odnaleźć dokumentu, który pozwoliłby przełożyć oznaczenia literowe na te przedziały. A ponieważ godzina była późna i urzędy stanu cywilnego były zamknięte, nie wchodziło w grę skontaktowanie się z którymś z nich i poproszenie o przesłanie klucza kodowego.

Pracownik stanął więc wobec konieczności odtworzenia tego klucza. Wyszedł z założenia, że małżeństwa zawierają na ogół osoby w zbliżonym wieku. W związku z tym postanowił zrekonfigurować tablicę, tak aby na przekątnej uzyskać największe liczebności. Pozwoliłoby to ustalić odpowiedniość między symbolami literowymi oznaczającymi te same przedziały wieku w wypadku mężczyzn i kobiet, co zawsze stanowiłoby krok do przodu. Okazało się jednak, że nie da się tego zrobić. Największa liczebność w wierszu P, czyli 47 312, odpowiada kolumnie W. Jednakże wiersz P zawiera również największą liczebność kolumny S. Nie można przez to rozstrzygnąć, czy wiersz P odpowiada kolumnie S, czy W. Największe liczebności w poszczególnych wierszach i kolumnach nie muszą wcale leżeć na przekątnej tablicy. Mężczyźni zawierają małżeństwo przeciętnie w nieco późniejszym wieku niż kobiety, więc pola o największych liczebnościach mogą leżeć nie tylko na przekątnej, lecz również bezpośrednio pod nią.

Pracownik zastanowił się więc, jak można inaczej rozumieć skłonność do zawarcia małżeństwa z osobą w zbliżonym wieku. Wyobraził sobie dwie kategorie mężczyzn – w wieku 20–24 lata i 25–29 lat. Wydało mu się oczywiste, że mężczyźni z pierwszej grupy żenią się średnio rzecz biorąc z kobietami młodszymi niż mężczyźni z drugiej grupy. Zauważył ponadto, że dotyczy to każdych dwóch kategorii wieku mężczyzn. Stąd wniosek, że wszystkie kategorie wieku mężczyzn są uporządkowane ze względu na średnie wieku kobiet, z którymi zawierają małżeństwo. Analogiczny wniosek dotyczy kobiet. Skłonność do znalezienia męża w podobnym wieku przejawia się tym, że kobiety z młodszych grup wiekowych wybierają mężczyzn przeciętnie młodszych, niż kobiety w starszych grupach wiekowych. Istota zjawiska nie zależy przy tym od sposobu kategoryzacji wieku. Nie jest ważne, czy przedziały wieku są

Tabela 5.8
Liczba mężczyzn i kobiet w kategoriach wieku dla małżeństw zawieranych w 2006 roku

[1] tablica uporządkowana według symboli kategorii wieku

wiek mężczyzny	wiek kobiety A	B	E	H	L	M	S	T	W	Z	ogółem
C	9	662	0	4	75	8709	41 329	0	8113	1	58 902
F	838	1416	1	399	1251	19	307	112	970	29	5342
G	252	8614	3	53	1231	483	7788	15	17 365	3	35 807
K	135	49	1936	318	55	3	16	710	16	904	4142
N	465	3843	6	173	1832	76	1199	50	3491	8	11 143
P	73	4895	0	15	515	2894	43 522	4	47 312	3	99 233
R	214	94	143	507	117	2	15	738	54	530	2414
U	453	254	74	863	326	6	42	770	157	220	3165
X	0	8	0	0	1	1280	769	0	68	0	2126
Y	790	635	21	874	664	11	114	388	371	115	3983
ogółem	3229	20 470	2184	3206	6067	13 483	95 101	2787	77 917	1813	226 257

[2] tablica uporządkowana według dopasowanych średnich

wiek mężczyzny	wiek kobiety M (100)	S (99)	W (98)	B (92)	L (82)	A (65)	H (47)	T (30)	Z (15)	E (0)	ogółem
X (100)	1280	769	68	8	1	0	0	0	0	0	2126
C (99)	8709	41 329	8113	662	75	9	4	0	1	0	58 902
P (98)	2894	43 522	47 312	4895	515	73	15	4	3	0	99 233
G (95)	483	7788	17 365	8614	1231	252	53	15	3	3	35 807
N (89)	76	1199	3491	3843	1832	465	173	50	8	6	11 143
F (79)	19	307	970	1416	1251	838	399	112	29	1	5342
Y (61)	11	114	371	635	664	790	874	388	115	21	3983
U (43)	6	42	157	254	326	453	863	770	220	74	3165
R (26)	2	15	54	94	117	214	507	738	530	143	2414
K (0)	3	16	16	49	55	135	318	710	904	1 936	4142
ogółem	13 483	95 101	77 917	20 470	6067	3229	3206	2787	1813	2184	226 257

Źródło danych: GUS 2007. W części [2] przy literowych symbolach kategorii podano w nawiasach dopasowane średnie w postaci znormalizowanej do przedziału <0, 100>.

jednakowej długości, ani czy wyodrębnione zostały w identyczny sposób dla mężczyzn i kobiet. Ten sam wiek metrykalny nie musi świadczyć o jednakowej atrakcyjności mężczyzny lub kobiety dla płci przeciwnej. Ważne jest wyłącznie to, aby zastosowane kategorie wieku korespondowały z postrzeganiem relacji wieku w fazie szukania kandydata na współmałżonka. Wtedy, jeśli osoby chcące zawrzeć małżeństwo **faktycznie** kierują się subiektywnie rozumianą zgodnością wieku, to średnie dla kategorii mężczyzn i kobiet **muszą** układać się w określonym porządku.

Pracownik doszedł więc do wniosku, że konsekwencję skłonności do zawierania małżeństw z osobą w zbliżonym wieku stanowi określony porządek średnich przypisanych kategoriom wieku mężczyzn i kobiet w rozpatrywanej tablicy. Przy czym kategoriom mężczyzn odpowiadałyby średnie wieku kobiet, z którymi zawarli małżeństwo, zaś kategoriom kobiet średnie wieku mężczyzn wybranych na męża. Pracownik nie miał wykształcenia socjologicznego, lecz do pracy w urzędzie trafił po ukończeniu akademii rolniczej. Dzięki temu uświadomił sobie, że natura rozpatrywanego problemu jest analogiczna, jak w wypadku klasyfikowania środowisk ze względu na gatunki występujących roślin. Środowiska – to kobiety wchodzące na rynek małżeński. Można uporządkować je ze względu na atrakcyjność dla mężczyzn. Atrakcyjność jest wypadkową wielu czynników, z których jednym z ważniejszych jest z pewnością wiek. Mężczyźni zaś, to poszczególne gatunki roślin, które adaptują się lepiej do pewnych środowisk, do innych zaś gorzej. Aby odtworzyć wzajemną odpowiedniość kategorii mężczyzn i kobiet zawierających małżeństwo można więc posłużyć się metodą dopasowania średnich.

Pracownik napisał więc w arkuszu kalkulacyjnym procedurę dopasowania średnich. Następnie ponumerował wiersze i kolumny tablicy kolejnymi liczbami od 1 do 10. Wiedział bowiem, że dla uzyskanego rozwiązania nie ma znaczenia, jakie wartości przyjmie się w punkcie startu. Po wykonaniu odpowiedniej liczby iteracji pracownik uznał, że dopasowywane średnie ustabilizowały się, po czym znormalizował je i przypisał wierszom oraz kolumnom. Na tej podstawie uporządkował wiersze i kolumny, otrzymując tablicę w postaci prezentowanej w części [2] tabeli 5.8.

Ostatni krok stanowiło sprawdzenie orientacji tablicy. Chodziło o rozstrzygnięcie, czy wierszom położonym wyżej i kolumnom z lewej strony odpowiadają osoby młodsze czy starsze. Porządek wartości skalowych jest bowiem wyznaczony z dokładnością do przekształcenia liniowego, czyli może być odwrócony bez naruszenia istoty metody. Liczebności brzegowe wierszy położonych w górnej części tablicy okazały się jednak przeciętnie rzecz biorąc wyższe niż wierszy położonych w dolnej części tablicy. Dowodzi to, że wiersze ułożone są w kolejności od osób najmłodszych do najstarszych. Małżeństwa

częściej zawierają bowiem ludzie młodsi niż starsi. Symbolowi X odpowiada więc przedział wieku „do 19 lat", symbolowi C przedział „20–24 lata" i tak dalej. W identyczny sposób pracownik ustalił znaczenie symboli przypisanych kategoriom wieku kobiet.

Przedstawiony przykład ilustruje odpowiedniość między kształtem związku w tablicy a porządkiem średnich dopasowanych do jej wierszy i kolumn. Nie rozstrzyga natomiast, czy na podstawie otrzymanych wartości średnich wnioskować można o wielkości dystansów między poszczególnymi wierszami i kolumnami. Wykorzystując dopasowane średnich, pracownik nie odwoływał się do ich wielkości, a jedynie do ich porządku. Wystarczyło to do wykonania powierzonego zadania. Badaczowi zjawisk społecznych nie musi wystarczać. Szkoda byłoby bowiem rezygnować z możliwości interpretacyjnych stwarzanych przez względne dystanse wyznaczone przez dopasowane średnie. Wartości przedstawione w części [2] tabeli 5.8 świadczą, że dystanse te są mniejsze w wypadku kategorii mężczyzn i kobiet zawierających małżeństwa w młodym wieku, zaś wyraźnie większe w wypadku małżeństw zawieranych przez osoby w starszym wieku. Powstaje pytanie, czy różnicom tym można nadać interpretację w języku badanego zjawiska.

5.6 Dopasowane średnie a podobieństwo profili

Metoda dopasowania średnich posiada pewne własności analogiczne do metody porównywania profili w tablicy. Wyobraźmy sobie, że profile szkół wybieranych przez dzieci mające ojców o wykształceniu wyższym i średnim są identyczne. W tej sytuacji średnie obliczone dla obu profili będą sobie równe. Metoda dopasowania średnich „sklei" więc obie kategorie wykształcenia ojca w jedną. Do analogicznego rezultatu doprowadziłoby obliczenie dla porównywanych profili omawianego w podrozdziale 5.2 wskaźnika różnic (wzór 5.1). Otrzymana wartość wyniosła by zero co oznacza, że między kategoriami nie ma żadnego dystansu. Czyli można je traktować jako takie same pod względem rodzaju szkoły wybieranej przez dziecko.

Zarówno omawiana własność, jak też przedstawione przykłady dopasowania średnich sugerują, że im dopasowane średnie są sobie bliższe, tym profile w tablicy są bardziej podobne. Należy jednak zastrzec, że intuicjom tym nie odpowiada ścisła zasada przełożenia dystansów między dopasowanymi średnimi na dystanse między profilami dla wierszy bądź kolumn. Tego rodzaju ekwiwalentność ma miejsce jedynie w wypadku niektórych tablic. Należą do nich wszystkie tablice, w których jedna z cech ma tylko dwie kategorie. Należą również niektóre z tablic o większych rozmiarach, aczkolwiek określenie

warunków, które o tym decydują, pozostawimy do rozdziału 6. W tym miejscu omówimy sposób rozumienia ekwiwalentności między podobieństwem dopasowanych średnich a podobieństwem profili w tablicach, w których jeden z wymiarów nie przekracza 2 (Gautam i Kimeldorf 1999).

Przykładem tego rodzaju tablicy jest – rozpatrywany w poprzednich rozdziałach – rozkład sposobu głosowania mężczyzn i kobiet w wyborach parlamentarnych w 2005 roku. Płeć jest cechą, która ma tylko dwie kategorie. Przypiszmy arbitralnie kobietom wartość skalową 100, zaś mężczyznom 0. W pierwszej iteracji procedura dopasowania średnich przypisze sposobom głosowania wartości, które odpowiadają odsetkom kobiet głosujących na poszczególne partie (tabela 5.9, część [1]). Można to prześledzić na przykładzie dowolnej kategorii. Rozważmy dla przykładu osoby głosujące na PiS. Wartość skalowa dla tej kategorii jest równa

$$\frac{100 \times 52{,}90 + 0 \times 47{,}10}{100{,}00} = 52{,}90$$

czyli odpowiada odsetkowi kobiet głosujących na PiS. W wypadku cechy dychotomicznej średnie przypisane kategoriom drugiej z cech są bowiem odsetkami wystąpienia tej z kategorii cechy dychotomicznej, której po znormalizowaniu przypisano wartość 100.

W wypadku tablicy zawierającej cechę dychotomiczną pierwsza iteracja jest zarazem ostatnią, którą warto wykonać. Następne powielać będą otrzymany wcześniej układ wartości. W drugiej iteracji obliczalibyśmy bowiem średnie dla cechy dychotomicznej. Po ich znormalizowaniu znów uzyskalibyśmy wartości 100 i 0, przez co w kolejnej iteracji otrzymalibyśmy dokładnie te same średnie dla sposobów głosowania, co w pierwszej.

Uporządkujmy wszystkie sposoby głosowania według wartości dopasowanych średnich (czyli według odsetków kobiet) od największej do najmniejszej, a następnie wartości te znormalizujmy do przedziału <0,100> (tabela 5.9, część [1]). Największa znormalizowana wartość – równa 100 – odpowiada kategorii osób, które w badaniu nie udzieliły odpowiedzi, na kogo oddały głos. W tej kategorii odsetek kobiet był bowiem najwyższy. Najmniejsza wartość skalowa – równa 0 – odpowiada z kolei kategorii osób głosujących na PSL, wśród których odsetek kobiet był najniższy. Otrzymane wartości skalowe warto zobrazować graficznie. W postaci tej przedstawione zostały na rycinie 5.5.

Obecnie policzmy różnice między profilami w kolumnach tabeli korzystając z wprowadzonego w podrozdziale 5.3 wskaźnika różnic. Wyniki zamieszczone zostały w części [2] tabeli 5.9. Pełną skalę dystansów wyznaczają dwie skrajne kategorie: osoby odmawiające podania swojego sposobu głosowania

Tabela 5.9
Odsetki i wartości skalowe dla sposobów głosowania w wyborach parlamentarnych we wrześniu 2005 roku
Europejski Sondaż Społeczny 2006

[1] uporządkowane udziały kobiet i mężczyzn wśród głosujących na poszczególne partie (w procentach) oraz dopasowane średnie w postaci znormalizowanej do przedziału <0, 100>

	sposób głosowania								
płeć	odmowa lub nie pamięta	LPR	pozostałe partie	SLD	PO	PiS	Samo-obrona	PSL	ogółem
kobiety	63,62	62,71	58,07	54,80	54,55	52,90	39,16	31,78	53,51
mężczyźni	36,38	37,29	41,93	45,20	45,45	47,10	60,84	68,22	46,49
ogółem	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
dopasowane średnie	100,00	97,12	82,55	72,29	71,49	66,33	23,17	0,00	

[2] wskaźniki różnic między rozkładami płci dla sposobów głosowania (w procentach)

	sposób głosowania								różnice wobec profilu brzegowego
sposób głosowania	odmowa lub nie pamięta	LPR	pozostałe partie	SLD	PO	PiS	Samo-obrona	PSL	
odmowa lub nie pamięta	0,00	0,92	5,56	8,82	9,08	10,72	24,46	31,84	10,11
LPR	0,92	0,00	4,64	7,90	8,16	9,80	23,55	30,92	9,20
pozostałe partie	5,56	4,64	0,00	3,27	3,52	5,16	18,91	26,28	4,56
SLD	8,82	7,90	3,27	0,00	0,26	1,90	15,64	23,02	1,29
PO	9,08	8,16	3,52	0,26	0,00	1,64	15,38	22,76	1,04
PiS	10,72	9,80	5,16	1,90	1,64	0,00	13,74	21,12	0,61
Samoobrona	24,46	23,55	18,91	15,64	15,38	13,74	0,00	7,38	14,35
PSL	31,84	30,92	26,28	23,02	22,76	21,12	7,38	0,00	21,73
różnice wobec profilu brzegowego	10,11	9,20	4,56	1,29	1,04	0,61	14,35	21,73	0,00

[3] dystanse sposobów głosowania w stosunku do głosujących na PSL w postaci znormalizowanej do przedziału <0, 100>

	sposób głosowania							
	odmowa lub nie pamięta	LPR	pozostałe partie	SLD	PO	PiS	Samo-obrona	PSL
dystanse znormalizowane	100,00	97,12	82,55	72,29	71,49	66,33	23,17	0,00

Źródło i opis danych podano w tabeli 2.9. Wartości wskaźników różnic podane w części [2] nad i pod przekątną są sobie równoważne. Kolorem szarym oznaczono pola odpowiadające dystansom między sąsiadującymi ze sobą sposobami głosowania.

oraz osoby głosujące na PSL. Jeśli tą ostatnią kategorię przyjmiemy jako dolny kraniec skali, to najniższy z wierszy w części [2] tabeli 5.9 – odpowiadający PSL – opisuje dystanse poszczególnych sposobów głosowania do dolnego krańca. Rozciągłość całej skali wynosi 31,84, co stanowić może podstawę normalizacji dystansów poszczególnych sposobów głosowania wobec PSL. W części [3] tabeli 5.9 przedstawiono te same wartości znormalizowane do przedziału <0,100>. Są one równe znormalizowanym wartościom dopasowanych średnich, prezentowanych w ostatnim wierszu części [1] tabeli 5.9. Otrzymany rezultat ilustruje fakt, że w wypadku tablicy zawierającej cechę dychotomiczną różnice między dopasowanymi średnimi są **równoważne** dystansom między profilami.

Warto zwrócić uwagę na fakt, że dystanse między sposobami głosowania spełniają w tym wypadku warunek addytywności. Czyli dystans między dowolnymi dwiema kategoriami jest równy sumie dystansów kategorii leżącymi między nimi. Na przykład, dystans między głosującymi na PSL oraz głosującymi na PiS wynosi 21,12 (tabela 5.9, część [2]). Między nimi leży kategoria głosujących na Samoobronę. Tym samym dystans między głosującymi na PSL a PiS jest równy sumie dystansów: między głosującymi na PSL i Samoobronę oraz między głosującymi na Samoobronę a PiS.

$$21{,}12 = 7{,}38 + 13{,}74$$

Warunek addytywności spełniają również dystanse obliczone w stosunku do kategorii ogółem. Dystanse te odzwierciedlają różnice między profilami płci osób głosujących na poszczególne partie a profilem brzegowym płci, to jest profilem dla ogółu badanych. Z ryciny 5.5 odczytać można, że profil brzegowy płci jest pomiędzy profilami dla osób głosującymi na PiS i na PO. Dystans między kategoriami osób głosujących na PiS i PO jest w tej sytuacji równy sumie dystansów ogółu respondentów w stosunku do głosujących na każdą z tych partii.

$$1{,}64 = 0{,}61 + 1{,}04$$

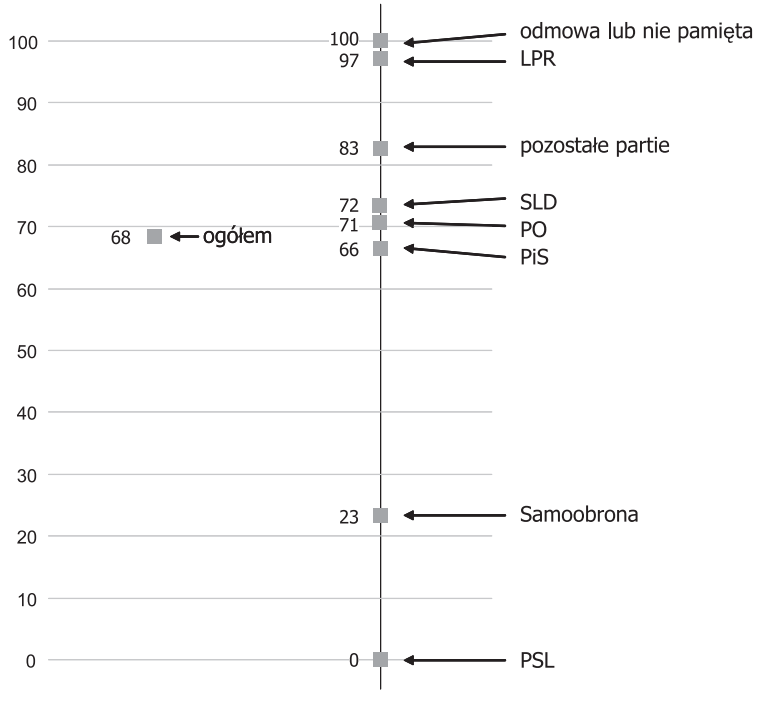
Spełnienie warunku addytywności stwarza możliwość przedstawienia układu dystansów w jedynym wymiarze. Rycina 5.5 przedstawia przez to **pełne informacje** na temat różnic profili płci w kategoriach osób głosujących na poszczególne partie. Jednowymiarowy układ dystansów jest konsekwencją tego, że cecha, dla której analizujemy profile, ma tylko dwie kategorie.

Analogiczna prawidłowość, poza szczególnymi przypadkami, nie zachodzi dla tablic o większych rozmiarach. Aby ustalić, jakie pociąga to konsekwencje dla sposobów analizowania tablic, przyjrzyjmy się bliżej wskaźnikom różnic między rozkładami wieku małżonka obliczonymi dla mężczyzn i kobiet (Ta-

bela 5.10). Uwagę ograniczymy do analizy wskaźników obliczonych dla kategorii wieku mężczyzn (część [1] tabeli 5.10), gdyż wnioski dla kobiet byłyby analogiczne.

Rycina 5.5
Wartości skalowe dla sposobów głosowania w wyborach parlamentarnych we wrześniu 2005 roku obliczone ze względu na udział kobiet

Europejski Sondaż Społeczny 2006



Przyjmijmy, że dolny kraniec skali dystansów wyznacza kategoria mężczyzn w wieku do 19 lat. Jej dystans do sąsiedniej kategorii – mężczyzn w wieku 20–24 lata – jest znaczny, gdyż wskaźnik różnicy dla porównywanych profili wynosi aż 45,4 procent. Prawie w połowie przypadków mężczyźni należący do obu kategorii zawierają małżeństwa z kobietami z różnych grup wiekowych. Dystans 19-latków do innych kategorii jest jeszcze większy. Rośnie tym bardziej, im porównujemy ich ze starszą kategorią mężczyzn. W wypadku mężczyzn ponad 60-letnich dystans wynosi już 98,7 procent. Tylko w jednym przypadku na sto wybranki mężczyzn z obu kategorii są w tym samym wieku.

Sumując wskaźniki dla sąsiednich kategorii wieku łatwo zauważyć, że nie spełniają one wymogu addytywności. Suma wskaźników z uwzględnieniem wszystkich kategorii pośrednich między grupą mężczyzn najmłodszych i najstarszych (pola oznaczone kolorem szarym) wynosi aż 283. Jest większa nie

Tabela 5.10
Wartości wskaźników różnic między profilami wieku żony w kategoriach mężczyzn oraz profilami wieku męża w kategoriach kobiet dla małżeństw zawartych w 2006 roku

[1] mężczyźni: wartości wskaźników różnic między profilami wieku żony (w procentach)

wiek	*wiek mężczyzny*									
mężczyzny	do 19	20–24	25–29	30–34	35–39	40–44	45–49	50–54	55–59	ponad 60
do 19	0,0	45,4	57,3	73,3	84,9	90,3	93,2	94,9	96,6	98,7
20–24		0,0	38,2	61,9	73,5	78,8	86,3	92,2	95,8	97,9
25–29			0,0	38,2	51,7	70,2	82,0	88,0	92,6	97,4
30–34				0,0	28,8	47,3	67,3	81,1	88,8	95,7
35–39					0,0	26,5	48,9	68,9	82,0	91,3
40–44						0,0	29,1	50,8	69,3	83,3
45–49							0,0	25,8	45,3	72,6
50–54								0,0	24,8	59,3
55–59									0,0	40,8
ponad 60										0,0

[2] kobiety: wartości wskaźników różnic między profilami wieku męża (w procentach)

wiek	*wiek kobiety*									
kobiety	do 19	20–24	25–29	30–34	35–39	40–44	45–49	50–54	55–59	ponad 60
do 19	0,0	29,8	63,6	70,8	85,8	93,0	96,9	98,5	98,9	99,4
20–24		0,0	33,8	62,8	80,3	87,9	96,0	97,5	98,7	99,3
25–29			0,0	44,0	63,5	83,2	91,3	95,5	97,2	98,8
30–34				0,0	39,2	63,3	80,4	88,5	92,5	96,6
35–39					0,0	35,5	60,8	74,4	83,0	92,4
40–44						0,0	30,6	54,7	68,3	84,5
45–49							0,0	26,9	53,4	78,7
50–54								0,0	27,2	63,2
55–59									0,0	38,8
ponad 60										0,0

Źródło danych: GUS 2007. Obliczono na podstawie liczebności w części [2] tabeli 5.8. Kolorem szarym oznaczono pola odpowiadające kategoriom sąsiadującym ze sobą pod względem wieku.

tylko od maksymalnej wielkości dystansu między mężczyznami 19-letnimi a ponad 60-letnimi, lecz wykracza również poza granice stosowanej miary, którą są różnice wyrażone w procentach. Układ dystansów między kategoriami mężczyzn znacznie więc odbiega od modelu addytywności. Ogranicza to możliwości analizy tablicy metodą porównywania profili. Człowiek ma bowiem potrzebę wyobrażania sobie dystansów w kategoriach linearnych. Jeśli nie dadzą się one sprowadzić do jednej osi, to ich obliczenie niewiele pomoże w zrozumieniu kształtu związku.

Powstaje w związku z tym pytanie, czy najbardziej popularna metoda analizy tablic, polegająca na porównywaniu profili w wierszach lub w kolumnach, nie zawiera w sobie immanentnej sprzeczności, która powoduje, że w wypadku przynajmniej niektórych tablic nie prowadzi do spójnych wniosków?

Aby rozstrzygnąć tę wątpliwość, wróćmy na chwilę do tej wersji rozpatrywanej tabeli, w której przedstawione zostały profile wyboru żony dla mężczyzn w poszczególnych kategoriach wieku (tabela 4.5, część [2]). Przyjmijmy, że chcemy na podstawie prezentowanych profili wyrobić sobie jedynie ogólny pogląd, na czym polega związek wieku mężczyzny i kobiety w momencie zawarcia małżeństwa. Czytając pierwszy wiersz tabeli widzimy, że aż 60 procent mężczyzn w wieku do 19 lat wybiera jako żonę kobietę w tym samym wieku, zaś dalsze 36 procent kobietę z następnej kohorty wiekowej. Czyli aż 96 procent wybiera kobietę w wieku do 24 lat, a więc nie tak bardzo odbiegającym od ich własnego wieku. Rozpatrzmy drugi wiersz tabeli. Aż 70 procent mężczyzn w wieku 20–24 lata żeni się z kobietami z tej samej kohorty, 14,8 procent z kobietami młodszymi o jeden przedział wieku, zaś 13,8 procent ze starszymi o jeden przedział. W sumie daje to prawie 100 procent. Rozkład wieku kobiet, z którymi żenią się mężczyźni w wieku 20–24 lata jest silnie skupiony wokół tego przedziału.

Nie będziemy kontynuować tego rozumowania dla dalszych wierszy, gdyż dwa pierwsze wystarczają aby uświadomić sobie, na czym w rzeczywistości polega porównywanie ze sobą profili w rozpatrywanej tablicy. Otóż mimochodem ograniczamy się do poszukiwania ich punktów skupienia. Na pozostałe fragmenty porównywanych profili nie zwracamy uwagi zadowalając się faktem, że obszar skupienia liczebności obejmuje prawie całość rozkładu w wierszu. A ponieważ kategorie w kolumnach tablicy mają interpretację w wielkościach liczbowych, gdyż są to przedziały wieku, to tak na prawdę **porównujemy ze sobą średnie** wieku żony w kategoriach wieku męża.

Sformułujmy raz jeszcze otrzymany wniosek, gdyż wydaje się on ważny dla zrozumienia sposobu korzystania przez badaczy z metod analizowania tablic. W tablicach, w których mniejszy z wymiarów przekracza 2, zestawienie ze sobą profili nie prowadzi do klarownego obrazu badanego związku. Obojętnie,

jaką miarą podobieństw czy różnic posłużymy się, otrzymany układ wskaźników nie musi sprowadzać się do jednego wymiaru. Dlatego badacze ograniczają się do porównywania jedynie **wybranych aspektów** profili w wierszach lub w kolumnach, szukając pól o największych liczebnościach (modalna profilu), czy też próbując zidentyfikować obszary pól skupiające największe liczebności (tendencja centralna profilu). Ignorują natomiast pozostałe pola porównywanych profili, zwłaszcza pola o niewielkich liczebnościach. Syntetycznego obrazu badanego zjawiska nie uda się uzyskać, gdy zbyt głęboko wejdzie się w szczegóły.

5.7 Dyskusja

W początkach rozwoju badań metody analizy tablic cechowo dążenie do syntezy. Tablica opisywana była za pomocą pojedynczego parametru, który charakteryzował ją jako całość. Znajdowało to uzasadnienie przede wszystkim w tym, że tablice miały niewielkie rozmiary. Wiele badanych zjawisk przedstawiano za pomocą tablic krzyżujących ze sobą cechy dychotomiczne, a w tym wypadku pojedynczy parametr w pełni opisuje kształt związku w tablicy. Jeszcze w latach pięćdziesiątych statystycy zajmowali się głównie udoskonalaniem mierników siły związku między cechami.

Stopniowo jednak pojedynczy parametr przestawał satysfakcjonować badaczy. Nie wiadomo, czy stało się to za sprawą eksplozji informacyjnej, czy też lawinowego wzrostu liczby prowadzonych badań – w każdym razie od statystyków zaczęto oczekiwać propozycji bardziej zaawansowanych narzędzi do interpretowania zależności w tablicach. Chodziło z jednej strony o narzędzia elastyczne, pozwalające zarówno pójść w głąb, jak też poprzestać na dość ogólnych wnioskach. Z drugiej zaś o narzędzia selektywne, pozwalające odsiać wnioski istotne od całej masy niewiele znaczących prawidłowości.

Rozważania w tym rozdziale stanowią rodzaj wstępu do problematyki poszukiwania narzędzi selektywnych. W miarę wszechstronnie starałem się przedstawić dość rudymentarny sposób wnioskowana, oparty na porównywaniu profili w wierszach lub w kolumnach tablicy. Nie bez przyczyny zagadnieniom tym poświęciłem tak wiele miejsca. Metody porównywania profili dają się bowiem zalgorytmizować, co starałem się wyjaśnić na przykładzie metody dopasowania średnich. Metoda startuje z punktu, który można wybrać przypadkowo. Odpowiada to sytuacji braku wiedzy na temat badanego zjawiska. W kolejnych iteracjach wiedza ta jest stopniowo odtwarzana z tablicy, która jest screenowana w pionie i w poziomie. Zaś na koniec otrzymujemy propozycję modelu badanego związku.

Metody zalgorytmizowane mogą być przydatne jedynie pod warunkiem, że logiką swojego działania powielają sposoby analizowania danych przez badaczy. Dlatego dość wszechstronnie starałem się omówić kwestię, na ile podobieństwo profili przedstawić można w postaci dystansów między kategoriami odpowiadającymi wierszom i kolumnom tablicy. Co do tego, że badacze analizują przedstawiane w tablicach zjawiska porównując odsetki w wierszach i w kolumnach, nie ma chyba żadnych wątpliwości. A to, czy są równie skłonni rozpatrywać różnice między profilami w terminach dystansów między obiektami przedstawionymi na prostej bądź na płaszczyźnie, zależy głównie od tego, na ile wzbogacić to może interpretację badanych zjawisk.

ROZDZIAŁ 6

Eksploracja istoty zjawiska

Rozdział wiąże dotychczas omawiane metody, które można nazwać aplikacyjnymi, z metodami heurystycznymi. Różnicę między obiema grupami metod wyobrazić sobie można przez analogię do różnicy między domową apteczką a lekarzem. W wypadku choroby korzysta się niekiedy z domowej apteczki i wybiera z niej lek, który może pomóc. Skuteczność takiej terapii zależy od tego, czy dobrany lek okaże się właściwy, o ile w ogóle jest on w domowej apteczce, oraz czy lek ten będzie właściwie dawkowany. Jest to funkcją dotychczasowych doświadczeń z radzeniem sobie z chorobą.

Lekarz przyjmie inną strategię. Przede wszystkim postawi diagnozę co do rodzaju choroby i jej intensywności. O ile jest dobrym lekarzem, to będzie starał się również ustalić, co spowodowało chorobę oraz czy występowała wcześniej. Dopiero na tej podstawie zaproponuje możliwości leczenia. Tryb ten okaże się tym skuteczniejszy, im lepiej pacjent kooperować będzie z lekarzem. Czy będzie w stanie pomóc zidentyfikować przyczyny, które spowodowały chorobę. Czy zyska przekonanie co do pozytywnych skutków proponowanych terapii.

Metody aplikacyjne pełnią funkcję domowej apteczki. Dostarczają narzędzi ułatwiających wgłębienie się w analizowaną tablicę, lecz pozostawiających badaczowi swobodę co do formułowanych wniosków. Ich trafność zależy od doświadczeń w posługiwaniu się tymi metodami – szczególnie w sytuacjach podobnych do analizowanej. Metody heurystyczne działają jak dobry lekarz. Diagnozują problem w fachowy sposób po czym inspirują do jego rozwiązania. Możliwe jest to dzięki temu, że zawierają zakumulowaną wiedzę dotyczącą zasad wyciągania wniosków z wyników badań. Wskazują badaczowi aspekty zjawiska, które warte są rozważenia. Rola badacza staje się przez to inna. Z identyfikacji zjawiska przenosi się na jego interpretację.

Rozdział rozpoczynam od omówienia kwestii relacji między prostotą modelu a dogłębnością uzyskanego wyjaśnienia (6.1). Kwestia ta staje się kluczowa, gdy do badacza należy wybór pomiędzy różnymi dostępnymi modelami.

W podrozdziale 6.2 wracam do problemu, który kończy rozważania rozdziału 5. Chodzi o znalezienie takiej miary dystansu profili, która spełniałaby

własność addytywności, a tym samym pozwoliłaby przedstawić te dystanse w jednym wymiarze. Jednowymiarowość rozwiązania jest warunkiem budowania intuicji związanych z kształtem badanego związku. Zaproponuję w tym celu miarę, która znana jest pod nazwą **dystansów chi-kwadrat** – prezentując jej własności i ograniczenia.

W podrozdziale 6.3 definiuję model związku, który nazywam **tablicą kanoniczną**. Model ten stanowi podstawę metod heurystycznych stosowanych w analizach wyników badań. Istota tego modelu jest poniekąd ukryta, gdyż na ogół stanowi element składowy innych modeli. Wiedza na temat jego kształtu może pomóc w trafnej interpretacji wyników uzyskanych za pomocą metod heurystycznych. W podrozdziale 6.4 przedstawię na przykładowych danych sposób wyodrębnienia tablicy kanonicznej a także zasady budowania modeli wyjaśniających badane zjawiska na bazie tej tablicy.

Podrozdział 6.5 zawiera omówienie własności jednego z tego rodzaju modeli, który nazwałem **modelem kanonicznym**. Uwagę skoncentruję na powiązaniach tego modelu z omawianymi w poprzednich rozdziałach metodami aplikacyjnymi. Przedstawię jego związki z metodą porównywania profili wierszy lub kolumn, z metodą dopasowania średnich, z koncepcją pomiaru siły zależności za pomocą pojedynczego wskaźnika, z identyfikacją modelu zjawiska za pomocą wskaźników Quételeta, a także z testowaniem hipotez statystycznych dotyczących występowania zależności w populacji, z której dobrano próbę.

W podrozdziale 6.6 przedstawię kolejną własność modeli kanonicznych, zwaną **hierarchicznością**. Pozwala ona zastosować tę samą strategię wyjaśniania badanego zjawiska w sposób rekurencyjny. To znaczy użyć jej jeszcze raz w tej samej formie do wyjaśnienia tych jego aspektów, które nie zostały wyjaśnione we wcześniejszych krokach.

Ostatni podrozdział, oznaczony 6.7, zawiera pełny wykład omawianego podejścia. Nazwałem je metodą **dekompozycji tablic**. Właściwie prezentacja ta powinna znaleźć się na początku rozdziału, gdyż wszystkie przedstawione w nim propozycje stanowią szczególne przypadki tej metody. Jej opis jest jednak dość formalny, toteż zdecydowałem się umieścić go na końcu. Czytelnikom preferującym ten rodzaj wykładu proponuję, aby rozpoczęli lekturę od podrozdziału 6.7. Chociaż utracą wtedy poczucie odkrywania rzeczy zaskakujących, to w zamian rozważania ułożą się w logiczną całość.

Zamieszczona w 6.8 dyskusja nawiązuje do kwestii większej popularności wśród badaczy niektórych metod analizy tablic, przy braku uznania dla innych. Stanowi to zarazem punkt wyjścia rozważań rozdziału 7.

Prezentowane w tym rozdziale metody wykraczają poza zestaw standardowych narzędzi, którymi na co dzień posługują się badacze. Z tego względu wykonanie stosownych obliczeń wymaga sięgnięcia po rzadziej stosowane

procedury i programy. Można je odnaleźć zarówno w większości najbardziej popularnych pakietów statystycznych, jak też w wyspecjalizowanym oprogramowaniu dostępnym w Internecie. Prezentując poszczególne metody nie będę jednak odwoływał się do tych możliwości, gdyż kluczowym celem jest interpretacja proponowanych rozwiązań. Natomiast wszelkie kwestie związane z wykonaniem obliczeń zostały przedstawione osobno, w postaci aneksu B zamieszczonego na końcu książki.

6.1 Reguła prostoty

W XIV wieku angielski filozof i teolog William z Ockham (1285–1349) sformułował znaną po dziś dzień zasadę, zwaną brzytwą Ockhama

> *bytów nie trzeba mnożyć ponad potrzeby.*

Zasada ta wydaje się wartościową dyrektywą skuteczności działania w wielu dziedzinach[1], w tym także w nauce i w badaniach. Elliott Sober, filozof nauki z uniwersytetu Wisconsin-Madison, uzasadnia to następująco (2001: 13)

> *W nauce niekiedy podejmuje się decyzje o wyborze jednej z konkurencyjnych hipotez kierując się ich prostotą.* [...] *Nie jest to niczym zaskakującym biorąc pod uwagę fakt, że metody naukowe często odzwierciedlają style myślenia obecne w życiu codziennym.* [...] *Ludzie wybierają prostsze z dwóch wyjaśnień z wielu powodów: ponieważ jest bardziej eleganckie, ponieważ jest łatwiejsze do zrozumienia lub do zapamiętania, bądź też ponieważ łatwiej je sprawdzić.*

Najczęściej przyjmowaną miarą prostoty modelu badanego zjawiska jest liczba jego parametrów (Keuzenkamp, McAleer i Zellner 2001: 3). Vladimir Vapnik, jeden ze światowych autorytetów w dziedzinie badań nad sztuczną inteligencją, sformułował to w postaci praktycznej dyrektywy (2006: 448)

> *wybierz funkcję z najmniejszą liczbą parametrów, która wyjaśnia obserwowane fakty.*

Sformułowanie to najlepiej chyba oddaje logikę konstruowania wyjaśnień przez badaczy, którzy do analizy zjawisk posługują się tablicami. Każda tablica posiada bowiem cechę zwaną **liczbą stopni swobody** (podrozdział 3.2). Okre-

[1] Niektóre środowiska stworzyły swoje własne odpowiedniki brzytwy Ockhama. W kręgach marketingowych popularna jest zasada KISS (*keep it simple, stupid*), czyli *nie komplikuj, głupku*. Sprawdza się ona szczególnie podczas przygotowywania prezentacji. Wśród polskich twórców oprogramowania komputerowego zasada KISS doczekała się nawet swojego odpowiednika w postaci akronimu BUZI (bez udziwnień zapisu, idioto).

śla ona liczbę wolnych parametrów, które można wykorzystać w celu opisania czy wyjaśnienia liczebności w polach tablicy otrzymanej w wyniku badania. Przypomnijmy, że gdy marginesy tablicy traktujemy jako zewnętrzne wobec badanego zjawiska, to jako liczbę stopni swobody przyjmuje się wielkość $(w - 1) * (k - 1)$, gdzie w oraz k oznaczają liczbę wierszy i kolumn. Wykorzystując je **wszystkie** można **dokładnie** odtworzyć obserwowane liczebności. Cała sztuka tkwi jednak w tym, aby liczebności te odtworzyć z akceptowalną precyzją, zaś stopni swobody zużyć jak najmniej. Kryterium minimalizacji liczby wykorzystanych stopni swobody utożsamia się z prostotą proponowanego modelu i preferuje w stosunku do sytuacji, w których bardziej złożony model wyjaśnia obserwowane zjawisko w większym stopniu (Fienberg 1977: 47).

Hegemonię tak rozumianej zasady prostoty najłatwiej prześledzić w obszarze zastosowań metod modelowania log-liniowego. Ograniczymy się do przykładu tablic, w których zakłada się odpowiedniość między kategoriami obu cech. Do grupy tej zalicza się tablice krzyżujące ze sobą cechy analogicznie zdefiniowane w wypadku par osób. Na przykład, wiek mężczyzny i kobiety w momencie zawarcia małżeństwa, wykształcenie męża i żony czy zawód ojca i syna. Omawiane tablice mają tyle samo wierszy i kolumn, czyli są kwadratowe.

Wyniki badań dowodzą, że liczebności w polach leżących na przekątnej głównej tego rodzaju tablic okazują się szczególnie wysokie, co interpretuje się jako homogamię małżeńską, dziedziczenie zasobów materialnych, transfer pozycji społecznej lub zawodowej z pokolenia na pokolenie czy też w inny podobny sposób, w zależności od charakteru cech uwzględnionych w tablicy. Zaproponowano w związku z tym, aby liczebności na przekątnej głównej wyjaśniać dokładnie, zaś co do pozostałych pól tablicy przyjąć, że ich wielkości wyznacza zasada niezależności. Tego rodzaju model nazwano quasi-niezależnością (Goodman 1965). Utworzenie modelu quasi-niezależności wymaga „poświęcenia" tylu stopni swobody, ile jest pól na przekątnej tablicy.

Okazuje się jednak, że model quasi-niezależności nie daje zadowalających szacunków dla liczebności w polach leżących poza przekątną główną (Duncan 1979; Hout 1983: 21–22). Wymaga to uzupełnienia modelu o reguły wyznaczania liczebności w tych polach tablicy. W tym miejscu nie będę prezentował wszystkich rozwiązań, jakie w tym zakresie zaproponowano (Goodman 1979). Ich przegląd można znaleźć w książce Domańskiego i Przybysza (2007: 70–87). Aby wyjaśnić zasadę prostoty odwołam się wyłącznie do jednego z nich, nazywanego modelem topologicznym (Featherman i Hauser 1978: 147–159; Domański i Przybysz 2007: 86–87). Właściwie nie jest to jeden model, lecz pewna klasa modeli. W ich wypadku badacz może świadomie decydować o tym, które pola lub które grupy pól mają specyficzny charakter, przez co wy-

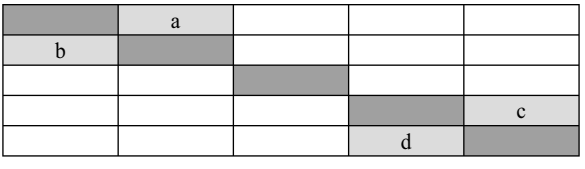
jaśnienie liczebności w tych polach warte jest poświęcenia jednego lub nawet kilku stopni swobody.

Jednym z modeli topologicznych jest tak zwany model **przeciwległych wierzchołków** (ang. *corners model*; Hout 1983: 23–25). Jego struktura przedstawiona została na rycinie 6.1. Ciemnoszarym kolorem oznaczono pola przekątnej głównej tablicy, które, jak przyjęliśmy wcześniej, są szacowane osobno. Oprócz nich wyodrębniono cztery dodatkowe pola, oznaczone literami a, b, c i d. Pola a i b tworzą wraz z dwoma przylegającymi do nich polami przekątnej głównej jeden z wierzchołków, zaś pola c i d wyznaczają drugi z wierzchołków w przeciwległym krańcu tablicy. Włączenie do modelu dodatkowych czterech pól odpowiada hipotezie, że liczebności w polach przylegających do krańców przekątnej głównej są z pewnego powodu większe. Na przykład, gdy tablica przedstawia wiek mężczyzny i kobiety w momencie zawarcia małżeństwa, to wysokie liczebności w polach oznaczonych a i b odpowiadają przekonaniu, że wśród osób najmłodszych szczególnie często zawierane są małżeństwa z osobami z sąsiedniej kategorii wieku. Analogicznie, wysokie liczebności w polach c i d odpowiadają temu samemu zjawisku w wypadku osób najstarszych. W modelu wierzchołków każde z czterech dodanych pól można potraktować osobno, dopasowując jego liczebność dokładnie do obserwowanej. Wtedy model zabiera dodatkowe 4 stopnie swobody. Można też przyjąć, że efekt zwiększonej liczebności jest taki sam w wypadku wszystkich 4 pól. Wtedy model wymaga tylko jednego dodatkowego stopnia swobody. Możliwe są też założenia pośrednie. Na przykład, że efekty a i b są jednakowe oraz że jednakowe są efekty c i d (symetria w obrębie wierzchołków). Taki z kolei model zmniejszy liczbę stopni swobody o 2.

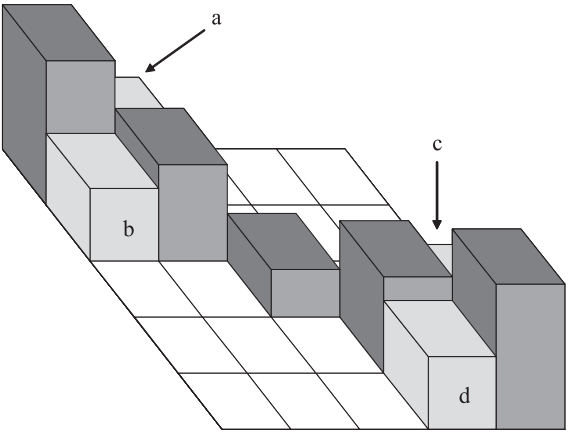
Model przeciwległych wierzchołków okazuje się niekiedy wyraźnie lepiej dopasowany do danych w porównaniu z modelem zakładającym jedynie specyfikę pól leżących na przekątnej głównej (Featherman i Hauser 1978: 152; Hout 1983: 23–25). Może okazać się więc zasadne poświęcenie dodatkowych stopni swobody na wyjaśnienie specyfiki pól tworzących wierzchołki tablicy. Oczywiście pod warunkiem, że dla pól tych można wskazać substantywny powód, dla którego należy spodziewać się w nich zwiększonych liczebności. Na przykład, badając wiek mężczyzn i kobiet zawierających małżeństwo, można sformułować hipotezę, że przestrzeń wyborów osób najmłodszych i najstarszych jest zawężona w stosunku do osób w średnim wieku. Za zasadnością tej hipotezy przemawiają dwa argumenty. Pierwszym jest efekt podłogi – w wypadku osób najmłodszych oraz efekt sufitu – w wypadku osób najstarszych. Powód ten nazwijmy **strukturalnym** i wyjaśnijmy na przykładzie rozpatrywanej w rozdziale 4 tablicy zawierającej rzeczywiste dane dotyczące wieku osób zawierających małżeństwo (tabela 4.5).

Rycina 6.1
Struktura przykładowego modelu topologicznego
Model wierzchołków dla tablicy 5 x 5

(a) rzut na płaszczyznę



(b) widok przestrzenny



Mężczyźni w wieku 20–24 lata mogą wziąć za żonę kobietę młodszą o jeden przedział wieku, nie mogą zaś zawrzeć małżeństwa z kobietą młodszą o dwa przedziały – gdyż nie zezwala na to prawo. Efekt podłogi polega więc na tym, że najniższy uwzględniony w tablicy przedział wieku kobiet zawiera wszystkie możliwości zawarcia małżeństwa z kobietą młodszą przez mężczyznę w wieku 20–24 lata. Owo ograniczenie powoduje, że należy spodziewać się zwiększonej liczby małżeństw dla tej kombinacji wieku mężczyzn i kobiet – co znajduje zresztą potwierdzenie w danych (najlepiej to widać w części [7] tabeli 4.5). Z kolei efekt sufitu w rozpatrywanej tabeli jest skutkiem sposobu grupowania wieku. Osoby w wieku ponad 60 lat zostały bowiem zgrupowane w jedną kategorię. Gdy więc kobieta w wieku 55–59 lat wychodzi za mąż za mężczyznę starszego, to jest on zawsze zaliczony do kategorii „60 lub więcej lat". A ponieważ kategoria ta obejmuje więcej niż 5 roczników demograficznych, to należy spodziewać się zwiększonej liczby małżeństw w polu odpowiadającym kobietom w wieku 55–59 lat i mężczyznom w wieku 60 lub więcej lat.

Drugi z powodów zwiększonych liczebności w polach tworzących wierzchołki tabeli ma charakter **psychologiczny**. Osoby najmłodsze większą część czasu spędzają w grupach rówieśniczych. Tym samym w ich świadomości pojawia się dystans wobec osób starszych. Osób takich znają mało, przez co nie wiedzą, jak wyglądać może wspólne życie z taką osobą. Partnera poszukują więc głównie wśród swoich rówieśników, co powinno skutkować zwiększeniem liczebności w polach wierzchołka tablicy obejmującego małżeństwa między osobami najmłodszymi. Z kolei osoby najstarsze mogą unikać zawierania małżeństw z osobami o wiele lat młodszymi z obawy, że nie podołają oczekiwaniom tych osób. Wraz z wiekiem możliwości funkcjonowania w wielu sferach ulegają ograniczeniu, co u ludzi starszych skutkować może ostrożnością i poszukiwaniem partnera w zbliżonym wieku.

Oba rodzaje barier, zarówno strukturalne, jak i psychologiczne, nie występują natomiast w wypadku ludzi w średnim wieku. Widać to wyraźnie w centralnych polach tabeli 4.5, gdzie nawet liczebności na przekątnej głównej nie odbiegają zbytnio od niezależności. W sumie można powiedzieć, że gdyby chcieć zastosować omawiany model wierzchołków do wyjaśnienia związku między wiekiem mężczyzn i kobiet zawierających małżeństwa, to bylibyśmy w stanie zgromadzić argumenty, że kształt związku jest zbliżony do przedstawionego na rycinie 6.1.

Nie bez powodu tak dużo uwagi poświęciliśmy kwestii mechanizmów, które uzasadniałyby adekwatność proponowanego modelu. Filozofowie nauki zgodnym chórem twierdzą bowiem, że zasada prostoty nie przekłada się na wskazówki co do wyboru teorii wyjaśniającej obserwowane fakty (Gernert 2009). Tymczasem w nauce i w badaniach pokutuje dość naiwne przekonanie,

że postępowanie zgodne z brzytwą Ockhama gwarantuje samo przez się uzyskanie wartościowych wyjaśnień (Keuzenkamp i inni 2001: 8–10).

Dlatego gdy brak nam wiedzy pozwalającej sformułować dobre uzasadnienie dla najprostszego modelu, po prostu zrezygnujmy z niego, nawet gdy jest dobrze dopasowany do wyników badania. Zastosujmy inny model – może bardziej złożony, a może prostszy lecz gorzej dopasowany do danych – lecz zarazem taki, w którym jesteśmy w stanie wskazać substantywne uzasadnienie dla każdego wprowadzonego parametru. Zwiększy to szanse, że zasięg sformułowanych wniosków nie ograniczy się do wyników konkretnego badania, lecz rzeczywiście dostarczy wiedzy na temat badanego zjawiska.

6.2 Wybór miary dystansów między profilami

Rozdział 5 zakończyliśmy wnioskiem, że struktury wielu tablic nie da się opisać za pomocą wskaźników dystansów między wierszami lub kolumnami. Można to zrobić wtedy, gdy jedna z cech ma tylko dwie kategorie, gdyż dystanse między profilami drugiej cechy są wówczas określone za pomocą różnic między średnimi. W tym podrozdziale będziemy kontynuować te rozważania. Odpowiemy na pytanie, w jaki sposób strukturę związku w tablicy opisać za pomocą dystansów między wierszami i kolumnami niezależnie od jej rozmiarów. Innymi słowy, czy analizując zróżnicowanie profili jesteśmy w stanie w zadowalającym stopniu opisać badane zjawisko niezależnie od tego, jakie rozmiary ma tablica.

Okazuje się, że do problemu łatwiej podejść modyfikując dotychczasowy sposób definiowania dystansów między wierszami bądź kolumnami tablicy. W rozdziale 5 posługiwaliśmy się w tym celu wskaźnikiem różnic między profilami (*dissimilarity index*), zdefiniowanego jako połowa sumy wartości bezwzględnych różnic odsetków w polach porównywanych profili (wzory 5.1 i 5.2). Tak określone dystanse w dość naturalny sposób odzwierciedlają intuicje związane z porównywaniem wierszy lub kolumn tablicy. Niestety jednak, z powodów wymienionych w ramce 4.2, dystanse te nie spełniają szeregu własności formalnych, co utrudnia przedstawienie ich wzajemnych relacji oraz określenie związków z innymi wielkościami w tablicy. Dlatego w większości metod analizy tablic przyjęto funkcję dystansu opartą nie na wartościach bezwzględnych różnic między wielkościami w porównywanych polach, lecz na kwadratach tych różnic. Oznaczmy tę funkcję dystansu symbolem δ (delta). W wypadku porównywania profili w dwóch wierszach tablicy, oznaczonych i_1 oraz i_2, wielkość dystansu między nimi jest równa (Greenacre 1994: 11–12; Blasius i Greenacre 2006b: 18)

$$\delta_{i_1 i_2} = \sqrt{\sum_{j=1}^{k} \frac{(pw_{i_1 j} - pw_{i_2 j})^2}{pb_j}} \tag{6.1}$$

Analogiczny wzór dla dystansu między profilami dwóch kolumn tablicy wygląda następująco

$$\delta_{j_1 j_2} = \sqrt{\sum_{i=1}^{w} \frac{(pk_{ij_1} - pk_{ij_2})^2}{pa_i}} \tag{6.2}$$

Na pierwszy rzut oka różnica w stosunku do wskaźników wcześniej stosowanych wydaje się dość zasadnicza. W praktyce jednak oba rodzaje wskaźników dystansu prowadzą do podobnych wniosków, o czym najlepiej przekonać się rozważając przykład.

W rozdziale 5 obliczaliśmy wielkości wskaźników różnicy między profilami dla tabeli przedstawiającej związek między wykształceniem ojca a wyborem szkoły ponadgimnazjalnej. W części [1] tabeli 6.1 przytoczone zostały uprzednio obliczone wskaźniki różnic między profilami wyboru szkoły dla kategorii wykształcenia ojca. W części [2] zamieszczone zostały wielkości dystansów między tymi samymi profilami obliczone według wzoru (6.1). Przede wszystkim zwraca uwagę, że te ostatnie wielkości są generalnie mniejsze. Wielkości wskaźników różnic profili wyrażały się w procentach i miały prostą interpretację jako suma nadwyżek odsetków w jednym profilu w stosunku do drugiego. Wartości wprowadzonej miary dystansu nie mają tak bezpośredniej interpretacji, przez co badacz może wyłącznie wierzyć, że odzwierciedlają różnice między profilami wierszy lub kolumn w takim sensie, jak chcielibyśmy to rozumieć.

Łatwo się jednak przekonać, że wielkości wprowadzonej miary dystansu dostarczają podobnego obrazu relacji między wierszami tablicy, jak wskaźniki różnic między profilami. W części [1] tabeli 6.1 największa różnica dotyczy profili szkół, do których uczęszczają dzieci ojców z wykształceniem wyższym i podstawowym. Różnica ta wynosi 67,28 procenta. W wypadku wprowadzonej obecnie miary dystansu największa jej wartość (1,416) również odpowiada porównaniu tych samych kategorii wykształcenia ojców. Podobnie jest w wypadku najmniejszej różnicy. Nie wszystkie jednak relacje są ściśle zgodne. Wielkość stosowanego wcześniej wskaźnika jest nieco większa przy porównywaniu ze sobą kategorii ojców o wykształceniu średnim i wyższym (34,43), niż gdy porównujemy ojców o wykształceniu podstawowym i średnim (32,85). W wypadku wprowadzonej obecnie miary jest odwrotnie. Pocieszające jest jedynie to, że w obu wypadkach zestawiane wielkości nie różnią się wiele od siebie (rząd ich wielkości jest podobny).

Tabela 6.1
Wartości dwóch miar dystansu między profilami wybranej szkoły ponadgimnazjalnej w kategoriach uczniów o różnym wykształceniu ojca
Badanie PISA 2006

[1] wskaźniki różnic profili (w procentach)

	wykształcenie ojca				
wykształcenie ojca	wyższe	średnie	zasadnicze zawodowe	podstawowe	do ogółu uczniów
wyższe	0,00	34,43	58,49	67,28	46,45
średnie		0,00	24,05	32,85	12,02
zasadnicze zawodowe			0,00	13,27	12,04
podstawowe				0,00	20,83
do ogółu uczniów	46,45	12,02	12,04	20,83	0,00

[2] dystanse profili (w jednostkach niemianowanych)

	wykształcenie ojca				
wykształcenie ojca	wyższe	średnie	zasadnicze zawodowe	podstawowe	do ogółu uczniów
wyższe	0,000	0,711	1,189	1,416	0,944
średnie		0,000	0,527	0,836	0,297
zasadnicze zawodowe			0,000	0,354	0,246
podstawowe				0,000	0,539
do ogółu uczniów	0,944	0,297	0,246	0,539	0,000

Obliczono na podstawie odsetków uczniów poszczególnych rodzajów szkół w kategoriach wykształcenia ojca podanych w części [2] tabeli 5.1. Zamieszczone w części [1] wielkości wskaźników różnic liczono na podstawie wzoru (5.1), natomiast wielkości dystansów w części [2] na podstawie wzoru (6.1).

Do tego wszystkiego dodajmy, że wielkości wprowadzonej miary dystansu nie są addytywne. Brak tej pożytecznej własności sygnalizowaliśmy w podrozdziale 5.3, próbując dokonać projekcji wskaźników różnic w jeden wymiar. Omawiana miara dystansu również nie spełnia tego warunku. Dystans między profilami wyboru szkoły przez dzieci ojców o wykształceniu podstawowym i wyższym wynosi 1,416. Jeśli ten dystans spróbujemy uzyskać, sumując dystanse między sąsiadującymi ze sobą kategoriami wykształcenia ojców (pola zaznaczone części [2] tabeli 6.1 kolorem szarym), to otrzymamy

$$0{,}711 + 0{,}527 + 0{,}354 = 1{,}592$$

Suma dystansów składowych jest więc większa od dystansu policzonego bezpośrednio. Warunek addytywności nie jest spełniony, toteż dystansów nie można w sposób spójny sprowadzić do jednej osi.

O zaletach wprowadzonej miary dystansu stanowią natomiast jej własności. W tym miejscu przedstawię jedną z nich, pozostawiając omówienie kolejnych do dalszej części rozdziału. Wartości dystansów obliczać można nie tylko między profilami dwóch dowolnych wierszy lub dwóch dowolnych kolumn tablicy, lecz również porównując profil wiersza czy kolumny z profilem rozkładu brzegowego. Wtedy wzór (6.1) sprowadzić można do postaci

$$\delta_{i\bullet} = \sqrt{\frac{1}{a_i}\sum_{j=1}^{k}\frac{\left(n_{ij} - e_{ij}\right)^2}{e_{ij}}} \tag{6.3}$$

gdzie występująca pod znakiem pierwiastka suma jest składnikiem wzoru na obliczanie wskaźnika χ^2. Jeśli więc podniesiemy do kwadratu wielkości dystansów profili poszczególnych wierszy od rozkładu brzegowego, a następnie każdy z nich zważymy przez potencjał danego wiersza, czyli przez jego liczebność brzegową a_i, to w sumie otrzymamy wartość wskaźnika χ^2 dla całej tablicy.

$$\chi^2 = \sum_{i=1}^{w} a_i * \delta_{i\bullet}^2 \tag{6.4}$$

Analogiczna prawidłowość zachodzi również dla kolumn

$$\chi^2 = \sum_{j=1}^{k} b_j * \delta_{\bullet j}^2 \tag{6.5}$$

Obie własności powodują, że wielkości określone za pomocą wzorów (6.1) i (6.2) nazywane są **dystansami chi-kwadrat**.

Fakt, że dystanse profili wierszy oraz kolumn w stosunku do rozkładów brzegowych sumują się do χ^2 oznacza zarazem, że chi-kwadrat zdekomponować można między poszczególne wiersze bądź kolumny tablicy. Zostało to zilustrowane w częściach [1] i [2] tabeli 6.2. W wypadku profili szkół wybieranych przez uczniów o różnych poziomach wykształcenia ojca najbardziej specyficzni okazują się uczniowie, których ojcowie mają wykształcenie wyższe. Jest to przede wszystkim wynikiem dużej wartości delty. Chociaż uczniów tych jest stosunkowo niewielu, ich udział w odstępstwach od modelu niezależności przekracza 50 procent. W podrozdziale 5.2 specyfika tej kategorii również przykuła naszą uwagę. Stwierdziliśmy wtedy, że 88 procent kształcących się w liceach ogólnokształcących stanowi wynik wyraźnie odbiegający od uczniów o pozostałych poziomach wykształcenia ojca. Wniosek potwierdza się w tym miejscu, mimo że zastosowany wskaźnik ma odmienną naturę.

Interesująco wygląda również dekompozycja χ^2 ze względu na profile wykształcenia ojców dzieci uczęszczających do poszczególnych rodzajów szkół. Dystanse liceów ogólnokształcących oraz szkół zasadniczych pod względem

wykształcenia ojców uczniów odbiegają w podobnym stopniu od profilu wykształcenia ojców wszystkich uczniów z badanego rocznika. Jednakże uczniów wybierających licea jest więcej. Dlatego też kategoria ta wyjaśnia ponad połowę wielkości wskaźnika χ^2.

Tabela 6.2
Wskaźniki dekompozycji χ^2 dla związku między wykształceniem ojca a wyborem szkoły ponadgimnazjalnej przez dziecko
Badanie PISA 2006

[1] dekompozycja na podstawie miar dystansu profili wyboru szkoły w kategoriach wykształcenia ojca wobec profilu dla wszystkich uczniów

wykształcenie ojca	a_i	$\delta_{i\bullet}$	$a_i * \delta^2_{i\bullet}$	udział w %
wyższe	454	0,944	404,4	52,5
średnie	1173	0,297	103,4	13,4
zasadnicze zawodowe	2126	0,246	128,7	16,7
podstawowe	461	0,539	134,1	17,4
ogółem	4214		770,6	100,0

[2] dekompozycja na podstawie miar dystansu profili wykształcenia ojców w kategoriach uczniów uczęszczających do poszczególnych rodzajów szkół wobec profilu dla wszystkich uczniów

rodzaj szkoły	b_j	$\delta_{\bullet j}$	$b_j * \delta^2_{\bullet j}$	udział w %
liceum ogólnokształcące	1746	0,479	399,9	51,9
technikum	1739	0,283	139,7	18,1
zasadnicza zawodowa	729	0,563	231,0	30,0
ogółem	4214		770,6	100,0

[3] dekompozycja na podstawie kwadratów średnich wskaźników Quételeta $n * w^2_{ij}$

wykształcenie ojca	liceum ogólno-kształcące	technikum	szkoła zasadnicza zawodowa	suma	udział w %
wyższe	236,4	108,2	59,8	404,4	52,5
średnie	40,9	1,9	60,6	103,4	13,4
zasadnicze zawodowe	74,3	28,7	25,7	128,7	16,7
podstawowe	48,3	1,0	84,8	134,1	17,4
suma	399,9	139,7	231,0	770,6	100,0
udział w %	51,9	18,1	30,0	100,0	

W rozdziale 4 zaproponowałem, aby odstępstwa od modelu niezależności w poszczególnych polach tablicy mierzyć za pomocą średniego wskaźnika Quételeta. Wykazałem również, że suma kwadratów tego wskaźnika dla wszystkich

pól tablicy sumuje się do chi-kwadrat, dzięki czemu sumę odstępstw od modelu niezależności można zdekomponować na poszczególne wiersze i kolumny tablicy. Z części [3] tabeli 6.2 wynika, że w wypadku analizowanego związku dekompozycja ta jest identyczna jak otrzymana na podstawie omawianej miary dystansu. Wynik ten nie jest przypadkowy. Wielkość dystansu między profilem *i*-tego wiersza a rozkładem brzegowym może być bowiem przedstawiona jako funkcja wskaźników Quételeta (Mirkin 2001: 116)

$$\delta_{i\bullet} = \sqrt{\frac{n}{a_i} * \sum_{j=1}^{k} q_{ij}^2} \tag{6.6}$$

Podstawienie wyrażenia (6.6) do wzoru (6.4) pozwala wykazać, że χ^2 dla całej tablicy stanowi sumę komponentów o postaci

$$n * \sum_{j=1}^{k} q_{ij}^2 \tag{6.7}$$

określonych dla jej poszczególnych wierszy. Analogiczną dekompozycję przedstawić można dla kolumn tablicy.

Podsumujmy wady i zalety wprowadzonej miary dystansu między profilami. W porównaniu z wskaźnikami różnic jej wartości nie mają klarownej interpretacji. Nie można więc przełożyć ich na język badanych zjawisk. W zamian metoda dostarcza dwóch korzyści. Po pierwsze, dystanse między profilami posłużyć mogą jako kryterium dekompozycji wskaźnika χ^2. Po drugie, dystanse dają się wyrazić jako funkcja wskaźników Quételeta. Stwarza to pomost między analizą profili a metodami opartymi na porównaniu pól tablicy z modelem niezależności.

6.3 Tablica kanoniczna

W rozdziale 3 wspomniałem o tym, że jedno z podejść do analizy tablic polega na przyjęciu pewnego modelu tablicy jako referencyjnego i porównaniu wyników badania z tym modelem. Od tego momentu staram się przekonać Czytelnika, że podstawową rolę w tym kontekście pełni model niezależności, który bywa traktowany zarówno jako samodzielny punkt odniesienia dla analizowanego związku, jak również stanowi składnik innych modeli. Na strategii tej oparte między innymi jest modelowanie log-liniowe, w którym model niezależności rozpatruje się w pierwszej kolejności, po czym nadbudowuje na nim właściwy model mający wyjaśnić specyfikę pól niewyjaśnioną przez niezależność. Analogiczny tok rozumowania proponowałem w rozdziale 4.

Polegał on na identyfikacji modelu związku składającego się z kategorii pól w różny sposób odbiegających od niezależności.

Omawiane sposoby postępowania zakładają aktywną rolę badacza w poszukiwaniu modelu związku. Uzasadnione jest to zwłaszcza w wypadku tablic pełniących centralną rolę w ramach rozpatrywanego problemu. Na przykład, dla badaczy procesów międzypokoleniowej transmisji pozycji społecznej rolę taką pełni tablica krzyżująca wskaźnik pochodzenia – najczęściej zawód jednego z rodziców – z analogicznie zdefiniowanym wskaźnikiem pozycji osiągniętej przez badanego. Co jednak zrobić, gdy z różnych powodów tablica nie może być poddana tak drobiazgowej eksploracji? Na przykład, mamy nadmiar danych i chcemy wybrać z nich przede wszystkim to, co w trafny bądź wręcz spektakularny sposób ilustruje istotę badanego zagadnienia. Bądź brakuje wiedzy teoretycznej, która pozwoliłaby sformułować model związku w tablicy. Jedną z możliwości jest poszukiwanie pól najbardziej odbiegających od niezależności, na przykład za pomocą skorygowanych reszt, którą to strategię przedstawiliśmy w podrozdziale 4.10. Dostarcza ona jednak dość wyrywkowej wiedzy ograniczonej do niektórych pól tablicy, konfigurujących się niekiedy w dość przypadkowy układ.

Potrzeba utworzenia modelu związku dla całej tablicy wydaje się więc dość oczywista. Chodzi o model, który cechowałby się podobną uniwersalnością, jak model niezależności, a którym można byłoby posługiwać się niejako „w ciemno" w sytuacjach, gdy brakuje pomysłów, jaki model związku zastosować. Mimo swojej oczywistości, problem nie został dotychczas *explicite* sformułowany. Swoją drogą, ten swoisty fenomen z pewnością prędzej czy później stanie się przedmiotem rozważań filozofów nauki. Dopóki jednak wyjaśnienie takie nie zostanie przedstawione, badacze – lepiej czy gorzej – zmuszeni są radzić sobie z problemem sami. Jedną z możliwości, z której coraz częściej korzysta się w analizach tablic, stwarza zastosowanie metody korespondencji (rozdział 7). Metoda korespondencji zakłada istnienie tego rodzaju modelu, aczkolwiek jest on dość głęboko ukryty w założeniach metody. Dlatego większość badaczy stosujących metodę korespondencji nie zdaje sobie sprawy z faktu, jaki model związku w rzeczywistości przyjmują jako punkt odniesienia.

Otóż modelem tym jest **tablica kanoniczna**. Zacznijmy może od intuicji związanych z jej kształtem. Na rycinie 6.2 przedstawiłem ją w dwóch wersjach, różniących się parametrem, o którym powiem dalej. Płaszczyzna obrazująca liczebności w tablicy w dwóch przeciwległych narożnikach wygięta jest w górę. Odpowiadają one tym fragmentom tablicy, których specyfika polega na wysokich liczebnościach. W dwóch pozostałych narożnikach płaszczyzna wygięta jest w dół. Odpowiada to obszarom tablicy, w których

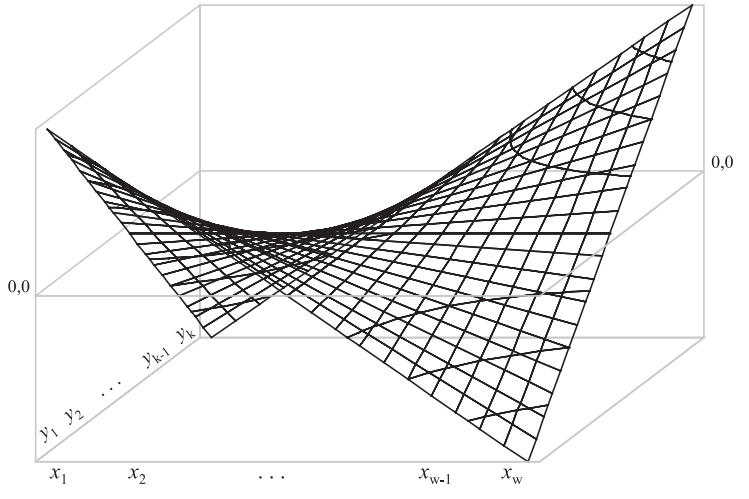
liczebności są mniejsze, niż należałoby się spodziewać. Charakterystyczną cechą prezentowanej płaszczyzny jest punkt przegięcia, który znajduje się w jej środku. Płaszczyzna ta przypomina przez to „siodło" (ang. *saddle*) i tak jest też niekiedy nazywana.

W podrozdziale 6.1 przedstawiłem przykład modelu związku w tablicy, który nazwałem modelem przeciwległych wierzchołków (rycina 6.1). Przypomina on swoim kształtem tablicę kanoniczną. Największe liczebności występują w dwóch przeciwległych narożnikach. Charakterystyczny jest też punkt przegięcia. Między modelem przeciwległych wierzchołków, a tablicą kanoniczną istnieje jednak dość zasadnicza różnica, gdyż reprezentują odmienne podejścia do analizy wyników badań. Pierwszy z modeli ma status hipotezy badawczej. Jego kształt jest pochodną wiedzy o badanej rzeczywistości. W podrozdziale 6.1 wymieniłem strukturalne i psychologiczne przyczyny, dla których należy spodziewać się zwiększonych liczebności w narożnikach tablicy przedstawiającej wiek mężczyzny i kobiety zawierających małżeństwo. Omawianą strategię stosuje się w modelowaniu log-liniowym. Zastosowany model odzwierciedla wiedzę bądź założenia przyjęte przez badacza.

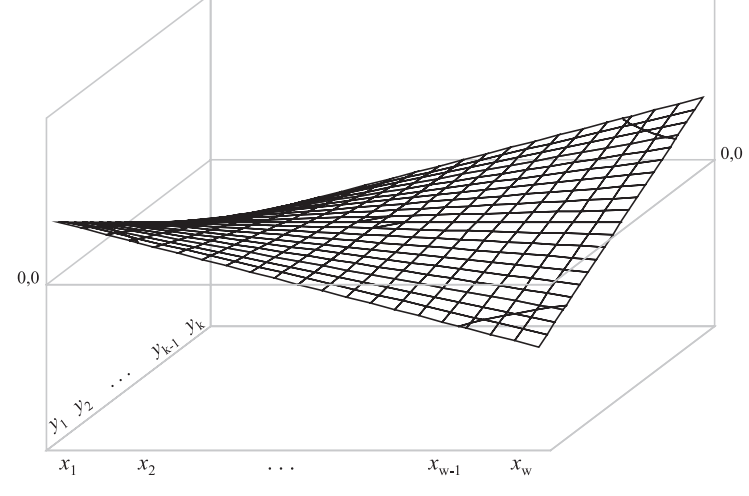
Tablica kanoniczna pełni odmienne funkcje. Stosowana jest w metodach eksploracyjnych, w których *a priori* nie przyjmuje się żadnych założeń co do kształtu badanego związku. Stosuje się ją jako model referencyjny zarówno wtedy, gdy model przeciwległych wierzchołków ściśle opisuje liczebności w tablicy, jak też wtedy, gdy tablica przypomina młody sosnowy las, to jest gdy pola o dużych liczebnościach występują na przemian z polami o liczebnościach niewielkich. Celem tego podejścia nie jest bowiem osiągnięcie dobrego dopasowania modelu do danych. Celem jest sprawdzenie, na ile wyniki badania wpisują się w model kanoniczny. Jeśli dobrze, to problem mamy praktycznie rozwiązany. Znamy bowiem kształt związku i uwagę możemy skupić na sformułowaniu substantywnego wyjaśnienia, dlaczego zjawisko wygląda tak, a nie inaczej. Natomiast gdy model kanoniczny okaże się źle dopasowany do danych, to mimo wszystko dostarczy pewnych korzyści. Przede wszystkim pozwoli ocenić, czy w ogóle da się utworzyć jakikolwiek model. Ponadto dostarczy wskazówek, w których fragmentach tablicy szukać pól specyficznych. Są to wystarczające informacje, aby podjąć decyzję, czy tablicą warto się dalej zajmować za pomocą bardziej wyrafinowanych metod.

Rycina 6.2
Graficzny obraz kształtu tablicy kanonicznej

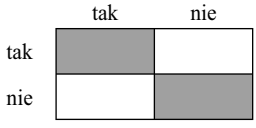
[1] ilustracja wysokiej wartości kanonicznej korelacji



[2] Ilustracja niskiej wartości kanonicznej korelacji



Heurystyczne zalety tablicy kanonicznej biorą się z jej prostoty. Jej kształt mówi jedynie tyle, że zjawisko ma strukturę wzajemnej relacji. Można to przedstawić na przykładzie tabeli obejmującej dwie cechy dychotomiczne.



„Tak" na pytanie pierwsze oznacza „tak" na pytanie drugie. Jeśli zaś „nie", to „nie". Jest to bodajże najprostszy konceptualnie model związku cech w tablicy.

6.4 Utworzenie tablicy kanonicznej

Oznaczmy tablicę kanoniczną dużą literą *C*, zaś liczebności w jej poszczególnych polach jako c_{ij}. Liczebności te są iloczynami czterech wielkości

$$c_{ij} = r\, e_{ij}\, x_i\, y_i \tag{6.8}$$

gdzie *r* jest stałą wspólną dla wszystkich pól, zwaną korelacją kanoniczną. Od wielkości tej stałej zależy na ile płaszczyzna obrazująca tablicę kanoniczną „wyciąga się" w górę lub w dół w narożnikach (rycina 6.2). Z kolei e_{ij} to dobrze nam znana liczebność w modelu niezależności. Odzwierciedla ona „wagę" danego pola. Jeśli jest większa, to szacowana liczebność c_{ij} w tablicy kanonicznej również będzie większa w odpowiedniej proporcji. Wielkości x_i oraz y_j to współrzędne kategorii odpowiadających wierszowi i kolumnie. Nazwiemy je współrzędnymi kanonicznymi. Gdy obie współrzędne przybierają wartości dodatnie bądź przybierają wartości ujemne, to płaszczyzna kanoniczna zaginać się będzie ku górze (rycina 6.2). Gdy natomiast jedna jest dodatnia, zaś druga ujemna, to płaszczyzna zaginać się będzie ku dołowi. Wykresy na rycinie 6.2 sporządzone zostały według formuły (6.8), z tym że dla przejrzystości przyjąłem wszystkie wartości e_{ij} jako jednakowe. Gdybym tego nie zrobił, to płaszczyzna miałaby wybrzuszenia w górę lub w dół, odpowiadające większym lub mniejszym wagom poszczególnych pól. Przyjęte założenie odpowiada sytuacji, w której rozkłady stanowiące marginesy obu cech byłyby równomierne. Wtedy bowiem wielkości e_{ij} dla wszystkich pól tablicy byłyby jednakowe. W praktyce nie ma jednak co liczyć na uzyskanie tak regularnego kształtu tablicy kanonicznej, jak to zostało przedstawione na rycinie 6.2.

Ideę tablicy kanonicznej można również wyrazić odwołując się do tabliczki mnożenia. Podstawmy do wzoru (6.8) wielkość e_{ij} jako $a_i b_j / n$ a następnie uporządkujemy wyrażenie nadając mu następującą postać

$$c_{ij} = \frac{r}{n}\left(x_i a_i\right) * \left(y_j b_j\right) \tag{6.9}$$

Czynnik *r*/*n* jest wspólny dla wszystkich pól, toteż pomińmy jego rolę. Wtedy liczebności tablicy kanonicznej interpretować można jako iloczyny pewnych

wielkości czy potencjałów, odpowiadających wierszom i kolumnom (we wzorze 6.9 wyodrębnionych nawiasami). Wielkości potencjałów kategorii cechy w wierszach i cechy w kolumnach zależą od ich liczebności brzegowych a_i i b_j – podobnie jak to ma miejsce w modelu niezależności (zob. 3.4). W stosunku do tego ostatniego w tablicy kanonicznej potencjał kategorii korygowany jest dodatkowo przez wartości współrzędnych kanonicznych x_i oraz y_j. Wkład kategorii obu cech do tablicy kanonicznej zależy więc nie tylko od tego, jak są liczne, lecz również od tego na ile „odbiegają" w górę lub w dół (ujemna bądź dodatnia wartość współrzędnej kanonicznej) wobec normy, którą stanowi model niezależności.

Analogie do tabliczki mnożenia prowadzą do przydatnych intuicji. Przypuśćmy, że pierwsza kategoria cechy umieszczonej w wierszach jest dwa razy mniej liczna od kategorii drugiej ($a_2 = 2a_1$). Pomimo tego obie kategorie mogą mieć jednakowy wkład w wielkości liczebności tablicy kanonicznej. Wystarczy bowiem, aby współrzędna kanoniczna odpowiadająca pierwszej kategorii była dwa razy większa niż współrzędna odpowiadająca drugiej kategorii ($x_1 = 2x_2$). Omawiana prawidłowość odpowiada sytuacjom występującym w tablicach konstruowanych na podstawie wyników badań, gdy w wierszach tablicy umieszczona zostaje cecha porządkująca badanych w pewnym wymiarze. Przyjmijmy, że cechą tą jest status społeczny. Należy się wówczas spodziewać, że pierwszej kategorii – usytuowanej w danym wymiarze najwyżej – odpowiadała będzie mniejsza liczebność niż kategorii drugiej, usytuowanej niżej. Aby elita była elitą, musi być mało liczna. Nie oznacza to jednak, że z tej racji musi mieć mniejszy wkład w zjawisko nierówności społecznych zobrazowane w tablicy. O ile faktycznie generuje duże odstępstwa od modelu niezależności – czyli od stanu równych szans – co znajduje wyraz w wyższych liczebnościach c_{ij} tablicy kanonicznej, to ujawni się to w postaci względnie wysokiej wartości współrzędnej kanoniczej. Ta ostatnia wielkość rekompensuje bowiem w modelu niewielką liczebność elity.

Przejdźmy obecnie do omówienia sposobu obliczenia parametrów tablicy kanonicznej. Wszystkie niewiadome w formule (6.8), to jest współczynnik korelacji kanonicznej r, a także współrzędne $x_1, x_2, ..., x_w$ oraz $y_1, y_2, ..., y_k$, szacuje się poprzez minimalizację kryterium

$$\sum_{i=1}^{w}\sum_{j=1}^{k}\frac{(d_{ij}-c_{ij})^2}{e_{ij}} \tag{6.10}$$

Z postaci kryterium (6.10) wynika, że liczebności c_{ij} w polach tablicy kanonicznej dobierane są w taki sposób, aby w **maksymalnym stopniu odtworzyć różnice** d_{ij} między liczebnościami obserwowanymi n_{ij} a liczebnościami e_{ij} pól modelu niezależności.

Na podstawie formuły (6.8) można też wyjaśnić, według jakich zasad ustalane są współrzędne x_i oraz y_j. Jeżeli w danym polu tablicy uzyskana w badaniu liczebność n_{ij} byłaby wyraźnie większa od liczebności e_{ij} – oszacowanej na mocy modelu niezależności – to odpowiadające temu polu współrzędne kanoniczne x_i oraz y_j muszą przyjąć odpowiednio duże wartości, tak aby liczebność c_{ij} pola tablicy kanonicznej była pod względem wielkości zbliżona do różnicy d_{ij}. Z kolei jeśli różnica d_{ij} byłaby niewielka, to współrzędne kanoniczne x_i oraz y_j powinny być bliskie zeru, tak aby ich iloczyn utworzył niewielką liczebność c_{ij}. Z kolei, gdy liczebność otrzymana w badaniu byłaby niższa od liczebności obliczonej na podstawie modelu niezależności, to jedna z wielkości x_i lub y_j powinna być ujemna. Ujemna jest bowiem wartość odtwarzanej różnicy d_{ij}.

Liczebności c_{ij} tablicy kanonicznej są więc dopasowywane nie do obserwowanych liczebności n_{ij}, lecz do różnic między tymi liczebnościami, a wielkościami e_{ij} w polach modelu niezależności. Tablica kanoniczna nie odtwarza więc liczebności uzyskanych w wyniku badania, lecz odtwarza je dopiero złożenie tablicy kanonicznej z modelem niezależności. Oznaczmy tablicę stanowiącą to złożenie literą M i nazwijmy **modelem kanonicznym**. Liczebności pól tego modelu wyrażają się wzorem

$$m_{ij} = e_{ij} + c_{ij} \tag{6.11}$$

Wróćmy do wcześniej omawianego związku między wyborem szkoły ponadgimnazjalnej a wykształceniem ojca i dla przykładu tego skonstruujmy tablicę kanoniczną. W tabeli 6.3 przedstawiono wartości współrzędnych dla obu cech oraz wartość korelacji kanonicznej. Na ich podstawie obliczono liczebności tablicy kanonicznej, podane w części [3] tabeli 6.4. Na przykład liczebność w polu odpowiadającemu uczniom liceów ogólnokształcących, których ojcowie mają wykształcenie wyższe, obliczono według wzoru (6.8) w podany niżej sposób

$$0{,}412 * 188 * 2{,}249 * 1{,}158 = 202$$

Uzyskanemu rozwiązaniu warto poświęcić parę słów komentarza. Model niezależności okazał się dość dobrze dopasowany do wyjściowej tablicy liczebności otrzymanych w badaniu (zostały one przedstawione w tabeli 5.1). Zamieszczone w części [2] tabeli 6.4 różnice między tymi liczebnościami a wielkościami pól w modelu niezależności należy w sumie ocenić jako niezbyt duże w kontekście łącznej liczby badanych osób. Wystarczyłoby, aby 704 uczniów (czyli 17%) zmieniło w odpowiedni sposób swoje decyzje co do wyboru szkoły, a mielibyśmy do czynienia z sytuacją niezależności. W sumie więc model niezależności poprawnie klasyfikuje 83 procent uczniów do właściwych pól tabeli.

Tabela 6.3
Parametry kanoniczne dla związku między wykształceniem ojca a wyborem szkoły
Badanie PISA 2006

cecha	kategoria	oznaczenie	współrzędne kanoniczne	
			standaryzowane	znormalizowane
wykształcenie ojca	wyższe	x_1	2,249	100,0
	średnie	x_2	0,669	53,9
	zasadnicze zawodowe	x_3	-0,593	17,2
	podstawowe	x_4	-1,182	0,0
rodzaj szkoły	liceum ogólnokształcące	y_1	1,158	100,0
	technikum	y_2	-0,628	26,6
	zasadnicza zawodowa	y_3	-1,275	0,0
	korelacja kanoniczna	r	0,412	

Na podstawie wielkości podanych w tabeli 5.1.

Największa nadwyżka wobec modelu niezależności – równa 211 osób – wystąpiła w wypadku uczniów, którzy wybrali licea ogólnokształcące, a których ojcowie mają wykształcenie wyższe. Tablica kanoniczna dość skutecznie odtwarza tę nadwyżkę, alokując do tego pola 202 osoby. Przekłada się to na wysokie współrzędne kanoniczne zarówno dla kategorii ojców z wykształceniem wyższym, jak też dla liceów ogólnokształcących. Drugą dość precyzyjnie odtworzoną wielkością jest niedobór w liceach ogólnokształcących uczniów, których ojcowie mają wykształcenie zasadnicze. Niedobór ten wynosi 256 osoby, z czego tablica kanoniczna wyjaśnia 250. Aby zlikwidować ten niedobór, procedura przypisała ujemną współrzędną kanoniczną kategorii ojców o wykształceniu zasadniczym. Iloczyn ujemnej wartości z dodatnią współrzędną dla liceów ogólnokształcących prowadzi do ujemnej liczebności w tablicy kanonicznej (wzór 6.8).

Ujemną współrzędną otrzymały również technika. W szkołach tych występuje bowiem niedobór uczniów, których ojcowie mają wykształcenie wyższe, zaś kategorii tej przypisana została współrzędna dodatnia. Występuje też spora nadwyżka uczniów, których ojcowie mają wykształcenie zasadnicze – a z kolei tym ostatnim przypisano współrzędną ujemną. Ujemna współrzędna dla techników pozwala więc poprawnie oszacować niedobory i nadwyżki uczniów o różnym poziomie wykształcenia ojca.

W części [4] przedstawione zostały liczebności modelu związku opartego na tablicy kanonicznej, określonego za pomocą wzoru (6.11). Warto zwrócić uwagę, że liczebności pól modelu kanonicznego sumują się do rozkładów brzegowych wyjściowej tablicy. Model kanoniczny stosować więc można wtedy,

Tabela 6.4
Liczebności modeli związku między wykształceniem ojca a wyborem szkoły
Badanie PISA 2006

wykształcenie ojca	rodzaj szkoły: liceum ogólno-kształcące	technikum	zasadnicza zawodowa	ogółem
	[1] liczebności modelu niezależności e_{ij}			
wyższe	188	187	79	454
średnie	486	484	203	1173
zasadnicze zawodowe	881	877	368	2126
podstawowe	191	190	80	461
ogółem	1746	1739	729	4214
	[2] różnice $d_{ij} = n_{ij} - e_{ij}$			
wyższe	211	-142	-69	0
średnie	141	-30	-111	0
zasadnicze zawodowe	-256	159	97	0
podstawowe	-96	14	82	0
ogółem	0	0	0	0
Minimalna liczba przesunięć: 704; odsetek osób poprawnie sklasyfikowanych: 83,3%				
	[3] liczebności tablicy kanonicznej c_{ij}			
wyższe	202	-109	-93	0
średnie	155	-84	-71	0
zasadnicze zawodowe	-250	135	115	0
podstawowe	-108	58	50	0
ogółem	0	0	0	0
	[4] liczebności modelu kanonicznego $m_{ij} = e_{ij} + c_{ij}$			
wyższe	390	78	-14	454
średnie	641	400	132	1173
zasadnicze zawodowe	631	1012	482	2126
podstawowe	83	249	129	461
ogółem	1746	1739	729	4214
	[5] różnice niewyjaśnione przez model kanoniczny ($n_{ij} - m_{ij}$)			
wyższe	9	-33	24	0
średnie	-14	54	-40	0
zasadnicze zawodowe	-6	24	-17	0
podstawowe	12	-45	33	0
ogółem	0	0	0	0
Minimalna liczba przesunięć: 155; odsetek osób poprawnie sklasyfikowanych: 96,3%				

Obliczono na podstawie liczebności podanych w tabeli 5.1

gdy zakłada się nadrzędną rolę marginesów wobec zjawiska przedstawionego w tablicy. W wypadku rozpatrywanego związku liczebności modelu kanonicznego okazały się dość dobrze dopasowane do danych wyjściowych. W części [5] podano różnice w tym zakresie. W sumie, do przeniesienia do innych pól tablicy pozostało jeszcze 155 uczniów. Oznacza to, że model kanoniczny, stanowiący złożenie niezależności i tablicy kanonicznej, pozwala poprawnie sklasyfikować 96,3 procent wszystkich badanych.

Zanim zakończymy omawianie przykładu, warto skomentować te z aspektów modelu kanonicznego, które mogą budzić niepokój. W tablicy liczebności modelu (część [4] tabeli 6.4) wielkość odpowiadająca uczniom szkół zasadniczych, których ojcowie mają wykształcenie wyższe, jest ujemna. Wynik ten rodzi wątpliwości co do sensowności uzyskanego rozwiązania, gdyż trudno jest znaleźć interpretację ujemnych wielkości w modelu odtwarzającym sytuację rzeczywistą.

Łatwo wyjaśnić, co stanowi powód uzyskania ujemnej liczebności w omawianym polu. W tablicy przedstawiającej wyniki badania (tabela 5.1) w polu tym znalazło się zaledwie 10 uczniów. Widocznie szkoły zasadnicze nie są zbyt chętnie wybierane przez młodzież z domów, w których rodzice mają wykształcenie wyższe. Gdyby wybór szkoły był niezależny od wykształcenia rodziców, to w polu tym powinno znaleźć się 79 uczniów. Niedobór wyniósł więc 69 uczniów. Procedura estymacji tablicy kanonicznej niedobór ten przeszacowała, ujmując z tego pola aż 93 uczniów. Po odjęciu tej wielkości od 79, liczebność modelu kanonicznego wyniosła minus 14 uczniów.

Warto zwrócić uwagę na fakt, że owo przeszacowanie nie miałoby miejsca, gdyby kategorii szkół zasadniczych przypisać mniejszą co do wartości bezwzględnej współrzędną kanoniczną. Wtedy jednak pogorszyłoby się dopasowanie nadwyżek i niedoborów w pozostałych polach tej kolumny tabeli. Na przykład, w szkołach zasadniczych niedobór uczniów, których ojcowie mają wykształcenie średnie bądź pomaturalne, wynosił pierwotnie 111 uczniów. Tablica kanoniczna pozwoliła zmniejszyć ten niedobór o 71 uczniów, czyli niedobór 40 z nich nadal pozostał niewyjaśniony. Gdyby przypisać szkołom zasadniczym współrzędną o mniejszej wartości bezwzględnej, to oszacowanie wielkości niedoboru w tym polu byłoby mniejsze niż 71 uczniów, a więc większa część pierwotnego niedoboru pozostałaby nie odtworzona. Podobna sytuacja miałaby miejsce w wypadku uczniów szkół zasadniczych, których ojcowie mają wykształcenie podstawowe. Z występującej pierwotnie nadwyżki 82 uczniów model umieścił w tym polu jedynie 50. Gdyby zmniejszyć bezwzględną wartość współrzędnej, to niewyjaśniona nadwyżka byłaby większa.

Każdy badany związek jest złożonym systemem naczyń połączonych. Nie należy więc oczekiwać, że jego strukturę uda się w pełni opisać za pomocą

modelu, który ma dostarczyć syntetycznego obrazu badanego zjawiska. Jeśli model do pewnych aspektów zjawiska dopasujemy lepiej, to na ogół do innych okaże się dopasowany gorzej. Rekomenduję więc, aby nie przywiązywać nadmiernej wagi do faktu uzyskania ujemnej liczebności w jednym z pól modelu. Nie to stanowi bowiem o jego wadach czy zaletach. W ostatecznym bilansie liczy się wyłącznie to, na ile zastosowanie danego modelu przybliża nas do zrozumienia i wyjaśnienia badanego zjawiska.

6.5 Własności modelu kanonicznego

Znajomość własności modelu kanonicznego nie tylko ułatwia interpretację wielkości obliczonych parametrów, jak to się dzieje w wypadku większości modeli. Rozwija także intuicje dotyczące rozumienia zjawisk przedstawianych w tablicach. Pod tym względem model kanoniczny jest wyjątkowy. Jest to konsekwencją wielorakich powiązań z najprostszymi i powszechnie stosowanymi metodami analizy tablic, takimi jak porównywanie profili wierszy i kolumn, badanie odstępstw liczebności od modelu niezależności, ocena siły związku na podstawie pojedynczego współczynnika, czy też wnioskowanie na temat kształtu związku w populacji na podstawie próby. Model kanoniczny nie tylko integruje wiele elementów tych podejść, lecz również stanowi punkt wyjścia dla bardziej wyrafinowanych metod analizy tablic.

6.5.1 Jednowymiarowość dystansów między profilami

Pod koniec rozdziału 5 zwróciliśmy uwagę na fakt, że wskaźniki dystansów między profilami w wierszach bądź kolumnami nie mają własności addytywności, gdy tablica ma więcej niż dwa wiersze i więcej niż dwie kolumny. W efekcie badacze zmuszeni są do zestawiania ze sobą jedynie niektórych pól porównywanych profili, najczęściej pól o największych liczebnościach. Porównanie ze sobą pełnych profili prowadzić bowiem może do niespójnych wniosków. Niebezpieczeństwo to zilustrowaliśmy porównując ze sobą profile wieku żon mężczyzn najmłodszych i najstarszych, a następnie próbując wyjaśnić stwierdzone różnice za pomocą zmian w profilach wieku żon w kolejnych kohortach wiekowych mężczyzn.

Warto więc podkreślić, że dystanse między profilami w modelu kanonicznym spełniają własność addytywności niezależnie od rozmiarów tablicy. Jednakże pod warunkiem, że jako funkcją dystansu posłużymy się wprowadzonym w podrozdziale 6.2 dystansem chi-kwadrat zamiast stosowanego w rozdziale 5 wskaźnika różnic między rozkładami.

Wielkość dystansu chi-kwadrat zależy przy tym w bardzo prosty sposób od współrzędnych kanonicznych. W wypadku porównywania ze sobą profili dwóch wierszy tablicy, które oznaczmy jako i_1 oraz i_2, dystans jest po prostu równy różnicy współrzędnych kanonicznych dla obu wierszy przemnożonej przez wielkość korelacji kanonicznej[2]

$$\delta_{i_1 i_2} = r * \left| x_{i_1} - x_{i_2} \right| \tag{6.12}$$

Różnica we wzorze (6.12) uwzględniona została w postaci wartości bezwzględnej. Oznacza to, że traktowana jest jako dystans między obiema współrzędnymi. Gdy współrzędne dla wszystkich wierszy zaznaczymy na pewnej osi, to różnice bezwzględne są długościami odcinków między zaznaczonymi punktami. Uświadamia to zarazem, że dystanse między profilami można sprowadzić do jednego wymiaru bądź – mówiąc inaczej – są one addytywne. W analogiczny sposób przedstawić można dystanse między profilami dla kolumn tablicy

$$\delta_{j_1 j_2} = r * \left| y_{j_1} - y_{j_2} \right| \tag{6.13}$$

Zilustrujmy omawianą własność, korzystając z modelu kanonicznego dla związku między wyborem szkoły a wykształceniem ojca. W części [1] tabeli 6.5 przedstawione zostały profile wyboru szkoły przez uczniów o różnym poziomie wykształcenia ojca obliczone na podstawie liczebności modelu kanonicznego, które uprzednio prezentowane były jako część [4] tabeli 6.4. Wielkości zostały podane nie w odsetkach, lecz w proporcjach – przez co w każdym wierszu sumują się do 1. W części [2] zamieszczono wielkości dystansów chi-kwadrat między profilami w kategoriach wykształcenia ojca. Każdy z tych dystansów obliczyć można na dwa sposoby. Pierwszy polega na podstawieniu porównywanych profili bezpośrednio do wzoru (6.1), który definiuje dystans chi-kwadrat, drugi zaś na skorzystaniu z wartości współrzędnych kanonicznych. Obliczmy pierwszym ze sposobów wielkość dystansu między profilami wyboru szkoły przez dzieci ojców o wykształceniu wyższym i podstawowym. Pamiętając jednakże, że dystanse obliczamy nie pomiędzy profilami tablicy otrzymanej w wyniku badania, lecz między profilami w modelu kanonicznym, stanowiącym pewne przybliżenie liczebności empirycznych

$$\delta_{14} = \sqrt{\frac{(0{,}859 - 0{,}180)^2}{0{,}414} + \frac{(0{,}172 - 0{,}539)^2}{0{,}413} + \frac{(-0{,}032 - 0{,}280)^2}{0{,}173}} = 1{,}415$$

[2] Dowód zamieszczony został w aneksie A.1.

Tę samą wielkość możemy obliczyć, korzystając z podanych w tabeli 6.3 współrzędnych kanonicznych

$$\delta_{14} = 0{,}412 * \left| 2{,}249 - (-1{,}182) \right| = 0{,}412 * 3{,}431 = 1{,}415$$

Pełny zestaw dystansów zamieszczony został w części [2] tabeli 6.5. Dla porządku można sprawdzić, że spełniają one warunek addytywności. Wyliczony powyżej bezpośredni dystans między profilami ojców o wykształceniu wyższym i podstawowym jest równy sumie dystansów profili między sąsiadującymi ze sobą kategoriami wykształcenia (pola zaznaczone kolorem szarym)

$$1{,}415 = 0{,}652 + 0{,}520 + 0{,}243$$

Z wielkości kanonicznych współrzędnych wynikają również dystanse między profilami wewnątrz tablicy a profilami brzegowymi. Współrzędne kanoniczne mają bowiem tę własność, że ich średnia jest równa 0. Zero odpowiada więc jak gdyby współrzędnej dla „przeciętnego" profilu, którym jest w tym wypadku profil brzegowy. Obliczmy na przykład dystans profilu wykształcenia ojców uczniów liceów ogólnokształcących wobec profilu wykształcenia ojców wszystkich uczniów. Od kanonicznej współrzędnej dla liceów ogólnokształcących (tabela 6.3) odejmujemy współrzędną dla ogółu uczniów, czyli zero, a następnie mnożymy wynik przez wartość korelacji kanonicznej

$$1{,}158 * 0{,}412 = 0{,}478$$

Wynik zgodny jest z wielkością dystansu podaną w części [3] tabeli 6.5. Obliczając analogiczne dystanse dla techników i szkół zasadniczych zawodowych należy pamiętać, że we wzorze (6.11) bierze się wartość bezwzględną z obliczonej różnicy. Dlatego oba dystanse będą dodatnie, mimo że odpowiadające im współrzędne kanoniczne są ujemne.

Dystanse między kategoriami obu cech dogodnie jest zobrazować na rysunku (rycina 6.3). Łatwo wtedy ocenić dystanse pod względem wielkości oraz odczytać, jakie są między nimi relacje. Na przykład w wypadku wykształcenia ojca największy dystans w profilach wyboru szkoły obserwuje się pomiędzy dziećmi ojców o wykształceniu średnim i wyższym. Dystanse są addytywne, więc można je dodawać. Stąd wiadomo, że dystans pomiędzy profilami wyboru szkoły przez dzieci ojców o wykształceniu średnim i podstawowym jest nieco większy od dystansu między kategoriami ojców o wykształceniu średnim i wyższym.

Dodatkowych korzyści dostarcza możliwość zestawiania ze sobą wielkości dystansów dla obu cech. W rozpatrywanym przykładzie zakres dystansów dla wykształcenia ojców jest wyraźnie większy od zakresu dystansów dla rodzajów szkół. Oznacza to, że skrajne kategorie wykształcenia ojców różnią się

bardziej pod względem profili wyboru szkół przez dzieci, niż licea ogólnokształcące różnią się od szkół zasadniczych kompozycją uczniów ze względu na wykształcenie ojca. Wniosek ten świadczy o pewnym aspekcie badanego zjawiska, na który nie zwróciliśmy uwagi, analizując je za pomocą wcześniej omawianych metod. Można sformułować go w ten sposób, że analizując

Tabela 6.5
Profile i dystanse między profilami w modelu kanonicznym dla związku między wykształceniem ojca a wyborem szkoły
Badanie PISA 2006

[1] profile wyboru szkoły w kategoriach wykształcenia ojca

	rodzaj szkoły ponadgimnazjalnej			
wykształcenie ojca	liceum ogólnokształcące	technikum	zasadnicza zawodowa	ogółem
wyższe	0,859	0,172	-0,032	1,000
średnie	0,547	0,341	0,112	1,000
zasadnicze zawodowe	0,297	0,476	0,227	1,000
podstawowe	0,180	0,539	0,280	1,000
ogółem	0,414	0,413	0,173	1,000

[2] dystanse chi-kwadrat między profilami wyboru szkoły w kategoriach wykształcenia ojca

	wykształcenie ojca				
wykształcenie ojca	wyższe	średnie	zasadnicze zawodowe	podstawowe	wobec profilu brzegowego
wyższe	0,000	0,652	1,172	1,415	0,928
średnie		0,000	0,520	0,763	0,276
zasadnicze zawodowe			0,000	0,243	0,245
podstawowe				0,000	0,487
wobec profilu brzegowego	0,928	0,276	0,245	0,487	0,000

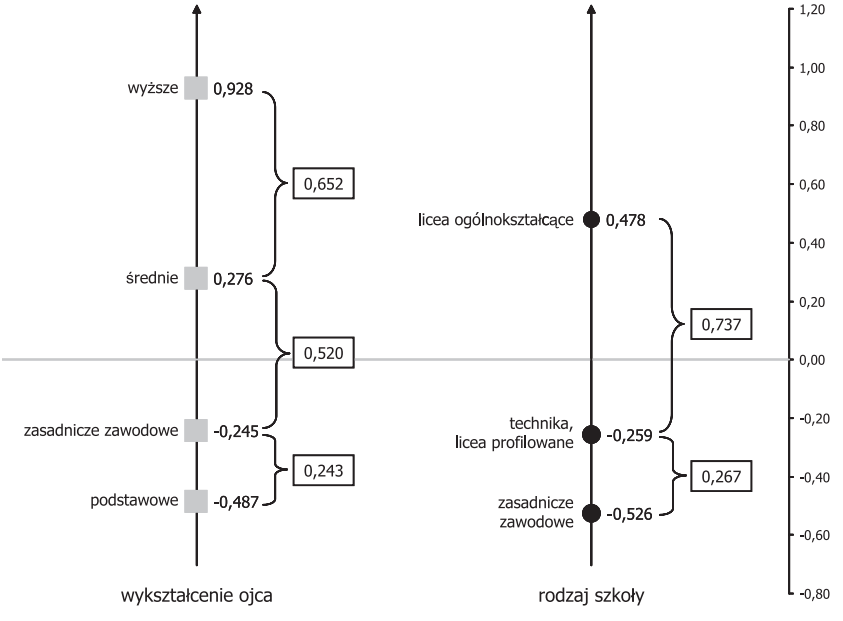
[3] dystanse chi-kwadrat między profilami wykształcenia ojca dla rodzajów szkół

	rodzaj szkoły ponadgimnazjalnej			
rodzaj szkoły ponadgimnazjalnej	liceum ogólnokształcące	technikum	zasadnicza zawodowa	wobec profilu brzegowego
liceum ogólnokształcące	0,000	0,737	1,003	0,478
technikum		0,000	0,267	0,259
zasadnicza zawodowa			0,000	0,526
wobec profilu brzegowego	0,478	0,259	0,526	0,000

Obliczono na podstawie liczebności w części [*4*] *tabeli 6.4.*

kompozycję społeczną uczniów poszczególnych szkół, nie jesteśmy w stanie odczytać, jak bardzo cechy społeczne różnicują szanse dostania się do nich. Stąd już krok do bardziej substantywnych pytań. Czy uczniowie i rodzice mechanizm ten zauważają i uwzględniają w swoich strategiach? Czy uczniowie szkół zasadniczych mają większe poczucie deprywacji, gdy już do tych szkół chodzą, czy też gdy oceniali swoje szanse, będąc jeszcze w gimnazjum?

Rycina 6.3
Dystanse między profilami kategorii wykształcenia ojca i rodzajów szkół
Badanie PISA 2006



Można też zestawiać ze sobą konkretne dystanse dla obu badanych cech. Licea ogólnokształcące w większym stopniu różnią się od techników profilami wykształcenia ojców uczniów, niż ojcowie o wykształceniu wyższym i średnim różnią się profilami wyboru szkoły przez ich dzieci. Być może to ostatnie porównanie nie przekonuje co do swojej przydatności dla budowania intuicji co do kształtu badanego zjawiska. Nie można jednak wykluczyć, że tego rodzaju porównania okażą się owocne w wypadku badania innych cech. Przykładem może być tablica krzyżująca wiek mężczyzny i kobiety w momencie ślubu. Porównanie dystansów między tymi samymi kategoriami wieku dla mężczyzn i kobiet może okazać się użyteczne dla zrozumienia zasad doboru małżeńskiego.

Na koniec pozostał do rozstrzygnięcia dość zasadniczy problem: trafności ustaleń dokonanych na podstawie analizy dystansów. Zostały one bowiem obliczone nie dla pełnych wyników badania, lecz na podstawie liczebności modelu kanonicznego.

Wydaje się, że trafność dokonanych ustaleń zależy przede wszystkim od tego, na ile liczebności modelu kanonicznego są zgodne z liczebnościami tablicy uzyskanej w wyniku badania. Potencjalnie może się zdarzyć, że za pomocą tablicy kanonicznej uda się dokładnie odtworzyć uzyskane liczebności. Wtedy omawiane dystanse w pełni opisywałyby zjawisko przedstawione w tablicy. Na ogół jednak model kanoniczny odtwarza liczebności empiryczne jedynie w pewnym stopniu. Aby ocenić, czy dzieje się to w stopniu wystarczającym, warto nie tylko kierować się ilościowym wskaźnikiem stopnia odtworzenia liczebności tablicy wyjściowej, lecz również istotą otrzymanych wniosków. Przede wszystkim – czy pozwalają zidentyfikować mechanizmy, które wydają się wartościowe dla zrozumienia samego zjawiska. Zjawisko może stanowić złożenie wielu mechanizmów, podczas gdy model kanoniczny pozwala zidentyfikować jedynie część z nich. Nie wyklucza to możliwości dalszej penetracji owej niewyjaśnionej części zjawiska. W podrozdziale 6.6 przedstawię, za pomocą jakich środków można tego dokonać.

6.5.2 Współrzędne kanoniczne a dopasowane średnie

W rozdziale 5 przedstawiłem rzadko stosowane podejście do analizy tablic, które nazwałem metodą dopasowania średnich. Przypomnę w skrócie, że metoda wywodzi się z botaniki i jej oryginalne zastosowania dotyczyły równoległego klasyfikowania środowisk ze względu na gatunki roślin, które w nich najczęściej występują, oraz gatunków roślin ze względu na środowiska, które stwarzają im korzystne warunki wzrostu. Metoda prowadzi do dość oryginalnej interpretacji związku między cechami w tablicy, która dostarczyć może wartościowego wglądu w badane zjawisko.

Metodę dopasowania średnich stosowaliśmy między innymi do analizy zależności między wykształceniem ojca a wyborem szkoły po ukończeniu gimnazjum. Otrzymane rezultaty, czyli średnie dopasowane do kategorii wykształcenia ojca oraz do poszczególnych rodzajów szkół, prezentowane były w tabeli 5.6. Otóż miło mi donieść, że owe dopasowane średnie są równoważne współrzędnym kanonicznym! W ostatniej kolumnie tabeli 6.3 współrzędne te podałem w postaci znormalizowanej do przedziału <0, 100>. W postaci tej są identyczne jak prezentowane uprzednio dopasowane średnie.

Otrzymana zbieżność wyników obu metod jest konsekwencją faktu, że obie mają tę samą matematyczną podstawę. Wyjaśnię to w podrozdziale 6.7,

ograniczając się w tym miejscu do stwierdzenia, że metoda dopasowania średnich stanowi podstawę działania komputerowych algorytmów znajdowania współrzędnych kanonicznych. Bodajże pierwszy możliwości te zauważył radziecki uczony O. V. Sarmanov, publikując w 1958 roku stosowny algorytm. Na algorytmie tym oparta jest procedura iteracyjna, z której korzystaliśmy w podrozdziale 5.4 do szacowania dopasowanych średnich. Jeśli średnie te potraktujemy jako współrzędne kanoniczne, to można pójść krok dalej i obliczyć wartość współczynnika korelacji między nimi, czyli wartość korelacji kanonicznej. W ostatniej rubryce tabeli 5.5 wartości te zostały obliczone po każdym kroku procedury. Jak widać, dość szybko pozwalają one uzyskać oszacowanie tego parametru z należytą dokładnością[3]. Faktyczna wartość korelacji kanonicznej – podana w tabeli 6.3 – wynosi bowiem w zaokrągleniu 0,412.

W wypadku wielu zagadnień dopasowane średnie okazują się przydatne do ustalenia kolejności kategorii cech umieszczonych w wierszach bądź w kolumnach tablicy. Ponieważ współrzędne kanoniczne są im równoważne, mogą pełnić tę samą funkcję. W 1971 roku dwójka amerykańskich badaczy, Sheila Klatzky i Robert Hodge, przedstawiła współrzędne kanoniczne dla tablicy, w której zawód ojca skrzyżowano z zawodem syna. Pozwoliło to ułożyć zawody wzdłuż wymiaru, który autorzy zinterpretowali jako wymiar statusu społecznego. Należy żałować, że tego typu analiz jest niewiele. Ustalenie kolejności kategorii stanowi bowiem spektakularny rezultat, gdy przedstawiane w tablicy cechy mają charakter jakościowy. Porządek wyznaczony przez współrzędne kanoniczne okazuje się na ogół niesprzeczny z wcześniejszą wiedzą na temat badanego zjawiska, a w wielu wypadkach prowadzi do odkrycia jego nowych aspektów, które trudno byłoby zidentyfikować za pomocą innych metod.

6.5.3 Korelacja kanoniczna a siła związku

W podrozdziale 3.12 przedstawiliśmy strategię badania związków w tablicach polegającą na obliczeniu pojedynczego parametru charakteryzującego jego siłę, zwaną też natężeniem związku. Na ogół strategia ta nie prowadzi do uzyskania pogłębionych wyjaśnień, gdyż zawartość informacyjna pojedynczego parametru jest niewielka. Niemniej jednak, w niektórych sytuacjach badawczych może być użyteczna. Zwłaszcza wtedy, gdy cel stanowi porównanie tego samego zjawiska w różnych kontekstach. Przypuśćmy, że chcielibyśmy porównać zależność osiągnięć edukacyjnych od pochodzenia w krajach europej-

[3] Sarmanov (1958) zauważył również, że jeśli po każdej kolejnej iteracji nie dokona się normalizacji otrzymanych średnich, to oszacowanie kwadratu korelacji kanonicznej można otrzymać, dzieląc wielkości współrzędnych uzyskanych w danym kroku przez wielkości współrzędnych otrzymanych w poprzednim kroku (Sawiński 1985: 40–41).

skich, dla których dostępne są stosowne dane. W tym celu skorzystać można by z wyników Europejskiego Sondażu Społecznego. W 2008 roku uczestniczyło w nim 31 krajów. Jeśli zadaniem jest prezentacja obrazu zjawiska w sposób syntetyczny, to przedstawianie szczegółów 31 tablic mija się z celem. Bardziej uzasadnione wydaje się opisanie każdego kraju za pomocą pojedynczej wartości liczbowej.

Powstaje pytanie, jaki parametr wybrać. Proponowana w podrozdziale 3.12 koncepcja pomiaru siły związku w tablicy kładła nacisk na operacyjną interpretację zastosowanego miernika. Chodziło o to, aby konkretną wartość, na przykład 0,4, wyrazić w języku badanego zjawiska. Nie jest to jednak jedyne kryterium, które uwzględnia się przy wyborze miernika. Aby przekazać pewne idee w skuteczny sposób, warto również wziąć pod uwagę przyzwyczajenia audytorium, do którego kierujemy komunikat.

Koncepcja interpretacji miar siły związku w terminach redukcji błędu przewidywania nie jest znana wszystkim badaczom. Co prawda miary oparte na tej koncepcji mają intuicyjną interpretację, lecz jest jedno „ale". Gdy odbiorcom opracowania zaserwujemy dawkę wskaźników opartych na tej koncepcji, to część osób, które nie mają doświadczenia w posługiwaniu się nimi, zmuszona będzie do podwójnego wysiłku. Po pierwsze, będzie musiała zrozumieć, jak interpretować proponowany wskaźnik, a po drugie, co mówią jego wartości na temat prezentowanego zjawiska.

Jest jednak pewien wskaźnik, co do którego można w ciemno przyjąć, że każdy się z nim zetknął. Wskaźnikiem tym jest współczynnik korelacji. Co prawda jego wartości nie mają bezpośredniego przełożenia na język badanego zjawiska, lecz mają za to inną ważną zaletę. Związana jest ona z intuicjami badacza co do oceny siły związku. Intuicje te biorą się zaś z doświadczeń, na które składa się wiedza zdobyta na studiach, lektura publikacji z danej dziedziny, czy wreszcie własna praca badawcza. Są one na tyle powszechne, że niekiedy wchodzą do kanonów wiedzy podręcznikowej. W wielu podręcznikach można znaleźć reguły interpretacji, które wyjaśniają, że współczynnik korelacji w granicach od 0,1 do 0,2 to bardzo słaby związek, od 0,2 do 0,3 to słaby związek, od 0,3 do 0,5 to średnio silny związek i tak dalej. Co prawda zawodowi statystycy z pewnym pobłażaniem traktują tego typu wskazówki, gdyż z formalnego punktu widzenia reguły takie nie mają sensu[4]. W prakty-

[4] Lissowski i inni (2008: 275) zwracają uwagę na fakt, że badacze wolą posługiwać się wielkością współczynnika korelacji zamiast kwadratem tej wielkości, mimo że interpretację w kategoriach przewidywania ma nie współczynnik korelacji, a jego kwadrat (jako odsetek wyjaśnionej wariancji przy przewidywaniu za pomocą regresji liniowej). Skutkiem jest to, że intuicje dotyczące natężenia badanego zjawiska buduje się na podstawie wartości, które *de facto* nie mają interpretacji.

ce badawczej są jednak stosowane, stanowiąc narzędzie **jakościowej oceny** natężenia zjawiska na podstawie wartości ilościowego wskaźnika. Ludzie po prostu myślą w ten sposób i prezentując wyniki badań warto to uwzględnić.

Na pierwszy rzut oka współczynnik korelacji nie stosuje się do tablic, gdyż krzyżuje się w nich różne rodzaje cech, w tym cechy jakościowe. Chyba że kategoriom w wierszach i w kolumnach tablicy przypisać pewne wartości liczbowe, co pozwoli traktować cechy jako ilościowe. Tego rodzaju propozycję przedstawił w 1952 roku australijski statystyk E. J. Williams[5]. Sformułował on problem następująco: znaleźć takie wielkości liczbowe x_1, x_2,..., x_{w-1}, x_w odpowiadające wierszom tablicy oraz takie wielkości y_1, y_2,..., y_{k-1}, y_k odpowiadające jej kolumnom, aby współczynnik korelacji Pearsona r_{XY} obliczony na podstawie tych wielkości osiągnął maksymalną możliwą wartość. Williams (1952) sformułował warunki, jakie muszą spełniać liczebności w tablicy, aby znalezienie rozwiązania było możliwe, podał postać tego rozwiązania, sposób obliczania zmaksymalizowanego współczynnika korelacji, a także rozkład jego statystyki.

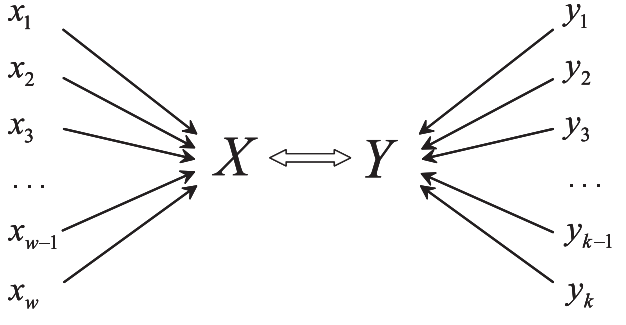
Propozycja Williamsa, aczkolwiek ciekawa, nie znalazła jednak wielu zwolenników (Sawiński 1985: 52). Zarówno w latach pięćdziesiątych, jak i jeszcze wiele lat później, w nauce dominowała koncepcja interpretacji mierników siły związku w kategoriach skuteczności przewidywania. Co prawda proponowany przez Williamsa współczynnik interpretację taką posiadał, lecz na tyle zawiłą, że w praktyce stawała się bezużyteczna (Sawiński 1979: 74–76). Najważniejszy powód braku zainteresowania propozycją Williamsa bierze się jednak stąd, że stanowi ona szczególny przypadek bardziej ogólnej metody – zwanej **analizą kanoniczną** (Kendall i Stuart 1979: 599–606; Sawiński 1985: 25–26). Pod względem formalnym nie wniosła więc wiele w stosunku do tego, co wiadomo było skądinąd.

Metoda analizy kanonicznej zalicza się do klasy metod wielozmiennowych. Jej celem jest określenie powiązania między dwiema grupami zmien-

[5] Zasługą Williamsa (1952) jest to, że jako pierwszy w klarowny sposób sformułował problem i przedstawił jego rozwiązanie, aczkolwiek korzenie metody sięgają początków XX wieku. Doszukać się ich można w pracach grupy badaczy skupionych wokół Karla Pearsona, którzy uważali, że zjawiska reprezentowane za pomocą tablic w rzeczywistości mają postać dwuwymiarowego rozkładu ciągłego. Starano się więc ustalić wpływ grupowania właściwości ciągłych na kształt związku w tablicy, koncentrując uwagę na punktach, które dzielą właściwość ciągłą na przedziały (Goodman 2000). Przy takim postawieniu problemu trudno jednak było dojść do wartościowych ustaleń. Williams reprezentował stanowisko, że zamiast opisywać przedziały – na które dzieli się zmienna ciągła – za pomocą ich granic, lepiej jest to zrobić za pomocą parametrów odpowiadających średnim wartościom zmiennej w poszczególnych przedziałach. Stąd wzięła się przedstawiona propozycja – aczkolwiek wcześniej pojawiały się prace, w których tablice analizowano właśnie w ten sposób (Maung 1941).

nych (Thompson 1984; Nosal 1987). Przypomina to regresję wielozmiennową, w której wartości pojedynczej zmiennej przewidujemy za pomocą grupy predyktorów (Lissowski i in. 2008: 347–397). W analizie kanonicznej zmienna przewidywana zostaje zastąpiona przez grupę zmiennych. Można to sobie wyobrazić jako sklejenie ze sobą dwóch modeli regresji wielozmiennowej, tak jak zostało to przedstawione na rycinie 6.4. W jednym modelu regresji zmienna X jest wyjaśniana przez predyktory $x_1, x_2,..., x_{w-1}, x_w$, zaś w drugim zmienna Y przez grupę predyktorów $y_1, y_2,..., y_{k-1}, y_k$. Podwójną strzałką przedstawiona została korelacja między owymi zmiennymi „przewidywanymi", zwana korelacją kanoniczną. Słowo „przewidywanymi" umieściłem w cudzysłowie, gdyż zmienne X oraz Y faktycznie nie istnieją. Są to jedynie konstrukty hipotetyczne, które reprezentują swoje grupy predyktorów. Tym samym korelację kanoniczną interpretuje się jako współczynnik korelacji między dwiema grupami zmiennych.

Rycina 6.4
Model analizy kanonicznej



Metodę kanoniczną można łatwo adaptować do analizy tablic. W tym celu wierszom i kolumnom wystarczy przypisać wartości tak zwanych zmiennych zero-jedynkowych[6]. Rozważmy to na przykładzie związku między wykształceniem ojca a wyborem szkoły. Pierwszy wiersz tabeli odpowiada badanym uczniom, których ojcowie mają wykształcenie wyższe. Zmienna zero-jedynkowa przypisana temu wierszowi przybierać więc będzie wartość 1 w wypadku uczniów, których ojcowie osiągnęli ten poziom wykształcenia, zaś wartość 0 w wypadku pozostałych uczniów. W analogiczny sposób zdefiniować można zmienne zero-jedynkowe dla pozostałych wierszy i kolumn tabeli. Następnie zmienne

[6] Kodowanie zero-jedynkowe nie jest jedynym sposobem przedstawienia cechy jakościowej w modelu analizy kanonicznej (Sawiński 1986: 271–315; Brzeziński 2007: 370–426).

przypisane wierszom tablicy traktujemy jako jedną grupę, zaś zmienne przypisane kolumnom jako drugą i obliczamy parametry modelu analizy kanonicznej[7].

Czytelnik może odczuwać pewną konfuzję, gdyż wcześniej mówiliśmy o modelu tablicy kanonicznej, zaś obecnie mówimy o metodzie analizy kanonicznej. Pora więc wyjaśnić, że w obu wypadkach chodzi o to samo podejście. Współczynniki regresji odpowiadające zmiennym zero-jedynkowym w metodzie analizy kanonicznej są tożsame z współrzędnymi kanonicznymi, które uprzednio obliczaliśmy dla tablicy kanonicznej. Również korelacja kanoniczna w metodzie analizy kanonicznej jest to ta sama wielkość, co stała r w modelu tablicy kanonicznej.

Do analizy kanonicznej nawiązałem z tego powodu, aby pokazać, że tablicę kanoniczną interpretować można w kategoriach metod wielozmiennowych, które stosuje się raczej do cech ilościowych. Na ile intuicje związane z analizą cech ilościowych mogą okazać się przydatne w analizie tablic – tego rozstrzygał nie będę. Nie można wykluczyć użyteczności takiej interpretacji, aczkolwiek wymaga ona sporej wyobraźni i doświadczenia w stosowaniu metod wielozmiennowych. Pozostaje natomiast faktem, że szczególnym przypadkiem analizy kanonicznej jest problem rozpatrywany przez Williamsa. A problem ten z całą pewnością dotyczy tablic.

Zastanówmy się wobec tego nad przydatnością zaproponowanej przez Williamsa miary siły związku. Stanowi ona rozszerzenie zakresu stosowania współczynnika korelacji na przypadek związku między cechami jakościowymi. Uważam, że w tym właśnie zawiera się kluczowa korzyść tej propozycji. Dzięki temu o związkach w tablicach można mówić w tym samym języku, w jakim prezentuje się związki między cechami ilościowymi. A ponieważ współczynnik korelacji jest miarą dobrze badaczom znaną, stąd łatwiej przekazać niektóre idee dotyczące zjawisk w tablicach.

Na zakończenie omawiania tej kwestii warto jeszcze wspomnieć, że korelacja kanoniczna – czyli proponowana przez Williamsa miara siły związku w tablicy – ma pewną użyteczną własność, której nie ma zwykły współczynnik korelacji. Mianowicie przybiera wartość 0 wtedy i tylko wtedy, gdy cechy są stochastycznie niezależne. Na przydatność tej własności zwracaliśmy uwagę,

[7] Jedna ze zmiennych zero-jedynkowych przypisanych wierszom jest zawsze funkcją pozostałych. Jeśli bowiem wiadomo, że badany uczeń nie należy do żadnego z pierwszych w-1 wierszy, to musi należeć do ostatniego. Analogiczna prawidłowość dotyczy kolumn. W modelu analizy kanonicznej uwzględnia się przez to o jedną zmienną mniej zarówno w grupie zmiennych odpowiadającej wierszom, jak i w grupie odpowiadającej kolumnom tablicy. Gdyby bowiem uwzględnić je wszystkie, to predyktory byłyby wzajemnie od siebie zależne, co uniemożliwiłoby znalezienie rozwiązania. Problem nie jest specyficzny dla analizy kanonicznej, lecz występuje również w analizie regresji.

omawiając w podrozdziale 3.12 możliwości oceny siły związku w tablicy za pomocą pojedynczej miary. Gdy stosowana miara własności tej nie posiada, to otrzymanie wartości bliskiej zeru nie gwarantuje, że związek niewiele odbiega od niezależności. Posiadanie omawianej własności przez korelację kanoniczną stanowi kolejny argument za stosowaniem tego wskaźnika w sytuacjach, gdy pojedynczy parametr wystarcza do opisania związku w tablicy.

6.5.4 Interpretacja wskaźników Quételeta

Stosunkowo niedawno zauważono, że tablica kanoniczna ma związki z wskaźnikami Quételeta (Mirkin 2001: 115–116). Przypomnę, że wskaźniki te zaproponowałem w rozdziale 4 jako narzędzie identyfikacji modelu związku w tablicy. Użyteczną własność wskaźników Quételeta stanowi to, że ich kwadraty można sumować po wierszach i kolumnach rozpatrywanej tablicy uzyskując w ten sposób wielkości określające wpływ poszczególnych kategorii obu cech na kształt związku w tablicy.

Tabela 6.6
Kwadraty wskaźników Quételeta dla pól modelu kanonicznego. Związek między wykształceniem ojca a wyborem szkoły ponadgimnazjalnej
Badanie PISA 2006

[1] kwadraty średnich wskaźników Quételeta

	rodzaj szkoły ponadgimnazjalnej			
wykształcenie ojca	liceum ogólno-kształcące	technikum	zasadnicza zawodowa	ogółem
wyższe	0,052	0,015	0,026	0,093
średnie	0,012	0,003	0,006	0,021
zasadnicze zawodowe	0,017	0,005	0,008	0,030
podstawowe	0,014	0,004	0,007	0,026
ogółem	0,095	0,028	0,048	0,170

[2] udziały kwadratów średnich wskaźników Quételeta (w procentach)

	rodzaj szkoły ponadgimnazjalnej			
wykształcenie ojca	liceum ogólno-kształcące	technikum	zasadnicza zawodowa	ogółem
wyższe	30,3	8,9	15,3	54,5
średnie	6,9	2,0	3,5	12,5
zasadnicze zawodowe	9,9	2,9	5,0	17,7
podstawowe	8,5	2,5	4,3	15,3
ogółem	55,6	16,3	28,1	100,0

Obliczono na podstawie modelu kanonicznego prezentowanego w części [*4*] *tabeli 6.3.*

Analogiczna prawidłowość zachodzi w wypadku liczebności modelu kanonicznego. W części [1] tabeli 6.6 przedstawione zostały wielkości kwadratów średnich wskaźników Quételeta dla poszczególnych pól tablicy tego modelu, zaś w części [2] ich procentowe udziały. Na podstawie tych ostatnich odczytać można, że o kształcie opisanej przez tablicę kanoniczną zależności wyboru szkoły od wykształcenia ojca aż w 30 procentach decyduje pole odpowiadające wyborowi liceum ogólnokształcącego przez dzieci ojców z wykształceniem wyższym. Ponadto, z tą kategorią ojców wiąże się ponad połowa wszystkich odstępstw od modelu niezależności wyjaśnianych przez tablicę kanoniczną. Podobnie, ponad połowę tych odstępstw wyjaśnia specyfika składu społecznego uczniów, którzy wybrali licea ogólnokształcące. Najmniejsze w tym kontekście znaczenie mają wybory szkolne uczniów, których ojcowie mają wykształcenie średnie. Po prostu wybory te w największym stopniu przypominają profil wyboru szkoły wśród całej młodzieży. Z analogicznego powodu najbardziej egalitarnym typem szkół są technika. Analiza wskaźników Quételeta prowadzi do tych samych ustaleń, co porównanie profili wyboru szkoły wśród uczniów o różnym wykształceniu ojca oraz porównanie profili wykształcenia ojców młodzieży uczęszczającej do różnych szkół.

W wypadku modelu kanonicznego kwadraty wskaźników Quételeta mają też tę własność, że ich suma dla całej tablicy odpowiada kwadratowi kanonicznej korelacji. W rozpatrywanym przykładzie suma ta wynosi 0,170, z czego pierwiastek jest równy 0,412. Odpowiada to wielkości kanonicznej korelacji podanej w tabeli 6.3. Omawiana własność dostarcza użytecznej interpretacji dla korelacji kanonicznej, o czym warto pamiętać, gdy korelację tę wykorzystuje się jako jedyny parametr opisujący tablicę (na przykład przy zestawianiu ze sobą analogicznych tablic dla wielu krajów). Gdy wartość korelacji budzi niepokój, na przykład jest z niezrozumiałych powodów zbyt niska bądź zbyt wysoka, można cofnąć się o krok i dokonać dekompozycji jej kwadratu na kwadraty wskaźników Quételeta w poszczególnych polach tablicy. Otrzymamy wtedy odpowiedź, które z pól w największym stopniu odpowiadają za otrzymaną wielkość.

6.5.5 Dekompozycja chi-kwadrat

Dla liczebności modelu kanonicznego obliczyć można wartość wskaźnika chi-kwadrat. W rozpatrywanym przykładzie wynosi ona 716,6. Można ją potraktować jako wartość statystyki χ^2 i wykorzystać do testowania hipotezy, że liczebności odtworzone za pomocą tego modelu pochodzą z populacji, w której wybór szkoły jest niezależny od wykształcenia ojca. W teście tym jako liczbę stopni swobody przyjmuje się wielkość (Goodman 1986: 249–250)

$$df = (w-2)(k-2) \tag{6.14}$$

gdzie w i k oznaczają liczbę wierszy i kolumn. W rozpatrywanym przykładzie liczba stopni swobody wynosi 2. Przyjmując poziom istotności α równy 0,05, odczytujemy w tablicach wartość krytyczną statystyki χ^2, która wynosi 5,991 (Zieliński 1972: 115). Hipotezę o niezależności cech w populacji należy więc odrzucić.

Chi-kwadrat ma w modelu kanonicznym jeszcze jedną interpretację. Mianowicie, gdy podzielimy tę wielkość przez liczebność próby, to otrzymamy kwadrat kanonicznej korelacji

$$r^2 = \frac{\chi^2}{n} \tag{6.15}$$

W rozpatrywanym przykładzie iloraz (6.15) wynosi 0,170, zaś pierwiastek z tej wielkości 0,412 – co jest równe podanej uprzednio wielkości kanonicznej korelacji. Omawiana własność oznacza, że w modelu kanonicznym zarówno korelacja kanoniczna, jak i chi-kwadrat stanowią dwie formy tego samego konceptu. Statystykę chi-kwadrat można więc stosować do testowania hipotezy, że w populacji wartość kanonicznej korelacji jest równa zero.

Wskaźnik chi-kwadrat można też wykorzystać do obliczenia udziału modelu kanonicznego w analizowanym związku. Dla tablicy liczebności otrzymanych w badaniu chi-kwadrat wynosi 770,6. Gdy przez wielkość tę podzielimy otrzymaną wyżej wartość chi-kwadrat dla modelu kanonicznego, to wynik wyniesie – w odsetkach – 93,0 procent. Wielkość tę można przyjąć jako miarę stopnia odtworzenia przez tablicę kanoniczną liczebności otrzymanych w badaniu. Uprzednio posługiwaliśmy się do tego celu odsetkiem jednostek poprawnie sklasyfikowanych do pól tablicy, która wyniosła 96,3 procenta. Oba wskaźniki mają różne zasady obliczania, toteż nie należy spodziewać się uzyskania identycznych wielkości. Niemniej jednak rząd obu wielkości powinien być zbliżony, więc otrzymane na ich podstawie wnioski – spójne. Zaletą odsetka jednostek poprawnie sklasyfikowanych jest jego bezpośrednia interpretacja, w kategoriach przemieszczania jednostek między polami tablicy. Z kolei chi-kwadrat ma związki z wielkością korelacji kanonicznej a także z wartościami wskaźników Quételeta dla pól tablicy.

6.6 Rekurencja, czyli wyjaśnienie za pomocą modelu kanonicznego tego, czego model kanoniczny nie wyjaśnił

Zaprezentowany w podrozdziale 6.4 model kanoniczny związku wykształcenia ojca z wyborem szkoły przez dziecko nie wyjaśnił tego związku w całości. Przypomnijmy, że pozwolił poprawnie sklasyfikować 96,3 procent badanych uczniów do właściwych pól tablicy. Odsetek ten uznaliśmy za wystarczający do zbudowania wartościowego wyjaśnienia badanego związku. Co jednak zrobić w wypadku, gdy interesują nas również te aspekty związku, których wyjaśnić się nie udało. Otóż okazuje się, że można je wyjaśnić tym samym sposobem. Tworząc dla nich kolejny model kanoniczny.

Niewyjaśnione przez model kanoniczny niedobory i nadmiary w poszczególnych polach prezentowane były w części [5] tabeli 6.4. Wielkości te możemy dodać do pól modelu niezależności, który stanowi punkt odniesienia dla wszelkich analiz z wykorzystaniem metod kanonicznych. Utworzy to tablicę przedstawioną w części [1] tabeli 6.7. Jej liczebności interpretować można w następujący sposób. Wyobraźmy sobie, że mechanizmy zidentyfikowane za pomocą uprzednio zastosowanej tablicy kanonicznej nie działają. To znaczy, ojcowie o wykształceniu wyższym wcale nie lokują swoich pociech częściej w liceach ogólnokształcących, szkoły zasadnicze nie mają najgorszej kompozycji uczniów ze względu na wykształcenie ojca i tak dalej. Wszystkie te mechanizmy zastąpione zostały niezależnością wybieranej szkoły od wykształcenia ojca. Obserwowane zjawisko obejmuje zaś wyłącznie te mechanizmy, których nie udało się odtworzyć za pomocą modelu kanonicznego.

Tabela 6.7
Liczebności związku między wykształceniem ojca a wyborem szkoły ponadgimnazjalnej niewyjaśnione za pomocą pierwszej tablicy kanonicznej
Badanie PISA 2006

	rodzaj szkoły			
wykształcenie ojca	liceum ogólno-kształcące	technikum	zasadnicza zawodowa	ogółem
	[1] liczebności wyjściowe $n_{ij}^{(1)}$			
wyższe	197	154	103	454
średnie	472	538	163	1173
zasadnicze zawodowe	875	901	350	2126
podstawowe	203	146	112	461
ogółem	1746	1739	729	4214

Tabela 6.7 (kontynuacja)

[2] różnice $d_{ij}^{(1)} = n_{ij}^{(1)} - e_{ij}$

wyższe	9	-33	24	0
średnie	-14	54	-40	0
zasadnicze zawodowe	-6	24	-17	0
podstawowe	12	-45	33	0
ogółem	0	0	0	0

Minimalna liczba przemieszczeń: 155; odsetek osób poprawnie sklasyfikowanych: 96,3%

[3] liczebności tablicy kanonicznej $c_{ij}^{(2)}$

wyższe	9	-33	24	0
średnie	-14	54	-40	0
zasadnicze zawodowe	-6	24	-17	0
podstawowe	12	-45	33	0
ogółem	0	0	0	0

[4] liczebności odtworzone $p_{ij}^{(2)} = e_{ij} + c_{ij}^{(2)}$

wyższe	197	154	103	454
średnie	472	538	163	1173
zasadnicze zawodowe	875	901	350	2126
podstawowe	203	146	112	461
ogółem	1746	1739	729	4214

$\chi^2 = 54{,}0$; odsetek osób poprawnie sklasyfikowanych: 100,0%

Model dla pierwszej tablicy kanonicznej przedstawia tabela 6.4.

Przyjmijmy więc, że liczebności przedstawione w części [1] tabeli 6.7 uzyskaliśmy w wyniku badania. Nic zatem nie stoi na przeszkodzie, aby do analizy tych wyników posłużyć się modelem kanonicznym. W tabeli 6.8 zamieszczone zostały otrzymane tą drogą współrzędne kanoniczne wraz z kanoniczną korelacją. Część [3] tabeli 6.7 przedstawia natomiast liczebności tablicy kanonicznej. Gdy porównamy je z przedstawionymi w części [2] różnicami dotychczas niewyjaśnionymi, to zauważymy, że są identyczne. Nowy model kanoniczny pozwala więc wyjaśnić badane zjawisko **do końca**.

Przyjrzyjmy się otrzymanym współrzędnym kanonicznym. Największe dodatnie wartości odpowiadają kombinacji wykształcenia podstawowego i szkoły zasadniczej. W poprzednio obliczonej tablicy kanonicznej sytuacja ta była niedoszacowana, co dotyczyło 33 uczniów. Wspólny mechanizm, decydujący o szansach kształcenia dzieci o niejednakowym wykształceniu ojców, nie był więc w stanie wyjaśnić szczególnie niskich szans kształcenia dzieci, których ojcowie mają wykształcenie podstawowe. Wyjaśnia je dopiero obecny model.

Rozważmy z kolei dwie najniższe współrzędne kanoniczne. Ponieważ są ujemne, ich iloczyn również wyjaśnia nadwyżkę. Jest to nadwyżka dzieci ojców o wykształceniu średnim, które wybrały technika. Utylitarna i przewidująca strategia. Technikum daje zawód, a zarazem pozostawia możliwość pójścia na studia. Warto zwrócić uwagę na fakt, że grupa ta jak ognia unika zawodówek. W odpowiednim polu tablicy kanonicznej deficyt wynosi 40 osób. Wydaje się uzasadnione, aby obecną tablicę kanoniczną interpretować jako przejaw strategii, które nie są częścią dominującego wzoru.

Tabela 6.8
Parametry kanoniczne związku między wykształceniem ojca a wyborem szkoły ponadgimnazjalnej w części niewyjaśnionej za pomocą pierwszej tablicy kanonicznej

Badanie PISA 2006

			współrzędne kanoniczne	
cecha	kategoria	oznaczenie	standaryzowane	znormalizowane
wykształcenie ojca	wyższe	x_1	1,541	83,5
	średnie	x_2	-0,969	0,0
	zasadnicze zawodowe	x_3	-0,236	24,4
	podstawowe	x_4	2,038	100,0
rodzaj szkoły	liceum ogólnokształcące	y_1	0,268	46,0
	technikum	y_2	-1,014	0,0
	zasadnicza zawodowa	y_3	1,776	100,0
	korelacja kanoniczna	r	0,113	

Parametry dla pierwszej tablicy kanonicznej podano w tabeli 6.5.

Metoda kanoniczna posiada tę szczególną własność, że dekomponuje tablicę zawierającą wyniki badania w sposób **hierarchiczny**. Oznacza to, że tablica kanoniczna utworzona jako pierwsza wyjaśnia maksimum tego, co możliwe. Czyli zgodnie z wzorem (6.10) – stara się maksymalnie dopasować estymowane liczebności do obserwowanych. Dopiero to, co nie zostanie wyjaśnione, może stanowić przedmiot dalszych analiz. Druga tablica kanoniczna jest dopasowywana do tego, czego nie wyjaśniła pierwsza. A ponieważ kryterium dopasowania nie ulega zmianie, znów wyjaśnieniu podlega maksimum tego, co możliwe.

Procedurę tworzenia kolejnych tablic kanonicznych można kontynuować aż do pełnego odtworzenia liczebności uzyskanych w badaniu. Liczba możliwych tablic kanonicznych zależy przy tym od rozmiarów rozpatrywanej tablicy. Jest ich zawsze o 1 mniej od liczby wierszy bądź liczby kolumn, w zależno-

ści od tego, która z tych wielkości jest mniejsza. Tablica krzyżująca zawód ojca z wyborem szkoły ma 4 wiersze i 3 kolumny. Kolumn jest mniej niż wierszy, czyli dokładne odtworzenie liczebności w tej tablicy wymaga dwóch tablic kanonicznych.

Druga z obliczonych tablic kanonicznych posiada te same własności co pierwsza. W szczególności, różnice między współrzędnymi odpowiadają wielkościom dystansów między profilami wierszy bądź kolumn w modelu utworzonym jako złożenie tablicy kanonicznej z niezależnością. Ponieważ współrzędne w drugiej tablicy kanonicznej odzwierciedlają specyfikę niektórych kategorii, stąd też nie układają się w tak spektakularne skale poziomów wykształcenia według „jakości" wyborów szkolnych dzieci czy też w ranking szkół ze względu na korzystne wykształcenie ojca. Dlatego też interpretacja dystansów między wierszami i kolumnami nie pełni tak ważnej roli, jak w wypadku pierwszej z obliczonych tablic kanonicznych. Tym niemniej, jak stwierdziliśmy wyżej, dystanse te stanowić mogą przejaw specyficznych strategii podejmowanych przez niektóre kategorie społeczne.

Każdą z kanonicznych tablic można scharakteryzować pod względem udziału w całości związku. Zgodnie z propozycją sformułowaną w podrozdziale 6.5.5, tę ostatnią wielkość można wyrazić jako iloraz wskaźnika chi-kwadrat do wielkości chi-kwadrat obliczonej dla tablicy liczebności uzyskanych w badaniu. W rozpatrywanym przykładzie wielkość χ^2 dla liczebności odtworzonych przez drugą tablicę kanoniczną wyniosła 54,0, co stanowi 7 procent wyjściowej wielkości równej 770,6. Gdy wielkość χ^2 dla tablicy liczebności odtworzonych podzielimy przez liczebność próby, to otrzymamy sumę kwadratów wskaźników Quételeta. Wyciągnięcie pierwiastka z tej wielkości da nam z kolei wartość korelacji kanonicznej.

6.7 Kanoniczna dekompozycja tablic: pełny wykład metody

Mój niedościgniony Mistrz, Tadeusz Krauze, wiele lat temu zwrócił uwagę na pewną manierę, którą zauważył w moich tekstach. Chodziło o to, że zaczynam od wniosków a kończę na założeniach. Tadeusz tłumaczył mi cierpliwie, że tekst naukowy ma swoją logikę zapewniającą mu przejrzystość, zaś jej nieprzestrzeganie stanowi wyraz braku szacunku dla czytelnika. Czytelnik ma bowiem prawo oczekiwać, że przez tok wywodu przejdzie w sposób przyjazny i efektywny. Czyli nie napotykając co krok wątpliwości – co z czego wynika i dlaczego.

Uwagę wziąłem sobie do serca, chociaż w późniejszym czasie wielokrotnie miałem okazję przekonać się, że ta niewątpliwie słuszna zasada nie zawsze

przekłada się na skuteczną strategię. Niekiedy bowiem umieszczenie założeń czy też nie daj Boże dużej liczby wzorów blisko początku tekstu powoduje, że kontakt czytelnika z tekstem kończy się w tym momencie.

Prezentowane do tego momentu ustalenia miały być może układ nieco chaotyczny, lecz pozwalały – przynajmniej w moim odczuciu – uwypuklić sens i korzyści z posługiwania się tablicami kanonicznymi do analizy związków między cechami. Nadszedł jednak czas, aby całą koncepcję przedstawić w sposób spójny. To też ma swoje zalety. Przede wszystkim rzuca więcej światła na samą ideę omawianego ujęcia, które nazywał będę dekompozycją tablic.

Ideę dekompozycji wyobrazić sobie można jako rozdzielenie badanego związku na warstwy, które odpowiadają różnym jego aspektom. Warstwami są tablice kanoniczne. Najwyższa warstwa obejmuje najbardziej znaczący aspekt badanego związku. To tak, jak komórki na górze plastrów w ulu zawierają miód najszlachetniejszy. Niekiedy eksploracja najwyższej warstwy wystarcza do zrozumienia i zinterpretowania badanego zjawiska. Jeśli zaś nie, to można wykorzystać warstwy położone niżej. Każda warstwa oznacza jednak utratę kolejnych stopni swobody. Dokonując interpretacji zjawiska na podstawie najwyższej warstwy musimy poświęcić ich $w + k - 3$. Jeśli chcemy skorzystać z drugiej warstwy, to pochłonie ona $w + k - 5$ dalszych i tak dalej. Co prawda koszt każdej kolejnej warstwy jest o 2 stopnie swobody niższy, lecz zarazem mniej wnosi ona do wyjaśnienia zjawiska. Mamy tu więc do czynienia z owym trade-offem, od omówienia którego rozpocząłem rozdział. Decydując się na najprostszy model, wyjaśnimy tylko jeden aspekt badanego zjawiska. Robimy to natomiast w sposób klarowny. Aby wyjaśnić więcej jego aspektów, musimy posłużyć się bardziej złożonym modelem. Wtedy jednak istotę modelu trudniej jest komukolwiek wyjaśnić. Stanowi to koszt, który ponosimy.

Przejdźmy do systematycznej prezentacji metody kanonicznej dekompozycji tablic. Przypomnijmy, że przez N oznaczaliśmy dotychczas tablicę liczebności obserwowanych o w wierszach oraz k kolumnach, zaś przez n_{ij} liczebności rozkładu łącznego tej tablicy. Model niezależności dla tej tablicy oznaczaliśmy zaś jako E, a jego pola jako e_{ij}.

W ogólnym przypadku liczebności tablicy N przedstawić można w podanej niżej postaci, która zwana jest **modelem korelacji kanonicznych** (Kendall i Stuart 1979: 602; Gilula i Haberman 1986: 780)

$$n_{ij} = e_{ij}\left(1 + \sum_{s=1}^{S} r^{(s)} x_i^{(s)} y_j^{(s)}\right) \tag{6.16}$$

gdzie

$$S = \min(w-1, k-1) \tag{6.17}$$

zaś pozostałe parametry oznaczają

$r^{(1)}, r^{(2)}, ..., r^{(s)}$	dodatnie wielkości, nazywane pierwszą, drugą, ..., s-tą korelacją kanoniczną
$x_1^{(s)}, x_2^{(s)}, \ldots, x_w^{(s)}$	s-ty zestaw współrzędnych kanonicznych przypisanych wierszom tablicy
$y_1^{(s)}, y_2^{(s)}, \ldots, y_k^{(s)}$	s-ty zestaw współrzędnych kanonicznych przypisanych kolumnom tablicy

Na współrzędne kanoniczne nałożone są dodatkowe warunki

$$\frac{1}{n}\sum_{i=1}^{w} x_i^{(s)} a_i = 0 \tag{6.18}$$

$$\frac{1}{n}\sum_{i=1}^{w} [x_i^{(s)}]^2 a_i = 1 \tag{6.19}$$

$$\frac{1}{n}\sum_{j=1}^{k} y_j^{(s)} b_j = 0 \tag{6.20}$$

$$\frac{1}{n}\sum_{j=1}^{k} [y_j^{(s)}]^2 b_j = 1 \tag{6.21}$$

gdzie a_1,..., a_w oraz b_1,..., b_k są odpowiednio liczebnościami brzegowymi wierszy i kolumn tablicy. Warunki te oznaczają, że średnia każdego zestawu współrzędnych kanonicznych jest równa 0, natomiast ich odchylenie standardowe jest równe 1. Wielkości $r^{(s)}$ określone są zaś jako współczynniki korelacji między odpowiadającymi sobie zestawami współrzędnych kanonicznymi obu zmiennych

$$r^{(s)} = \frac{1}{n}\sum_{i=1}^{w}\sum_{j=1}^{k} x_i^{(s)} y_j^{(s)} n_{ij} \tag{6.22}$$

Standardowa formuła na współczynnik korelacji (Lissowski i inni 2008: 274) upraszcza się w tym wypadku do podanej we wzorze (6.22), ponieważ współrzędne kanoniczne mają średnie 0 i odchylenia standardowe 1.

Oszacowanie wielkości występujących w modelu korelacji kanonicznych możliwe jest pod warunkiem, że wielkość pierwszej korelacji będziemy maksymalizować. Pełny proces dochodzenia do rozwiązania zostanie tu pominięty (Kendall i Stuart 1979: 601–604; Sawiński 1979: 99–103; 1985: 34–50). Wy-

mienię jedynie jego kolejne kroki. W pierwszej fazie wyliczane są wartości korelacji kanonicznych. Otrzymuje się je jako pierwiastki algebraicznego równania stopnia *S*, skonstruowanego na podstawie tablicy *N*. Są więc one zależne wyłącznie od liczebności otrzymanych w badaniu. Pierwiastki te, czyli korelacje kanoniczne, porządkuje się następnie od największej do najmniejszej

$$1 \geq r^{(1)} \geq r^{(2)} \geq \ldots \geq r^{(S)} \geq 0 \tag{6.23}$$

Po znalezieniu wielkości korelacji kanonicznych oblicza się dla każdej z nich oba zestawy współrzędnych. Z własności rozwiązania wynika, że zestawy współrzędnych odpowiadające danej korelacji spełniają następujące równania (Kendall i Stuart 1979: 603)

$$y_j^{(s)} = \frac{1}{r^{(s)} b_j} \sum_{i=1}^{w} x_i^{(s)} n_{ij} \tag{6.24}$$

$$x_i^{(s)} = \frac{1}{r^{(s)} a_i} \sum_{j=1}^{k} y_j^{(s)} n_{ij} \tag{6.25}$$

co oznacza, że w ramach każdego zestawu współrzędnych kanonicznych – współrzędne $y_j^{(s)}$ odpowiadające kolumnom tablicy są średnimi warunkowymi współrzędnych $x_i^{(s)}$ przypisanym jej wierszom i odwrotnie. Należy podkreślić, że nie jest to przyjmowane założenie, lecz własność rozwiązania. Wyjaśnia ona, dlaczego procedura dopasowania średnich prowadzi do otrzymania kanonicznych współrzędnych. Sarmanov (1958) wykazał, że obie metody są równoważne. Podał również iteracyjny algorytm pozwalający szacować te wielkości, z którego korzystaliśmy w rozdziale 5.

Własnością rozwiązania jest również to, że poszczególne zestawy współrzędnych kanonicznych są od siebie niezależne w tym sensie, że współczynniki korelacji między zestawami $X^{(s)}$ oraz $X^{(s')}$, dla dowolnych $s \neq s'$, są równe 0 (analogicznie dla zestawów współrzędnych kanonicznych *Y*). Własność tę nazywa się **ortogonalnością** zestawów współrzędnych kanonicznych. Dzięki tej własności metoda ma charakter hierarchiczny. Pierwsza para współrzędnych kanonicznych maksymalizuje korelację dla liczebności otrzymanych w badaniu, druga maksymalizuje korelację dla tych aspektów tablicy empirycznej, które nie zostały opisane przez pierwszą tablicę kanoniczną, i tak dalej.

Rozwiązanie modelu kanonicznych korelacji można przedstawiać w różnej postaci, korzystając z wzajemnych zależności między liczebnościami wyjściowej tablicy *N*, korelacjami kanonicznymi i zestawami współrzędnych kanonicznych. W książce proponuję ujęcie, które wiele lat temu nazwałem metodą

kanonicznej dekompozycji tablic (Sawiński i Domański 1984, 1989). Sprowadza się ono do tego, że tablica liczebności empirycznych N dekomponowana jest na ciąg tablic kanonicznych. Liczebności tablicy kanonicznej zdefiniowałem w sposób podany we wzorze (6.8). Formułę (6.16) można wtedy przedstawić w następującej postaci

$$n_{ij} = e_{ij} + c_{ij}^{(1)} + c_{ij}^{(2)} + \ldots + c_{ij}^{(S)} \tag{6.26}$$

co można zapisać jako sumę tablic

$$N = E + C^{(1)} + C^{(2)} + \ldots + C^{(S)} \tag{6.27}$$

W ujęciu tym tablica obserwowanych liczebności dekomponuje się na model niezależności oraz kolejne warstwy, którymi są tablice kanoniczne. Tablice te określają jednak wyłącznie odstępstwa od modelu niezależności, toteż nie mogą służyć jako modele zjawiska. Modelami tymi są natomiast tablice $M^{(s)}$ postaci

$$M^{(s)} = E + C^{(s)} \tag{6.28}$$

gdzie $s = 0, 1, 2,..., S$, które wcześniej nazywaliśmy **modelami kanonicznymi**. Przy czym modelowi $M^{(0)}$ (wyjściowemu) nie odpowiada żadna tablica kanoniczna. Jest on równoważny modelowi niezależności E.

Z modeli zdefiniowanych w podany sposób korzystaliśmy we wcześniejszych rozważaniach, dokonując dekompozycji tablicy przedstawiającej związek wykształcenia ojca z wyborem szkoły przez dziecko. Modele te mają również interpretację w kategoriach dekompozycji wskaźnika chi-kwadrat dla tablicy . Wielkość tego wskaźnika można bowiem przedstawić jako (Kendall i Stuart 1979: 606)

$$\chi^2(N) = \chi^2(M^{(1)}) + \chi^2(M^{(2)}) + \ldots + \chi^2(M^{(s)}) \tag{6.29}$$

Dekompozycję tę wykorzystać można do oceny, jaką część związku wyjaśniają poszczególne tablice kanoniczne. Analogiczne wyrażenie skonstruować można dla kanonicznych korelacji. Suma ich kwadratów jest bowiem równa wartości wskaźnika χ^2 dla tablicy liczebności otrzymanych w badaniu podzielonej przez sumę liczebności w tej tablicy, czyli przez n (Kendall i Stuart 1979: 606)

$$\chi^2(N)/n = (r^{(1)})^2 + (r^{(2)})^2 + \ldots + (r^{(s)})^2 \tag{6.30}$$

Dekompozycje (6.29) i (6.30) są sobie równoważne, co wynika z faktu, że każda z korelacji kanonicznych jest współczynnikiem korelacji obliczonym dla odpowiadającego jej modelu kanonicznego[8]

[8] Dowód w aneksie A.2.

$$r^{(s)} = r(M^{(s)}) \tag{6.31}$$

Własność ta oznacza między innymi, że liczebności modelu kanonicznego można interpretować tak, jak gdyby stanowiły wynik badania. Pozwala to odrębnie zilustrować każdy z aspektów zjawiska opisanych przez poszczególne tablice kanoniczne.

Dekompozycję skonstruować również można dla kwadratów wskaźników Quételeta, określonych wzorem (4.9). Oznaczmy sumę kwadratów tych wskaźników dla całej tablicy jako Q. Wówczas dekomponuje się ona na wielkości odpowiadające poszczególnym modelom kanonicznym

$$Q(N) = Q(M^{(1)}) + Q(M^{(2)}) + \ldots + Q(M^{(s)}) \tag{6.32}$$

Nie wnosi to jednak żadnej nowej interpretacji, gdyż suma wskaźników Quételeta jest równa wielkości wskaźnika χ^2 podzielonej przez n, czyli przez łączną liczbę jednostek uwzględnionych w tablicy. Opisana dekompozycja prowadzi więc do analogicznego rozdzielenia związku między tablice kanoniczne, jak podana w (6.29) dekompozycja χ^2. Niemniej, w wypadku kwadratów wskaźników Quételeta dekompozycja zachodzi również dla **każdego z wierszy** i dla **każdej z kolumn**[9]. Niech $Q_{i\bullet}$ oznacza sumę kwadratów wskaźników Quételeta dla i-tego wiersza. Wówczas

$$Q_{i\bullet}(N) = \sum_{s=1}^{S} Q_{i\bullet}(M^{(s)}) \tag{6.33}$$

Analogiczna prawidłowość zachodzi dla kolumn

$$Q_{\bullet j}(N) = \sum_{s=1}^{S} Q_{\bullet j}(M^{(s)}) \tag{6.34}$$

Możliwość zdekomponowania sumy kwadratów wskaźników Quételeta między wiersze i kolumny tablicy poszerza możliwości interpretacyjne uzyskane dzięki zastosowaniu modeli kanonicznych. Pozwala bowiem odpowiedzieć na pytanie, w jakim stopniu każda z kategorii odpowiadających wierszom i kolumnom uczestniczy, czy wyjaśnia aspekty zjawiska opisane za pomocą poszczególnych modeli kanonicznych. Analizując wpływ wykształcenia ojca na wybór szkoły stwierdziliśmy na tej podstawie, że najbardziej dominujący jego aspekt – zidentyfikowany przez pierwszy model kanoniczny – dotyczy przede wszystkim ojców o wykształceniu wyższym i związany jest z wybieraniem przez ich dzieci liceów ogólnokształcących.

[9] Dowód w aneksie A.3.

Sumy kwadratów wskaźników Quételeta dla poszczególnych modeli kanonicznych są zarazem równe kwadratom korelacji kanonicznych[10]

$$Q(M^{(s)}) = (r^{(s)})^2 \tag{6.35}$$

Dostarcza to kolejnej interpretacji wielkości pól w modelach kanonicznych. Pozwala bowiem zidentyfikować te pola modelu, które w największym stopniu odpowiedzialne są za uzyskaną wartość korelacji.

Nie należy też zapominać o jeszcze jednej własności modeli kanonicznych. Różnice współrzędnych kanonicznych dla wierszy bądź kolumn określają wielkości dystansów między profilami wierszowymi lub kolumnowymi, gdy wielkości te wyrażone są w postaci dystansów chi-kwadrat (wzory 6.1 i 6.2). Wielkość dystansu między profilami dowolnych wierszy i_1 oraz $i_2$wynosi wówczas[11]

$$\delta_{i_1 i_2}(M^{(s)}) = r^{(s)} * \left| x_{i_1}^{(s)} - x_{i_2}^{(s)} \right| \tag{6.36}$$

Analogiczny wzór przedstawić można dla kolumn j_1 oraz j_1

$$\delta_{j_1 j_2}(M^{(s)}) = r^{(s)} * \left| y_{j_1}^{(s)} - y_{j_2}^{(s)} \right| \tag{6.37}$$

Z postaci wzorów (6.36) i (6.37) wynika, że dystanse chi-kwadrat dla modeli kanonicznych są addytywne, co pozwala zobrazować je graficznie w jednym wymiarze. Własność ta stanowi o istocie metody analizy tablic zwanej analizą korespondencji (rozdział 7).

Omawiane ujęcie, nazwane dekompozycją tablic, w literaturze w zasadzie nie występuje – przynajmniej w przedstawionej postaci. W tych nielicznych wypadkach, gdy do analizy tablic wykorzystuje się kanoniczne korelacje, modele związku definiowane są nieco inaczej. Konstruuje się je jako sumę t pierwszych tablic kanonicznych. Oznaczmy tego rodzaju modele jako $P^{(t)}$ (literę P przyjąłem od słowa „przewidywanie"). Wtedy

$$P^{(t)} = E + C^{(1)} + C^{(2)} + \ldots + C^{(t)} \tag{6.38}$$

gdzie $1 \leq t \leq S$. Tego rodzaju ujęcie proponowano w literaturze wiele razy (Eckart i Young 1936; Lancaster 1969; Goodman 1986; Gilula i Haberman 1986). Model, w którym wykorzystuje się wszystkie tablice kanoniczne, nazywany jest „nasyconym". Jest on równoważny modelowi (6.26). Z kolei model ograniczony do dwóch pierwszych tablic kanonicznych stanowi punkt wyjścia analizy korespondencji (Mirkin 2001: 115–116).

[10] Dowód w aneksie A.4.
[11] Dowód w aneksie A.1.

Modele klasy P spełniają szereg dalszych relacji, które można wykorzystać dla interpretacji rozwiązania kanonicznego. Gdy model tego rodzaju złożymy z kolejną tablicą kanoniczną, to otrzymamy następny model tej klasy

$$P^{(t-1)} + C^{(t)} = P^{(t)} \tag{6.39}$$

dla $t = 1, 2,..., S$. Odpowiada to intuicji składania ze sobą kolejnych „warstw" w model związku. Oznaczmy jako $D^{(t)}$ wielkości w polach tablicy N niewyjaśnione przez dany model

$$D^{(t)} = N - P^{(t)} \tag{6.40}$$

Pozwala to sformułować alternatywną interpretację dla kryterium tworzenia tablicy kanonicznej, jako minimalizację wielkości

$$\sum_{i-1}^{w}\sum_{j=1}^{k}\frac{\left(d_{ij}^{(t-1)} - c_{ij}^{(t)}\right)^2}{e_{ij}} \rightarrow \min \tag{6.41}$$

Kolejna tablica kanoniczna dopasowywana jest w sensie powyższego kryterium do różnic liczebności, które nie zostały wyjaśnione przez wcześniejsze tablice kanoniczne. Do podanej interpretacji odwoływaliśmy się już wcześniej (wzór 6.10). Ma ona bardziej naturalny charakter niż kryterium maksymalizacji współczynnika korelacji.

Liczba stopni swobody dla modeli należących do klasy $P^{(t)}$ wynosi (Goodman 1986: 249–250)

$$df = (w - t - 1)(k - t - 1) \tag{6.42}$$

Gdy $t = 0$, czyli gdy model obejmuje wyłącznie niezależność, to wtedy wielkość (6.41) sprowadza się do $(w - 1)(k - 1)$, czyli do liczby stopni swobody tablicy N. Z podanej formuły wyprowadzić można wzór dla liczby stopni swobody wiązanych przez modele kanoniczne $M^{(s)}$. Wynosi ona

$$df = w + k - 2s - 1 \tag{6.43}$$

Należy jednak zastrzec, że nie jest to wielkość, którą można stosować przy konstruowaniu testów statystycznych dla tych modeli. Dla $s \geq 2$ testy takie nie są bowiem określone, gdyż dla rozważanych sytuacji nie można stworzyć modelu badania reprezentacyjnego. Wynika to z własności analitycznych stosowanej metody (Kendall i Stuart 1979: 606). Drugiej tablicy kanonicznej nie można wyodrębnić bez uprzedniego wyodrębnienia pierwszej. Nie ma więc sensu testowanie wyłącznie hipotezy, że wartość drugiej korelacji kanonicznej jest w populacji równa zeru, bez powiązania jej z wartością pierwszej korelacji.

Wzór (6.43) może więc służyć wyłącznie do określenia liczby stopni swobody traconych przy dołączaniu do modelu kolejnej tablicy kanonicznej – czyli kosztu, który musimy ponieść w związku z komplikacją modelu. W praktyce przydatny okazuje się również wzór na liczbę stopni swobody wiązanych przez model klasy $P^{(t)}$. Wynosi ona

$$tw + tk - t(t + 2) \tag{6.44}$$

Wielkość tę otrzymać również można poprzez odjęcie od $(w - 1)(k - 1)$, czyli od liczby stopni swobody tablicy N, wielkości określonej w równaniu (6.42) – to jest liczby stopni swobody modelu $P^{(t)}$. Przykładowo, gdy model obejmuje tylko jedną tablicę kanoniczną, to jego „koszt" jest równy $w + k - 3$ stopnie swobody. Model skonstruowany na podstawie dwóch pierwszych tablic kanonicznych wymaga zaś poświęcenia $2w + 2k - 8$ stopni swobody. Do wielkości tych będziemy odwoływać się oceniając modele proponowane w ramach metody korespondencji, przedstawionej w kolejnym rozdziale.

6.8 Dyskusja

Metoda dekompozycji wyróżnia się wśród innych metod eksploracyjnych przede wszystkim tym, że pozwala ocenić koszty związane z podjęcia określonej decyzji. Wiadomo bowiem, jaką część zjawiska wyjaśniają poszczególne warstwy modelu, czyli kolejne tablice kanoniczne. Dzięki temu można zdecydować się na taki poziom wyjaśnienia, jaki uznamy za satysfakcjonujący. Stwarza to możliwość pogodzenia dwóch, nie zawsze zbieżnych ze sobą celów: uzyskania prostego, a zarazem akceptowalnego wyjaśnienia badanego zjawiska.

Użyteczność metody dekompozycji ujawnia się przede wszystkim w projektach, które mają charakter eksploracyjny. Metoda jest narzędziem uniwersalnym, o rozbudowanym arsenale własności analitycznych, co sprzyja wszechstronnej analizie badanego zjawiska i ułatwia trafne zidentyfikowanie tych aspektów, które decydują o jego kształcie. Uniwersalność metody oznacza jednak, że całkowicie abstrahuje ona od *meritum* badanego problemu.

Metoda dekompozycji przez wiele lat nie zyskała wśród badaczy takiego uznania, do jakiego predysponują ją stwarzane możliwości. Przyczyny tego stanu rzeczy wiążą się przede wszystkim z pragmatycznymi ograniczeniami metody. W naukach społecznych przez wiele lat obowiązywał paradygmat popperowski, zgodnie z którym badania powinny przede wszystkim służyć weryfikacji teorii. Metoda kanoniczna, jako eksploracyjna, nie nadaje się do tego celu. Z tego zapewne powodu większą popularność zyskały w latach siedem-

dziesiątych i osiemdziesiątych techniki wywodzące się z modelowania log-liniowego. Pozwalały one konstruować modele oparte bezpośrednio na założeniach teoretycznych, a następnie testować ich spójność z wynikami badań.

Inny powód, który ograniczał zainteresowanie badaczy metodą kanoniczną, wiąże się z dominującymi schematami analizy danych. Można zaryzykować twierdzenie, że w drugiej połowie XX wieku o zjawiskach społecznych myślano przede wszystkim w kategoriach międzykrajowych analiz porównawczych. Wysiłek skierowany był na identyfikację mechanizmów obecnych w jednakowej formie w różnych społeczeństwach, niezależnie od ich ustroju politycznego czy też poziomu rozwoju ekonomicznego i społecznego. Podstawowy schemat analizy danych wyglądał w tej sytuacji następująco: obraz tego samego zjawiska – na przykład międzypokoleniowej ruchliwości zawodowej – zestawiano dla pewnej liczby badanych krajów, starając się znaleźć wspólny model, który wyjaśniałby kształt zjawiska w każdym z nich. Omawiany schemat wyobrazić sobie można jako zestaw identycznie skonstruowanych tablic wypełnionych w wyniku badań, w których szuka się elementów wspólnych. Istota problemu nie odpowiada więc założeniom analizy kanonicznej, która byłaby w tym wypadku ukierunkowana na identyfikację specyficznych elementów badanego zjawiska w tablicach odpowiadających poszczególnym krajom.

Nie można jednak wykluczyć, że o braku zainteresowania metodą kanoniczną zadecydowały też czynniki, które niewiele mają wspólnego z rzeczywistą przydatnością tej metody do analizy zjawisk społecznych. Przyczyny te związane są z czymś, co można nazwać modą na określone metody analityczne. Na powstawanie tego rodzaju preferencji mają bez wątpienia wpływ ośrodki akademickie wiodące w danej dziedzinie. W latach siedemdziesiątych i osiemdziesiątych dominujące znaczenie w dziedzinie analizy tablic miały ośrodki amerykańskie (Agresti 2002: 627). Był to uniwersytet w Chicago, z którym związani między innymi byli: William Kruskal, Leo Goodman, Shelby Haberman, Clifford Clogg czy Zvi Gilula. Drugim takim ośrodkiem był Harvard, z którym związane były takie osoby, jak Frederick Mosteller, William Cochran, czy Stephen Fienberg. Ostatni z wymienionych wraz z Yvonne Bishop i Paulem Hollandem, są autorami jednego z bardziej popularnych podręczników na temat metod analizy tablic, na którym, jak można sądzić, wychowały się pokolenia badaczy (Bishop, Fienberg i Holland 1975). Wkład Goodmana jest z kolei tak przebogaty, że chociażby pobieżne jego omówienie wymagało osobnej ramki (4.2).

W okresie, o którym piszę, wytworzyła się moda na modelowanie log-liniowe. Nie należy się temu specjalnie dziwić biorąc pod uwagę fakt, że było to podejście atrakcyjne zarówno dla samych twórców metod, jak też ich użytkowników. Po pierwsze, było podejściem nowym, a rozwijanie nowych propozy-

cji zawsze dostarcza satysfakcji. Po drugie, obejmowało zróżnicowany zakres problemów i stosowanych schematów badawczych. A każdy problem i każdy schemat wymagał propozycji osobnego modelu, który w twórczy i formalnie elegancki sposób byłby w stanie przedstawić istotę badanego zjawiska.

Inne metody były w tym czasie marginalizowane, a jeśli już o nich wspominano, to w kontekście panującej mody. Analizę kanoniczną reprezentanci szkoły amerykańskiej wzięli na warsztat dopiero w połowie lat osiemdziesiątych. Absolutny autorytet w dziedzinie analizy tablic, czyli Leo Goodman, nie poświęcił jej jednak nigdy odrębnego tekstu. Zestawiał ją na ogół z modelem log-liniowym, własnego zresztą pomysłu, określanego mianem *association model*. Argumentując zarazem, że ten ostatni jest bardziej oszczędny (ang. *parsimonious*), gdyż angażuje mniejszą liczbę stopni swobody (Goodman 1986).

Zdaję sobie sprawę z tego, że wyjaśnianie preferencji dotyczących metod naukowego poznania w kategoriach panującej mody brzmi trochę niepoważnie. Nie potrafię jednak wskazać bardziej przekonujących argumentów. Pozostaje faktem, że proponowane w tym rozdziale podejście lokuje się gdzieś na peryferiach nauki, biorąc pod uwagę liczbę zastosowań czy też liczbę poświęconych mu publikacji. Czy oznacza to, że jego przydatność w wyjaśnianiu zjawisk jest rzeczywiście niewielka? Wielokrotnie zadawałem sobie to pytanie, analizując wyniki badań za pomocą przedstawionych metod. Uzyskiwane rezultaty dawały mi poczucie, że odkrywam znaczące prawidłowości. Nie chciałbym jednak powoływać się na własne doświadczenia, nawet te, które poparte są publikacjami. Nie uważam bowiem, aby był to skuteczny sposób przekonania kogokolwiek, że metoda kanonicznej dekompozycji tablic sprawdza się w wypadku dużo szerszego zakresu zagadnień, niż wynikałoby to z jej miejsca w nauce. W zamian zaproponuję coś innego. W następnym rozdziale oddam głos badaczom, którzy wyszli poza kanony panującej mody. Opowiedzą oni o tym, jak metoda prezentowana w tym rozdziale z Kopciuszka przeobraziła się w Królewnę.

ROZDZIAŁ 7

Obrazy zjawisk

Metodę omawianą w tym rozdziale nazwano trochę nieszczęśliwie. Gdy wymieniam jej nazwę podczas rozmowy z kimś niewtajemniczonym, to mój rozmówca natychmiast wyciąga wniosek, że pracuję w Instytucie Badań Literackich. Nie chodzi tu jednak o nieznane listy Adama Mickiewicza do Maryli Wereszczakówny, lecz o metodę badania tablic zwaną **analizą korespondencji**. Polega ona na wizualizacji podobieństwa profili wierszy i kolumn w tablicy, czyli sposobu, w jaki wzajemnie sobie odpowiadają (korespondują) kategorie obu cech.

Rozpocznę od przedstawienia historycznych korzeni metody. Wyjaśniają one wiele, gdy chodzi o miejsce analizy korespondencji wśród współcześnie stosowanych metod (7.1). Następnie przedstawię jej związki z omawianą w poprzednim rozdziale analizą kanoniczną (7.2). Są one bardzo bliskie, co pozwoli zaoszczędzić wyjaśnień dotyczących formalnych podstaw analizy korespondencji. W zamian, *gros* rozważań zostanie poświęconych omówieniu specyficznych elementów metody oraz przykładów jej zastosowań. Rozpoczniemy od przedstawienia zasad interpretacji wyników w postaci graficznej (7.3). Forma ta stanowi podstawową zaletę metody, pozwalającą spojrzeć na złożoność badanego zjawiska w sposób syntetyczny. W podrozdziale 7.4 przedstawię specyficzną terminologię stosowaną przez badaczy posługujących analizą korespondencji. Terminologia ta nawiązuje do mechaniki klasycznej, co poszerza zakres możliwości interpretacji uzyskanego rozwiązania. Podrozdział 7.5 poświęcony został kwestii relacji między uzyskanymi wynikami a liczebnościami, na których oparte są wnioski. Wbrew pozorom nie jest to problem wyłącznie statystyczny. Ten sam wynik wymagać może zupełnie innej interpretacji w zależności od społecznego kontekstu badanego zjawiska. Pokażę to na przykładach fikcyjnych badań, w których wśród respondentów znaleźli się monarcha i premier.

Podrozdział 7.6 rozpoczyna analizę przykładów pochodzących z rzeczywistych badań. Każdy z nich posłuży do pokazania pewnych korzyści, które trudno byłoby uzyskać za pomocą innych metod analizy tablic. I tak, w pod-

rozdziale 7.6 dokonam analizy składu społecznego studentów uczelni w Polsce międzywojennej. Omówię zarazem kwestię znaczenia, jakie dla interpretacji wyników ma sposób utworzenia wykresu. Podrozdział 7.7 poświęcony zostanie analizie sposobów, w jaki cytują się nawzajem badacze w nauce i w marketingu. Przykład ten pozwoli określić granice użyteczności dwuwymiarowego rozwiązania. W podrozdziale 7.8 za cel postawiłem sobie przedstawienie kluczowej korzyści, jaką stanowi użyteczność rozwiązania graficznego w wypadku, gdy cechy w tablicy mają znaczną liczbę kategorii. Do realizacji tego celu posłużę się danymi dotyczącymi wykorzystania mediów na europejskim rynku reklamy w 2008 roku.

W podrozdziale 7.9 omówię problemy związane z prezentacją w tablicy cechy ilościowej. Z sytuacją taką badacze mają do czynienia dość często. Niektóre dane statystyczne są dostępne jedynie w pogrupowanej postaci (na przykład wiek). Za pomocą prekategoryzowanych narzędzi zbiera się wiele informacji w badaniach sondażowych. Na przykład tak pyta się z reguły o dochód. Okazuje się, że analiza korespondencji może wiele pomóc w ustaleniu istoty związków tego rodzaju cech z innymi cechami. Ustalenia te stanowić mogą punkt wyjścia dla posłużenia się w dalszych analizach metodami właściwymi dla cech ilościowych – jak regresja czy analiza wariancji. Pokażę to na przykładach związku inteligencji z wykształceniem oraz wykształcenia z dochodami.

W ostatnim podrozdziale (7.10) przedstawię przykład zastosowania analizy korespondencji do analizy tablic obejmujących więcej niż dwie cechy. Skorzystam w tym celu raz jeszcze z omawianych już chyba na wszystkie sposoby danych na temat zależności wyboru szkoły od wykształcenia ojca. Tym razem dane te uzupełnione zostaną o wykształcenie matki. W pierwszej kolejności przedstawię rozwiązanie oparte na intensywnie obecnie rozwijanym modelu tak zwanych tablic łączonych. Przez wielu badaczy tablice łączone postrzegane są jako naturalne rozszerzenie analizy korespondencji na przypadek więcej niż dwóch cech. Propozycja ta pozostaje w pewnej sprzeczności z istotą analizy korespondencji, której dziedzinę od początku stanowiły tablice krzyżujące ze sobą dwie cechy, co umożliwiało prezentację związku w sposób odwołujący się do intuicji i wyobraźni. Zarazem nawiążę do innych możliwości uwzględnienia w konwencjonalnej tablicy więcej niż dwóch cech.

W podsumowaniu (7.11) przedstawię swój punkt widzenia w następującej kwestii. Czy w czasach, gdy przekaz obrazkowy coraz częściej wypiera przekaz słowny, uzyskiwane w analizie korespondencji obrazy zjawisk zastąpią tradycyjne tabele.

Podobnie jak w rozdziale 6, prezentację istoty proponowanych metod oddzielimy od kwestii wykonania stosowanych obliczeń. Oprogramowanie pozwalające obliczyć wymagane wielkości przedstawione zostało w aneksie B.

7.1 Historia analizy korespondencji

Pozwolę sobie rozpocząć od dykteryjki. W 1986 roku broniłem na Uniwersytecie Warszawskim rozprawy doktorskiej. Według ówczesnych przepisów obronę poprzedzał egzamin, aczkolwiek wiadomo było, że chodzi wyłącznie o formalność. Komisja doskonale zorientowana była, czym doktorant się zajmuje. Nie obawiałem się więc zupełnie, że ktoś z członków zada mi pytanie, które sprawiłoby mi trudności. Podczas egzaminu wszystko szło zgodnie ze scenariuszem aż do momentu, gdy nieżyjąca już profesor Antonina Kłoskowska zadała mi pytanie: co sądzi Pan o propozycjach metodologów francuskich dotyczących analizy związków w tabelach? Mimo że treść pytania obejmowała dziedzinę, w której czułem się ekspertem – przyznam szczerze, że zgłupiałem. Zaraz, zaraz... Jacy metodolodzy francuscy? Słyszałem chyba o wszystkich, ale nie o francuskich... Było sympatycznie, a zrobiło się nieprzyjemnie. Nie byłem w stanie nic tu wykombinować.

Mimo udanej obrony kac po tym zdarzeniu pozostał. W wolnej chwili poszedłem więc do czytelni, rozłożyłem przed sobą roczniki *Revue française de sociologie* i rozpocząłem studiowanie prac metodologów francuskich. Szybko jednak zniechęciłem się. Takie nazwiska jak Florens, Grémy, Prévot czy Merllié niewiele mi mówiły. Z treści artykułów też niewiele wynikało. Czysta geometria – punkty, wektory, kąty, odległości. Doszedłem do wniosku, że to nie dla mnie.

W początkach lat dziewięćdziesiątych pojechałem za granicę na konferencję, której tematem była bodajże archiwizacja danych. Na konferencji było parę osób, które reprezentowały branżę badań marketingowych. Wymieniały one między sobą uwagi na temat pewnej metody analizy tablic o francuskim rodowodzie, którą nazywali „analizą korespondencji". Że prosta, uniwersalna, dogłębna, intuicyjna, przydatna i tak dalej. Starałem się nie zdradzić swojej kompletnej ignorancji na ten temat. Odczułem jednak tak duży dysonans, że po powrocie do kraju prosto z lotniska udałem się do biblioteki. Okazało się, że jest parę pozycji na ten temat, na co wcześniej nie zwróciłem uwagi. Wtedy zrozumiałem, o co pytała mnie profesor Kłoskowska na egzaminie. Zrozumiałem co gorsza, że metoda analizy tablic, którą wtedy intensywnie promowałem w publikacjach czy na wykładach – to właśnie owa analiza korespondencji!

Powyższą dykteryjkę przytoczyłem z tego powodu, gdyż przypomina ona sposób, w jaki świat zaznajomił się i przekonał do analizy korespondencji (por. Clausen 1998: vii). Francuskie korzenie tej metody wiąże się z osobą Jean-Paula Benzécriego – profesora statystyki na Uniwersytecie Paryskim. Jest on dość ortodoksyjny w swoich poglądach na temat metod naukowego poznania. Uważa bowiem, że badając stworzoną przez Boga rzeczywistość, nie wolno

z góry przyjmować żadnych idei co do sposobu jej konstrukcji. Poznanie powinno być wolne od apriorycznych założeń czy też arbitralnego spojrzenia badacza (Van Meter i in. 1994: 128–130; Górniak 2000: 115–116).

Wszystko zaczęło się od tego, że w latach sześćdziesiątych XX wieku Benzécri zainteresował się możliwościami interpretacji danych w tablicach za pomocą pojęć geometrycznych, a także zastosowaniami do analizy tablic metod wielozmiennowych. Jego prace z tego okresu zostały w 1973 wydane w formie książki (Benzécri i in. 1973), której drugi tom nosił nazwę *L'Analyse des correspondances*. Słowo *correspondances* w tytule zostało użyte w liczbie mnogiej. Chodziło bowiem o jego matematyczne znaczenie jako synonimu relacji między obiektami. Do użyteczności propozycji Benzécriego szybko przekonali się badacze. Zaowocowało to nie tylko sporą liczbą ich zastosowań w analizach wyników badań, lecz również kilkoma książkami na temat samej metody (zob. Van Meter i inni 1994: 130).

Wszystkie te prace – podobnie jak samo dzieło Benzécriego – dostępne były po francusku. Francuzi nie dbali w tamtych czasach o to, aby cokolwiek promować poza Francją. W efekcie nie były prawie zupełnie znane badaczom w innych krajach. Nie były też znane w amerykańskich ośrodkach uniwersyteckich, które w latach siedemdziesiątych i osiemdziesiątych wyznaczały standardy metodologiczne w wielu dziedzinach, w tym także w obszarze metod analizy tablic. Być może przez to francuskie propozycje stopniowo odeszłyby w zapomnienie, jak dzieje się to niekiedy z wieloma wartościowymi ideami. Wsparcie przyszło jednak z zupełnie nieoczekiwanej strony. W 1979 roku francuski socjolog Pierre-Félix Bourdieu (1930–2002) wydał książkę *La Distinction, Critique sociale du jugement*. Książka ta stała się wydarzeniem nie tylko we Francji. Już pięć lat później przetłumaczono ją i udostępniono amerykańskim czytelnikom. Została też wydana w wielu innych krajach, w tym również w Polsce (Bourdieu 2006). W 1998 roku International Sociological Association uznało ją za jedną z dziesięciu najważniejszych pozycji w dziedzinie socjologii, jakie ukazały się w XX wieku.

Bourdieu sformułował tezę, że kapitał kulturowy wyznacza style życia. Do wykazania jej spójności z wynikami badań analiza korespondencji nadawała się idealnie. Bourdieu skonstruował szereg tabel, w których krzyżował klasę społeczną z wskaźnikami gustów i preferencji, a także z zachowaniami w określonych sytuacjach. Metoda korespondencji pozwoliła zwizualizować te zależności, co dawało niezwykle sugestywny efekt. Właściwie trudno byłoby wymyślić lepszy sposób promocji metody, niż zrobił to Bourdieu (Blasius 1994: 25–26).

Książka Bourdieu wywołała wzrost zainteresowania analizą korespondencji w sposób lawinowy. Dotarło ono również do amerykańskich uniwersytetów. Już w 1986 roku sam Leo Goodman uwzględnił tę metodę wśród innych

podejść do analizy tablic, które proponował badaczom (Goodman 1986). Rok później artykuł na ten temat zamieścił na łamach *American Journal of Sociology,* zachęcając w tytule, że pisze o metodach nowych (Goodman 1987). Wywołało to wzrost zainteresowania metodą u socjologów.

Amerykańscy akademicy we wszystkim muszą być jednak najlepsi. Bez trudu zauważyli, że propozycje ich francuskich kolegów *de facto* pokrywają się z metodą analizy kanonicznej (prezentowaną w rozdziale 6). Odkurzyli więc parę artykułów, z których wynikało, że metoda została zaproponowana wiele lat wcześniej, nim zainteresował się nią Benzécri. Poszli przede wszystkim tropem artykułu szkockiego statystyka M. O. Hilla (1974), którego już sam tytuł świadczył, że pewne sprawy przeoczono (*Correspondence analysis: A neglected multivariate method*). Hill w swoim artykule powoływał się co prawda na Benzécriego[1], lecz korzeni metody doszukał się w pracach Hirschfelda (1935), Fishera (1940) i Maunga (1941). Cytował również Williamsa (1952), którego wkład przedstawiliśmy w podrozdziale 6.5.3.

Ostatecznie Leo Goodman rozstrzygnął, a jego autorytetu nikt nie śmiałby podważyć, że metoda korespondencji miała następującą genezę. Hirschfeld wpadł na pomysł przypisania wierszom i kolumnom tablicy wartości skalowych – między którymi policzył korelację. Fisher niezależnie zrobił to samo, opierając się na odmiennej koncepcji optymalizacji rozwiązania. Zaś Maung rozwinął podejście Fishera, korzystając z kanonicznych korelacji. Omawiane podejście należy więc nazwać metodą „Hirschfelda–Fishera–Maunga" (np. Goodman 1996: 409, 2000: 203)[2].

Już wcześniej pisałem, jak bardzo wpływową postacią jest Alan Agresti (zob. ramka 3.1). Jest on autorem znanych podręczników z dziedziny analizy danych kategorialnych, dających przegląd chyba wszystkiego, co na ten temat powiedziano (Agresti 1984, 1990, 2002, 2007; Agresti i Finlay 2008). Agresti z wyraźną rezerwą odnosi się jednak do metody korespondencji. W podręczniku wydanym w 1984 roku napisał o niej co następuje (Agresti 1984: 227)[3]

> *Finally, there are methods not considered in this book, such as the graphical method called correspondence analysis (see Benzécri 1976), that provide yet alternative views of the data.*

[1] Hill odwołuje się nie do fundamentalnej pracy Benzécriego i in. (1973), lecz do artykułu z 1969 roku opublikowanego w języku angielskim w książce wydanej przez Academic Press. Powodem ignorowania prac Benzécriego przez badaczy amerykańskich nie była więc wyłącznie bariera językowa.

[2] Tę samą propozycję odnaleźć można również w innych pracach autora.

[3] Jak widać z przytoczonego cytatu, Alan Agresti nie orientował się zbyt dobrze w pracach Benzécriego. Książka, na którą się powołuje, wydana była bowiem nie w 1976, a w 1973 roku. W późniejszych wydaniach Agresti skorygował ten błąd.

To i tak dużo, gdyż w kolejnej książce z 1990 roku nie napisał nic na ten temat. Co się jednak odwlecze, to nie uciecze. W 2002 roku Agresti wydał rozszerzone i uzupełnione wydanie swojego fundamentalnego dzieła *Categorical Data Analysis*. W 732-stronicowej książce metodzie korespondencji poświęcił niecałe trzy strony (Agresti 2002: 382–384). Co ciekawe, wynika z nich, że twórcą metody był nie kto inny, jak Leo Goodman. Omawiając analizę korespondencji, Agresti nie cytuje bowiem nikogo innego. No, chyba że ktoś zajrzy do sekcji „Notes" owego rozdziału. Znajdzie tam dodatkowe informacje na temat historii metody i jej zastosowań, które zacytuję w całości (Agresti 2002: 399)

> *Correspondence analysis gained popularity in France under the influence of Benzécri (see, e.g., 1973). Goodman (1996) attributed its origins to H. O. Hartley, publishing under his original German name (Hirschfeld, 1935). Greenacre (1993) related it to the singular value decomposition of a matrix.*

Nauka nie znosi jednak próżni. Brak zainteresowania metodą korespondencji wśród amerykańskich akademików wypełniony został licznymi opracowaniami w tej dziedzinie, które zaczęły pojawiać się jak grzyby po deszczu w różnych częściach świata. W wydanej w 1994 roku pracy zbiorowej pod tytułem *Correspondence Analysis in the Social Sciences* (Greenacre i Blasius 1994) zaproszeni do współpracy autorzy reprezentowali takie kraje jak: Hiszpania, Niemcy, Francja, Holandia, Słowenia, Kanada i USA.

Siłę analizy korespondencji stanowi przede wszystkim pluralizm jej umiejscowienia wśród międzynarodowej społeczności badaczy (Falguerolles 2008: 27). Znalazła wsparcie nie tylko w najbardziej liczących się na świecie centrach akademickich, lecz również w wiodących ośrodkach badawczych, które decydują o kierunkach rozwoju metodologii badań społecznych. Nie chciałbym w tym miejscu wymieniać najważniejszych z tych ośrodków, gdyż godziłoby to w pluralistyczną istotę metody. Zwłaszcza że jej współczesny obraz kształtują nie tylko prace teoretyczne, lecz również – a może przede wszystkim – tysiące zastosowań. Według Erica Beha (2004) z roku na rok wzrasta liczba publikowanych artykułów na temat samej metody bądź jej zastosowań.

Fenomen analizy korespondencji bierze się między innymi stąd, że stanowi ona niezwykle pojemną platformę wymiany myśli. Obejmuje z jednej strony prezentację i analizę wyników badań z różnych dziedzin, o różnorodnej tematyce, realizowanych za pomocą szerokiej gamy technik badawczych. Z drugiej strony inspiruje do poszukiwań w sferze samej metodologii analizy danych. Już w tej chwili platforma analizy korespondencji obejmuje badanie powiązań między wieloma cechami, zagadnienia budowy skal czy inne przydatne techniki analizy danych, jak na przykład *conjoint* (Górniak 2000; Blasius i Greenacre

2006a). Można przewidywać, że liczba podejść do analizy danych rozpatrywanych w ramach modelu korespondencji będzie nadal rosnąć.

Na koniec warto wspomnieć o jeszcze jednej sprawie. Przez prawie czterdzieści lat różnych zawirowań metoda korespondencji nie zatraciła swojego głównego przesłania. Jak wyraził je Fionn Murtagh (2007: 275) – profesor informatyki na Uniwersytecie Londyńskim – *the data is King*. Idea, która wywodzi się z metafizycznych przekonań Benzécriego, zrobiła w badaniach ogromną furorę. Zaś jedno z jej dzieci – analiza korespondencji – wciąż zaskakuje nowymi możliwościami.

7.2 Czym różni się analiza korespondencji od analizy kanonicznej

Współczesna analiza korespondencji nie jest pojedynczą metodą, lecz całą grupą technik służących wizualizacji danych (Górniak 2000; Blasius i Greenacre 2006a). Ograniczymy się do omówienia jedynie najprostszej techniki z tej grupy, od której zresztą omawiane podejście zapoczątkowało swój rozwój. Jest to tak zwana **prosta analiza korespondencji** (ang. *simple*), której celem jest wizualizacja związku przedstawionego w tablicy. Przy czym, aby wywód uprościć, określenie „prosta" będziemy pomijać.

Przy tak ograniczonym zakresie rozważań na pytanie sformułowane w tytule podrozdziału odpowiemy, że od strony praktycznych zastosowań wszelkie istniejące różnice wolno jest zignorować. W analizie korespondencji skorzystać bowiem można z aparatu analitycznego metody kanonicznej – i tak się na ogół czyni. Z aparatem tym mieliśmy okazję zapoznać się w rozdziale 6. Przedstawione pojęcia dystansu chi-kwadrat, współrzędnych kanonicznych, korelacji kanonicznej oraz dekompozycji chi-kwadrat wystarczają w zupełności, aby analizę korespondencji stosować w praktyce.

Model korespondencji to taki wariant analizy kanonicznej, w którym bierze się pod uwagę dwie pierwsze tablice kanoniczne. Korzystając z symboli wprowadzonych w rozdziale 6, liczebności $\hat{n}_{ij}$ w poszczególnych polach tablicy estymowane za pomocą modelu korespondencji przedstawić można jako

$$\hat{n}_{ij} = e_{ij}\left(1 + r^{(1)}x_i^{(1)}y_j^{(1)} + r^{(2)}x_i^{(2)}y_j^{(2)}\right) + d_{ij}^{(2)} \tag{7.1}$$

W formule (7.1) istotne jest między innymi to, że liczebności estymowane odbiegać mogą od liczebności n_{ij}, stanowiących wynik badania. Różnicami są przedstawione na końcu formuły wielkości $d_{ij}^{(2)}$, czyli to, co pozostaje niewyjaśnione przez dwie pierwsze tablice kanoniczne. Dopiero wtedy, gdy wielkości te pozostają znaczne, formułuje się pytanie, czy model korespondencji można uznać za należycie dopasowany do danych. Jeśli nie, to jedną z dróg

wyjścia stanowi uwzględnienie większej liczby tablic kanonicznych – wystarczającej dla osiągnięcia zadowalającej dokładności oszacowań. Rozwiązaniami takimi nie będziemy się jednak zajmować. Zatracają one bowiem najważniejszą korzyść metody, jaką stanowi łatwość prezentacji wyników za pomocą rysunku na płaszczyźnie. Co prawda, gdy rozwiązanie uwzględnia więcej niż dwa wymiary, to wtedy rysunek również można sporządzić metodą rzutowania wyników na płaszczyznę. Wymaga to jednak dodatkowych rozstrzygnięć, przez co sprawa zaczyna się komplikować. Do kwestii tej powrócimy w podrozdziale 7.7.

Ograniczenie się do konstatacji, że analiza korespondencji jest równoważna modelowi dwóch pierwszych tablic kanonicznych, zubożyłoby jednak możliwości korzystania z metody. Społeczność zwolenników analizy korespondencji wypracowała bowiem swój własny język opisu otrzymywanych wyników. Stosowana terminologia kładzie nacisk na dodatkowe możliwości interpretacyjne, do których nie przywiązywano tak dużej roli w poetyce modeli kanonicznych. Język ten, a także możliwości interpretacyjne, jakie stwarza, warto więc omówić.

Terminologia metody korespondencji wywodzi się z prac Benzécriego. Jego celem było skojarzenie proponowanych pojęć z intuicjami dotyczącymi relacji przestrzennych znanych z innych dziedzin, głównie z mechaniki. Dlatego też liczebności brzegowe wierszy i kolumn nazwał **masami**, zaś różnice między profilami wierszy i kolumn – **dystansami** pomiędzy tymi masami. Zgodnie z mechaniką newtonowską, im większa masa punktu materialnego, tym bardziej oddziałuje on na pozostałe punkty, które w tym wypadku odpowiadają kategoriom cech w tablicy. Zarazem, im dystans między masami jest większy, tym oddziaływanie mniejsze.

Od Benzécriego pochodzi również pojęcie **bezwładności** (ang. *inertia*). Jako całkowitą bezwładność badanego zjawiska – czyli układu liczebności w polach tablicy – przyjmuje się sumę kwadratów korelacji kanonicznych

$$\sum_{s=1}^{S}\left(r^{(s)}\right)^2 \tag{7.2}$$

gdzie S jest liczbą tablic kanonicznych możliwych do utworzenia, czyli $\min(w-1, k-1)$. Całkowita bezwładność odpowiada tablicy otrzymanej w wyniku badania, nie zaś jej modelowi przyjętemu w analizie korespondencji, który – jak przyjęliśmy – ograniczony jest do dwóch pierwszych tablic kanonicznych. Jak wykazaliśmy w rozdziale 6, zdefiniowaną tak bezwładność – równą co do wielkości χ^2 / n – zdekomponować można na wiele sposobów. Między poszczególne tablice kanoniczne, między wiersze i kolumny w mode-

lach składających się z dowolnej liczby tablic kanonicznych, a także między poszczególne pola tablicy (zob. wzory od 6.30 do 6.35). Na gruncie analizy korespondencji twierdzi się, że zdekomponować ją również można między dystanse pomiędzy wierszami bądź kolumnami (Greenacre 1994: 12). Bezwładność stanowi więc uniwersalną płaszczyznę oceny znaczenia poszczególnych elementów tablicy – to jest jej pojedynczych pól, poszczególnych wierszy, kolumn czy też dystansów między nimi – zarówno w wypadku obserwowanego, jak też modelowanego kształtu związku.

7.3 Zasady interpretacji rozwiązania w postaci graficznej

Zasady te najlepiej wyjaśnić na konkretnym przykładzie. W tym celu skorzystamy z omawianego wcześniej związku między wykształceniem ojca a wyborem szkoły ponadgimnazjalnej przez dziecko[4]. Tabela 7.1 przedstawia cztery różne sposoby wypełnienia tej tablicy – wszystkie oparte na danych fikcyjnych. Posłużenie się danymi fikcyjnymi ma tę zaletę, że pozwalają one w przejrzysty sposób przedstawić zależności między omawianymi pojęciami[5]. We wszystkich czterech przykładach przyjęto, że liczba uczniów uczęszczających do poszczególnych rodzajów szkół jest jednakowa, zaś badaniu podlegało po 100 uczniów z każdego toru kształcenia. Część [1] tabeli 7.1 obrazuje bardzo słaby związek między rozważanymi cechami. Gdyby wybór szkoły nie zależał od wykształcenia ojca, to w każdym polu tabeli należałoby się spodziewać po około 25 uczniów. Obserwowane liczebności niewiele odbiegają od tych wielkości. Tym samym profile wykształcenia ojca wśród uczniów poszczególnych rodzajów szkół są do siebie zbliżone.

W części [2] tabeli 7.1 zarysowuje się pewien związek między wykształceniem ojca a wyborem szkoły. W liceach ogólnokształcących jest nieco więcej uczniów, których ojcowie mają wykształcenie wyższe niż w pozostałych rodzajach szkół. Z kolei w zawodówkach jest nieco więcej uczniów, których ojcowie mają wykształcenie podstawowe. Porównanie tych samych liczebności w części [3] tabeli wskazuje, że prezentuje ona jeszcze silniejszy związek. Natomiast w części [4] mamy do czynienia ze skrajnie zróżnicowanymi pro-

[4] Przykład i sposób jego interpretacji oparte są na przykładzie omawianym przez Greenacre'a (1994: 12–14).

[5] Należy zaznaczyć, że prezentowane dane fikcyjne nie replikują marginesów autentycznej tabeli, podczas gdy marginesy te określone są w sposób zewnętrzny wobec badanego zjawiska (wyznaczają jego ramy). Opisane modele nie mogą więc służyć jako punkt odniesienia dla zjawiska, które omawialiśmy wcześniej. Celem jest tu wyłącznie ilustracja związków między pojęciami stosowanymi w analizie korespondencji.

filami. Ojcowie prawie wszystkich uczniów liceów ogólnokształcących mają wykształcenie wyższe. Na drugim biegunie mamy szkoły zasadnicze grupujące uczniów, których ojcowie mają wykształcenie podstawowe. Uczniowie, których ojcowie mają wykształcenie średnie bądź zasadnicze, w przeważającej większości trafiają natomiast do techników.

Tabela 7.1
Liczebności i miary bezwładności dla przykładów zależności między wykształceniem ojca a wyborem szkoły ponadgimnazjalnej

Dane fikcyjne

	liczebności				miary bezwładności			
wykształcenie ojca	Liceum ogólnokształcące	Technikum	Szkoła zasadnicza	ogółem	Liceum ogólnokształcące	Technikum	Szkoła zasadnicza	ogółem
[1] prawie jednakowe profile								
wyższe	26	24	23	73	0,0004	0,0000	0,0002	0,0006
średnie	25	26	25	76	0,0000	0,0001	0,0000	0,0001
zasadnicze	25	26	26	77	0,0001	0,0000	0,0000	0,0001
podstawowe	24	24	26	74	0,0001	0,0001	0,0002	0,0004
ogółem	100	100	100	300	0,0005	0,0001	0,0005	0,0012
[2] słabo zróżnicowane profile								
wyższe	34	29	19	82	0,0054	0,0003	0,0085	0,0142
średnie	29	27	21	77	0,0014	0,0002	0,0028	0,0045
zasadnicze	21	25	27	73	0,0015	0,0001	0,0010	0,0026
podstawowe	16	19	33	68	0,0065	0,0020	0,0157	0,0242
ogółem	100	100	100	300	0,0149	0,0026	0,0280	0,0455
[3] silnie zróżnicowane profile								
wyższe	55	14	8	77	0,1117	0,0177	0,0405	0,1700
średnie	29	33	11	73	0,0030	0,0103	0,0244	0,0376
zasadnicze	11	36	31	78	0,0288	0,0128	0,0032	0,0449
podstawowe	5	17	50	72	0,0501	0,0068	0,0939	0,1508
ogółem	100	100	100	300	0,1937	0,0476	0,1620	0,4033
[4] skrajnie zróżnicowane profile								
wyższe	98	1	0	99	0,4268	0,1034	0,1100	0,6402
średnie	2	49	0	51	0,0441	0,2008	0,0567	0,3016
zasadnicze	0	45	5	50	0,0556	0,1606	0,0272	0,2433
podstawowe	0	5	95	100	0,1111	0,0803	0,3803	0,5717
ogółem	100	100	100	300	0,6376	0,5451	0,5742	1,7568

Na rycinie 7.1 przedstawione zostały graficzne obrazy omawianych związków uzyskane metodą analizy korespondencji. Czarnymi kółkami oznaczono rodzaje szkół, zaś białymi kwadracikami poziomy wykształcenia ojców. Na

wszystkich czterech obrazach przyjęto konwencję polegającą na tym, że szkoły zostały umieszczone w tych samych punktach wykresu. Przy czym punkty te zostały dobrane w specjalny sposób. Mianowicie obrazują fikcyjną sytuację, w której wszyscy uczniowie uczęszczający do szkół danego rodzaju nie różniliby się poziomem wykształcenia ojców. Jak wiemy z poprzednich rozdziałów, nie odpowiada to sytuacji rzeczywistej. Ułatwia natomiast zrozumienie sposobu prezentacji układu liczebności tablicy za pomocą narzędzi analizy korespondencji.

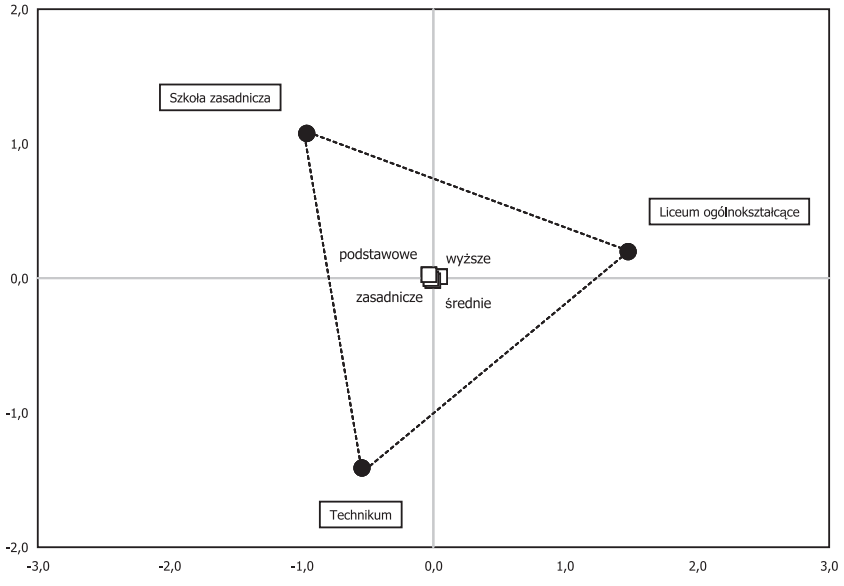
Rozważmy rycinę oznaczoną jako [a]. Obrazuje ona bardzo słabe zróżnicowanie profili w tablicy. Jak widać, punkty obrazujące profile odpowiadające kategoriom wykształcenia ojca zgrupowały się w środku wykresu, zwanego też jego **punktem centralnym** (ang. *centroid*). Środek ten odpowiada przecięciu pionowej i poziomej linii – wykreślonych w kolorze szarym – nazywanych **osiami** wykresu. Generalnie, im punkt reprezentujący wiersz lub kolumnę tablicy leży bliżej punktu centralnego, tym odpowiadający mu profil podobny jest do profilu brzegowego (marginesu) danej cechy, zwanego też **profilem przeciętnym**. I odwrotnie, im bardziej punkt jest odległy od punktu centralnego, tym profil wiersza bądź kolumny różni się od profilu przeciętnego. Miarą owych różnic profili jest **dystans chi-kwadrat**, omówiony w podrozdziale 6.2.

Na rycinie [b] punkty odpowiadające profilom wierszy tablicy występują w pewnym oddaleniu od punktu centralnego. Na rycinie [c] oddalenie to jest już znaczne. Przy czym widać wyraźnie, że każdy z punktów przesunął się w kierunku tego z wierzchołków trójkąta utworzonego przez rodzaje szkół, które wykazują przewagę uczniów o danym rodzaju wykształcenia ojców. Z położenia punktów wynika więc, że dzieci ojców mających wykształcenie wyższe uczą się przede wszystkim w liceach ogólnokształcących. Na wykresie odpowiadający im punkt znalazł się najbliżej wierzchołka oznaczającego licea ogólnokształcące. Dzieci ojców o wykształceniu średnim i zasadniczym kształcą się głównie w technikach, lecz między tymi dwiema grupami jest pewna różnica. Gdy ojciec ma wykształcenie średnie, to część dzieci wybiera licea ogólnokształcące, zaś gdy ojciec ma wykształcenie zasadnicze, to syn czy córka częściej wybierają szkoły zasadnicze. Na wykresie punkt odpowiadający kategorii ojców o wykształceniu średnim znalazł się przez to między czarnymi kółkami oznaczającymi licea i technika, zaś punkt dla ojców o wykształceniu zasadniczym między kółkami oznaczającymi technika i szkoły zasadnicze.

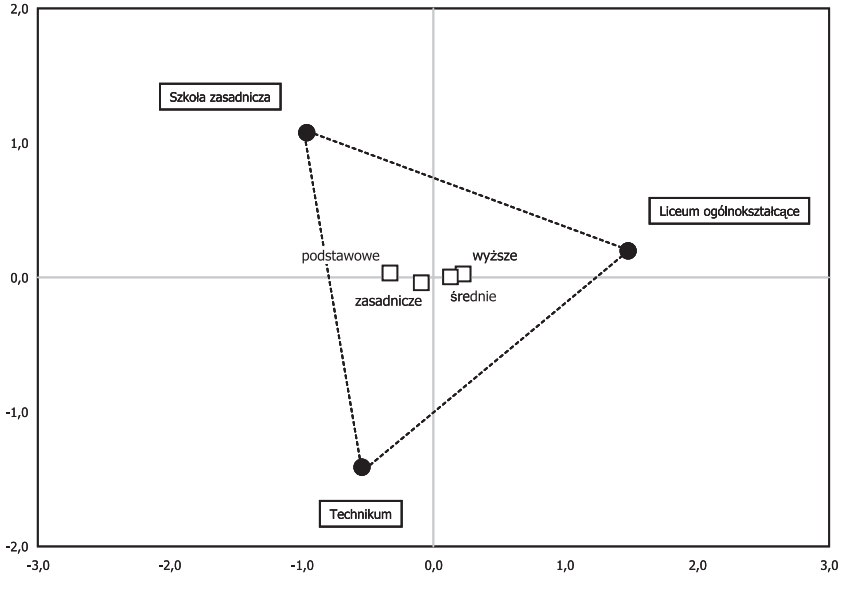
Reguły wyznaczające położenie punktów na wykresie wyobrazić sobie też można w następujący sposób (por. Greenacre 1994: 14). Przypuśćmy, że punkt odpowiadający kategorii wykształcenia ojców jest pewnym materialnym przedmiotem, który za pomocą trzech sprężyn zamocowano do trzech wierzchołków trójkąta. Sprężyny te wykonane są z drutu różnej grubości, co

Rycina 7.1
Graficzny obraz profili wyboru szkoły ponadgimnazjalnej przez dzieci o różnym wykształceniu ojca dla przykładów z tabeli 7.1. Dane fikcyjne

[a] prawie jednakowe profile

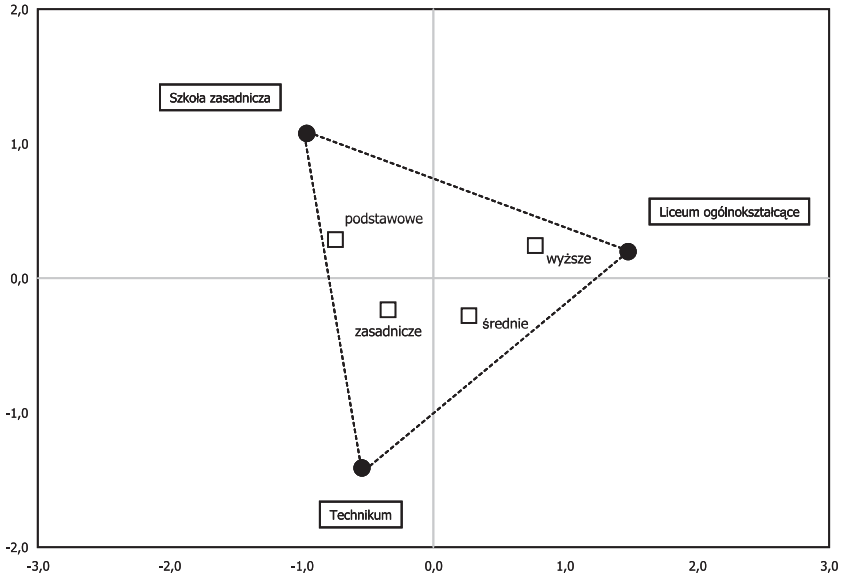


[b] słabo zróżnicowane profile

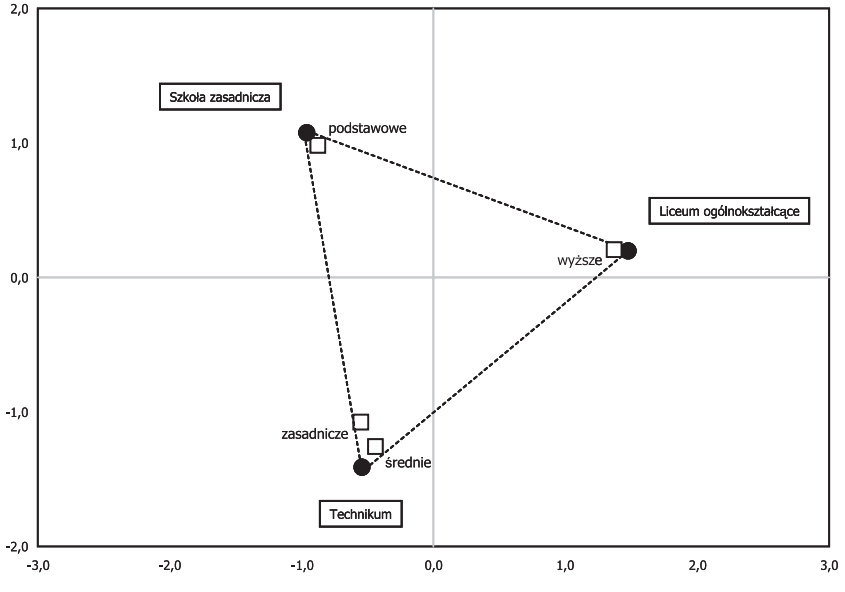


Rycina 7.1: kontynuacja

[c] silnie zróżnicowane profile



[d] skrajnie zróżnicowane profile



daje im niejednakową sprężystość. Grubość drutu zależy zaś w sposób rosnący do tego, ile dzieci mających ojców o tym wykształceniu uczy się w szkołach poszczególnych rodzajów. Rozpatrzmy dzieci ojców o wykształceniu zasadniczym (cały czas analizujemy wykres [c] na rycinie 7.1). Najwięcej z nich wybrało technika, nieco mniej szkoły zasadnicze. Sprężyny przymocowane do wierzchołków trójkąta odpowiadających technikom i szkołom zasadniczym wywierać więc będą większą siłę od sprężyny wiążącej przedmiot symbolizujący kategorię ojców z wierzchołkiem oznaczającym licea. W efekcie położenie przedmiotu na planszy ustali się w sposób podany na rycinie. Znajdzie się on najbliżej techników, nieco dalej wobec szkół zasadniczych, zaś zdecydowanie najdalej w stosunku do liceów ogólnokształcących.

Wykres [d] obrazuje skrajną sytuację, w której wykształcenie ojca praktycznie przesądza o wyborze szkoły. Prawie wszystkie dzieci ojców o wykształceniu wyższym kształcą się w liceach ogólnokształcących. Punkt odpowiadający tej kategorii ojców leży więc bardzo blisko wierzchołka reprezentującego ten rodzaj szkół. Gdyby wszystkie dzieci ojców o wykształceniu wyższym szły do liceów ogólnokształcących, to oba punkty na wykresie pokryłyby się. Podobną sytuację obserwujemy w pozostałych narożnikach trójkąta. W narożniku reprezentującym technika znalazły się aż dwa punkty, odpowiadające kategoriom ojców o wykształceniu średnim i zasadniczym. Jak bowiem wynika z tabeli 7.1, dzieci mające ojców w obu kategoriach w zdecydowanej większości wybierają technika. Przy czym punkt odpowiadający wykształceniu zasadniczemu leży nieco bliżej wierzchołka szkół zasadniczych, gdyż 5 na 50 dzieci z tej kategorii wybrało nie technika, lecz szkoły zasadnicze.

Dwie spośród omawianych własności graficznej postaci rozwiązania okazują się szczególnie przydatne podczas interpretacji rzeczywistych danych, więc warto je powtórzyć.

Po pierwsze, im bardziej punkt odpowiadający wierszowi tablicy leży bliżej punktu odpowiadającego pewnej kolumnie, tym bardziej profil wiersza skupiony jest w danej kategorii cechy. W omawianym przykładzie, im punkt odpowiadający kategorii wykształcenia ojca (wierszowi tablicy) leżał bliżej punktu odpowiadającemu danej kolumnie (rodzajowi szkół), tym większy odsetek uczniów o danym wykształceniu ojca uczęszczał do szkół danego rodzaju. Własność ta zachodzi również w drugą stronę, aczkolwiek omawiany przykład tego nie pokazuje. Punkty odpowiadające kolumnom zostały bowiem ustalone z góry, w sposób wspólny dla czterech analizowanych tablic[6].

[6] Omawiana własność zachodzi w sposób ścisły jedynie w wypadku wykresów sporządzanych w ramach tak zwanej normalizacji niesymetrycznej (Greenacre i Hastie 1987; Górniak 2000: 119–126; Greenacre 2006: 62–63). Wykresy na rycinie 7.1 spełniają ten warunek, nato-

Druga z godnych zapamiętania własności sprowadza się do tego, że im bardziej punkty reprezentujące wiersze lub kolumny odsunięte są od środka wykresu, tym przedstawiony w tablicy związek jest silniejszy. Formułując tę własność niejako odwrotnie, można powiedzieć, że im punkty leżą bliżej środka, tym bardziej profile są do siebie podobne, a więc układ liczebności w tablicy przypomina model niezależności.

7.4 Bezwładność a graficzna postać rozwiązania

Druga z wymienionych wyżej własności związana jest z pojęciem bezwładności. Bezwładność posiada bowiem interpretację odzwierciedlającą ową własność, zaczerpniętą zresztą z mechaniki. Ściśle rzecz biorąc, interpretacji takiej dostarcza wprowadzone w mechanice pojęcie **momentu bezwładności** (Greenacre 1994: 12). Dla układu *n* punktów materialnych moment bezwładności *I* określa się w mechanice jako sumę momentów bezwładności wszystkich tych punktów względem ustalonej osi obrotu

$$I = \sum_{i=1}^{n} m_i d_i^2 \tag{7.3}$$

gdzie m_i oznacza masę *i*-tego punktu, zaś d_i jego odległość od osi obrotu.

Parametry rozwiązania w analizie korespondencji zachowują się w sposób analogiczny, jak fizyczne wielkości we wzorze (7.3). Wróćmy na chwilę do wykresu [a] z ryciny 7.1. Punkty odpowiadające wierszom skupione są blisko środka wykresu, który stanowi odpowiednik osi całego układu. Odpowiednikiem występujących we wzorze (7.3) mas są wielkości kategorii wykształcenia ojców, nazywane w analizie korespondencji również masami. Odpowiednikiem dystansów są zaś odległości punktów od środka układu. Ze względu na fakt, że odległości te są niewielkie, bezwładność całego układu jest również niewielka. Bezwładności łączne rozpatrywanych tablic odczytać można w prawym panelu tabeli 7.1. Dla tabeli [1] bezwładność wynosi 0,0012, dla tabeli [2] wynosi zaś 0,0455. Na wykresie [b] widać, że punkty odsunęły się nieco od środka, więc gdyby był to fizyczny układ punktów materialnych, to jego bezwładność –

miast wszelkie dalsze wykresy prezentowane w tym rozdziale sporządzone będą w ramach normalizacji symetrycznej – preferowanej w praktycznych zastosowaniach metody. Należy zaznaczyć, że w analizie korespondencji nie definiuje się dystansu między profilem wiersza a profilem kolumny tablicy. To, że na wykresie punkt odpowiadający wierszowi może leżeć blisko punktu odpowiadającego kolumnie wynika stąd, że oba punkty są podobnie zorientowane wobec punktu centralnego i osi wykresu. Bardziej pogłębione omówienie problemu można znaleźć w podanej literaturze.

w sensie newtonowskim – byłaby większa niż układu z wykresu [a]. Jeszcze większą bezwładność ma układ z wykresu [c], a największą – z wszystkich rozpatrywanych – układ z wykresu [d].

Gdyby punkty zostały maksymalnie odsunięte od środka, na ile jest to tylko możliwe, to wtedy pokrywałyby się z wierzchołkami trójkąta. W tablicy odpowiada to sytuacji, gdy w każdym wierszu jest tylko jedna liczebność niezerowa. W rozpatrywanym przykładzie oznaczałoby to, że wykształcenie ojca determinuje rodzaj szkoły dla dziecka. Gdyby tak skonstruowaną tablicę dodać do tabeli 7.1 i obliczyć jej bezwładność, czyli wielkość χ^2 / n, to wyniosłaby ona 2. Maksymalna bezwładność jest bowiem zawsze równa $\min(w - 1, k - 1)$, niezależnie od kształtów rozkładów brzegowych cech przedstawionych w tablicy.

Analogie między bezwładnością w mechanice a bezwładnością liczoną w analizie korespondencji nie są przypadkowe. Te ostatnie wielkości można bowiem otrzymać z wzoru (7.3), podstawiając jako d_i odległości punktów na wykresie od środka układu, zaś jako masy częstości brzegowe kategorii, czyli a_i / n bądź b_j / n. Obliczona w ten sposób bezwładność poszczególnych punktów podana została w tabeli 7.1 w rubryce ogółem (dla wierszy) w wypadku każdej z czterech tablic. Na wykresie [c] punkty odpowiadające wykształceniu podstawowemu i wyższemu są bardziej odległe od środka, niż dwa pozostałe. A ponieważ wszystkie punkty na tym wykresie mają zbliżone masy (zob. liczebności brzegowe wierszy w części [3] tabeli 7.1), stąd bezwładności kategorii wykształcenia wyższego i podstawowego są większe, niż średniego i zasadniczego.

Bezwładności punktów nie zawsze odczytać można z wykresu. Na wykresie [d] wszystkie kategorie wykształcenia leżą w podobnej odległości od środka. Ponieważ jednak liczebności kategorii wykształcenia średniego i zasadniczego są około dwukrotnie mniejsze niż dwóch pozostałych, stąd ich bezwładności są również około dwukrotnie mniejsze (relacja ta wynika z wzoru 7.3). Bezwładność poszczególnych punktów ma znaczenie dla interpretacji zjawiska obrazowanego na wykresie. Mając bowiem dwie kategorie w podobny sposób odbiegające od przeciętnego profilu, większe znaczenie jesteśmy skłonni przypisywać tej z nich, która jest bardziej liczna. Jeśli więc bezwładności punktów nie wynikają z ich konfiguracji na wykresie, to zawsze warto je dodatkowo podać.

Nawet jednak, gdy bezwładności punktów wynikają z wykresu, gdyż na przykład skądinąd wiadomo, że kategorie mają podobną wielkość, to zawsze dla poprawnej interpretacji rozwiązania konieczne jest uwzględnienie **bezwładności całkowitej** układu. Wykresy w analizie korespondencji sporządza się bowiem w taki sposób, aby wygodnie było odczytać różnice w położeniu

punktów. Jeśli więc punkty skupione są ściśle wokół środka układu (tak jak na wykresie [a] ryciny 7.1), to zwiększa się odpowiednio skalę wykresu, aby rozgęścić punkty. Może doprowadzić to do błędnej konkluzji, że różnice między profilami są wyraźne, podczas gdy związek w tablicy niewiele różni się od niezależności i najrozsądniej byłoby przyjąć, że właśnie model niezależności najlepiej wyjaśnia rozważane zjawisko. Aby uchronić odbiorcę wyników opracowania przed niebezpieczeństwem wyciągnięcia nietrafnych wniosków, badacze posługujący się analizą korespondencji przyjęli niepisaną umowę, że wielkość bezwładności całkowitej stanowi absolutne minimum tego, co należy podać w wypadku każdego wykresu.

Lista parametrów, które pomagają poprawnie zinterpretować rozwiązanie graficzne, nie ogranicza się zresztą do bezwładności całkowitej. Inne wielkości przydatne do tego celu omówimy na konkretnych przykładach w dalszej części rozdziału. W tym miejscu ograniczymy się do stwierdzenia, że graficzny obraz związku należy traktować jedynie jako punkt wyjścia do jego interpretacji. W wielu wypadkach daje on syntetyczny wgląd w istotę badanego zjawiska, aczkolwiek na ogół nie przedstawia wszystkich jego ważnych aspektów. W szczególności, z wykresu nie można odczytać wielkości bezwładności całkowitej, która *de facto* odpowiada pojęciu siły związku między cechami. Trudno byłoby znaleźć badacza, który uznałby ten aspekt związku za nieważny. Dlatego wykresy w analizie korespondencji uzupełnia się zawsze zestawem wielkości liczbowych, koniecznych do poprawnej interpretacji rozwiązania.

7.5 Wyniki analizy korespondencji nie odzwierciedlają liczebności

Pisząc książkę, zastanawiałem się, czy nie ograniczyć tego podrozdziału do jednego zdania. Albo nawet do samego tytułu! Przypuszczam, że zwróciłbym w ten sposób uwagę Czytelników na jedno z najpoważniejszych niebezpieczeństw, na jakie narażony jest badacz interpretujący dane. Niebezpieczeństwo nie ogranicza się zresztą do analizy korespondencji, lecz dotyczy większości metod analizy danych. Wynika bowiem z emocjonalnego i społecznego kontekstu, jaki analizie danych towarzyszy. Jak wiadomo, zajęcie to wiąże się z wysiłkiem i zaangażowaniem, przez co wymaga pozytywnego bodźcowania. Każdy badacz oczekuje więc nagród w postaci odkrycia prawidłowości w rozpatrywanym zjawisku. Gdy natura zjawiska jest taka, że brak w niej prawidłowości widocznych na pierwszy rzut oka, to badacz szuka dalej. Zaczyna się denerwować, jest podirytowany, ale szuka. Zwłaszcza, gdy czuje na karku oddech klienta, który zapłacił za badanie. Nasz świat jest bowiem tak zorganizowany, że nie można zleceniodawcy badania po prostu po-

wiedzieć: „w rezultacie przeprowadzonego badania nie stwierdzono żadnych prawidłowości dotyczących badanego zjawiska".

W danych zawsze uda się odnaleźć pewne prawidłowości, niezależnie od przedmiotu badania, techniki gromadzenia danych, czy wielkości próby. Chodzi jednak o to, czy odkryte prawidłowości informują o substantywnych aspektach badanego zjawiska. Nie wnikając zbyt głęboko w to, co właściwie oznacza „substantywny aspekt" – bo o tym można byłoby napisać osobną książkę – przyjmijmy roboczo, że chodzi o prawidłowość, która nie ogranicza się do zrealizowanego badania. Czyli, jeśli zleceniodawca zamówi takie samo badanie w innej firmie, to znaleziona prawidłowość również się w nim ujawni.

Najbardziej znaną strategią ograniczania ryzyka w omawianej kwestii jest odwołanie się do aparatu wnioskowania statystycznego. Badacz pisze w raporcie, że otrzymana prawidłowość wydaje się ciekawa, lecz biorąc pod uwagę wielkość próby, jest statystycznie nieistotna. Czytając raport zleceniodawca badania myśli sobie – szkoda, że nie miałem środków na realizację badania na większej próbie. Wtedy miałbym większą pewność co do owej prawidłowości. Ale cóż, pozostaje mi zaryzykować i przyjąć, że stwierdzona prawidłowość rzeczywiście ma miejsce. Zwłaszcza, że firma badawcza nic lepszego nie znalazła.

W wielu sytuacjach aparat wnioskowania statystycznego niewiele więc pomaga w rozstrzyganiu, czy stwierdzone prawidłowości są godne uwagi. Ciężar odpowiedzialności wraca tym samym do badacza. Pozostaje mu w pocie czoła podejmować dziesiątki decyzji, czy to, co znalazł, stanowi ów pożądany i przez wszystkich wyczekiwany *insight*. Gdy owo coś oparte jest na większej liczbie badanych jednostek, to wtedy budzi większe zaufanie. Jednakże statystyczna podstawa to nie wszystko. Na ogół ważniejsza okazuje się owa „substantywność" uzyskanego wyniku. Czy odkryta prawidłowość dotyka istoty badanego zjawiska, czy też wyłącznie jego peryferiów.

Zilustrować to można przykładem, do którego odwoływałem się już w innym miejscu (Sawiński i Domański 1986: 68–69). Przyjmijmy, że ustrój pewnego kraju jest monarchią konstytucyjną, przy czym konstytucja określa, że monarchą zostaje najstarsze dziecko monarchy. W kraju tym postanowiono przeprowadzić badanie międzypokoleniowej ruchliwości zawodowej. Jego cel stanowiło oszacowanie liczebności w polach tablicy, w której zawód badanej osoby skrzyżowany byłby z zawodem jej ojca w wieku, gdy badana osoba miała 14 lat. Wyniki wcześniej prowadzonych badań wykazywały niezbicie, że zróżnicowanie zawodowe w tym kraju sprowadza się do podziału na trzy warstwy: urzędników, robotników i rolników. Tabelę ruchliwości zdecydowano się więc skonstruować na bazie tego podziału. Badanie przeprowadzono na 1000-osobowej próbie wylosowanej spośród dorosłych mieszkańców. Po

zebraniu wyników okazało się, że wśród badanych osób znalazł się monarcha! Zaburzyło to nieco założenia badania, gdyż nie bardzo było wiadomo, jak go zaklasyfikować. Monarcha to jednak monarcha, w związku z tym utworzono dla niego osobną kategorię. Doprowadziło to do uzyskania tablicy ruchliwości w wersji przedstawionej w części [1] tabeli 7.2.

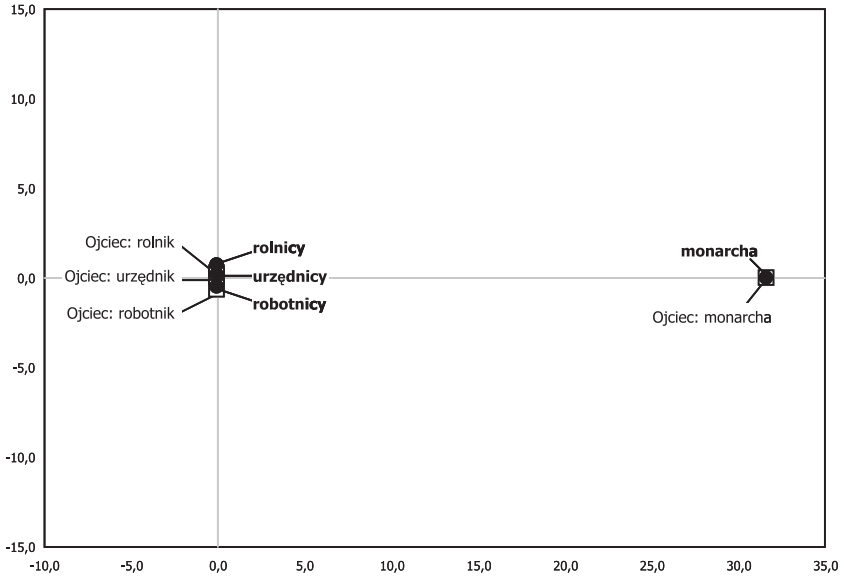
Aby odtworzyć wzory ruchliwości międzypokoleniowej badacze postanowili posłużyć się analizą korespondencji. Obliczyli współrzędne kanoniczne a następnie sporządzili wykres (rycina 7.2, wykres [a]). Uzyskany obraz nieco ich zaskoczył. Właściwie, to całe obserwowane zróżnicowanie sprowadziło się do dystansu między monarchą, a pozostałymi kategoriami zawodowymi. Obliczyli więc bezwładności dla wierszy, kolumn i poszczególnych pól tablicy, a dodatkowo ich procentowe udziały w bezwładności łącznej (części [2] i [3] tabeli 7.2). Wyniki te świadczyły, że z pozycją monarchy wiąże się 86 procent zróżnicowania profili przynależności zawodowej mieszkańców kraju w zależności od pozycji zajmowanych przez ich ojców. No cóż, stwierdzili badacze, widocznie tak dzieje się w sytuacji, gdy pewne pozycje są ustawowo dziedziczone. Ich znaczenie dla kształtu struktury społecznej staje się przez to znacząco większe, niż pozostałych międzypokoleniowych przepływów.

Tabela 7.2
Liczebności i bezwładność w tablicy międzypokoleniowej ruchliwości zawodowej w fikcyjnym społeczeństwie o ustroju monarchii konstytucyjnej

	zawód badanej osoby				
zawód ojca	monarcha	urzędnicy	robotnicy	rolnicy	ogółem
		[1] liczebności			
monarcha	1	0	0	0	1
urzędnicy	0	119	60	20	199
robotnicy	0	100	250	50	400
rolnicy	0	80	190	130	400
ogółem	1	299	500	200	1000
		[2] wielkości bezwładności			
monarcha	0,9980	0,0003	0,0005	0,0002	0,9990
urzędnicy	0,0002	0,0595	0,0157	0,0099	0,0852
robotnicy	0,0004	0,0032	0,0125	0,0113	0,0274
rolnicy	0,0004	0,0131	0,0005	0,0313	0,0453
ogółem	0,9990	0,0761	0,0292	0,0526	1,1569
		[3] udziały bezwładności (w procentach)			
monarcha	86,27	0,03	0,04	0,02	86,36
urzędnicy	0,02	5,14	1,36	0,85	7,37
robotnicy	0,03	0,28	1,08	0,97	2,37
rolnicy	0,03	1,13	0,04	2,70	3,91
ogółem	86,36	6,58	2,52	4,54	100,00

Rycina 7.2
Obraz międzypokoleniowej ruchliwości w fikcyjnym społeczeństwie

[a] z uwzględnieniem dziedziczenia pozycji monarchy



[b] po zaliczeniu premiera do urzędników

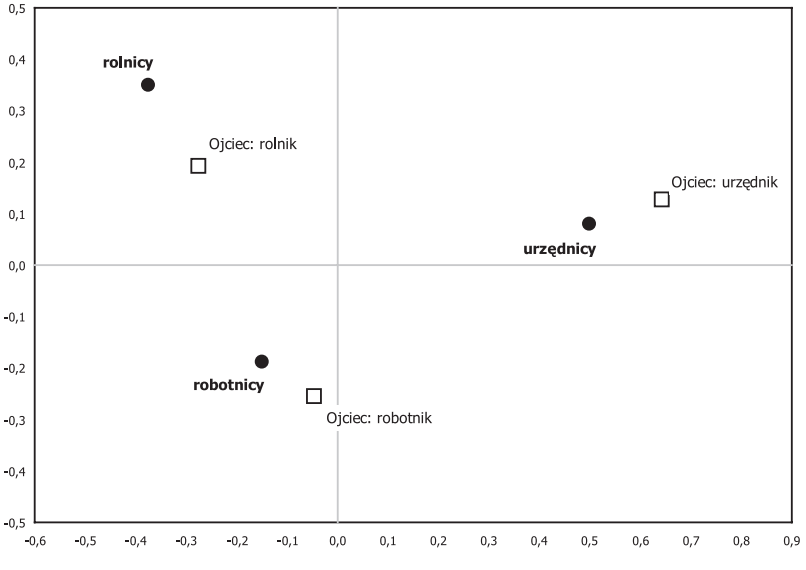


Tabela 7.3
Liczebności, profile i bezwładność w tablicy międzypokoleniowej ruchliwości zawodowej w fikcyjnym społeczeństwie po włączeniu stanowiska premiera do urzędników

	zawód badanego			
zawód ojca	urzędnicy	robotnicy	rolnicy	ogółem
	[1] liczebności			
urzędnicy	120	60	20	200
robotnicy	100	250	50	400
rolnicy	80	190	130	400
ogółem	300	500	200	1000
	[2] profile wierszy (w procentach)			
urzędnicy	60	30	10	100
robotnicy	25	63	13	100
rolnicy	20	48	33	100
ogółem	30	50	20	100
	[3] profile kolumn (w procentach)			
urzędnicy	40	12	10	20
robotnicy	33	50	25	40
rolnicy	27	38	65	40
ogółem	100	100	100	100
	[4] wielkości bezwładności			
urzędnicy	0,0600	0,0160	0,0100	0,0860
robotnicy	0,0033	0,0125	0,0113	0,0271
rolnicy	0,0133	0,0005	0,0313	0,0451
ogółem	0,0767	0,0290	0,0525	0,1582
	[5] udziały bezwładności (w procentach)			
urzędnicy	37,93	10,12	6,32	54,37
robotnicy	2,11	7,90	7,11	17,12
rolnicy	8,43	0,32	19,76	28,50
ogółem	48,47	18,34	33,19	100,00

Przenieśmy się teraz do innego kraju, którego ustrój jest demokracją parlamentarną. W kraju tym przeprowadzono badanie według identycznego schematu. Po zgromadzeniu wypełnionych ankiet okazało się, że wśród badanych znalazł się premier rządu. Badacze mieli podobny zgryz – jak go zaklasyfikować. Po dyskusji ustalili, że bardziej uzasadnione jest utworzenie dla niego osobnej kategorii, niż zaliczenie do urzędników. Dotychczasowe badania wykazywały bowiem, że warstwa urzędników jest dość homogeniczna. Dołączenie do nich premiera zaburzyłoby ten porządek.

Podczas wypełniania pól tablicy ruchliwości okazało się, że ojciec badanego premiera był również premierem! Nie powinno to zresztą specjalnie dziwić. Po pierwsze, konstytucja nie wyklucza takiej możliwości. A po drugie, jest to racjonalny wybór. Stanowisko premiera powierzono kandydatowi, który od dziecka mógł przyglądać się, jak wygląda rozwiązywanie problemów państwa. Trudno byłoby znaleźć kogoś bardziej doświadczonego. Otrzymana tablica międzypokoleniowej ruchliwości przypominała w rezultacie omawianą wcześniej tablicę dla monarchii konstytucyjnej. Podobnie wyglądał też wykres uzyskany metodą korespondencji. Czy na jego podstawie należy wnioskować, że wzory międzypokoleniowej ruchliwości w tym kraju sprowadzają się do dystansu między pozycją premiera a resztą społeczeństwa?

W tym miejscu dochodzimy do tego, co należy rozumieć przez istotę badanego zjawiska. Zasada dziedziczenia pozycji monarchy jest substantywnym elementem międzypokoleniowej ruchliwości. Można jedynie żałować, że zasada ta zdominowała rozwiązanie w analizie korespondencji, przez co z wykresu niewiele można dowiedzieć się na temat zasad ruchliwości w innych warstwach. Natomiast dziedziczenie stanowiska premiera nie jest zasadą ogólną. Jest wynikiem, który w badaniu uzyskano wyłącznie z tego powodu, że akurat na premiera wybrano kandydata, którego ojciec był również premierem. Nie oznacza to, że w kolejnych wyborach na stanowisko premiera nie zostanie wybrany kandydat, którego ojciec był rolnikiem, robotnikiem, czy urzędnikiem. Jeśli tak się stanie, to nie zmieni to **istoty** międzypokoleniowej ruchliwości zawodowej w tym kraju. Obraz ruchliwości otrzymany po utworzeniu dla premiera osobnej kategorii należy więc uznać za nieadekwatny. Zdominowany został przez jednostkowy wynik, który nie odzwierciedla istoty zjawiska.

Lepiej uzasadnione byłoby utworzenie tablicy ruchliwości, w której premier zostałby zaklasyfikowany do urzędników. Tablica w tej postaci przedstawiona została w tabeli 7.3, zaś uzyskany dla niej wykres korespondencji prezentujemy jako część [b] ryciny 7.2. Spróbujmy zinterpretować ten wykres śledząc zarazem, czy relacje między liczebnościami w tak utworzonej tablicy odpowiadają relacjom między punktami na wykresie.

Po włączeniu premiera do urzędników kategorii tej odpowiada 54 procent zróżnicowania profili w wypadku ojców, zaś 48 procent w pokoleniu badanych osób. Różnica bierze się stąd, że urzędnicy-ojcowie alokują większy odsetek swoich dzieci do kategorii urzędników, niż wynosi odsetek urzędników wśród badanych, których ojcowie również byli urzędnikami. Znajduje to potwierdzenie na wykresie. Punkt odpowiadający ojcom-urzędnikom leży dalej od środka wykresu, niż punkt odpowiadający warstwie urzędników w aktualnej strukturze zawodowej.

Warto też zwrócić uwagę, że punkty grupują się w pary obejmujące tą samą warstwę w pokoleniu ojców oraz w pokoleniu badanych osób. Bliskość punktów w parach odzwierciedla znaczną samorekrutację w ramach warstw, co wyraża się względnie wysokimi bezwładnościami na przekątnej głównej tablicy. Para obejmująca robotników jest bliżej środka wykresu niż pozostałe pary, gdyż profile pochodzenia wśród robotników są bardziej zbliżone do przeciętnych profili pochodzenia, niż w pozostałych kategoriach, a zarazem profil pozycji warstwowych, do których dochodzą dzieci robotników jest najbardziej zbliżony do profilu całego społeczeństwa. W rozpatrywanym społeczeństwie robotnicy w obu pokoleniach zajmują więc pozycję centralną – to jest pomiędzy dwiema skrajnymi kategoriami, jakimi są urzędnicy i rolnicy.

Stosunkowo blisko środka wykresu lokuje się kategoria ojców-rolników. Z kategorii tej nastąpił bowiem znaczny odpływ do dwóch pozostałych warstw, gdyż udział rolników w strukturze społeczeństwa zmniejszył się z 40 do 20 procent. Odpływ ten spowodował upodobnienie się profilu przynależności warstwowej osób, których ojcowie byli rolnikami, do struktury warstwowej całego społeczeństwa. W porównaniu z ojcami-rolnikami, punkt odpowiadający badanym rolnikom jest odsunięty od środka wykresu dalej. Wynika to stąd, że aż 65 procent rolników ma ojców rolników, przez co profil pochodzenia rolników jest specyficzny.

Podsumowując, należy stwierdzić, że wykres uzyskany w analizie korespondencji trafnie odzwierciedla układ przepływów w rozpatrywanej tablicy międzypokoleniowej ruchliwości zawodowej. Mimo że przykładowe dane mają charakter fikcyjny, zostały one skonstruowane w sposób zbliżony do rzeczywistych tablic ruchliwości. Otrzymane wnioski są przez to spójne z wnioskami uzyskiwanymi przez badaczy w różnych krajach, analizujących tego rodzaju tablice za pomocą różnych zresztą metod (Sawiński i Domański 1986, 1989).

Przykład ten uświadamia zarazem, że podczas analiz tablic szczególną uwagę należy zwrócić na kategorie o niewielkich liczebnościach. Blasius i Greenacre (2006b: 20) ujęli ten problem następująco

> *kategorie o względnie niewielkich proporcjach (to jest niewielkich masach) mają tendencję do zajmowania na wykresie skrajnych pozycji [outlying], gdyż ich profile różnią się na ogół znacznie od przeciętnych* [...] *Z tego powodu w analizie korespondencji należy szczególnie uważnie przyglądać się kategoriom o względnie niewielkich proporcjach. Jeżeli wykazują zbyt duży wkład do rozwiązania, to należy je łączyć z innymi kategoriami w sposób zachowujący substantywny sens*

Rekomendacja ta jest z pewnością użyteczna jako dyrektywa ogólna. Nie wyklucza jednak, że w pewnych wypadkach pozostawienie owych kategorii

Ramka 7.1
Michael John Greenacre: związki muzyki z analizą korespondencji

Michael John Greenacre urodził się w 1951 roku w Cape Town w Afryce Południowej. W wieku 5 lat rozpoczął naukę gry na pianinie, a potem na gitarze. Po ośmiu latach nauki muzyki klasycznej zaczął grać w zespołach muzycznych najpierw w rodzinnym mieście, a następnie w Pretorii. Występował tam w kwartecie „The Smooth Ones" grającym muzykę taneczną. W tym czasie był zarówno kompozytorem, wykonawcą, jak też pisał teksty. W późniejszych latach zajął się głównie aranżacją tematów klasycznych. Pisał również muzykę do sztuk teatralnych. W 1994 roku Greenacre przeniósł się do Katalonii, gdzie koncertował w klubach, na festiwalach letnich, a także występował na okolicznościowych koncertach inaugurujących wystawy artystyczne. Niektóre z jego wystąpień miały niekonwencjonalny charakter. Koncertem rozpoczął międzynarodowy warsztat w dziedzinie modelowania statystycznego na uniwersytecie w Barcelonie w lipcu 2007 roku. Wspomagany przez wokalistkę Gurdeep Stephens zagrał utwory Ellingtona, Gershwina, tradycyjne pieśni katalońskie a także własne kompozycje. W tekście jednej z nich dowcipnie opowiedział o problemach uczestników kursu ze zrozumieniem modeli statystycznych...

Do najbardziej znanych kompozycji Greenacre'a należy *The Millenium Song*, będący aranżacją Preludium c-moll opus 28 nr 20 Fryderyka Chopina. Co prawda Polakom może nie wydawać się najlepszym pomysłem połączenie tej akurat inspiracji, znanej u nas jako *Marsz żałobny*, ze świętowaniem przełomu tysiącleci. Jednakże aranżacja jest ciekawa, łącząc w sobie brzmienie klasyczne z rockiem, jazzem i techno. Równie interesujący jest głęboko humanistyczny tekst utworu. Greenacre zawczasu zadbał o to, aby tekst ten przetłumaczony został na kilka języków, co umożliwiło 1 stycznia 2000 roku emisję utworu w narodowych wersjach przez stacje radiowe różnych krajów. Do utworu nakręcony został teledysk nadawany przez stacje telewizyjne.

Naukowa kariera Greenacre'a przebiegała równie barwnie. Tytuł M. Sc. uzyskał w dziedzinie informatyki w 1972 roku. Miał wtedy 21 lat. Jeszcze w okresie studiów rozpoczął pracę jako asystent na University of South Africa w Pretorii. W latach 1974–1978 przebywał w Paryżu, gdzie kontynuował studia na Université Pierre et Marie Curie. Zakończył je rozprawą doktorską pod tytułem „Quelque méthodes objectives de représentation graphique d'un tableau de données". Jego promotorem był nie kto inny, jak Jean-Paul Benzécri. Być może wyjaśnia to, dlaczego Greenacre stał się orędownikiem metod wizualizacji danych, a w szczególności analizy korespondencji. Jest autorem ponad 50 artykułów na ten temat a także dwóch książek: *Theory and Applications of Correspondence Analysis* (1984) oraz *Correspondence Analysis in Practice* (2007). Jest też współredaktorem dwóch prac zbiorowych (Greenacre i Blasius 1994; Blasius i Greenacre 2006a), które uznawane są za kamienie milowe na drodze rozwoju tej metody.

Od 1994 roku Greenacre jest profesorem Uniwersytetu Pompeu Fabra w Barcelonie, gdzie na wydziale ekonomii prowadzi zajęcia ze statystyki. Na arenie międzynarodowej znany i ceniony jest jako organizator konferencji poświęconych teorii i zastosowaniom analizy korespondencji. Metodę propaguje również jeżdżąc po świecie, prowadząc warsztaty na ten temat, a także występując na licznych konferencjach. W 2002 roku wygłosił między innymi referat podczas konferencji na Uniwersytecie Jagiellońskim.

Więcej informacji na temat naukowych dokonań Michaela Greenacre'a znaleźć można na jego stronie internetowej http://www.econ.upf.edu/~michael/. Tam również znajdują się linki do jego kompozycji muzycznych, wydanych płyt oraz fragmentów koncertów.

o niewielkich proporcjach uzasadnione jest właśnie względami substantywnymi. Tak było w wypadku, gdy wśród badanych znalazł się monarcha. Niewielka proporcja oznaczać jednak może niewielką liczebność. Gdy badanie realizowane jest na próbie dobranej z populacji, to istnieje ryzyko, że profil dla kategorii o niewielkiej liczebności oszacowany został niewłaściwie. Mówiąc inaczej, wynik związany z daną kategorią ma charakter przypadkowy i nie powtórzy się w innych badaniach tej samej populacji.

Analiza korespondencji nie stanowi więc panaceum na odwieczny dylemat badaczy, czy uzyskany wynik można traktować jako substantywną prawidłowość, czy też nie. Jak sygnalizowałem wcześniej, problem ma charakter ogólny i dotyczy większości metod analizy danych. Analiza korespondencji wiąże się ponadto z tym niebezpieczeństwem, że na wykresie nie są uwidocznione wielkości kategorii. Jeśli więc badacz ograniczy uwagę do wykresu, to niektóre z wniosków mogą nie mieć realnych odpowiedników na poziomie badanego zjawiska. Dlatego tak ważna jest równoległa analiza liczebności oraz wskaźników bezwładności. Chroni to przed wyciąganiem nieuzasadnionych wniosków, lecz z drugiej strony eliminować może szereg wniosków o charakterze substantywnym, jeśli uzna się je jako oparte na zbyt małych liczebnościach. Te ostatnie wnioski są zaś w wypadku wielu badań najcenniejsze.

Umiejętność odróżniania nieuzasadnionych wniosków od wniosków substantywnych jest z całą pewnością wyrazem najwyższych kompetencji badawczych. Z jednej strony jest funkcją doświadczenia w analizowaniu wyników badań, z drugiej zaś – a może przede wszystkim – pochodną wiedzy badacza na temat analizowanego zjawiska.

7.6 Wybór skal do prezentacji rozwiązania. Pochodzenie społeczne studentów różnych kierunków studiów w roku akademickim 1928/29

Problem wyboru skal dla graficznej prezentacji położenia punktów reprezentujących kategorie obu cech nazywany jest też problemem normalizacji. W analizie korespondencji stosuje się w tym zakresie kilka rozwiązań, które pozwalają uwypuklić nieco inne aspekty badanego związku. W książce stosować będziemy normalizację zwaną **symetryczną** lub **kanoniczną** (Górniak 2000: 119; Blasius i Greenacre 2006b: 14) i jej omówieniu poświęcimy ten podrozdział.

Rozważania zilustrujemy przykładowymi danymi dotyczącymi składu społecznego studentów wydziałów wyższych uczelni w Polsce w roku akademickim 1928/29 (GUS 1931). Dane te wybrałem z dwóch względów. Po pierwsze, analiza korespondencji wykazuje szczególną przydatność w sytua-

cjach, gdy niewiele wiemy o badanym zjawisku. Jak się za chwilę przekonamy, zarówno ówczesne kierunki studiów, jak też podziały studentów według pochodzenia brzmią dziś dość egzotycznie. Po drugie, dane te stanowiły już wcześniej przedmiot analizy (Charszewski 1931), której celem było ustalenie, jak silne było w owych czasach zróżnicowanie społeczne wśród studentów. Autor analizy wykonał sporą pracę, grupując szczegółowe dane dotyczące pochodzenia studentów w pięć klas społecznych, a następnie analizując je za pomocą tablicy, w której obliczał profile procentowe w wierszach i kolumnach. Interesujące więc będzie, czy zastosowanie do tych samych danych analizy korespondencji pozwoli wzbogacić wnioski, które uzyskano 80 lat temu za pomocą najprostszej z metod analizy tablic.

Dla tablicy, w której autor skrzyżował pochodzenie społeczne studentów z wydziałem, na którym studiowali, wykonałem analizę kanoniczną w sposób opisany w rozdziale 6. W tabeli 7.4 przedstawiam jej wyniki. Dla ich interpretacji użyteczne są zasady zaliczania pochodzenia studentów do poszczególnych klas społecznych, podane w nocie pod tabelą. Zasady wyodrębnienia kierunków studiów odzwierciedlają sposób organizacji szkolnictwa wyższego w tym czasie. Dla interpretacji wyników istotny może być fakt, że wydziały sklasyfikowane jako teologiczne obejmowały uczelnie i kierunki przygotowujące do stanu kapłaństwa.

Współrzędne kanoniczne, podane w części [2] i [3] tabeli 7.4, odpowiadają współrzędnym omawianym w rozdziale 6. Przypomnijmy, że wielkości współrzędnych kanonicznych podaje się najczęściej w postaci standaryzowanej, to znaczy średnia każdego zestawu współrzędnych jest równa 0, zaś odchylenie standardowe 1 (wzory 6.18–6.21). Wielkości w tej postaci nie używa się jednak do sporządzania wykresów w przyjętej normalizacji, gdyż pierwszy i drugi wymiar kanoniczny miałyby takie same wagi – ze względu na jednakowe odchylenia standardowe. Tymczasem znaczenie drugiego wymiaru jest mniejsze, gdyż w analizie kanonicznej pierwszy wyjaśnia maksimum zróżnicowania, które jest w stanie opisać pojedynczy wymiar. Ilustrują to wielkości korelacji kanonicznych podane w części [1] tabeli 7.4. Do sporządzenia wykresu w analizie korespondencji używa się więc współrzędnych, które nazywane są **głównymi**. Są to wielkości standaryzowanych współrzędnych kanonicznych przemnożone przez wielkości korelacji kanonicznych dla danego wymiaru (Blasius i Greenacre 1994: 64–65; Górniak 2000: 119). Współrzędne te zachowują porządek współrzędnych standaryzowanych oraz względne dystanse pomiędzy kolejnymi współrzędnymi. Mają natomiast mniejszy rozrzut niż oryginalne współrzędne standaryzowane.

Na rycinie 7.3 przedstawiony został obraz korespondencji dla rozpatrywanego związku wykreślony z wykorzystaniem współrzędnych głównych

Tabela 7.4
Wyniki analizy kanonicznej składu społecznego studentów kierunków (wydziałów) szkół wyższych w Polsce w roku akademickim 1928/29

[1] korelacje kanoniczne i ich udziały w procentach

wymiar rozwiązania	korelacja	kwadrat korelacji	udział w %	skumulowany udział w %
1	0,1950	0,0380	62,8	62,8
2	0,1426	0,0203	33,6	96,4
3	0,0449	0,0020	3,3	99,7
4	0,0126	0,0002	0,3	100,0
w sumie		0,0606	100,0	

[2] liczba i odsetki studentów poszczególnych kierunków oraz współrzędne kanoniczne w pierwszym i drugim wymiarze

			współrzędne kanoniczne		współrzędne główne	
kierunek studiów	liczba studentów	udział w procentach	wymiar 1	wymiar 2	wymiar 1	wymiar 2
teologia	548	1,9	6,933	0,486	1,352	0,069
medycyna	7546	26,2	0,077	0,627	0,015	0,089
technika	3605	12,5	0,072	-0,611	0,014	-0,087
filozofia	6413	22,3	0,039	-0,528	0,008	-0,075
rolnictwo	1557	5,4	-0,330	-2,036	-0,064	-0,290
handel	3109	10,8	-0,336	-1,347	-0,066	-0,192
prawo	5826	20,2	-0,540	1,327	-0,105	0,189
sztuki plastyczne	211	0,7	-0,881	1,065	-0,172	0,152
ogółem	28 815	100,0				

[3] liczba i odsetki studentów z poszczególnych klas społecznych oraz współrzędne kanoniczne w pierwszym i drugim wymiarze

			współrzędne kanoniczne		współrzędne główne	
klasa społeczna	liczba studentów	udział w procentach	wymiar 1	wymiar 2	wymiar 1	wymiar 2
chłopi	3097	10,7	2,784	-0,623	0,543	-0,089
robotnicy przemysłowi	443	1,5	0,488	2,677	0,095	0,382
pozostali robotnicy	2032	7,1	0,488	2,302	0,095	0,328
drobnomieszczaństwo	13 854	48,1	-0,383	0,447	-0,075	0,064
burżuazja	9389	32,6	-0,482	-1,078	-0,094	-0,154
ogółem	28 815	100,0				

Tabela 7.4 (kontynuacja)

[4] wielkości bezwładności

klasy społeczne	kierunek studiów: teologia	medycyna	technika	filozofia	rolnictwo	handel	prawo	sztuki plastyczne	ogółem
chłopi	0,0284	0,0000	0,0002	0,0001	0,0000	0,0000	0,0035	0,0003	0,0326
robotnicy przemysłowi	0,0004	0,0000	0,0000	0,0003	0,0004	0,0003	0,0016	0,0001	0,0031
pozostali robotnicy	0,0007	0,0023	0,0014	0,0002	0,0019	0,0018	0,0010	0,0000	0,0093
drobnomieszczaństwo	0,0021	0,0000	0,0000	0,0003	0,0003	0,0002	0,0020	0,0000	0,0049
burżuazja	0,0034	0,0006	0,0001	0,0006	0,0022	0,0021	0,0019	0,0000	0,0108
ogółem	0,0349	0,0030	0,0016	0,0014	0,0048	0,0045	0,0099	0,0004	0,0606

[5] udziały bezwładności (w procentach)

klasy społeczne	kierunek studiów: teologia	medycyna	technika	filozofia	rolnictwo	handel	prawo	sztuki plastyczne	ogółem
chłopi	47,0	0,0	0,3	0,1	0,0	0,1	5,8	0,5	53,8
robotnicy przemysłowi	0,6	0,0	0,0	0,5	0,7	0,6	2,6	0,1	5,1
pozostali robotnicy	1,1	3,8	2,3	0,3	3,1	3,0	1,7	0,1	15,3
drobnomieszczaństwo	3,4	0,0	0,0	0,5	0,5	0,3	3,3	0,1	8,1
burżuazja	5,6	1,0	0,1	1,0	3,6	3,4	3,1	0,0	17,8
ogółem	57,7	4,9	2,7	2,4	7,9	7,4	16,4	0,7	100,0

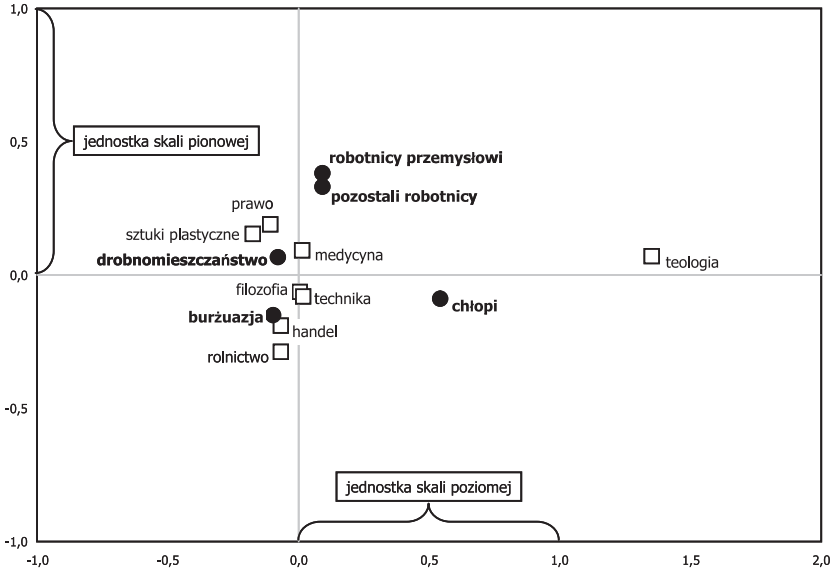
Skład klas społecznych (Charszewski 1931)

chłopi	klasyfikowani niezależnie od wielkości gospodarstwa rolnego (w tym gospodarstwa powyżej 50 ha)
robotnicy przemysłowi	robotnicy fabryczni i przemysłowi
pozostali robotnicy	robotnicy zatrudnieni w handlu, w komunikacji, w służbach państwowych i samorządowych; wyrobnicy
drobnomieszczaństwo	rzemieślnicy; urzędnicy państwowi i samorządowi, urzędnicy w przemyśle, w handlu, w bankach; nauczyciele; oficerowie i podoficerowie wojska
burżuazja	obszarnicy; przemysłowcy; kupcy; wolne zawody; kapitaliści i rentierzy

podanych w części [2] i [3] tabeli 7.4. Sporządzając wykres, należy zadbać o to, aby jednostki skali poziomej i pionowej rysunku były jednakowe. Ze względu na fakt, że prostokąt stanowiący wykres częściej orientuje się poziomo niż pionowo, współrzędne pierwszego wymiaru najlepiej jest wykreślić na osi poziomej, zaś współrzędne drugiego na pionowej. Pozwala to na bardziej równomierne rozłożenie punktów na wykresie. Konwencję tę przyjmiemy we wszystkich dalszych prezentacjach wyników analizy korespondencji.

Analizowana tablica ma wymiary 5 na 8, stąd też pełna jej dekompozycja określona jest przez 4 wymiary kanoniczne o kolejno malejących korelacjach. Suma kwadratów wszystkich tych korelacji jest równa 0,0606, która to wielkość określa zarazem bezwładność całkowitą rozpatrywanej tablicy ($= \chi^2 / n$). Pierwsze pytanie, na które zawsze trzeba odpowiedzieć, dotyczy stopnia wyjaśnienia bezwładności całkowitej przez dwa pierwsze wymiary kanoniczne. Wykres ogranicza się bowiem do tych dwóch wymiarów. Udział ten określa suma kwadratów dwóch pierwszych korelacji kanonicznych zrelatywizowana do bezwładności całkowitej. W naszym wypadku wielkość ta wynosi 96,4 procenta (tabela 7.4, część [1]), co stanowi podstawę do uznania dwuwymiarowego obrazu jako należytej reprezentacji związku w tablicy.

Rycina 7.3
Skład społeczny studentów poszczególnych kierunków (wydziałów) szkół wyższych w Polsce w roku akademickim 1928/29. Obraz korespondencji



Bezwładność całkowitą odczytać też można z części [4] tabeli 7.4, jako sumę bezwładności wierszy, kolumn, czy pól wnętrza tablicy. Przed analizą wykresu warto zawsze przyjrzeć się wielkościom bezwładności dla wierszy i kolumn, co ułatwia jego interpretację. Chodzi przede wszystkim o sprawdzenie, czy niektóre z wierszy bądź kolumn nie absorbują zbyt wiele bezwładności, co spowodować by mogło opisany w poprzednim podrozdziale efekt „monarchy" – to jest zdominowania wykresu przez pojedynczą kategorię o niewielkiej masie. Wielkości bezwładności wyrażają się ułamkami dziesiętnymi o niewielkich wartościach, stąd też wygodniej wyprocentować je do ich sumy, czyli do bezwładności całkowitej tablicy. Bezwładności w tej postaci przedstawione zostały w części [5] tabeli 7.4. Wyraźnie widać, że prawie połowa badanego zjawiska – to jest 47 procent bezwładności całkowitej tablicy – sprowadza się do wyboru przez dzieci chłopów kierunków teologicznych. Warto mieć to na uwadze interpretując wykres.

W tym miejscu winien jestem czytelnikom jeszcze jedną podpowiedź. Prezentując wyniki analizy korespondencji w tabeli warto uporządkować je według wielkości współrzędnych pierwszego wymiaru kanonicznego. Konwencję tę zastosowałem zarówno w tabelach [2] i [3] – prezentujących masy wierszy i kolumn i odpowiadające im współrzędne kanoniczne – jak też w tabelach [4] i [5], gdzie przedstawione zostały tablice bezwładności. Łatwiej wtedy zrozumieć schemat związku, który w analizie korespondencji odpowiada w największym stopniu strukturze pierwszej tablicy kanonicznej. Nałożenie jej obrazu (podrozdział 6.3, rycina 6.2) na tak uporządkowaną tablicę (na przykład na układ wierszy i kolumn w tabeli [4]) pozwala wnioskować, że synowie i córki chłopów i robotników – czyli kategorii społecznych po jednej stronie hierarchii – wybierają przede wszystkim takie kierunki jak teologia, medycyna czy technika. Tablica kanoniczna „wygina się" bowiem w tym narożniku ku górze w stosunku do płaszczyzny niezależności. Podobnie dzieje się w przeciwległym narożniku tablicy. Studenci wywodzący się z drobnomieszczaństwa i burżuazji mają skłonność do wybierania takich kierunków jak sztuki plastyczne, prawo, handel, czy rolnictwo. W pozostałych dwóch narożnikach tablica kanoniczna wygina się ku dołowi. Czyli studenci o pochodzeniu chłopskim czy robotniczym są niedoreprezentowani na kierunkach preferowanych przez drobnomieszczaństwo i burżuazję, zaś studenci wywodzący się z dwóch ostatnich kategorii rzadziej wybierają studia teologiczne, medyczne czy techniczne.

Ostatnim ustaleniem, którego należy dokonać przed interpretacją wykresu, jest ocena siły badanego związku. Określa ją wielkość bezwładności całkowitej, która dla rozpatrywanej tablicy wynosi 0,0606. To niewiele, biorąc pod uwagę fakt, że maksymalna możliwa wielkość tego wskaźnika wynosi $\min(w - 1, k - 1)$, czyli w tym wypadku 4. W analizach zjawisk społecznych

maksymalną możliwą wielkość traktuje się jednak jako mało realistyczną podstawę oceny wielkości faktycznie uzyskanej. Z różnych bowiem powodów badane związki są nie osiągają swojej maksymalnej możliwej siły[7]. Dlatego jako punkt odniesienia przyjmuje się raczej bezwładność całkowitą innych związków, najlepiej koncepcyjnie podobnych do badanego. W tym rozdziale analizowaliśmy do tej pory dwa związki: wybór szkoły w zależności od wykształcenia ojca (tabela 7.1) oraz ruchliwość międzypokoleniową w fikcyjnym społeczeństwie (tabela 7.3). W pierwszym wypadku podobną wielkość (0,0455) uzyskaliśmy dla wersji [2], którą określiliśmy jako słabo zróżnicowane profile. W wypadku międzypokoleniowej ruchliwości uzyskana bezwładność całkowita wyniosła 0,1582, czyli ponaddwukrotnie więcej niż dla tablicy wyboru kierunków studiów w zależności od pochodzenia. Rozpatrywany związek należy więc ocenić jako słaby. Innym kryterium, którym można się posłużyć, jest wielkość pierwszej korelacji kanonicznej. Wynosi ona 0,1950, zaś jej kwadrat 0,0380. Gdybyśmy rozpatrywali związek między dwiema cechami ilościowymi, na przykład między liczbą lat nauki a wysokością zarobków, to przy tej wielkości korelacji pierwsza z tych cech wyjaśniałaby 3,8 procenta zróżnicowania drugiej. Czyli niewiele.

Po dokonaniu tych wszystkich ustaleń, a zwłaszcza mając na uwadze fakt, że rozpatrujemy słaby związek, przejdźmy do interpretacji układu punktów na wykresie 7.3. Większość z nich skupiona jest wokół centralnego punktu wykresu, odpowiadającego przecięciu osi poziomej i pionowej. Oznacza to, że profile obliczone dla tych kategorii w niewielkim stopniu odbiegają od profili przeciętnych. Weźmy na przykład profil wyboru kierunku studiów wśród młodzieży po pochodzeniu drobnomieszczańskim. Z położenia punktu wynika, że jest on prawie identyczny, jak profil kierunku studiów wśród wszystkich studentów. Mówiąc inaczej, jest zgodny ze strukturą miejsc na poszczególnych kierunkach oferowanych przez system szkolnictwa wyższego. Trudno w każdym razie byłoby twierdzić, że drobnomieszczaństwo jest kategorią w jakimkolwiek sensie uprzywilejowaną, gdy chodzi o wybór kierunku studiów.

Studentów wywodzących się z burżuazji cechuje pod tym względem pewna specyfika. Nieco częściej kształcą się oni w szkołach handlowych, rolniczych i technicznych. Punkt odpowiadający pochodzeniu burżuazyjnemu odbiega bowiem od punktu centralnego wykresu w tę samą stronę, co punkty reprezentujące wymienione kierunki studiów. Rezultatu takiego należało się spodziewać, gdyż do kategorii burżuazji zaliczono właścicieli ziemskich, przemysłowców oraz kupców. Studia mają więc służyć uzyskaniu kwalifikacji do

[7] Jednym z powodów jest wzajemne niedopasowanie marginesów (Sawiński 1984). Innym „rozmycie" związku na skutek niedokładności w pomiarze obu cech (Sawiński 1988).

zarządzania w przyszłości majątkiem czy firmą rodziców. Biorąc jednak pod uwagę niewielką odległość punktu odpowiadającego burżuazji od środka wykresu, należy stwierdzić, że omawiana specyfika zarysowuje się w sumie dość słabo. Przy czym wydaje się, że nawet tak niewielkiej specyfiki nie można traktować jako pochodnej uprzywilejowania burżuazji w dostępie do pewnych kierunków studiów, lecz raczej jako element racjonalnego wyboru.

Obie kategorie robotników odbiegają od środka wykresu nieco dalej, przy czym oba punkty leżą bardzo blisko siebie. Więc niezależnie od tego, czy ojciec należy do przemysłowej klasy robotniczej, czy do pozostałych jej odłamów, profile wyboru kierunku studiów przez dziecko są prawie takie same. Różnią się one zarazem od profili dla drobnomieszczaństwa czy burżuazji. Synowie i córki robotników nieco częściej wybierają studia prawnicze i medyczne. Czy świadczy to o ich społecznym upośledzeniu? Trudno rozstrzygnąć. W każdym razie warto wziąć pod uwagę fakt, że prestiż zawodu lekarza oraz prawnika był w owych czasach wysoki (Buławski 1932).

Kategorią o najbardziej specyficznych wyborach kierunku studiów przez dzieci jest chłopstwo. Przedstawiający je punkt odsunął się od pozostałych w stronę teologii, którą z kolei reprezentuje punkt najbardziej z wszystkich odseparowany na wykresie. Ostatni z wyników oznacza, że profil pochodzenia studentów teologii najbardziej odbiega od struktury pochodzenia wszystkich studentów. Z części [5] tabeli 7.4 odczytać można, że teologia ma największą bezwładność zarówno spośród wszystkich kierunków studiów, jak też większą niż dowolna z kategorii pochodzenia. Czynnik ten odpowiedzialny jest za jej odseparowanie na wykresie. Wysoką bezwładność ma również kategoria studentów o pochodzeniu chłopskim. Czy oznacza to, że studenci o tym pochodzeniu wybierają przede wszystkim teologię, gdyż rola księdza stanowi dla nich jedyny kanał awansu?

Aby to rozstrzygnąć należy porównać masy obu kategorii. W części [2] tabeli 7.4 podano, że teologię studiuje zaledwie 2 procent studentów, zaś z części [3] można odczytać, że studenci pochodzenia chłopskiego stanowią prawie 11 procent wszystkich studiujących. Nie ma więc takiej możliwości, aby wszyscy studenci pochodzenia chłopskiego kształcili się na księży! Może to zrobić najwyżej co piąty. Dążenie chłopskich synów do stanu kapłaństwa z pewnością jest częstsze, niż w innych kategoriach. Punkt odpowiadający chłopstwu leży bowiem na wykresie bliżej teologii niż dowolny z punktów obrazujących pozostałe kategorie pochodzenia. Gdyby na teologii studiowali wyłącznie synowie chłopscy, a zarazem gdyby wszyscy synowie chłopscy szli wyłącznie na teologię, to wtedy oba punkty na wykresie musiałyby się pokryć. Stworzyłaby się sytuacja analogiczna, jak w badaniu międzypokoleniowej ruchliwości, które objęło monarchę (rycina 7.2[a]). Teologia wraz z chłopstwem stanowiłyby

jeden biegun wykresu, zaś pozostałe kierunki studiów oraz pozostałe kategorie pochodzenia zbiłyby się dość ściśle w drugi biegun. Na rozważanym rysunku tak nie jest. Dowodzi to, że młodzież pochodzenia chłopskiego wybiera w większości kierunki studiów takie, jak młodzież wywodząca się z innych środowisk. Co nie wyklucza, że na teologii, na której kształci się niewielki odsetek studentów, większość z nich jest pochodzenia chłopskiego[8].

Podsumujmy najważniejsze wnioski otrzymane z oglądu wykresu, wzbogaconego o analizę niektórych parametrów ilościowych, jak bezwładności czy masy kategorii. Generalnie rzecz biorąc, dominuje podobieństwo wyboru kierunku studiów przez młodzież o różnym pochodzeniu. Pewną specyfiką cechują się jedynie kierunki teologiczne, które w większym stopniu wybiera młodzież chłopska. Specyfika ta obejmuje połowę obserwowanego związku między pochodzeniem a wyborem kierunku studiów. Jednakże na kierunkach teologicznych kształci się zaledwie 2 procent populacji studentów, toteż nie ma to wpływu na siłę związku, który pozostaje bardzo słaby.

Na zakończenie omawiania przykładu przedstawmy wnioski, które wyciągnął w 1931 roku z tych samych danych Adam Charszewski. Analizował on tablicę wyprocentowaną w obie strony. Kategorie obu cech ułożone były w tablicy według liczby studentów. W wypadku kategorii pochodzenia oznacza to kolejność: burżuazja, drobnomieszczaństwo, chłopi, proletariat, zaś w wypadku kierunków studiów kolejność: medycyna, filozofia, prawo, technika, handel, rolnictwo, teologia, sztuki plastyczne. Tym samym w lewym górnym narożniku tablicy, czyli w miejscu najbardziej eksponowanym (zob. podrozdział 2.7), znalazły się „klasy posiadające" oraz kierunki humanistyczne. Doprowadziło to autora do następujących wniosków (Charszewski 1931)

> [...] *prawo – to kwalifikacje na kierowników aparatu władzy, na najwyższe stanowiska w administracji państwowej, sądownictwie, na przywódców burżuazyjnych partii politycznych. Nauki tzw. handlowe i agronomia – to przysposobienie do kierowania aparatem produkcji, do sprawowania bezpośredniej komendy kapitału nad pracą. Prawo, nauki handlowe i agronomia – to monopol burżuazji. Wydziały filozoficzne wytwarzają nauczycieli, teologiczne – księży, techniczne – personel techniczny w produkcji – słowem: wykonawców woli burżuazji. Tutaj burżuazja jest bardziej łaskawa, tu „ustępuje miejsca" proletariatowi.*

Analiza korespondencji pozwala spojrzeć na dane bez takich emocji.

[8] W rzeczywistości studenci pochodzenia chłopskiego stanowili 50 procent studentów kierunków teologicznych, zaś na teologii kształciło się 9 procent studentów pochodzenia chłopskiego (Charszewski 1931).

7.7 Granice dwuwymiarowości rozwiązania. Czyli, kto kogo cytuje

We wprowadzeniu do 89 woluminu *Journal of Econometrics* redaktorzy Tom Wansbeek i Michel Wedel (1999) przedstawili analizę sposobu wzajemnego cytowania się przez autorów czasopism ekonometrycznych i marketingowych. Do analizy wybrali 3 wiodące czasopisma ekonometryczne, dwa marketingowe, uzupełniając je o jedno czasopismo ekonomiczne (*Journal of Finance*) i jedno metodologiczne (*Psychometrika*). Korzystając z *Social Cita-*

Tabela 7.4
Liczebności i udziały bezwładności cytowań wybranych czasopism ekonometrycznych i marketingowych (Wansbeek i Wedel 1999)

[1] liczebności

	cytaty z							
czasopisma	ECTR	JE	JAE	JF	JMR	MS	PM	ogółem
ECTR	1777	228	12	53	0	0	0	2070
JE	1789	1283	126	172	0	0	0	3370
JAE	636	317	104	37	0	0	0	1094
JF	625	108	3	2723	0	0	0	3459
JMR	68	14	0	0	1439	451	136	2108
MS	118	16	0	0	642	606	30	1412
PM	7	0	0	0	48	10	1310	1375
ogółem	5020	1966	245	2985	2129	1067	1476	14888

[2] udziały bezwładności (w procentach)

	cytaty z							
czasopisma	ECTR	JE	JAE	JF	JMR	MS	PM	ogółem
ECTR	4,5	0,0	0,0	0,9	0,8	0,4	0,6	7,2
JE	1,0	4,3	0,2	1,0	1,3	0,7	0,9	9,5
JAE	0,5	0,6	1,1	0,4	0,4	0,2	0,3	3,6
JF	0,7	0,7	0,1	16,2	1,3	0,7	0,9	20,7
JMR	1,6	0,7	0,1	1,1	11,7	1,6	0,1	16,9
MS	0,7	0,4	0,1	0,8	2,6	6,9	0,2	11,7
PM	1,2	0,5	0,1	0,8	0,3	0,2	27,5	30,5
ogółem	10,3	7,2	1,8	21,1	18,5	10,6	30,5	100,0

Oznaczenia tytułów:

ECTR	*Econometrica*
JE	*Journal of Econometrics*
JAE	*Journal of Applied Econometrics*
JF	*Journal of Finance*
JMR	*Journal of Marketing Research*
MS	*Marketing Science*
PM	*Psychometrika*

tions Index sporządzili tablicę, której zawartość prezentujemy jako część [1] tabeli 7.4. Czasopisma zostały ułożone w kolejności według dziedzin. Najpierw wyszczególniono czasopisma ekonometryczne oraz *Journal of Finance*, a następnie marketingowe uzupełnione o *Psychometrikę*. Liczebności w tym układzie wskazują wyraźnie, że wzajemne cytowanie się występuje głównie wewnątrz dziedzin. Wrażenie robi zwłaszcza blok zer w prawym górnym narożniku tabeli.

A teraz proszę Czytelniku odgadnąć, na ile grup na wykresie podzieli rozważne czasopisma analiza korespondencji? Oczywiście, bez zaglądania na wykres. Specjalnie poprosiłem Wydawcę, aby umieścił go dwie strony dalej.

Dotychczas każdy, komu zaproponowałem ten problem do rozwiązania, był przekonany, że analiza korespondencji podzieli czasopisma na dwie grupy. Wydaje się to logiczne biorąc pod uwagę konfigurację liczebności w tabeli 7.4. Tymczasem na wykresie (rycina 7.4) wyraźnie wyodrębniają się trzy takie grupy!

Tabela 7.5
Wyniki analizy korespondencji struktury wzajemnych cytowań dla wybranych czasopism ekonomicznych i marketinowych

[1] korelacje kanoniczne i ich udziały w procentach

wymiar rozwiązania	korelacja	kwadrat korelacji	udział [%]	skumulowany udział [%]
1	0,9704	0,9417	38,2	38,2
2	0,8770	0,7691	31,2	69,3
3	0,7754	0,6012	24,4	93,7
4	0,2917	0,0851	3,5	97,1
5	0,2384	0,0568	2,3	99,4
6	0,1216	0,0148	0,6	100,0
w sumie		2,4687	100,0	

[2] Współrzędne główne w trzech pierwszych wymiarach

	miejsce cytowania			pochodzenie cytatu		
czasopisma	1 wymiar	2 wymiar	3 wymiar	1 wymiar	2 wymiar	3 wymiar
ECTR	0,6157	0,0458	0,6936	0,5885	0,0331	0,5629
JE	0,6331	0,0583	0,7171	0,6279	0,0521	0,7741
JAE	0,6302	0,0548	0,7624	0,6516	0,0662	0,9169
JF	0,7434	0,2068	-1,2648	0,7558	0,2207	-1,4067
JMR	-1,2329	-1,1158	-0,1315	-1,2404	-1,2225	-0,1466
MS	-1,0908	-1,2176	-0,0870	-1,1932	-1,3030	-0,1346
PM	-1,8398	2,1851	0,0644	-1,8226	2,0660	0,0558

Rozwinięcia tytułów podano w tabeli 7.4

Próbę wyjaśnienia tego nieoczekiwanego wyniku rozpocznijmy od analizy podanych w części [2] tabeli 7.4 udziałów poszczególnych pól w bezwładności całkowitej. Ponad jedna czwarta (27,5 procenta) bezwładności całkowitej badanego związku związana jest z cytatami w ramach *Psychometriki*. Aż 1310 cytatów w *Psychometrice*, spośród 1476 objętych analizą, odwoływało się do artykułów zamieszczonych w tymże czasopiśmie. Druga kolejna wielkość bezwładności dotyczy cytatów wewnątrz *Journal of Finance*. Trzecia zaś wewnętrznych odwołań w ramach *Journal of Marketing Research*. Suma bezwładności trzech rozważanych pól tablicy stanowi 55 procent bezwładności całkowitej.

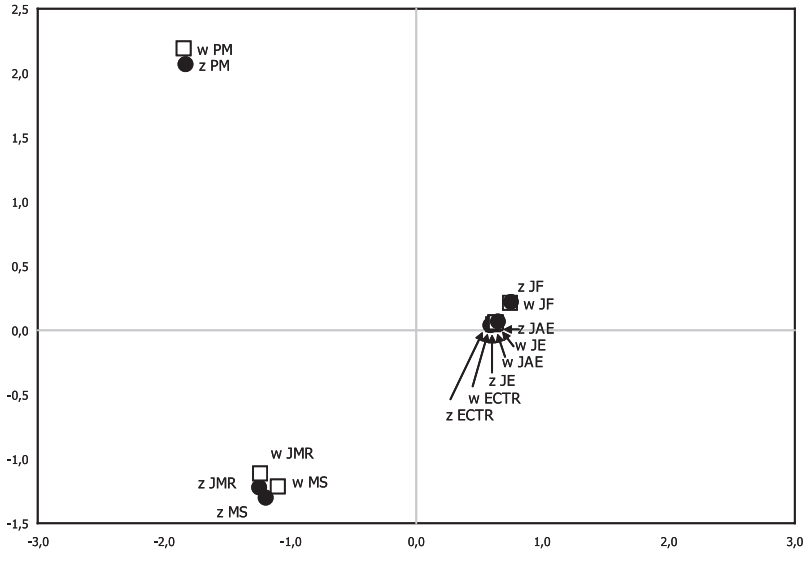
Metoda korespondencji zorientowana jest na wyjaśnienie największych odstępstw od modelu niezależności. Wierzchołek pierwszej tablicy kanonicznej wyznaczony wobec tego został przez *Psychometrikę*. Ilustrują to wielkości współrzędnych głównych pierwszego wymiaru podane w części [2] tabeli 7.5. Na przeciwległym wierzchołku znalazł się *Journal of Finance*. Ma on również wysoką bezwładność, gdy chodzi o liczbę wewnętrznych cytatów, a zarazem autorzy artykułów w obu czasopismach nie cytują się wzajemnie (zero w obu wypadkach). Wokół tych dwóch tytułów ułożyły się pozostałe. Jeśli zrzutować punkty na wykresie na pierwszy wymiar (poziomy), to otrzyma się podział analizowanych czasopism na ekonomiczne i marketingowe odpowiadający intuicjom, do których prowadzi ogląd liczebności w tablicy.

Jednakże pierwszy wymiar wyjaśnia zaledwie 38 procent bezwładności całkowitej (tabela 7.5, część [1]). To zbyt mało, aby uznać, że rozwiązanie w należyty sposób obrazuje liczebności tablicy. Należy więc rozpatrzyć drugi wymiar, który wyjaśnia dalsze 31 procent. Wielkości współrzędnych głównych świadczą, że wymiar ten kontrastuje tytuły w ramach grupy czasopism nieekonomicznych. Autorzy artykułów w *Journal of Marketing Research* i w *Marketing Science* dość często cytują się wzajemnie, co odróżnia oba czasopisma od *Psychometriki* – mającej odmienny profil tematyczny.

Podobną funkcję pełni trzeci wymiar. Kontrastuje on *Journal of Finance* wobec trzech czasopism ekonometrycznych (niska ujemna wartość współrzędnej kanonicznej dla JF *versus* w miarę wysokie wartości dodatnie dla ECTR, JE i JAE, przy bliskich zeru współrzędnych dla pozostałych czasopism). Przy czym wymiaru tego nie można pominąć, gdyż wyjaśnia on 24 procent bezwładności całkowitej. W sumie, trzy pierwsze wymiary wyjaśniają jej 94 procent. Dopiero trójwymiarowy model można więc uznać za dostatecznie dobrze dopasowany do danych. Dzieli on rozpatrywane czasopisma nie na dwie, lecz na cztery grupy. Do pierwszej należą trzy czasopisma ekonometryczne, do drugiej dwa czasopisma marketingowe, natomiast osobne grupy stanowią *Journal of Finance* i *Psychometrika*. Układu czterech grup w trzech wymiarach nie da się jednak zobrazować graficznie na płaszczyźnie. Dlatego rozwiązanie na

rycinie 7.4 przedstawia tylko trzy skupienia punktów. Punkt odpowiadający czwartej grupie, którą stanowi *Journal of Finance*, należy bowiem wyobrazić sobie jako położony „w głębi" rysunku.

Rycina 7.4
Struktura cytatów dla wybranych czasopism ekonomicznych i marketingowych
Obraz korespondencji



Nota: Rozwinięcia tytułów podano w tabeli 7.4.

Jak więc uniknąć sytuacji, w których analiza korespondencji prowadzi do nieadekwatnego obrazu badanego zjawiska. Po pierwsze, zawsze warto przyjrzeć się bezwładnościom kolejnych wymiarów kanonicznych. Gdy dwa pierwsze wyjaśniają w sumie niewiele bezwładności całkowitej bądź gdy trzeci wykazuje znaczny udział w tej bezwładności, to wtedy dwuwymiarowy wykres nie obrazuje wszystkich istotnych elementów badanego zjawiska. W tego rodzaju sytuacjach można zrezygnować z wykresu i poprzestać na rezultatach analizy kanonicznej. Jak dowodzi omawiany przykład – pozwala to poprawnie zidentyfikować prawidłowości w danych. Można też sporządzić wykres w przestrzeni trójwymiarowej. Do tego celu posłużyć się warto wyspecjalizowanym oprogramowaniem graficznym, które pozwala rotować otrzymany wykres. Z tego rodzaju programów korzystają między innymi architekci, co pozwala im spojrzeć na projektowany budynek z różnych stron.

Dodawanie kolejnych wymiarów do modelu korespondencji nie zawsze stanowi właściwy sposób postępowania. Z rozdziału 6 wynika, że model uwzględniający dwa wymiary wymaga poświęcenia $2w + 2k - 8$ stopni swobody (wzór 6.44). To wcale nie jest **niski koszt** stworzenia sobie podstaw do wyjaśnienia badanego zjawiska! Godzimy się na niego wyłącznie z tego powodu, że uzyskany graficzny obraz sprzyja budowaniu twórczych interpretacji. Jeśli więc nie można zaakceptować modelu dwuwymiarowego ze względu na jego niewystarczające dopasowanie do danych, to wtedy trzeba powiedzieć sobie – trudno. Struktura badanego zjawiska jest na tyle złożona, że nie da się go po prostu narysować. A następnie rozważyć inne sposoby podejścia do problemu. Użyteczna okazać się może – omówiona w rozdziale 4 – metoda identyfikacji schematu związku. Zbudowany na jej podstawie log-liniowy model tablicy okazuje się niekiedy należycie dopasowany do danych, angażując przy tym niewielką liczbę stopni swobody.

7.8 Korzyści w wypadku dużej liczby kategorii. Ilustracja na przykładzie europejskiego rynku reklamy

Kluczową korzyścią analizy korespondencji jest możliwość zobrazowania badanego zjawiska za pomocą konfiguracji punktów na wykresie. Z poprzedniego podrozdziału wiemy, że jest sens to robić, gdy dwa pierwsze wymiary kanoniczne wyjaśniają należyty odsetek bezwładności całkowitej tablicy. Z wcześniejszych rozważań wynika natomiast, że liczebności w tablicy powinny być dostatecznie duże, tak aby któryś z profili nie okazał się specyficzny na skutek określonej konfiguracji danych w badanej próbie. Jeśli oba wymagania są spełnione, to wtedy analiza korespondencji dostarcza korzyści, które trudno odnaleźć w innych metodach analizy tablic. Sprawdza się przez to w sytuacjach, gdy tablica zawiera dużą liczbę wierszy lub kolumn.

Duża liczba wierszy bądź kolumn występuje na ogół w tablicach, w których zestawia się ze sobą pewne zjawisko dla różnych krajów. *Europejski Sondaż Społeczny* realizowany jest w ponad 30 krajach, *International Social Survey Programme* (ISSP) w 45, *Program Międzynarodowej Oceny Umiejętności Uczniów* (PISA) w ponad 50. Wielkości prób w każdym z krajów są na tyle duże, że można przyjąć oszacowania profili badanego zjawiska za wiarygodne. Analiza korespondencji wydaje się więc idealnym narzędziem zestawiania ze sobą wyników dla różnych krajów. Nie jest więc niczym zaskakującym, że większość przykładów zastosowań analizy korespondencji zamieszczanych w podręcznikach, encyklopediach czy innych opracowaniach poświęconych metodom analizy danych – prezentuje zalety metody na przykładzie porównań

międzykrajowych (zob. np. Greenacre 1994; Blasius i Greenacre 2006b; Nenadić i Greenacre 2006).

Aby nie powielać przykładów pochodzących z najczęściej wykorzystywanych w tym celu badań przedstawię porównanie międzykrajowe z zupełnie innej dziedziny. Chodzi o podział budżetu reklamowego między media w krajach europejskich. Skorzystam z danych zbieranych i udostępnianych przez *World Advertising Research Center* (www.warc.com). Przedstawiona analiza obejmie 25 krajów europejskich, dla których dostępne były dane za 2008 rok dotyczące podziału wydatków reklamowych między następujące media: telewizję, prasę, radio, outdoor (czyli reklamę zewnętrzną), kino i Internet. Analizowana tablica ma więc 25 wierszy i 6 kolumn. Wielkości w polach tablicy wyrażone są wielkością wydatków w milionach dolarów.

Omówienie wyników analizy korespondencji rozpocznijmy od wskaźników bezwładności. Bezwładność całkowita tablicy równa jest 0,1381, czyli mamy do czynienia ze średnio silnym związkiem. Mówiąc inaczej, między krajami nie występują różnice w strukturze wydatków, które można byłoby uznać za skrajne. Jeśli chcielibyśmy dowiedzieć się, co to właściwie oznacza, to można obliczyć i porównać profile wydatków na media w poszczególnych krajach. Medium o największych wydatkach w Europie stanowi prasa. Najniższe odsetki budżetów reklamowych alokowane są w prasie w Rosji (22 procent), w Serbii (25 procent) i w Polsce (26 procent), zaś najwyższe w Szwajcarii i w Austrii (po 61 procent), oraz w Finlandii (62 procent). Drugim w kolejności medium jest telewizja. Wydatki wahają się od 16 procent w Szwajcarii, 18 w Danii i 19 w Finlandii, do 50 w Rosji, 52 w Turcji oraz 54 w Serbii. Aczkolwiek może wydawać się, że wymienione różnice są znaczne, to należy wziąć pod uwagę fakt, że dotyczą krajów – poza Rosją – których budżety reklamowe mają niewielki udział w wydatkach w całej Europie. Mówiąc językiem analizy korespondencji, masy tych krajów są niewielkie, a przez to ich udział w bezwładności całkowitej również jest niewielki. W niewielkim więc stopniu „podwyższają" siłę związku.

Największy w Europie rynek niemiecki obejmuje 19 procent wydatków wszystkich rozpatrywanych krajów. Niewiele ustępuje mu rynek brytyjski (18 procent), zaś kolejne – francuski i włoski – obejmują odpowiednio 11,5 oraz 9 procent sumy europejskich budżetów reklamowych. Rynki te w największym stopniu wyznaczają przeciętny profil wykorzystania mediów w Europie. Zarazem – poza francuskim – w ich wkład do bezwładności całkowitej badanego zjawiska wynosi po około 10 procent. Największy wkład do bezwładności ma jednak Rosja (26 procent) – piąty kraj w Europie pod względem wysokości wydatków. Specyfika Rosji nie wynika więc z wielkości jej budżetu, lecz bierze się raczej z odmiennej struktury wydatków.

Tabela 7.6
Wyniki analizy korespondencji podziału wydatków reklamowych między media w krajach europejskich w 2008 roku

[1] korelacje kanoniczne i ich udziały w procentach

wymiar rozwiązania	korelacja	kwadrat korelacji	udział [%]	skumulowany udział [%]
1	0,2874	0,0826	59,8	59,8
2	0,1681	0,0282	20,5	80,3
3	0,1408	0,0198	14,4	94,7
4	0,0766	0,0059	4,3	98,9
5	0,0389	0,0015	1,1	100,0
w sumie		0,1381	100,0	

[2] Wydatki reklamowe dla krajów, współrzędne główne oraz bezwładności

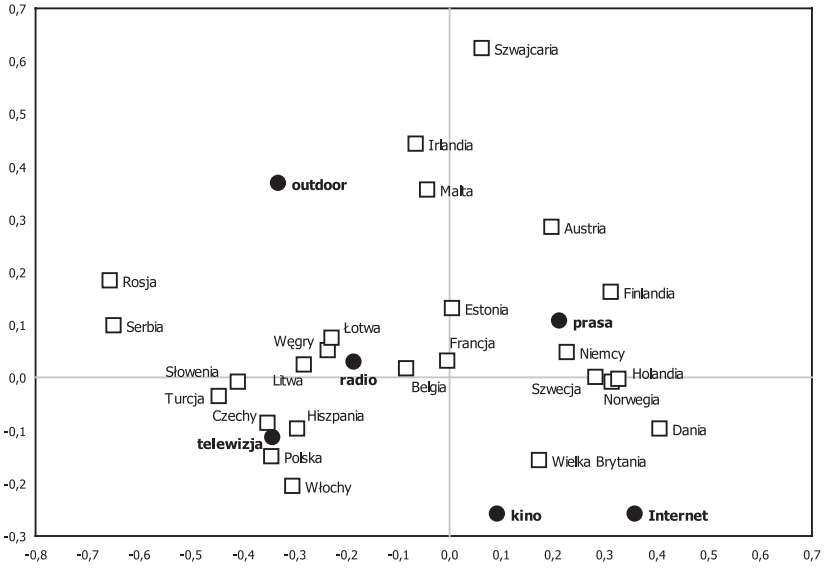
	wydatki reklamowe		współrzędne główne		bezwładność	
kraje	wielkość [mln USD]	udziały [%]	wymiar 1	wymiar 2	wielkości	udziały [%]
Rosja	10 359	7,0	-0,6568	0,1832	0,0360	26,1
Serbia	302	0,2	-0,6486	0,0977	0,0009	0,7
Turcja	2 491	1,7	-0,4453	-0,0362	0,0043	3,1
Słowenia	460	0,3	-0,4082	-0,0086	0,0006	0,5
Czechy	1 114	0,8	-0,3509	-0,0859	0,0011	0,8
Polska	3 359	2,3	-0,3433	-0,1495	0,0035	2,6
Włochy	13 501	9,1	-0,3023	-0,2062	0,0165	12,0
Hiszpania	10 160	6,9	-0,2943	-0,0980	0,0088	6,4
Litwa	230	0,2	-0,2817	0,0234	0,0002	0,1
Węgry	1 202	0,8	-0,2354	0,0507	0,0005	0,4
Łotwa	206	0,1	-0,2270	0,0754	0,0001	0,1
Belgia	4 000	2,7	-0,0825	0,0165	0,0018	1,3
Irlandia	2 312	1,6	-0,0654	0,4431	0,0037	2,7
Malta	42	0,0	-0,0420	0,3561	0,0000	0,0
Francja	17 062	11,5	-0,0031	0,0320	0,0022	1,6
Estonia	163	0,1	0,0062	0,1313	0,0000	0,0
Szwajcaria	3 659	2,5	0,0621	0,6239	0,0102	7,4
Wielka Brytania	26 802	18,1	0,1737	-0,1569	0,0161	11,6
Austria	4 367	3,0	0,1982	0,2846	0,0049	3,6
Niemcy	28 569	19,3	0,2270	0,0480	0,0117	8,5
Holandia	6 317	4,3	0,2816	0,0008	0,0041	3,0
Finlandia	2 259	1,5	0,3128	0,1617	0,0026	1,9
Norwegia	2 762	1,9	0,3140	-0,0083	0,0019	1,4
Szwecja	3 788	2,6	0,3277	-0,0028	0,0029	2,1
Dania	2 490	1,7	0,4067	-0,0962	0,0032	2,3
ogółem	147 975	100,0			0,1381	100,0

Tabela 7.6 (kontynuacja)

[3] Wydatki reklamowe w podziale na media, współrzędne główne oraz bezwładności

	wydatki w 2008 roku		współrzędne główne		bezwładności	
media	wielkość [mln USD]	udziały [%]	wymiar 1	wymiar 2	wartości	udziały [%]
telewizja	46 079	31,1	-0,3422	-0,1141	0,0416	30,1
outdoor	10 421	7,0	-0,3316	0,3689	0,0271	19,6
radio	7 603	5,1	-0,1846	0,0292	0,0079	5,7
kino	1 019	0,7	0,0917	-0,2574	0,0026	1,9
prasa	62 451	42,2	0,2120	0,1077	0,0268	19,4
Internet	20 402	13,8	0,3575	-0,2584	0,0322	23,3
ogółem	147 975	100,0			0,1381	100,0

Rycina 7.5
Podział wydatków reklamowych między media w krajach europejskich w 2008 roku
Obraz korespondencji



Wykres prezentowany na rycinie 7.5 dostarcza globalnego obrazu struktury wydatków na media w krajach europejskich. Na reklamie telewizyjnej opiera się rynek w krajach takich jak Polska, Czechy, Hiszpania czy Włochy. Z kolei grupa krajów obejmująca Niemcy, Holandię i kraje skandynawskie

korzysta przede wszystkim z prasy jako głównego medium komunikacji reklamowej. Radio wykorzystuje się częściej na Węgrzech i w krajach nadbałtyckich. Internet i kino to media specyficzne, wykorzystywane na szerszą skalę tylko w niektórych krajach. Najbliżej modelu reklamy opartego na Internecie i kinie znajdują się Wielka Brytania i kraje skandynawskie. Outdoor jest medium wspomagającym, niezwiązanym w sposób ścisły z żadnym pojedynczym krajem. Kierunek, w jakim odbiega punkt reprezentujący outdoor od środka wykresu informuje, że outdoor powinien mieć spory udział przede wszystkim na rynku szwajcarskim i rosyjskim, w Irlandii oraz na Malcie. Należy jeszcze dodać, że krajem, w którym wykorzystanie mediów w reklamie jest najbardziej zbliżone do przeciętnej europejskiej, jest Francja. Niewiele odbiega od niej Belgia, aczkolwiek z rysunku wynika, że ma nieco większy udział radia. Kraje o najbardziej specyficznych udziałach mediów w wydatkach reklamowych to Szwajcaria (znaczny udział outdooru), Rosja i Serbia (duże udziały telewizji i outdooru). Z rysunku odczytać też można specyfikę poszczególnych mediów. Radio wykorzystuje się w podobnym stopniu w wielu krajach. Reprezentujący je punkt leży bowiem najbliżej środka wykresu. Największe różnice między krajami dotyczą zaś Internetu i outdooru.

Można powiedzieć, że na podstawie wykresu udało się odtworzyć w miarę bogaty i dość wszechstronny obraz wykorzystania mediów na europejskim rynku reklamowym, mimo że porównanie objęło aż 25 krajów. Jeśli ze względu na cel badania obraz ten byłby niewystarczający, to bez wątpienia stanowi dogodny punkt wyjścia dla bardziej pogłębionych analiz.

7.9 Cecha ilościowa w analizie korespondencji. Jak wykształcenie wiąże się z inteligencją i dochodami

Z dotychczasowych rozważań wynika, że analiza korespondencji znajduje przede wszystkim zastosowanie w wypadku tablic, w których obie cechy mają naturę jakościową. Metoda dostarcza bowiem wartości skalowych dla kategorii każdej z cech, które to wartości pozwalają uporządkować kategorie według natury związku w tablicy. W praktyce występują jednak i takie sytuacje, w których jedna lub obie cechy mają charakter ilościowy. Tablicę zaś stosuje się jako narzędzie analizy związku między nimi, gdyż wartości jednej lub obu cech ilościowych są pogrupowane. Pogrupowanie wartości może być konsekwencją posłużenia się podczas badania kafeterią przedziałów wielkości danej cechy. Rozwiązanie takie stosuje się między innymi w badaniach sondażowych, pytając respondentów o dochód. Niekiedy badacz nie ma wpły-

wu na formę danych, na których musi się oprzeć. Na ogół dzieje się tak, gdy korzysta z danych publikowanych, bądź z innego źródła danych zastanych, w którym wartości cechy ilościowej są pogrupowane (na przykład wiek w tabelach statystycznych).

Zastosowanie analizy korespondencji w wypadku tablic zawierających cechy ilościowe prowadzi niekiedy do trudności polegających na tym, że porządek współrzędnych okazuje się niezgodny z porządkiem kategorii wyznaczonym przez wartości cechy. Próby poradzenia sobie z tym problemem doprowadziły do zaproponowania tak zwanych **ograniczonych** modeli korespondencji (ang. *constrained* bądź *restricted*), które pozwalają między innymi zachować wymagany porządek obliczanych współrzędnych (Gilula i Haberman 1988; Ritov i Gilula 1993; Böckenholt i Takane 1994). Propozycje te są jednak dyskusyjne, gdyż naruszają istotę modelu korespondencji, opartą na idei wzajemnego dopasowania średnich. Dlatego uwagę skupimy na innym podejściu, które ideę tę zachowuje. Oparte jest ono na grupowaniu kategorii, które naruszają porządek wynikający z istoty badanego zjawiska. Pokażemy zarazem przydatność proponowanego podejścia do analizy wyników badań sondażowych.

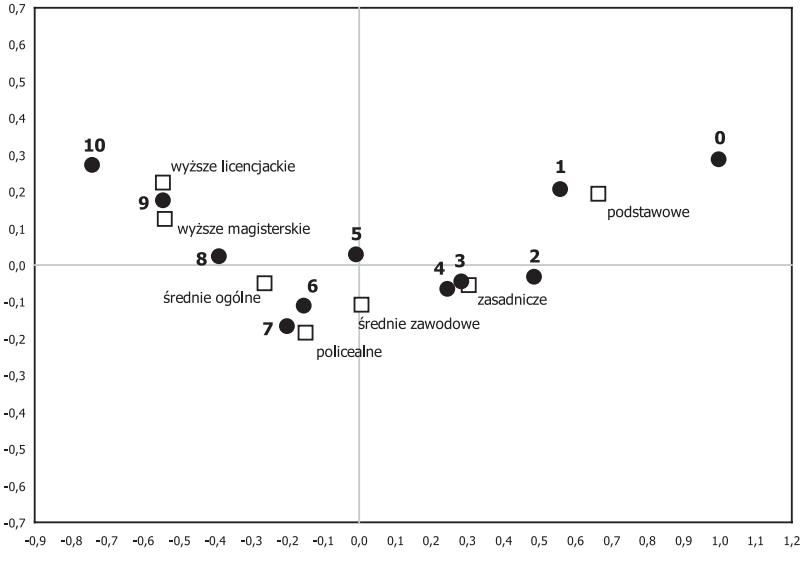
Właściwe narzędzie opisu związku między cechą jakościową a ilościową stanowi metoda zwana **regresją średnich** (Lissowski, Haman i Jasiński 2008: 240–242). Dla każdej kategorii cechy jakościowej oblicza się średnią wartości cechy ilościowej dla osób, które należą do danej kategorii, a następnie średnie te traktuje jako parametry opisujące poszczególne kategorie. Analizując związek poziomu wykształcenia z zarobkami można powiedzieć, że osoby o wykształceniu podstawowym zarabiają średnio określoną kwotę, osobom o wykształceniu zasadniczym odpowiada również pewna kwota średnich zarobków i tak dalej, dla każdej z kategorii wykształcenia. Siłę tak rozumianego związku między obiema cechami wyraża miara zwana **stosunkiem korelacyjnym** (Lissowski i in. 2008: 252–257), która jest ilorazem wariancji owych średnich zrelatywizowanej do pełnej wariancji zmiennej, dla której średnie są liczone. Im bardziej zróżnicowane są średnie odpowiadające kategoriom cechy jakościowej (mówiąc potocznie – wykształcenie bardziej różnicuje zarobki), tym wartość stosunku korelacyjnego jest większa. A ponieważ zwykły współczynnik korelacji jest szczególnym przypadkiem stosunku korelacyjnego (Lissowski i in. 2008: 278–283), stąd wartość stosunku korelacyjnego zestawić można z wartością korelacji kanonicznej. Stwarza to możliwość uwzględnienia regresji średnich oraz analizy korespondencji we wspólnym modelu.

Dla ilustracji tego ujęcia posłużmy się przykładem związku między wykształceniem a wynikami testu Ravena. Dane pochodzą z piątej edycji projek-

tu PolPan[9] zrealizowanej w roku 2008. Respondenci wypełniali test Ravena w wersji składającej się z 10 zadań. Pozwoliło to na skonstruowanie tablicy, w której wyniki testu skrzyżowane zostały z wykształceniem. Przy czym każdy z możliwych wyników, to jest od 0 do 10 poprawnie rozwiązanych zadań, potraktowany został jako osobna kategoria. Wyniki analizy kanonicznej prezentujemy w tabeli 7.7, zaś obraz korespondencji tej tablicy na rycinie 7.6.

Rozpocznijmy od przyjrzenia się wykresowi. Kategorie wyników testu ułożyły się wzdłuż poziomej osi zgodnie z liczbą poprawnie rozwiązanych zadań. W pierwszej chwili nie wydaje się, aby było w tym coś nieoczekiwanego. Warto jednak uświadomić sobie, że procedura obliczeniowa „nie wie", ile poprawnie rozwiązanych zadań odpowiada poszczególnym kategoriom wyników. Jedyne co robi, to dopasowuje średnie. Kategoriom wyników testu Ravena dopasowuje średnie wykształcenia, zaś kategoriom wykształcenia średnie wyniki testu. Uzyskany rezultat traktować można jako kolejny argument za tym, że metoda wprowadzona na gruncie botaniki sprawdza się w wyjaśnianiu zjawisk społecznych.

Rycina 7.6
Wykształcenie a liczba poprawnie rozwiązanych zadań w teście Ravena
Obraz korespondencji



[9] Projekt PolPan jest ogólnopolskim panelem prowadzonym od 1988 roku w cyklu pięcioletnim (Słomczyński i in. 1989; Słomczyński 2002). W każdej kolejnej edycji panel uzupełniany jest losową próbą osób w wieku 20–24 lata, dzięki czemu przekrój badanych osób jest zgodny ze strukturą wieku populacji.

Tabela 7.7
Wyniki analizy kanonicznej punktacji testu Ravena w kategoriach wykształcenia
Badanie PolPan 2008

[1] korelacje kanoniczne i ich udziały w procentach

wymiar rozwiązania	korelacja	kwadrat korelacji	udział w %	skumulowany udział w %
1	0,3853	0,1484	83,8	83,8
2	0,1268	0,0161	9,1	92,9
3	0,0809	0,0065	3,7	96,6
4	0,0567	0,0032	1,8	98,4
5	0,0429	0,0018	1,0	99,5
6	0,0308	0,0010	0,5	100,0
w sumie		0,1771	100,0	

[2] liczba i odsetki osób o różnym wykształceniu oraz współrzędne kanoniczne oraz bezwładności

			współrzędne główne		bezwładności	
wykształcenie	liczba osób	udziały [%]	wymiar 1	wymiar 2	wartości	udziały [%]
podstawowe	212	13,4	0,6648	0,1929	0,0649	36,7
zasadnicze zawodowe	331	20,9	0,3049	-0,0561	0,0227	12,8
średnie ogólne	232	14,6	-0,2616	-0,0518	0,0138	7,8
średnie zawodowe	404	25,5	0,0085	-0,1078	0,0045	2,5
policealne	87	5,5	-0,1467	-0,1840	0,0041	2,3
wyższe licencjackie	98	6,2	-0,5438	0,2236	0,0235	13,3
wyższe magisterskie	220	13,9	-0,5389	0,1242	0,0435	24,6
ogółem	1584	100,0				100,0

[3] liczba i odsetki dobrze rozwiązanych zadań, współrzędne kanoniczne oraz bezwładności

			współrzędne główne		bezwładności	
liczba poprawnie rozwiązanych zadań	liczba osób	udziały [%]	wymiar 1	wymiar 2	wielkości	udziały [%]
0	42	2,7	0,9990	0,2876	0,0290	16,4
1	96	6,1	0,5584	0,2055	0,0219	12,4
2	129	8,1	0,4852	-0,0342	0,0200	11,3
3	161	10,2	0,2835	-0,0448	0,0100	5,6
4	184	11,6	0,2464	-0,0653	0,0087	4,9
5	218	13,8	-0,0083	0,0278	0,0009	0,5
6	216	13,6	-0,1525	-0,1110	0,0077	4,4
7	191	12,1	-0,1997	-0,1667	0,0112	6,3
8	173	10,9	-0,3879	0,0238	0,0172	9,7
9	100	6,3	-0,5437	0,1745	0,0209	11,8
10	74	4,7	-0,7399	0,2705	0,0294	16,6
ogółem	1584	100,0				100,0

Tabela 7.8
Średnia liczba poprawnie rozwiązanych zadań w teście Ravena oraz współrzędne pierwszego wymiaru kanonicznego dla osób o różnym wykształceniu
Badanie PolPan 2008

wykształcenie	średnia liczba rozwiązanych zadań	średnie standaryzowane	standaryzowane współrzędne kanoniczne
podstawowe	3,58	-1,71	-1,73
zasadnicze zawodowe	4,45	-0,82	-0,79
średnie ogólne	5,93	0,69	0,68
średnie zawodowe	5,24	-0,01	-0,02
policealne	5,64	0,40	0,38
wyższe licencjackie	6,60	1,38	1,41
wyższe magisterskie	6,62	1,40	1,40
ogółem	5,25	0,00	0,00

To, że kategorie ułożyły się na wykresie w kolejności od 0 do 10, świadczy zarazem o tym, że test jest dobrze skonstruowany. Jego punktacja jest trafna wobec kryterium zewnętrznego, jakie w tym wypadku stanowi wykształcenie. Dodatkowym argumentem jest w tym wypadku wysoki udział pierwszego wymiaru kanonicznego w opisie związku punktacji testu z wykształceniem, który wynosi prawie 84 procent. Warto też zauważyć, że wynik równy „5 poprawnie rozwiązanych zadań", stanowiący punkt środkowy skali punktacji, leży blisko centralnego punktu wykresu. Profil wykształcenia osób poprawnie rozwiązujących 5 z 10 zadań jest więc zbliżony do profilu przeciętnego, czyli do profilu wykształcenia wszystkich badanych. Dowodzi to z kolei, że stosowana wersja testu Ravena jest dobrze zrównoważona.

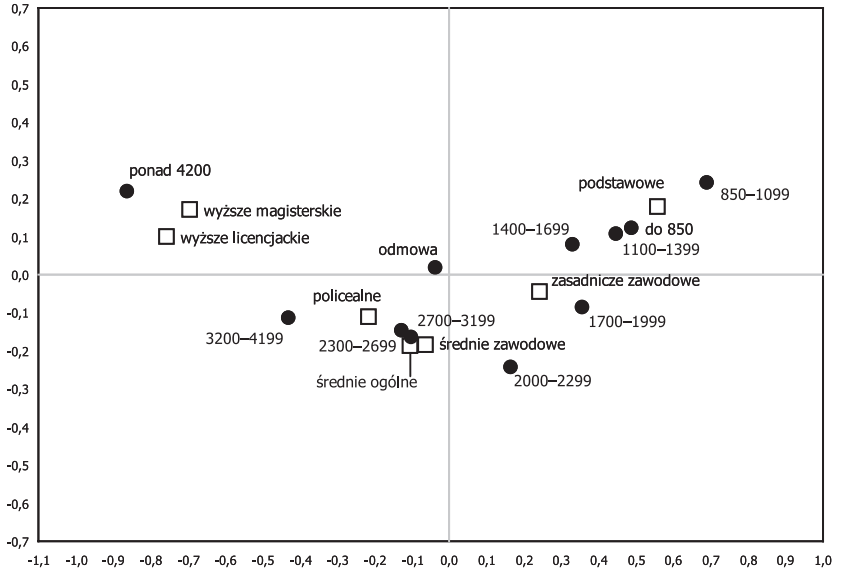
Układ kategorii wykształcenia jest również godny uwagi. Osoby o wykształceniu średnim ogólnym uzyskują w teście lepsze wyniki od osób kończących średnie szkoły zawodowe, a także od osób, które po skończeniu szkoły średniej uczyły się dalej w szkołach policealnych. W wypadku ostatniej kategorii dłuższe kształcenie nie przekłada się więc na wyższy wynik testu. Uwagę zwraca też podobne położenie punktów odpowiadających osobom, które skończyły studia pierwszego stopnia (licencjackie a także inżynierskie), oraz osobom, które zdobyły wykształcenie na studiach magisterskich (kategoria ta obejmuje również osoby, które kształciły się dalej na studiach podyplomowych oraz osoby, które doszły do stopni naukowych). Wydawać by się mogło, że studia drugiego stopnia kończą w większym stopniu osoby o wyższej inteligencji ogólnej, którą mierzy test Ravena. Okazuje się, że nie jest to prawdą.

O ukończeniu studiów drugiego stopnia decydują więc inne czynniki – prawdopodobnie związane z nierównościami pochodzeniowymi.

Do tych samych wniosków można dojść inną drogą. Wyniki testu są cechą ilościową, więc można obliczyć średnie liczby poprawnie rozwiązanych zadań dla osób należących do poszczególnych kategorii wykształcenia. Obliczone w ten sposób średnie zamieszczone zostały w tabeli 7.8. Przeciętne wyniki osób kończących licea ogólnokształcące są wyższe, niż osób kończących średnie szkoły zawodowe bądź szkoły policealne. Przeciętne wyniki osiągane przez osoby o obu rodzajach wykształcenia wyższego są podobne.

Aby porównać otrzymane średnie z wynikami korespondencji należy przekształcić je do postaci standaryzowanej, tak aby ich średnia była równa 0, zaś odchylenie standardowe było równe 1. W postaci tej mogą być zestawione wprost ze współrzędnymi kanonicznymi w wersji standaryzowanej. Oba porównywane zestawy przedstawione zostały w dwóch ostatnich kolumnach tabeli 7.8. Jak widać, praktycznie są one równoważne. Potwierdza to obliczona między nimi korelacja, która wynosi 0,9999. Identyczne wnioski otrzymamy więc niezależnie od tego, czy do analizy związku między wykształceniem a wynikami testu posłużymy się metodą średnich, czy też metodą korespondencji.

Rycina 7.7
Wykształcenie a miesięczne dochody gospodarstwa domowego
Europejski Sondaż Społeczny 2008



Wyniki testu Ravena ze względu na swój sposób konstrukcji wyrażają się wielkościami liczbowymi. Inaczej rzecz się ma, gdy cechę ilościową o skategoryzowanych wartościach stanowią odpowiedzi na pytanie zadane w badaniu. Wtedy na ogół powstaje problem nadania wartości liczbowej kategorii odmów, a także przypisania wartości przedziałom skrajnym, gdy są one jednostronnie otwarte. Tego rodzaju przykład przedstawia rycina 7.7 (parametry liczbowe rozwiązania podano w tabeli 7.9). Na wykresie dokonano projekcji związku między wykształceniem a miesięcznymi dochodami gospodarstwa domowego. Dane pochodzą z Europejskiego Sondażu Społecznego 2008, w którym o dochody pytano za pomocą pytania skategoryzowanego. Odmowy podania dochodu stanowiły 16 procent wszystkich odpowiedzi. Jest to zresztą niewielki odsetek jak na badania sondażowe. Skrajne przedziały były jednostronnie otwarte (do 850 zł i ponad 4200 zł). Do tego należy jeszcze dodać, że długości poszczególnych przedziałów były niejednakowe.

Rezultaty analizy korespondencji informują przede wszystkim o tym, że kategoria odmów lokuje się blisko punktu centralnego wykresu. Można więc przyjąć, że odmowy rozłożone są podobnie w poszczególnych kategoriach wykształcenia, a więc nie mają wpływu na wnioski dotyczące badanego zjawiska. Udział bezwładności związanej z tą kategorią wynosi zaledwie 0,5 procenta, czyli jest pomijalny w kontekście całkowitej bezwładności tablicy. Kategorię odmów można więc wyskalować, przyjmując dla niej średnią dochodów obliczoną dla osób, które swoje dochody podały.

Drugi wniosek z wyników analizy korespondencji dotyczy porządku kategorii dochodowych. Ich ułożenie wzdłuż osi poziomej nie we wszystkich fragmentach jest zgodne z porządkiem odpowiadających im kwot. Rozpatrzmy prawy kraniec skali obejmujący grupy o najniższych dochodach. Najbardziej skrajne położenie zajmuje kategoria „850–1099 zł", zaś kategoria „do 850 zł" – nominalnie odpowiadająca niższym dochodom – umiejscowiła się blisko dochodów „1100–1399 zł". Jakie mogą być powody tej konfiguracji? Nie można wykluczyć, że osoby o dochodach do 850 złotych mają przeciętnie wyższy poziom wykształcenia od osób, które osiągają dochody w przedziale od 850 do 1099 złotych. Gdyby mieć uzasadnienie dla takiej hipotezy, to wtedy otrzymaną konfigurację należałoby pozostawić, zastanawiając się wyłącznie nad tym, jaką pojedynczą kwotę należałoby przypisać osobom, które deklarują dochód poniżej 850 złotych.

Zasadność tej hipotezy wydaje się jednak wątpliwa. Większość przedziałów dochodowych układa się zgodnie z wykształceniem, zaś związek wyodrębnia się w wyraźny sposób. Bezwładność całkowita wynosi 0,2264, to jest nawet więcej niż w wypadku związku wykształcenia z wynikami testu Ravena. W miarę wysoka jest też pierwsza korelacja kanoniczna (0,4330), co w tym

Tabela 7.9
Wyniki analizy kanonicznej związku wykształcenia respondenta z miesięcznym dochodem gospodarstwa domowego
Europejski Sondaż Społeczny 2008

[1] korelacje kanoniczne i ich udziały w procentach

wymiar rozwiązania	korelacja	kwadrat korelacji	udział [%]	skumulowany udział [%]
1	0,4330	0,1875	82,8	82,8
2	0,1481	0,0219	9,7	92,5
3	0,0831	0,0069	3,1	95,6
4	0,0754	0,0057	2,5	98,1
5	0,0515	0,0027	1,2	99,2
6	0,0417	0,0017	0,8	100,0
w sumie		0,2264	100,0	

[2] liczba i odsetki osób o różnym wykształceniu, współrzędne kanoniczne oraz bezwładności

			współrzędne główne		bezwładności	
wykształcenie	liczba osób	udziały [%]	wymiar 1	wymiar 2	wartości	udziały [%]
podstawowe	262	20,5	0,5578	0,1768	0,0709	31,3
zasadnicze zawodowe	338	26,5	0,2431	-0,0446	0,0197	8,7
średnie ogólne	91	7,1	-0,1043	-0,1879	0,0068	3,0
średnie zawodowe	255	19,9	-0,0631	-0,1855	0,0087	3,9
policealne	68	5,3	-0,2149	-0,1104	0,0072	3,2
wyższe licencjackie	61	4,7	-0,7552	0,0979	0,0306	13,5
wyższe magisterskie	205	16,1	-0,6936	0,1687	0,0824	36,4
ogółem	1279	100,0				100,0

[3] liczba i odsetki osób w przedziałach dochodowych, współrzędne kanoniczne oraz bezwładności

			współrzędne główne		bezwładności	
miesięczny dochód gospodarstwa	liczba osób	udziały [%]	wymiar 1	wymiar 2	wielkości	udziały [%]
do 850 zł	62	4,9	0,4884	0,1219	0,0137	6,0
850—1099 zł	71	5,6	0,6902	0,2399	0,0307	13,5
1100—1399 zł	116	9,1	0,4465	0,1067	0,0199	8,8
1400—1699 zł	123	9,6	0,3293	0,0776	0,0139	6,1
1700—1999 zł	93	7,3	0,3554	-0,0853	0,0128	5,7
2000—2299 zł	104	8,1	0,1634	-0,2423	0,0075	3,3
2300—2699 zł	103	8,0	-0,1274	-0,1466	0,0059	2,6
2700—3199 zł	110	8,6	-0,1026	-0,1648	0,0051	2,2
3200—4199 zł	143	11,2	-0,4308	-0,1153	0,0235	10,4
ponad 4200 zł	148	11,6	-0,8627	0,2184	0,0923	40,8
odmowa	207	16,1	-0,0379	0,0179	0,0012	0,5
ogółem	1279	100,0				100,0

wypadku oznacza, że wykształcenie wyjaśnia około 19 procent zróżnicowania dochodów gospodarstw. To sporo, zważywszy, że w części gospodarstw respondent nie musi być osobą najwięcej zarabiającą, a nawet może nie mieć żadnego wkładu w dochód gospodarstwa. Istotne jest również to, że udział pierwszego wymiaru kanonicznego w bezwładności całkowitej jest wysoki, gdyż wynosi 83 procent. To prawie tyle samo, co w wypadku testu Ravena, który był specjalnie konstruowany z myślą o jednowymiarowości.

Warto więc rozważyć i inne czynniki, które spowodować mogły, że punkty w dolnej części skali ułożyły się w porządku niezgodnym z odpowiadającymi im kwotami. Jednym z takich powodów mogą być błędy związane z reprezentacyjnym charakterem badania. Najniższy przedział dochodowy wskazały 62 osoby, zaś kolejny (850–1099 zł) 71 osób. Nie są to liczne podpróby. Jednakże testowanie hipotezy, że w populacji obu punkty leżą w odwrotnej kolejności jest pracochłonne, gdyż wymaga posłużenia się metodami repróbkowania (ang. *bootstrap*; zob. Maddala 2006: 665–669; Moore, McCabe i Craig 2007), przy czym każda powtórzona próba wymagałaby szacowania od nowa parametrów modelu korespondencji. Dlatego w praktyce rzadko idzie się tą drogą (Lebart 2006). W każdym razie nie można wykluczyć, że w badanej populacji kolejność obu kategorii jest odwrotna.

Istnieje też trzecia możliwość. Odwrócony porządek kategorii może być wynikiem sposobu postrzegania i przedstawiania swoich dochodów przez respondentów. Na wykresie kategoria „do 850 zł" jest przesunięta w kierunku punktu centralnego wobec kategorii sąsiedniej. Oznacza to, że dochód taki wskazywały osoby o bardziej zróżnicowanym wykształceniu, niż w wypadku przedziału dochodowego „850–1099 zł". Przyczyny wskazywania szczególnie niskich dochodów nie muszą więc być związane z wykształceniem. W grę wchodzić może podeszły wiek, zamieszkiwanie samotnie, trudna sytuacja życiowa – na przykład na skutek rozwodu, bądź też nieregularność uzyskiwanych przychodów. Wymienione grupy badanych postrzegać mogą swój dochód jako niski, czy niewystarczający, a stąd wskazywać najniższą kategorię w przedstawionej kafeterii.

W każdym razie niezależnie od przyczyn uzyskany układ trzech najniższych kategorii nie jest spójny z ideą skorelowania dochodów z wykształceniem. Dlatego zasadne staje się zgrupowanie tych kategorii w jedną. Podobny należałoby postąpić w wypadku dwóch innych kategorii: „2300–2699 zł" oraz „2700–3199 zł". Odpowiadające im na wykresie punkty właściwie pokrywają się. Oznacza to, że podobne są profile wykształcenia osób osiągających dochody w każdym z tych przedziałów. Ich połączenie nie powinno więc zaburzyć kształtu badanego zjawiska. Z tych samych powodów zasadne jest połączenie kategorii „1400–1699 zł" oraz „1700–1999 zł". Co prawda na wykresie dzieli

je pewien dystans, lecz w pierwszym wymiarze, który zdecydowanie więcej wyjaśnia niż drugi, różnica między nimi jest minimalna.

Tabela 7.10
Wyniki analizy kanonicznej związku wykształcenia respondenta z miesięcznym dochodem w gospodarstwie domowym. Pogrupowane kategorie dochodów
Europejski Sondaż Społeczny 2008

[1] korelacje kanoniczne i ich udziały w procentach

wymiar rozwiązania	korelacja	kwadrat korelacji	udział [%]	skumulowany udział [%]
1	0,4305	0,1853	86,7	86,7
2	0,1431	0,0205	9,6	96,3
3	0,0751	0,0056	2,6	99,0
4	0,0382	0,0015	0,7	99,7
5	0,0263	0,0007	0,3	100,0
6	0,0080	0,0001	0,0	100,0
w sumie		0,2137	100,0	

[2] liczba i odsetki osób o różnym wykształceniu, współrzędne kanoniczne oraz bezwładności

			współrzędne główne		bezwładności	
wykształcenie	liczba osób	udziały [%]	wymiar 1	wymiar 2	wartości	udziały [%]
podstawowe	262	20,5	0,5530	0,1804	0,0695	32,5
zasadnicze zawodowe	338	26,5	0,2398	-0,0578	0,0178	8,3
średnie ogólne	91	7,1	-0,0990	-0,1442	0,0038	1,8
średnie zawodowe	255	19,9	-0,0588	-0,1692	0,0065	3,0
policealne	68	5,3	-0,2074	-0,1542	0,0046	2,1
wyższe licencjackie	61	4,7	-0,7521	0,1144	0,0299	14,0
wyższe magisterskie	205	16,1	-0,6931	0,1559	0,0815	38,1
ogółem	1279	100,0			0,2137	100,0

[3] liczba i odsetki osób w przedziałach dochodowych, współrzędne kanoniczne oraz bezwładności

			współrzędne główne		bezwładności	
miesięczny dochód gospodarstwa	liczba osób	udziały [%]	wymiar 1	wymiar 2	wielkości	udziały [%]
do 1399 zł	249	19,5	0,5256	0,1507	0,0587	27,5
1400–1999 zł	216	16,9	0,3408	0,0107	0,0214	10,0
2000–2299 zł	104	8,1	0,1647	-0,2351	0,0075	3,5
2300–3199 zł	212	16,6	-0,1140	-0,1621	0,0091	4,2
3200–4199 zł	143	11,2	-0,4299	-0,1227	0,0235	11,0
ponad 4200 zł	148	11,6	-0,8641	0,2170	0,0923	43,2
odmowa	207	16,1	-0,0378	0,0203	0,0012	0,6
ogółem	1279	100,0			0,2137	100,0

Po połączeniu wskazanych kategorii analiza korespondencji została wykonana ponownie, zaś jej wyniki przedstawione w tabeli 7.10 oraz na rycinie 7.8. Jak widać, obraz związku wyklarował się. Obecnie kategorie dochodów układają się ściśle w porządku rosnącym. Pozostaje więc przejść do ostatniego zadania, jakim jest przypisanie pojedynczych kwot kategoriom dochodów.

Optymalne wartości skalowe dla kategorii dochodów wyznaczone są przez współrzędne główne pierwszego wymiaru. Optymalne w tym sensie, że maksymalizują siłę związku. Wobec tego wystarczy zamienić współrzędne te na kwoty. Przedstawimy dwa sposoby rozwiązania tego problemu, które, jak okaże się dalej, prowadzą do podobnego rezultatu. Punktem wyjścia obu sposobów jest określenie pojedynczych kwot dla wszystkich przedziałów, które nie są skrajne. Ponieważ badanie nie dostarcza informacji na temat rozkładu dochodów wewnątrz przedziałów, to jako reprezentujące je kwoty najrozsądniej jest przyjąć środki przedziałów. Kwoty te zostały podane w kolumnie [1] tabeli 7.11. Tabela ta ma więcej wierszy, niż wynosi liczba skalowanych kategorii dochodu, gdyż między kategoriami umieszczono wiersze przedstawiające dystanse między nimi (oznaczone symbolem „|").

Rycina 7.8
Wykształcenie a miesięczny dochód w gospodarstwie domowym
Pogrupowane kategorie dochodów
Europejski Sondaż Społeczny 2008

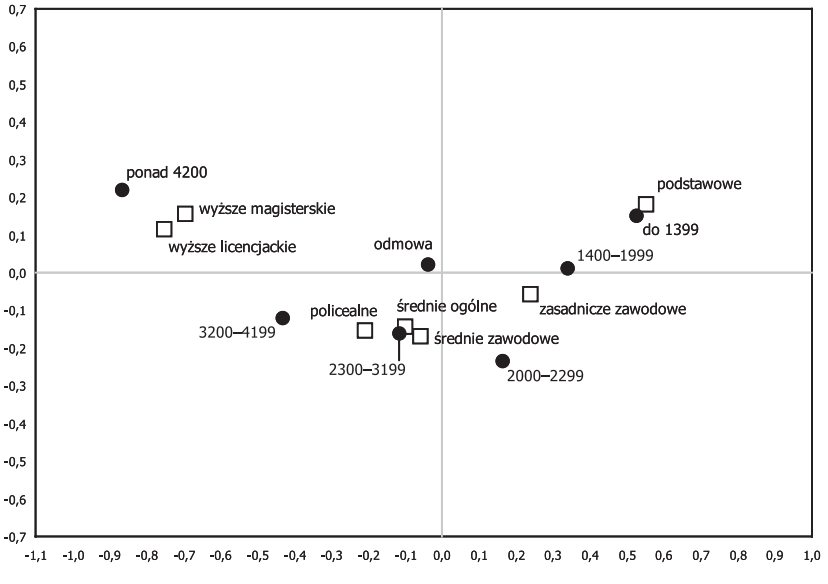


Tabela 7.11
Wielkości ilustrujące sposób estymacji wartości skalowych dla dwóch skrajnych przedziałów dochodu
Europejski Sondaż Społeczny 2008

kategorie dochodu	środki przedziałów [1]	dystanse między środkami [2]	współrzędne główne pierwszego wymiaru [3]	różnice współrzędnych [4]	estymacje różnic między środkami [5]	estymowane kwoty dla przedziałów [6]
do 1399 zł			-0,5256			1228
|				0,1848	472	
1400–1999 zł	1700		-0,3408			1700
|		450		0,1761		
2000–2299 zł	2150		-0,1647			2150
|		600		0,2788		
2300–3199 zł	2750		0,1140			2750
|		950		0,3159		
3200–4199 zł	3700		0,4299			3700
|				0,4342	1306	
ponad 4200 zł			0,8641			5006
odmowa						2566

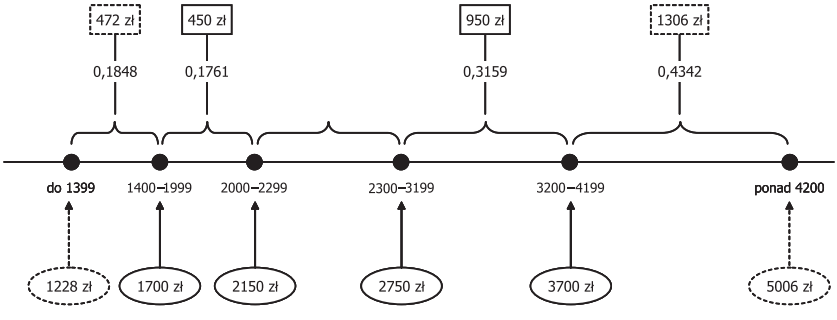
Pierwszy ze sposobów polega na przypisaniu przedziałom, które nie są skrajne, kwot stanowiących ich środki. W sytuacji tej oszacowania wymagają jedynie wartości dla przedziałów skrajnych oraz kwota, którą należy przypisać kategorii odmów.

Zacznijmy od najniższej kategorii „do 1399 zł". Jedyną informacją, na której można się oprzeć, jest otrzymana metodą korespondencji wartość współrzędnej głównej (kolumna [3] tabeli 7.11). Przyjmijmy, że dystans owej współrzędnej do współrzędnej kolejnej kategorii (1400–1999 zł) wyznacza dystans między kwotami przypisanymi tym kategoriom. Sytuację tę prezentuje rycina 7.9. Różnica między środkami przedziałów „1400–1999 zł" oraz „2000–2299 zł" wynosi 450 złotych. Kwocie tej odpowiada dystans między współrzędnymi głównymi równy 0,1761. Z kolei różnica między wartościami współrzędnych głównych odpowiadających dwóm kategoriom najniższych dochodów wynosi 0,1848. Jeśli założymy, że różnice kwot zachowują proporcje różnic współrzędnych głównych, to wtedy różnica kwot przypisanych dwóm najniższym przedziałom będzie równa

$$\frac{0{,}1848}{0{,}1761} * 450zł = 472zł$$

Po odjęciu tej wielkości od środka przedziału „1400–1999 zł" otrzymujemy wartość skalową dla najniższej kategorii dochodów – równą 1228 złotych. Obliczona w analogiczny sposób wartość skalowa dla najwyższego przedziału dochodów „ponad 4200 zł" wynosi 5006 złotych. Po określeniu obu wielkości obliczyć można średnią dochodów dla respondentów, którzy dochody podali. Średnia ta wynosi 2566 złotych. Zgodnie z wcześniejszym wnioskiem dotyczącym położenia kategorii odmów blisko centralnego punktu obrazu korespondencji, zasadne jest przypisanie tej kategorii jako wartości średniej zarobków.

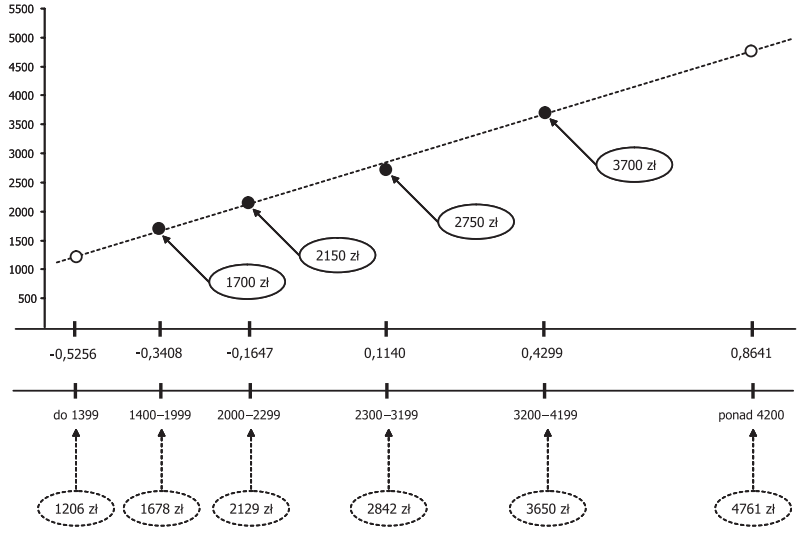
Rycina 7.9
Schemat skalowania skrajnych kategorii dochodów według dystansów współrzędnych głównych



Druga z metod wyznaczania wartości skalowych opiera się na założeniu, że różnice w ramach każdej pary tych wartości proporcjonalne są do dystansów między współrzędnymi głównymi. Do zakotwiczenia przypisanych kwot w wynikach badania skorzystać można, tak jak poprzednio, ze środków przedziałów. Przyjmijmy owe środki jako wartości zmiennej wyjaśnianej w modelu regresji liniowej, w którym rolę predyktora pełnią współrzędne główne. Omawiany schemat postępowania zilustrowany został na rycinie 7.10. Do czterech znanych środków przedziałów dochodowych – zaznaczonych na rycinie czarnymi kółkami – dopasowany został model regresji liniowej. Model ten pozwala obliczyć wartości skalowe dla dwóch skrajnych przedziałów. Na rycinie estymowane wartości skalowe zostały zaznaczone białymi kółkami, zaś wielkości odpowiadających im kwot podane na dole ryciny. Model pozwala również wyznaczyć skorygowane wartości skalowe dla przedziałów, których środki były znane. W rozważanym przypadku skorygowane wartości nie różnią się wiele od oryginalnych, toteż w gruncie

rzeczy jest obojętne, które z nich ostatecznie przyjąć. Wielkości wyliczone za pomocą regresji mają jednak tę własność, że są ściśle proporcjonalne do współrzędnych głównych. Daje to pewność, że uczynione zostało wszystko, aby zachować istotę analizowanego związku. Należy jeszcze zaznaczyć, że wartość skalowa dla kategorii odmów również może być wyznaczona metodą regresji. Dla kategorii tej, tak jak dla każdej innej, znana jest bowiem wartość współrzędnej głównej.

Rycina 7.10
Schemat skalowania kategorii dochodu metodą regresji współrzędnych głównych



Wartości skalowe otrzymane obiema metodami zebrane zostały w części [1] tabeli 7.12. Zgodność między nimi jest nadspodziewanie wysoka, o czym świadczy wartość korelacji równa 0,9975. Należy jedynie odnotować, że metoda skalowania skrajnych kategorii rozciągnęła nieco skalę w górę, przypisując wyraźnie wyższą wartość kategorii dochodów ponad 4200 złotych.

Znalezione wartości skalowe posłużyć mogą do obliczenia średnich dochodów w gospodarstwach osób o różnym poziomie wykształcenia. Wielkości te przedstawione zostały w części [2] tabeli 7.12. Średnie otrzymane w oparciu o zestaw wartości uzyskanych metodą regresji są na ogół nieco niższe od średnich opartych na drugim zestawie wartości, niemniej jednak różnice te są niewielkie. Oznacza to, że wyznaczają prawie analogiczne dystanse między kategoriami wykształcenia. Korelacja między oboma zestawami średnich wy-

nosi 0,9999, czyli obie metody wyznaczania wartości skalowych prowadzą w tym wypadku do nierozróżnialnych interpretacji badanego zjawiska. Skonstruowany w ten sposób jednowymiarowy model związku między wysokością dochodu w gospodarstwie a wykształceniem badanej osoby zobrazowano na rycinie 7.11.

Tabela 7.12
Wyznaczone dwiema metodami wartości skalowe dla kategorii dochodów gospodarstw przy uwzględnieniu jako kryterium wykształcenia badanych oraz średnie dochody w gospodarstwach osób o różnym poziomie wykształcenia
Europejski Sondaż Społeczny 2008

[1] wartości skalowe dla kategorii dochodów gospodarstw

miesięczny dochód gospodarstwa	skalowanie skrajnych kategorii według dystansów kanonicznych	regresja współrzędnych głównych	różnice
do 1399 zł	1228	1206	-22
1400–1999 zł	1700	1678	-22
2000–2299 zł	2150	2129	-21
2300–3199 zł	2750	2842	92
3200–4199 zł	3700	3650	-50
ponad 4200 zł	5006	4761	-245
odmowa	2566	2647	81

korelacja: 0,9975

[2] średni dochód w gospodarstwach osób o różnym poziomie wykształcenia dla dwóch zestawów wartości skalowych

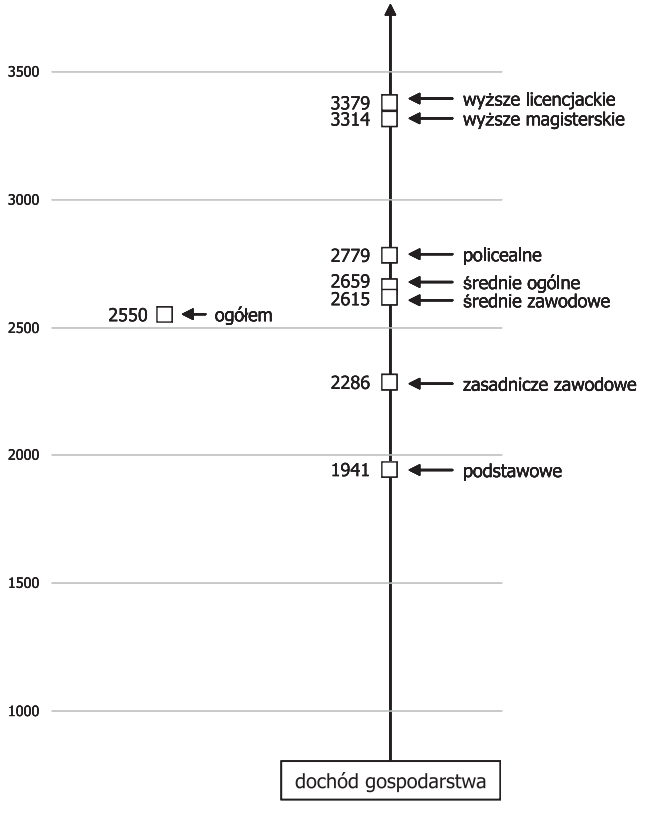
miesięczny dochód gospodarstwa	skalowanie skrajnych kategorii według dystansów kanonicznych	regresja współrzędnych głównych	różnice
podstawowe	1940	1941	1
zasadnicze zawodowe	2286	2286	0
średnie ogólne	2679	2659	-20
średnie zawodowe	2623	2615	-8
policealne	2794	2779	-15
wyższe licencjackie	3426	3379	-47
wyższe magisterskie	3374	3314	-60
ogółem	**2566**	**2550**	**-15**

korelacja: 0,9999

W tabeli 7.13 zestawiono etapy przekształcania wyników analizy korespondencji w model średnich. Bezwładność całkowitą pierwotnej tablicy potraktujmy jako miarę ilości informacji niezbędnych do najbardziej pełnego – jak to jest możliwe – opisu związku otrzymanego w badaniu. W pierwszym etapie przekształcania tej tablicy zdecydowaliśmy się na pogrupowanie niektórych

oryginalnych przedziałów dochodowych. Po wykonaniu tej operacji zachowało się 94,4 procenta pierwotnej bezwładności rozpatrywanego związku. Następnie ograniczyliśmy opis związku do jednego wymiaru. Po wykonaniu tej operacji zachowało się 86,7 procenta bezwładności z poprzedniego etapu. Przypisanie wartości skalowych kategoriom dochodów nie spowodowało żadnej utraty informacji w wypadku stosowania metody regresji, gdyż przypisane wartości równoważne były współrzędnym kanonicznym. W wypadku posłużenia się drugą metodą utrata informacji była pomijalna. W sumie można przyjąć, że pełny proces przekształcania modelu korespondencji w model średnich zachował w rozważanym przypadku około czterech piątych pierwotnej bezwładności czy – mówiąc inaczej – zawartej w wynikach badania wiedzy

Rycina 7.11
Średnie dochody w gospodarstwach osób o różnym wykształceniu
Europejski Sondaż Społeczny 2008



o kształcie związku w tablicy. Przy czym większą utratę bezwładności spowodowało ograniczenie modelu do pojedynczego wymiaru, niż pogrupowanie wybranych kategorii dochodów.

Tabela 7.13
Miary bezwładności zachowanej przez kolejne modele związku między wykształceniem a dochodami w gospodarstwa domowego

etap	opis modelu związku	rodzaj miary	wartość miary	odsetek zachowanej bezwładności wobec modelu: poprzedniego [%]	odsetek zachowanej bezwładności wobec modelu: początkowego [%]
0	model korespondencji: oryginalny zestaw kategorii dochodów	bezwładność całkowita	0,2264	–	–
1	model korespondencji: po pogrupowaniu niektórych kategorii	bezwładność całkowita	0,2137	94,4	94,4
2	ograniczenie modelu do pierwszego wymiaru kanonicznego	kwadrat korelacji kanonicznej	0,1853	86,7	81,8
3a	model średnich (skalowanie metodą regresji)	kwadrat stosunku korelacyjnego	0,1853	100,0	81,8
3b	model średnich (skalowanie wartości skrajnych)	kwadrat stosunku korelacyjnego	0,1844	99,5	81,4

Omówiony przykład pokazuje przydatność analizy korespondencji również w sytuacjach, gdy badamy związek między cechą jakościową a ilościową. Pozwala ona określić, co tracimy, przechodząc na prostszy konceptualnie jednowymiarowy model badanego związku. Jest także źródłem wskazówek co do sposobu grupowania i skalowania oryginalnych kategorii, tak aby zachować istotę badanego zjawiska.

7.10 Uwzględnienie w tablicy więcej niż dwóch cech

Na zakończenie przedstawię sposoby uwzględnienia w tablicy więcej niż dwóch cech. O tej dość paradoksalnej możliwości nie mówiliśmy wcześniej z tego względu, że ingerowałaby w logikę interpretacji tablic, które z natury rzeczy stanowią narzędzie analizy i prezentacji zależności między dwiema cechami. Niemniej jednak, gdy logikę tę omówiliśmy już wszechstronnie, nie grozi nam niebezpieczeństwo pomieszania propozycji w tym zakresie z meto-

dami zorientowanymi na analizę tablic wielowymiarowych, które można chociażby znaleźć w ramach modelowania log-liniowego (Lissowski 1984). Podejście prezentowane w tym miejscu ma z gruntu rzeczy odmienny charakter. Zachowuje bowiem logikę analizy tablic dwuzmiennowych, któremu poświęcona jest książka. Kolejne cechy wprowadza się zaś do modelu w specyficzny sposób, jako komponenty jednej lub obu cech przedstawionych w konwencjonalnej tablicy.

Jedno z rozwiązań w tym zakresie zaproponowano w ramach analizy korespondencji. Otóż są nim tak zwane **tablice łączone** (ang. *stacked*). Analiza tablic łączonych nie różni się niczym od omawianej dotychczas prostej analizy korespondencji. Sprawa sprowadza się do sposobu konstrukcji tych tablic. Polega on na sklejaniu ze sobą pewnej liczby konwencjonalnych tablic, co możliwe jest pod warunkiem, że posiadają one wspólną cechę w wierszach bądź w kolumnach (Blasius 1994; Greenacre 2006: 49–61).

Zasadę budowy i analizy tablic łączonych wyjaśnijmy na przykładzie. W wielu miejscach tej książki rozważaliśmy związek między wykształceniem ojca a wyborem przez dziecko szkoły ponadgimnazjalnej. Wiemy już chyba wszystko na temat tego związku. Ustalenia te zakwestionować jednak można w bardzo prosty sposób. Wystarczy zadać pytanie: dlaczego wzięliśmy pod uwagę wykształcenie ojca? Czy zasadne jest wykluczenie możliwości, że cechy matki również wywierają istotny wpływ na decyzje dziecka w tym wieku. A co w sytuacjach, gdy matka ma od ojca wyższy poziom wykształcenia – na przykład, matka ma wykształcenie średnie, zaś ojciec podstawowe. Czy o aspiracjach rodziców wobec dziecka decydować będzie i w tym wypadku wykształcenie ojca?

Tablica łączona pozwala w jednej analizie uwzględnić cechy obojga rodziców. W części [1] tabeli 7.14 przedstawiona została zasada jej budowy. Do uprzedniej tablicy, skonstruowanej na podstawie wykształcenia ojca, doklejona została od dołu identyczna tablica krzyżująca wybór szkoły z wykształceniem matki. Tym samym każde z badanych dzieci liczone jest dwukrotnie.

Wyniki analizy korespondencji dla otrzymanej w ten sposób tablicy łączonej zamieszczone zostały w tabeli 7.15. Dla porównania, przedstawiono w niej również wyniki obliczone dla konwencjonalnej tablicy, uwzględniającej jedynie wykształcenie ojca. Rozwiązania w postaci graficznej dla obu tablic przedstawione zostały na rycinie 7.12. Porównanie tych wykresów pozwala zauważyć, że układ punktów reprezentujących drogi edukacyjne jest podobny. Kategorie wykształcenia ojców i matek – na wykresie sporządzonym dla tablicy łączonej – połączyły się zaś w pary. Pary te zajmują pozycje zbliżone do położenia punktów odpowiadających kategoriom ojców w modelu korespondencji sporządzonym na podstawie tablicy uwzględniającej jedynie wykształcenie ojca.

Utworzenie tablicy łącznej nie zmieniło więc obrazu badanego zjawiska, omawianego wielokrotnie we wcześniejszych rozdziałach. Nie zmieniło też jego charakterystyk ilościowych. W obu wypadkach podobna jest bezwładność całkowita, podobny jest udział pierwszego wymiaru w tej bezwładno-

Tabela 7.14
Liczebności i bezwładności dla tablicy łączonej – przedstawiającej wybór szkoły wśród uczniów o różnych poziomach wykształcenia ojców i matek
Badanie PISA 2006

[1] liczebności

	rodzaj szkoły			
	liceum ogólnokształcące	technikum	szkoła zasadnicza	ogółem
wykształcenie ojca				
wyższe	399	45	10	454
średnie	627	454	92	1173
zasadnicze	625	1036	465	2126
podstawowe	95	204	162	461
wykształcenie matki				
wyższe	502	98	8	608
średnie	839	709	179	1727
zasadnicze	330	752	399	1481
podstawowe	75	180	143	398
ogółem	3492	3478	1458	8428

[2] udziały w bezwładności całkowitej (w procentach)

	rodzaj szkoły			
	liceum ogólnokształcące	technikum	szkoła zasadnicza	ogółem
wykształcenie ojca				
wyższe	14,4	6,6	3,6	24,6
średnie	2,5	0,1	3,7	6,3
zasadnicze	4,5	1,7	1,6	7,8
podstawowe	2,9	0,1	5,2	8,2
ogółem (ojcowie)	**24,3**	**8,5**	**14,0**	**46,9**
wykształcenie matki				
wyższe	15,1	5,7	5,5	26,2
średnie	1,3	0,0	2,9	4,2
zasadnicze	8,0	2,0	4,8	14,8
podstawowe	3,0	0,1	4,9	7,9
ogółem (matki)	**27,3**	**7,7**	**18,1**	**53,1**
ogółem (ojcowie i matki)	51,7	16,2	32,1	100,0

ści. Podobne są wielkości pierwszych korelacji kanonicznych, które traktować można jako miarę siły związku. Podobne są wreszcie wartości współrzędnych głównych. Właściwie jedyna korzyść wiąże się z uzyskaniem pewności, że uwzględnienie wykształcenia matki nie zmienia obrazu badanego zjawiska.

Zastanówmy się więc, czy owa korzyść jest warta podjętego trudu utworzenia tablicy łączonej i wykonania dla niej analizy korespondencji. Aby spojrzeć na problem z szerszej perspektywy przydatny okazać się może pewien wtręt historyczny. W latach boomu badawczego, który rozpoczął się po II wojnie

Tabela 7.15
Wyniki analizy korespondencji dla tablicy łączonej – przedstawiającej wybór szkoły wśród uczniów o różnych poziomach wykształcenia ojców i matek oraz dla tablicy konwencjonalnej, prezentującej wybór szkoły w kategoriach wykształcenia ojca
Badanie PISA 2006

[1] korelacje kanoniczne i ich udziały w procentach

	wykształcenie ojca i matki (tablica łączona)			tylko wykształcenie ojca (tablica konwencjonalna)		
wymiar rozwiązania	korelacja	kwadrat korelacji	udział [%]	korelacja	kwadrat korelacji	udział [%]
1	0,4287	0,1837	94,2	0,4124	0,1701	93,0
2	0,1068	0,0114	5,8	0,1132	0,0128	7,0
bezwładność całkowita		0,1951	100,0		0,1829	100,0

[2] współrzędne główne

	wykształcenie ojca i matki		tylko wykształcenie ojca	
	wymiar 1	wymiar 2	wymiar 1	wymiar 2
wykształcenie ojca				
wyższe	-0,9209	0,2068	-0,9276	0,1744
średnie	-0,2796	-0,1000	-0,2759	-0,1097
zasadnicze	0,2435	-0,0353	0,2446	-0,0267
podstawowe	0,4952	0,2135	0,4874	0,2308
wykształcenie matki				
wyższe	-0,8349	0,1108		
średnie	-0,1760	-0,0957		
zasadnicze	0,4051	0,0093		
podstawowe	0,5318	0,2117		
rodzaj szkoły				
liceum ogólnokształcące	-0,4921	0,0330	-0,4776	0,0304
technikum	0,2540	-0,1105	0,2592	-0,1148
zasadnicza zawodowa	0,5728	0,1848	0,5257	0,2011

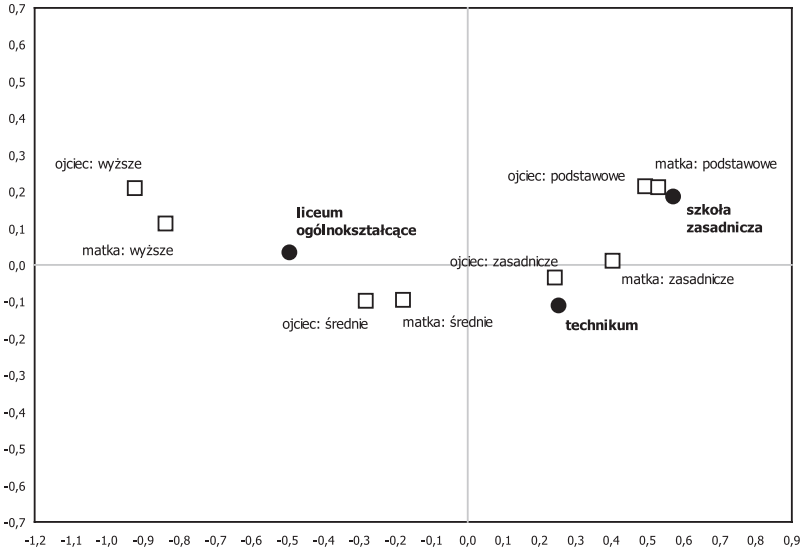
światowej, szczególną popularność zyskały metody analiz wielozmiennowych. Wynikało to z przeświadczenia badaczy, że skoro pomiar w badaniach – zwłaszcza sondażowych – jest niedoskonały, to włączenie do modelu wielu wskaźników tego samego zjawiska pozwala wypowiadać się na jego temat z większą dozą pewności (Jencks i in. 1972: 331–332; Sawiński 1988). Apogeum tego kierunku myślenia przypada na lata siedemdziesiąte, gdy do modeli regresji starano się wprowadzać tyle predyktorów, ile tylko pytań zadano na dany temat. W modelach warunkowania osiągnięć przez pochodzenie do standardów należało uwzględnianie wykształcenia każdego z rodziców z osobna, wskaźników pozycji zawodowej ojca, analogicznych wskaźników dla matki, dochodu rodziców, wyposażenia gospodarstwa domowego w różne dobra w dzieciństwie, liczby książek w domu i tak dalej (Bielby, Hauser i Fetherman 1977; Jencks i in. 1979). Każdy z tych czynników podnosił wartość korelacji między pochodzeniem a osiągniętą pozycją w wymiarze edukacyjnym czy zawodowym. Jednocześnie jednak, z każdym kolejno dołączanym czynnikiem tak zwane „czyste" wpływy poszczególnych predyktorów malały – na skutek istniejących między nimi powiązań. W efekcie uzyskiwano ilościowe oszacowanie łącznego wpływu wszystkich uwzględnionych predyktorów, bez możliwości ustalenia wpływu każdego z nich z osobna. Mówiąc inaczej, bez możliwości identyfikacji mechanizmów, które decydują o istocie zjawiska będącego przedmiotem badania.

Nie bez przyczyny więc począwszy od lat siedemdziesiątych datuje się swoisty renesans tablic jako narzędzia badania zjawisk, który trwa do chwili obecnej. Na początku prace nastawione na odświeżenie tej metody szły w kierunku budowania narzędzi analizy tablic wielowymiarowych (Bishop, Fienberg i Holland 1975; Haberman 1978). Po pewnym czasie przekonano się jednak, że parametry uzyskiwanych rozwiązań są trudne do interpretacji. A już szczególnie beznadziejną sprawą stała się prezentacja uzyskanych tą drogą wyników wobec audytoriów nieobeznanych z logiką metod analizy tablic wielowymiarowych.

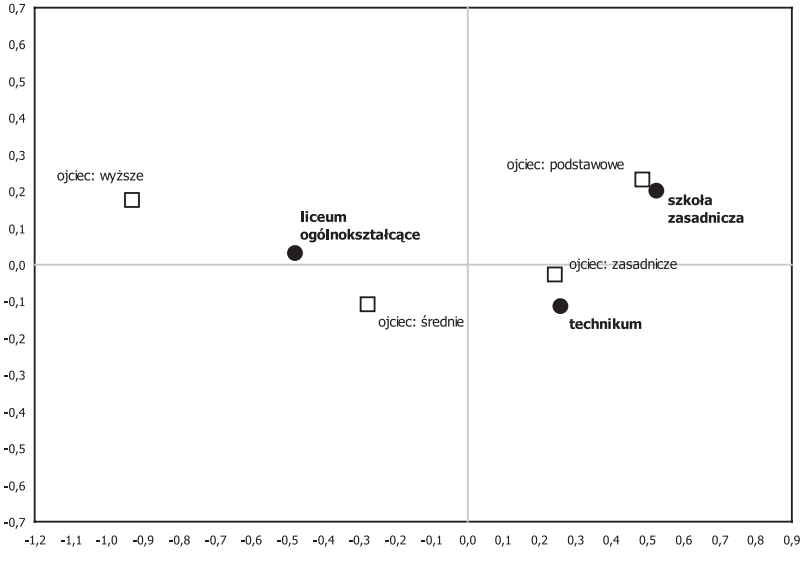
Stąd między innymi wziął się sukces analizy korespondencji. Pozwalała ona przedstawić dane o prostej, tabelarycznej strukturze w równie prosty, graficzny sposób, odwołujący się do wyobraźni i intuicji. Nie rozwiązuje to oczywiście, bo rozwiązać nie może, problemów badania zjawisk w całej ich złożoności. Co więcej, wymaga od badacza znacznego wysiłku i odpowiedzialności związanej z wyborem pojedynczych czynników, **na przykładzie** których wyjaśniona zostanie istota zjawiska. Jeśli więc badacz na podstawie swojej wiedzy i doświadczenia uzna, że wykształcenie ojca pozwoli lepiej czy łatwiej wyjaśnić badane prawidłowości, to zapewne odwoła się do tego czynnika, nie zaś – na przykład – do wykształcenia matki (Domański 2009: 46).

Rycina 7.12
Wykształcenie ojca i matki a wybór szkoły ponadgimnazjalnej. Badanie PISA 2006

[a] Tablica łączona: uwzględniono zarówno wykształcenie ojca, jak i wykształcenie matki



[b] Tablica konwencjonalna: uwzględniono jedynie wykształcenie ojca



Prezentowany przykład zastosowania tablic łączonych uświadamia, że obrazy zjawisk uzyskane przy uwzględnieniu większej liczby czynników nie we wszystkich wypadkach wnoszą nową jakość do konstruowanych wyjaśnień. Wystarczy porównać wykres [a] z wykresem [b] na rycinie 7.12. Oba w gruncie rzeczy odzwierciedlają tę samą prawidłowość, przy czym wykres dla tablicy konwencjonalnej przedstawia ją w prostszy sposób. Nie oznacza to, że tablice łączone nie stanowią użytecznego narzędzia analizy danych w wypadku wielu problemów. Chociażby w sytuacji, gdy opinie czy postawy badamy za pomocą więcej niż jednego wskaźnika. Blasius i Greenacre (2006b: 21–27) zilustrowali to przykładem analizy tożsamości narodowej, którą w badaniu ISSP zoperacjonalizowano za pomocą siedmiu pytań. Odpowiedzi na te pytania umieścili w boczku tablicy, zaś w jej główce pięć czynników, które wyjaśniać mogły udzielane odpowiedzi (płeć, stan cywilny, wiek, wykształcenie i kraj). W rezultacie otrzymali tablicę łączoną składającą się z 35 tablic składowych. Tego rodzaju przykładów wskazać można więcej. Wielu badaczy uważa, że tablice łączone, dzięki możliwości uwzględnienia w jednym modelu więcej niż dwóch cech, stanowią przyszłość analizy korespondencji.

Metoda tablic łączonych nie jest jedynym rozwiązaniem pozwalającym uwzględnić w tablicy więcej niż dwie cechy. Możliwości stwarza również wyodrębnienie jednej bądź obu cech w taki sposób, aby każda z nich już sama w sobie stanowiła złożenie pewnej liczby komponentów zjawiska. Rozpoczynając ten podrozdział, postawiliśmy pytanie, czy uwzględnienie wykształcenia matki – oprócz wykształcenia ojca – zmieni otrzymany wcześniej obraz wpływu domu rodzicielskiego na wybór szkoły ponadgimnazjalnej. Problem ten można przeformułować, wprowadzając konstrukt odpowiadający wykształceniu **obojga rodziców** – stanowiący złożenie wykształcenia ojca i matki. Cecha taka miałaby następujące kategorie: „oboje rodzice mają wykształcenie wyższe", „matka ma wykształcenie wyższe, ojciec średnie", „matka ma wykształcenie średnie, ojciec wyższe" i tak dalej. Pozostając przy czterech wyjściowych poziomach wykształcenia uzyskamy tą drogą 16 kombinacji wykształcenia obojga rodziców.

Tak zdefiniowaną cechę umieścić można w boczku tablicy, a następnie tablicę tę poddać analizie korespondencji. Wyniki przeprowadzonej w ten sposób analizy prezentujemy w tabeli 7.16. Zwraca uwagę fakt, że niektóre kombinacje poziomów wykształcenia wystąpiły w badanej próbie rzadko. Na przykład, tylko w wypadku dwóch uczniów ojciec ma wykształcenie wyższe a matka podstawowe. Podane w przedostatniej kolumnie tabeli wielkości bezwładności świadczą jednak, że specyfika kategorii o niewielkich liczebnościach nie ma znaczącego wpływu na obraz związku. Wielkości te są bowiem nieduże. Pozwala to przejść do analizy wykresu, który przedstawiony został na rycinie 7.13.

Tabela 7.16
Wyniki analizy korespondencji wyboru szkoły ponadgimnazjalnej przez uczniów o różnych kombinacjach wykształcenia ojca i matki
Badanie PISA 2006

[1] korelacje kanoniczne i ich udziały w procentach

wymiar rozwiązania	korelacja	kwadrat korelacji	udział w %	skumulowany udział w %
1	0,4862	0,2364	91,9	91,9
2	0,1449	0,0210	8,1	100,0
w sumie		0,2574	100,0	

[2] liczba i odsetki uczniów o różnych kombinacjach wykształcenia ojca i matki, współrzędne główne oraz bezwładności

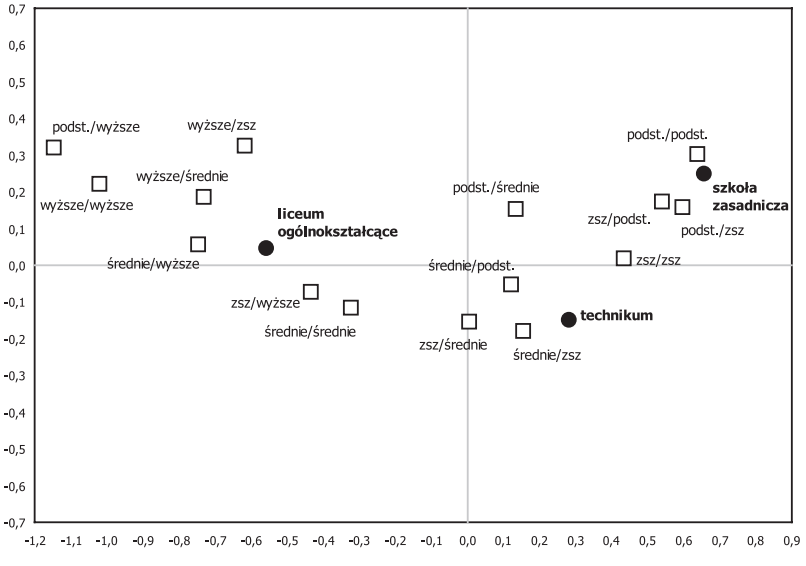
wykształcenie ojca	wykształcenie matki	liczba uczniów	udziały [%]	współrzędne główne wymiar 1	współrzędne główne wymiar 2	bezwładności wielkości	bezwładności udziały [%]
podstawowe	wyższe	4	0,1	-1,1452	0,3194	0,0013	0,5
wyższe	wyższe	304	7,2	-1,0202	0,2209	0,0786	30,5
średnie	wyższe	200	4,7	-0,7448	0,0547	0,0265	10,3
wyższe	średnie	140	3,3	-0,7309	0,1855	0,0189	7,3
wyższe	zasadnicze	8	0,2	-0,6170	0,3242	0,0009	0,4
zasadnicze	wyższe	100	2,4	-0,4328	-0,0732	0,0046	1,8
średnie	średnie	693	16,4	-0,3208	-0,1159	0,0191	7,4
zasadnicze	średnie	794	18,8	0,0070	-0,1554	0,0046	1,8
wyższe	podstawowe	2	0,0	0,1035	1,0187	0,0005	0,2
średnie	podstawowe	43	1,0	0,1232	-0,0531	0,0002	0,1
podstawowe	średnie	100	2,4	0,1359	0,1521	0,0010	0,4
średnie	zasadnicze	237	5,6	0,1562	-0,1797	0,0032	1,2
zasadnicze	zasadnicze	1044	24,8	0,4340	0,0174	0,0467	18,2
zasadnicze	podstawowe	188	4,5	0,5405	0,1737	0,0144	5,6
podstawowe	zasadnicze	192	4,6	0,5984	0,1561	0,0174	6,8
podstawowe	podstawowe	165	3,9	0,6382	0,3022	0,0195	7,6
ogółem		4214	100,0			0,2574	100,0

[3] liczba i odsetki uczniów w poszczególnych rodzajach szkół, współrzędne główne oraz bezwładności

rodzaj szkoły	liczba uczniów	udziały [%]	współrzędne główne wymiar 1	współrzędne główne wymiar 2	bezwładności wielkości	bezwładności udziały [%]
liceum ogólnokształcące	1746	41,4	-0,5568	0,0463	0,1294	50,3
technikum	1739	41,3	0,2834	-0,1508	0,0425	16,5
zasadnicza zawodowa	729	17,3	0,6575	0,2489	0,0855	33,2
ogółem	4214	100,0			0,2574	100,0

Położenie na wykresie poszczególnych rodzajów szkół ponadgimnazjalnych jest w zasadzie analogiczne, jak w wypadku analizy przeprowadzonej za pomocą tabeli łączonej. Podobna jest też konfiguracja kategorii wykształcenia rodziców. Zaletą obecnej prezentacji jest jednak to, że potrafimy odpowiedzieć na pytanie: co dzieje się w sytuacjach, gdy rodzice mają **rozbieżne** poziomy wykształcenia? Okazuje się, że w większości wypadków decyduje poziom wykształcenia tego z rodziców, który jest rezultatem dłuższej nauki. Po lewej stronie wykresu zgromadziły się punkty odpowiadające sytuacjom, gdy przynajmniej jedno z rodziców ma wykształcenie wyższe. Punkty w środku wykresu odpowiadają kategoriom uczniów, których rodzice mają co najwyżej wykształcenie średnie. Natomiast po prawej stronie wykresu leżą punkty reprezentujące rodziny, w których ojciec i matka mają wykształcenie zasadnicze lub podstawowe.

Rycina 7.13
Kombinacje wykształcenia ojca i matki a wybór szkoły ponadgimnazjalnej
Badanie PISA 2006



W oznaczeniach kategorii wykształcenie podano w układzie: wykształcenie ojca / wykształcenie matki.

Omawiany przykład dowodzi, że konwencjonalne tabele, krzyżujące ze sobą tylko dwie cechy, pozwalają budować wartościowe wyjaśnienia również wtedy, gdy w analizie uwzględnić trzeba więcej czynników.

7.11 Tabela czy obraz: podsumowanie

Znaleziska w grocie Lasceaux są dowodem, że ludzie za pomocą obrazów wyrażali swoje myśli i odczucia zanim jeszcze wynaleziono pismo. Przez wiele lat formułowanie komunikatów za pomocą znaków okazywało się jednak bardziej efektywne. Aż nadeszła era współczesna. Komputery i pojemne informacyjnie kanały przekazu sprawiły, że nasza kultura staje się znów kulturą obrazkową.

Nastąpiło to w sumie nie tak dawno. W Polsce *Windowsy* wprowadzone zostały dopiero w latach dziewięćdziesiątych. Wcześniej, jak pewnie pamięta wielu Czytelników, z komputerami komunikowaliśmy się za pomocą komend tekstowych. Aby skopiować plik, należało napisać „Copy", a następnie wpisać lokalizacje i nazwy kopiowanego i docelowego pliku. To strasznie dużo niepotrzebnej roboty. Obecnie wystarczy przenieść myszką ikonkę pliku.

Czy oznacza to zmierzch tablic, jako sposobu analizy i prezentacji danych? Chyba byłoby zbyt pochopnie ferować tego typu wyroki. Z dwóch zasadniczych powodów. Po pierwsze, mimo że graficzna prezentacja zjawisk powstała na długo przed wprowadzeniem pisma, nie przeszła swojej historycznej próby. Zawsze pozostawała w cieniu komunikowania się za pomocą znaków. Po drugie, pojemność informacyjna graficznej prezentacji danych zawsze pozostaje ograniczona, co wiąże się z naturą procesów percepcyjnych. Na ogół nie udaje się nią objąć wszystkich interpretacji, które chcielibyśmy zakomunikować.

Źródeł przekazów graficznych wyjaśniać nie trzeba. Chociażby wspomniana grota Lasceaux zawiera tysiące obrazków, za pomocą których ich twórcy pragnęli utrwalić nie tylko fakty z życia, lecz również zasady organizacji społeczności w której żyli, a także tak ulotne kwestie, jak wyobrażenia czy emocje. Gdy w XVIII wieku o zjawiskach społecznych zaczęto myśleć w kategorii statystycznych prawidłowości, ich istotę, odkrywaną na ogół za pomocą tablic, nie raz próbowano zobrazować graficznie. Szereg rozwiązań w tym zakresie przytacza Falguerolles (2008), którego artykuł poświęcony jest zresztą źródłom inspiracji, z których czerpała szkoła Benzécriego. W XVIII wieku francuski matematyk i filozof Jean Le Rond d'Alembert zaproponował użycie trójkąta równobocznego do przedstawiania powiązań między zjawiskami[10]. Wykorzystywano go później wielokrotnie, między innymi do analizy zjawisk ze sfery polityki czy ekonomii. Żyjący również w XVIII wieku Johann Hein-

[10] Na gruncie analizy korespondencji z propozycji d'Alemberta korzysta się przy wyjaśnianiu zasad interpretacji dystansów od punktu centralnego wykresu (Greenacre 1994: 12–14). Skorzystaliśmy z niej również w przykładzie zamieszczonym w 7.3 (rycina 7.1).

rich Lambert stosował z kolei metodę typograficzną. Polegała ona na eliminowaniu z tablicy niewielkich liczebności aż do momentu, gdy pozostałe ułożą się w znaczący wzór.

W drugiej połowie XIX wieku zapanowała swoista moda na przedstawianie zjawisk społecznych w formie graficznej. W okresie tym wprowadzono stosowane po dzień dzisiejszy mapy, na których różne kolory oznaczają różne natężenie zjawiska w ramach danego podziału terytorialnego. Wprowadzono również szeroko obecnie stosowane diagramy: kołowe, słupkowe i wiele innych. Towarzyszyła temu refleksja teoretyczna, która doprowadziła do wyodrębnienia trzech funkcji, jakie pełnią graficzne prezentacje zjawisk. Pierwszą i najważniejszą jest sama demonstracja zjawiska. Komunikat powinien być prosty i wyrazisty, gdyż adresowany jest do szerokiego audytorium. Druga z funkcji służy kontroli. Graficzna forma pomaga twórcy prezentacji zidentyfikować przypadkowe elementy, które powinny być uśrednione bądź w inny sposób wyeliminowane, gdyż nie dotyczą istoty przedstawianego problemu. I wreszcie trzecią funkcją jest inspiracja. Graficzna forma powinna ułatwiać zrozumienie również i tych aspektów zjawiska, które wyrażają się nie tylko poprzez wielkości liczb, lecz również poprzez relacje między tymi liczbami (Falguerolles 2008: 23–24).

Trzecia, z wyodrębnionych jeszcze w XIX wieku funkcji, wydaje się tym, czego przede wszystkim oczekują badacze od metod wizualizacji danych. Historia pokazała jednak, że w praktyce trudno to osiągnąć. Graficzna prezentacja danych łatwo staje się bowiem źródłem nadużyć. Opisał to ze swoistym wdziękiem Darrell Huff (1954) w znanej książeczce *How to Lie with Statistics*. Falguerolles (2008: 25) formułuje tezę, że ryzyko niewłaściwej interpretacji graficznego obrazu jest konsekwencją swoistego automatyzmu instytucji publikujących dane (na przykład urzędów statystycznych). Sprowadzają one swoją rolę do przekładania na format graficzny liczb, które uzyskano w badaniu. W efekcie publikują obrazy w formie wyjałowionej z elementów charakterystycznych dla poszczególnych zjawisk.

Zdaniem Falquerolles'a, sukces analizy korespondencji polega właśnie na tym, że podejście to wyszło poza doktrynę przekładania liczb na obrazy. Proponowane rozwiązania graficzne nie pretendują do roli jedynych narzędzi interpretacji badanych zjawisk, lecz ogniskują uwagę odbiorcy na tym, co stanowi o ich istocie. Prezentując metodę, wielokrotnie podkreślałem, że szereg istotnych elementów badanego zjawiska nie jest uwidocznionych na wykresie, zaś ich pominięcie prowadzić może do błędnych wniosków. Kształt wykresu nie informuje bowiem ani wielkości badanej próby, ani o sile zależności między cechami. Istnieje ponadto ryzyko zdominowania rozwiązania przez jedną lub kilka kategorii obejmujących niewielką liczbę badanych. Interpretując gra-

ficzny obraz badanego zjawiska, pozostaje więc równolegle śledzić zależności między wielkościami liczbowymi w tablicy.

Wypada pogodzić się z tym, że komunikacja między ludźmi staje się w coraz większym stopniu zdominowana przez obrazy, ikony, zdjęcia, czy filmy. Systematycznie wypierają one formy przekazu tekstowego. Komunikowania wyników badań zmiany te jeszcze nie objęły. Przynajmniej na razie.

Zakończenie

Świat przedstawiony jest w księdze, którą masz przed oczami. Nie możesz jej jednak odczytać, gdy nie znasz języka i nie rozpoznajesz znaków, za pomocą których została napisana

Galileo Galiei[1]

Tę ponadczasową mądrość przytaczam nie po to, aby próbować kogokolwiek przekonać, że stosowanie prezentowanych w książce metod stanowi warunek zrozumienia zjawisk społecznych. Nie zamierzam nikogo stawiać pod ścianą za pomocą parafrazy hasła reklamowego Suzuki: „Albo analiza korespondencji, albo nic". Skłaniałbym się raczej ku stanowisku, które dość przewrotnie ujął George Edward Box, twórca departamentu statystyki na Uniwersytecie Wisconsin-Madison (Box i Draper 1987: 424)

> (...) *all models are approximations. Essentially,* ***all models are wrong, but some are useful*** [podkr. *Z.S.*]

Ludovic Lebart, współtwórca dziedziny nazwanej *text mining*, wyjaśnia to w następujący sposób. *Pure exploration is rare*. Pewna wiedza na temat zjawiska jest bowiem dostępna, zanim jeszcze zdecydujemy się je badać. To, co faktycznie robi się w badaniach, stanowi rodzaj splotu podejścia eksploracyjnego i konfirmacyjnego (Lebart 2008: 3–4). Aby coś odkryć trzeba wiedzieć więcej, niż odczytać można z danych. Więcej, niż uda się wyrazić za pomocą dowolnego modelu.

Współcześni badacze kładą duży nacisk na potrzebę uzyskania *insight*'u, odkrycia prawidłowości, której nikt wcześniej nie zauważył. Stąd biorą się zapewne poszukiwania nowych podejść, które okazać się mają zbawienne dla badaczy nerwowo szukających odkryć. Modły do *data mining* słychać obecnie na każdym kongresie badawczym. Odkrycia nie są jednak dostępne jak jabłka na bazarze. Aby dokonać odkrycia, nie wystarczy posłużyć się odpowiednią metodą (Murtagh 2008: 22).

[1] Cytuję za: Murtagh (2008: 1). Tłumaczenie własne z języka angielskiego.

Książka ta nie uczestniczy w wyścigu szczurów. Jej celem nie jest dostarczenie narzędzi, które mogłyby okazać się ratunkiem dla badaczy nienadążających z interpretowaniem wyników kolejnych badań. Jeżeli już mówić o zaletach prezentowanego podejścia, to należałoby raczej wymienić jego generyczny charakter. Łączy ono w ramach jednej platformy – opartej na analizie kanonicznej – różne sposoby wyciągania wniosków z danych prezentowanych za pomocą tablic. Integruje doświadczenia badaczy próbujących zrozumieć zjawiska: w konserwatywnej Anglii, w oszczędnej Szkocji, w liberalnych Stanach, w ekscentrycznej Francji, w pragmatycznej Kanadzie, w dalekiej Australii, w artystycznej Hiszpanii, w egzotycznej Afryce Południowej, w psychodelicznej Holandii, w poukładanych Niemczech, jak również badaczy zniewalanych przez lata w Związku Radzieckim. Wymieniłem jedynie kraje, których przedstawicieli w książce cytuję, gdyż ich propozycje stworzyły podwaliny prezentowanego podejścia. Lista krajów, w których rozwiązania te stosuje się w praktyce, byłaby z pewnością dłuższa. Trudno bowiem wskazać badacza, który analizując tablice, nie porównuje odsetków bądź nie próbuje ustalić, w których polach obserwowane liczebności są szczególnie wysokie.

Owa codzienna praktyka to najważniejszy z filarów umiejętności wyciągania trafnych wniosków z wyników badań. Wcale nie mniej ważny od poszukiwania nowych metod, czy nawet ich kreowania. Lata praktyki wiążą każdego badacza z pewnymi tylko metodami, a mianowicie tymi, które ocenia jako przydatne, co do których ma przekonanie czy które po prostu dobrze opanował. W książce starałem się wykazać, że powszechnie stosowane i w sumie dość proste metody analizy tablic kryją w sobie głębokie pokłady różnorodnych interpretacji, intuicji czy skojarzeń, po które warto sięgnąć. Na ogół bowiem lepiej sprawdza się doskonalenie tego, co już zna się z wcześniejszych doświadczeń, niż poszukiwanie na gwałt czegoś *extra*, co może pomóc rozwiązać problem. Księgę, o której mówi Galileusz, w istocie rzeczy nosi w sobie każdy badacz.

* * *

Ostatnie słowo to prawo autora każdej książki. Powiem więc na koniec, że mam ciepły, emocjonalny stosunek do większej liczby metod badania zjawisk za pomocą tablic, niż byłem w stanie w tej książce opisać. Mimo że zburzyć to może mozolnie konstruowany holistyczny obraz proponowanego podejścia, o dwóch z tych metod chciałbym powiedzieć przynajmniej po parę słów.

Pierwsza z metod dotyczy modelowania zawartości tablic. Opiera się ona na algorytmie zaproponowanym w latach czterdziestych przez amerykańskiego statystyka Williama Edwardsa Deminga (1900–1993). Sam Deming to ciekawa postać, której w zasadzie powinienem poświęcić osobną ramkę. Mimo że

z wykształcenia był matematykiem, jego nazwisko wiąże się z rozwojem teorii zarządzania. Jest autorem znanej definicji jakości pracy jako ilorazu włożonego wysiłku do poniesionych kosztów. Po wojnie Deming znalazł się w okupowanej przez Amerykanów Japonii, gdzie przez kilka lat zajmował się wprowadzaniem statystycznych standardów jakości w japońskich przedsiębiorstwach. Wzrost produktywności tych przedsiębiorstw okazał się tak duży, że w latach zimnej wojny gospodarka japońska zaczęła zagrażać amerykańskiej. Deming wrócił w związku z tym do Stanów, gdzie swoją wiedzą wspomagać zaczął amerykańskie przedsiębiorstwa, co robił aż do śmierci w wieku 93 lat. Do bardziej spektakularnych sukcesów Deminga zalicza się wprowadzenie zarządzania jakością u Forda, co pozwoliło firmie na początku lat osiemdziesiątych zażegnać poważny kryzys. Z zadłużonego przedsiębiorstwa Ford stał się firmą najlepiej prosperującą na rynku i przez szereg następnych lat stanowił wizytówkę amerykańskiej branży motoryzacyjnej.

Wróćmy jednak do dokonań Deminga w dziedzinie analizy danych. W artykule opublikowanym w 1940 roku przedstawił on algorytm, który pozwalał zmodyfikować zawartość tablicy w taki sposób, aby jej liczebności wewnętrzne sumowały się do z góry ustalonych marginesów (Deming i Stephan 1940; Ireland i Kullback 1968; Fienberg 1970; Deming 1984). Warto podkreślić, że jako kryterium dopasowania przekształconej tablicy do oryginalnych danych Deming zaproponował minimalizację wskaźnika chi-kwadrat, z którego to kryterium wielokrotnie korzystaliśmy w metodach opisanych w książce. Pod względem analitycznym metodę Deminga bez wątpienia można zaliczyć do proponowanego podejścia. W swojej warstwie koncepcyjnej wybiega ona jednak poza to, co w książce prezentuję. Nie dotyczy bowiem zjawiska w takiej postaci, w jakiej stanowi ono przedmiot badania, lecz pewnej jego wizji, która mogłaby zaistnieć, gdyby marginesy tablicy – czyli ramy zjawiska – miały inny kształt.

Wydaje się, że podejście takie stanowić może atrakcyjną płaszczyznę analizy i rozumienia badanych zjawisk. Przykładowo, stwarza możliwość pokazania, jak mogłaby wyglądać homogamia małżeństwa w społeczeństwie, w którym rozkłady wykształcenia mężczyzn i kobiet byłyby identyczne. Obecnie, jak wiadomo, kobiety są lepiej wykształcone niż mężczyźni, stąd część kobiet jest niejako „zmuszona" wyjść za mąż za mężczyznę o niższym wykształceniu (Domański i Przybysz 2007, 2009). Można też dzięki metodzie Deminga pokazać, jak wyglądałyby osiągnięcia zawodowe absolwentów wyższych uczelni, gdyby struktura miejsc pracy byłaby w stanie wessać wszystkich, którzy zdecydowali się zdobyć wykształcenie wyższe. Niedopasowanie obu „marginesów" postrzega się jako jedną z kluczowych barier efektywnego funkcjonowania współczesnych społeczeństw (Arum, Gamoran i Shavit 2007; Sawiński 2009). Mimo że tego rodzaju hipotetyczne modele mogłyby stanowić dogodną

platformę analizy wielu zjawisk, w praktyce nie stosuje się ich. Może poza nielicznymi wyjątkami, które w historii rozwoju metod zawsze się znajdą (zob. Mosteller 1968: 8–11).

Mimo braku zastosowań metody Deminga do tworzenia hipotetycznych modeli zjawisk, znalazła ona praktyczne zastosowanie podczas ważenia wyników badań. Na niej opiera się bowiem tak zwane **ważenie wieńcowe** (ang. *rim weighting*).

W konwencjonalnej metodzie ważenia wielkości wag ustalane są poprzez podzielenie wielkości populacyjnych przez liczebności otrzymane w badaniu. Przestaje to być możliwe w sytuacji, gdy niektóre z otrzymanych liczebności są zerowe ze względu na to, że w badanej próbie nie ma ani jednego respondenta o danej kombinacji cech. Ważenie wieńcowe pozwala wyznaczyć wielkości wag w taki sposób, aby marginesy populacyjne zostały zachowane mimo pustych niektórych pól w tablicach krzyżujących ze sobą cechy uwzględnione podczas ważenia (Sawiński 2007b). Dzięki tej własności ważenie wieńcowe zyskało niezwykłą popularność. Szczególnie w czasach, gdy *response rate* obniża się gwałtownie (Stoop 2005). Ważenie wieńcowe stosuje się przede wszystkim w badaniach marketingowych, w tym w największych projektach dotyczących konsumpcji mediów. Na przykład, według tej metody synchronizuje się wielkości poszczególnych kategorii widowni w badaniach telemetrycznych. Konieczność udostępniania wyników w cyklu dziennym, której towarzyszy ryzyko nieściągnięcia danych z części gospodarstw z przyczyn technicznych powodują, że metodę Deminga w badaniach telemetrycznych stosuje się powszechnie.

Druga metoda, o której chciałbym w zakończeniu książki wspomnieć, również nawiązuje do marginesów tablicy. Jej genezy doszukać się można w ustaleniach badaczy zajmujących się międzypokoleniową ruchliwością zawodową, dokonanych jeszcze w latach 50. i 60 (Sawiński 1981). Zauważono wtedy, że liczebności we wnętrzu tablicy nie mogą rozkładać się w sposób dowolny właśnie ze względu na niejednakowe marginesy obu cech. Nazwano to efektem **niedopasowania struktur**. Na przykład, wszyscy synowie rolników nie mogą zostać rolnikami, gdyż udział rolników w strukturze zawodowej maleje z pokolenia na pokolenie. Część synów rolników podlega więc ruchliwości wymuszonej. Przemieszczenia takie przez wiele lat postrzegano jako awans społeczny. Po części nie są one jednak wynikiem wrodzonych predyspozycji czy uzyskania kwalifikacji pozwalających zająć wyższą pozycję społeczną, lecz po prostu odzwierciedlają fakt, że dla kolejnego pokolenia nie ma odpowiedniej liczby miejsc pracy w rolnictwie.

Niedopasowanie struktur prowadzi do pytania: jak powinny układać się liczebności w tablicy, aby maksymalnie zapewnić zgodność między zasoba-

mi – opisanymi w wierszach tablicy – a osiągnięciami jednostek, które przedstawiane są w kolumnach. W tablicy ruchliwości miedzypokoleniowej byłaby to zgodność między pochodzeniem – operacjonalizowanym na ogół poprzez zawód ojca – a osiągniętą pozycją społeczną, wyrażoną poprzez zawód badanego. W ten właśnie sposób sformułował problem Tadeusz Krauze z Hofstra University w Nowym Jorku, nazywając ową maksymalną zgodność **alokacją merytokratyczną**. W semestrze zimowym 1982/83 profesor Krauze wygłosił na Uniwersytecie Warszawskim cykl wykładów, których przedmiotem była owa koncepcja. Na podstawie tych wykładów napisałem niewielką książeczkę (Sawiński 1984), w której przedstawiam założenia i korzenie koncepcji, opisuję sposób utworzenia alokacji merytokratycznej na podstawie marginesów tablicy, a także omawiam związki koncepcji z innymi metodami analizy tablic. Prezentację koncepcji można również znaleźć w publikacjach poświęconych jej zastosowaniom do analizy zjawisk społecznych – głównie ze sfery sprawiedliwej dystrybucji dóbr (Wesołowski i Krauze 1981; Słomczyński 1983, 1989; Krauze i Słomczyński 1985; Słomczyński, Krauze i Peradzyński 1988; Wang 2002; Kunovich i Słomczyński 2007).

Zarówno koncepcja dopasowania zawartości tablicy do z góry ustalonych marginesów, jak też koncepcja alokacji merytokratycznej mają pewną cechę wspólną. Mianowicie postulowane w nich modele są hipotetycznymi modelami badanych zjawisk. Modelami, które w rzeczywistości empirycznej nie zdarzają się i służą wyłącznie jako punkt odniesienia do analizowania tablic uzyskanych w wyniku badań. Jest to podobna funkcja, jaką pełni model niezależności, zwany też w socjologii modelem równych szans.

Badacze jak ognia unikają jednak rozpatrywania badanych zjawisk poprzez pryzmat nierzeczywistych modeli. Pisałem o tym w rozdziale 3, zwracając uwagę na fakt, że nawet model niezależności – stanowiący *de facto* matematyczną podstawę większości metod analizy tablic – nie jest uznawany przez wielu badaczy za dogodną platformę analizy realnych zjawisk. Badacze wolą stosować modele, które możliwie dokładnie wyjaśniają czy odtwarzają liczebności w tablicy otrzymanej w wyniku badania. Można zadać pytanie, czy sami sobie nie zawężają w ten sposób zakresu możliwych wniosków? Czy wyjście poza paradygmat analizy danych w postaci otrzymanej w badaniu nie prowadziłoby w wielu wypadkach do nowego, świeżego spojrzenia na badane zjawisko? Czy nie sprzyjałoby odkrywaniu tak pożądanych *insight*'ów?

W książce nie podjąłem się udzielenia odpowiedzi na te pytania. Jeśli jednak zachęci ona Czytelnika do własnych poszukiwań, to będę miał satysfakcję, że spełniła swoje cele.

Warszawa, 15 grudnia 2009

ANEKS A

Dowody wybranych własności

W aneksie zamieszczone zostały jedynie dowody tych własności rozpatrywanych modeli, których dowody trudno wskazać w literaturze przedmiotu.

A.1 Dystans chi-kwadrat między profilami dwóch wierszy (kolumn) w modelu kanonicznym jest proporcjonalny do różnicy współrzędnych kanonicznych

Dowód przeprowadzę tylko dla przypadku dwóch wierszy, gdyż dla kolumn jest analogiczny. Należy dowieść, że

$$\delta_{i_1 i_2} = r^{(s)} * \left| x_{i_1}^{(s)} - x_{i_2}^{(s)} \right| \tag{A.1}$$

gdzie $\delta_{i_1 i_2}$ oznacza dystans chi-kwadrat między profilami wierszy i_1 oraz i_2 dany wzorem (6.1), $x_{i_1}^{(s)}$ oraz $x_{i_2}^{(s)}$ są współrzędnymi kanonicznymi s-tego zestawu, zaś $r^{(s)}$ jest s-tą korelacją kanoniczną. Porównywane profile liczone są w obrębie tablicy liczebności odtworzonych (modelu kanonicznego) $M^{(s)}$ utworzonej na podstawie tablicy kanonicznej $C^{(s)}$. W zapisie dowodu przyjmę następujące uproszczenia: pominę nadskrypty s określające, o które rozwiązanie kanoniczne chodzi oraz zastąpię i_1 indeksem 1, zaś i_2 indeksem 2.

Wielkość szacowanego dystansu wynosi więc

$$\delta_{12} = \sqrt{\sum_{j=1}^{k} \frac{\left(\frac{m_{1j}}{a_1} - \frac{m_{2j}}{a_2} \right)^2}{\frac{b_j}{n}}} \tag{A.2}$$

Na początku przekształćmy wyrażenie w nawiasach

$$\left(\frac{m_{1j}}{a_1} - \frac{m_{2j}}{a_2} \right) = \frac{1}{a_1 a_2} \left(a_2 m_{1j} - a_1 m_{2j} \right)$$

Podstawmy do niego wielkości m_{1j} i m_{2j} korzystając z wzorów (6.11) i (6.8)

$$= \frac{1}{a_1 a_2}\left[a_2 e_{1j}\left(1+rx_1y_j\right) - a_1 e_{2j}\left(1+rx_2y_j\right)\right]$$

następnie podstawmy $e_{ij} = a_i b_j / n$

$$= \frac{1}{a_1 a_2}\left[\frac{a_1 a_2 b_j}{n}\left(1+rx_1y_j\right) - \frac{a_1 a_2 b_j}{n}\left(1+rx_2y_j\right)\right]$$

$$= \frac{b_j}{n}\left[\left(1+rx_1y_j\right) - \left(1+rx_2y_j\right)\right]$$

$$= \frac{r}{n}\left(x_1 - x_2\right) b_j\, y_j \tag{A.3}$$

Wynik podstawmy do (A.2), dzieląc jednocześnie licznik i mianownik przez b_j / n

$$\delta_{12} = \sqrt{\sum_{j=1}^{k} r^2 \left(x_1 - x_2\right)^2 \frac{b_j}{n} y_j^2} \tag{A.4}$$

Przenosząc przed znak sumy czynniki niezależne od j, otrzymamy

$$\delta_{12} = \sqrt{r^2 \left(x_1 - x_2\right)^2 \sum_{j=1}^{k} \frac{b_j}{n} y_j^2} \tag{A.5}$$

Wyrażenie objęte znakiem sumy jest zgodnie z wzorem (6.21) równe 1. Wyciągając pierwiastek z pozostałego wyrażenia otrzymujemy szukaną wielkość

$$\delta_{12} = r\left|x_1 - x_2\right| \tag{A.6}$$

■

A.2 Korelacja w modelu kanonicznym $M^{(s)}$ jest równa korelacji $r^{(s)}$ w tablicy liczebności obserwowanych

Zgodnie z (6.28) liczebności tablicy $M^{(s)}$ wyrażają się jako

$$m_{ij}^{(s)} = e_{ij} + c_{ij}^{(s)} \tag{A.7}$$

W dalszych przekształceniach pominę nadskrypty s oznaczające, o którą tablicę i korelację kanoniczną chodzi. Tablica M ma te same marginesy, co tablica liczebności obserwowanych N. W jej wypadku współrzędne kanoniczne spełniają więc warunki (6.18)–(6.21). Oznaczmy przez $r(M)$ wartość obliczanego

współczynnika korelacji dla tablicy *M*. Podstawiając wyrażenie (A.7) do wzoru na współczynnik korelacji (6.22) otrzymamy

$$r(M) = \frac{1}{n}\sum_{i=1}^{w}\sum_{j-1}^{k} x_i y_j (e_{ij} + c_{ij})$$

$$= \frac{1}{n}\left(\sum_{i=1}^{w}\sum_{j=1}^{k} x_i y_j e_{ij} + \sum_{i-1}^{w}\sum_{j=1}^{k} x_i y_j c_{ij}\right) \tag{A.8}$$

podstawiając $c_{ij} = re_{ij}x_i y_j$ (wzór 6.x), a następnie $e_{ij} = a_i b_j / n$ (wzór 3.10), otrzymamy

$$= \frac{1}{n}\sum_{i=1}^{w} a_i x_i \left[\sum_{j=1}^{k}\frac{b_j}{n} y_j\right] + r\left\{\sum_{i=1}^{w}\frac{a_i}{n}x_i^2\right\}\left\{\sum_{j=1}^{k}\frac{b_j}{n}y_j^2\right\} \tag{A.9}$$

Wyrażenie w nawiasie kwadratowym odpowiada średniej współrzędnych kanonicznych cechy *Y*, a więc zgodnie z przyjętymi założeniami (6.20) jest równe 0. Tym samym pierwszy składnik sumy w (A.9) zeruje się. Z kolei wyrażenia w nawiasach klamrowych drugiego składnika oznaczają wariancje współrzędnych kanonicznych obu zmiennych, czyli zgodnie z (6.19) i (6.21) każde z tych wyrażeń jest równe 1. W rezultacie

$$r(M) = r \tag{A.10}$$

czyli jest równe wartości korelacji kanonicznej dla tablicy *N*, co kończy dowód.

■

A.3 Suma kwadratów wskaźników Quételeta w dowolnym wierszu bądź kolumnie tablicy *N* jest sumą analogicznych wielkości obliczoną po wszystkich modelach kanonicznych $M^{(s)}$

Mamy udowodnić, że dla każdego $i = 1, 2, ..., w$

$$Q_{i\bullet}(N) = \sum_{s=1}^{S} Q_{i\bullet}(M^{(s)}) \tag{A.11}$$

gdzie *S* jest liczbą modeli kanonicznych określoną wzorem (6.17). Analogicznie, dla każdego $j = 1, 2, ..., k$

$$Q_{\bullet j}(N) = \sum_{s=1}^{S} Q_{\bullet j}(M^{(s)}) \tag{A.12}$$

Dowód przeprowadzimy dla sumy w wierszu *i*. Dowód dla sumowania w kolumnach jest analogiczny. Dla dowolnego modelu kanonicznego mamy

$$Q_{i\bullet}(M) = \sum_{j=1}^{k} \frac{c_{ij}^2}{a_i b_j} \tag{A.13}$$

Zastępując c_{ij} wyrażeniem (6.8) a następnie podstawiając $e_{ij} = a_i b_j / n$ otrzymamy

$$Q_{i\bullet}(M) = \sum_{j=1}^{k} \frac{\left(r a_i b_j x_i y_j\right)^2}{n^2 a_i b_j} \tag{A.14}$$

Następnie wyprowadźmy stałe oraz czynniki zależne tylko od *i* przed znak sumowania po *j*

$$Q_{i\bullet}(M) = r^2 \frac{a_i}{n} x_i^2 \left(\sum_{j=1}^{k} \frac{b_j}{n} y_j^2 \right) \tag{A.15}$$

Wyrażenie w nawiasach na mocy (6.21) równe jest 1. Formuła upraszcza się więc do postaci

$$Q_{i\bullet}(M) = r^2 \frac{a_i}{n} x_i^2 \tag{A.16}$$

Należy zauważyć, że wyrażenie po prawej stronie znaku równości jest zawsze dodatnie. Zsumujmy te wyrażenia po wszystkich modelach kanonicznych

$$Q_{i\bullet}(N) = \sum_{s=1}^{S} Q_{i\bullet}(M^{(s)}) = \frac{a_i}{n} \sum_{s=1}^{S} \left(r^{(s)}\right)^2 \left(x_i^{(s)}\right)^2 \tag{A.17}$$

a następnie zsumujmy po wszystkich wierszach tablicy *N*

$$\sum_{i=1}^{w} \frac{a_i}{n} \sum_{s=1}^{S} \left(r^{(s)}\right)^2 \left(x_i^{(s)}\right)^2 \tag{A.18}$$

w formule (A.18) zmieńmy kolejność sumowania

$$= \sum_{s=1}^{S} \left(r^{(s)}\right)^2 \left[\sum_{i=1}^{w} \frac{a_i}{n} \left(x_i^{(s)}\right)^2 \right] \tag{A.19}$$

Wyrażenie w nawiasach kwadratowych na mocy (6.19) jest równe 1 dla każdego *s*. W rezultacie dostajemy sumę kwadratów korelacji kanonicznych

$$= \sum_{s=1}^{S} \left(r^{(s)}\right)^2 = Q(N) \tag{A.20}$$

która na mocy (6.30) równa jest sumie kwadratów wskaźników Quételeta dla tablicy *N*. Ze względu na fakt, że w (A.18) sumowaliśmy dodatnie wielkości, to dla każdego $i = 1, 2,..., w$ spełnione jest (A.11)

■

A.4 Suma kwadratów wskaźników Quételeta w modelu kanonicznym $M^{(s)}$ jest równa kwadratowi korelacji kanonicznej $r^{(s)}$

Mamy udowodnić, że

$$Q(M^{(s)}) = \left(r^{(s)}\right)^2 \tag{A.21}$$

gdzie *Q* zgodnie z oznaczeniami przyjętymi w podrozdziale 6.7 oznacza sumę kwadratów wskaźników Quételeta (wzór 4.9) po wszystkich polach tablicy.

$$Q(M) = \sum_{i=1}^{w} \sum_{j=1}^{k} \frac{d_{ij}^{\ 2}}{a_i b_j} \tag{A.22}$$

W dowodzie pominiemy nadskrypt *s* oznaczający, o który model kanoniczny chodzi. Zgodnie z (6.28) liczebności modelu kanonicznego *M* wyrażają się jako

$$m_{ij} = e_{ij} + c_{ij} \tag{A.23}$$

więc różnica d_{ij} w (A.22) jest równa

$$d_{ij} = m_{ij} - e_{ij} = c_{ij} \tag{A.24}$$

Podstawmy tę wielkość do (A.22)

$$Q(M) = \sum_{i=1}^{w} \sum_{j=1}^{k} \frac{c_{ij}^2}{a_i b_j} \tag{A.25}$$

Zastępując c_{ij} wyrażeniem (6.8) a następnie podstawiając $e_{ij} = a_i b_j / n$, otrzymamy

$$Q(M) = \sum_{i=1}^{w} \sum_{j=1}^{k} \frac{\left(r a_i b_j x_i y_j\right)^2}{n^2 a_i b_j} = \sum_{i=1}^{w} \sum_{j=1}^{k} r^2 \frac{a_i}{n} x_i^2 \frac{b_j}{n} y_j^2 \tag{A.26}$$

Wyprowadźmy czynniki zależne tylko od *i* przed znak sumowania po *j*

$$Q(M) = r^2 \sum_{i=1}^{w} \frac{a_i}{n} x_i^2 \left(\sum_{j=1}^{k} \frac{b_j}{n} y_j^2 \right) \tag{A.27}$$

Wyrażenie w nawiasach zgodnie z (6.21) równe jest 1. Po jego wyeliminowaniu wyrażenie pod znakiem sumy po *i* staje się równe 1 (na mocy 6.19). W rezultacie

$$Q(M) = r^2 \tag{A.28}$$

■

ANEKS B

Sposób wykonania obliczeń

Niektóre z proponowanych w tej książce modeli i wskaźników wykraczają poza zestaw najczęściej stosowanych narzędzi analizy tablic. Dlatego w aneksie tym przedstawiam wskazówki dotyczące wykonania odpowiednich obliczeń. Dotyczą one podanych niżej modeli oraz wskaźników.

rodzaj wskaźnika lub modelu	podrozdziały, w których omówiona została budowa lub interpretacja wskaźnika lub modelu
wskaźniki Quételeta	4.5–4.8, 6.5.4, 6.7
model dekompozycji chi-kwadrat	4.7–4.8, 6.5.4–6.5.5, 6.7
dopasowane średnie	5.4–5.6, 6.5.2, 6.7
dystanse chi-kwadrat między profilami	6.2, 6.5.1, 6.7
korelacje kanoniczne	6.4, 6.5.3, 6.7
współrzędne kanoniczne	6.3, 6.5.2, 6.7, 7.6
modele kanoniczne (na bazie tablicach kanonicznych)	6.3–6.7
bezwładność i jej dekompozycja (równoważna dekompozycji chi-kwadrat)	7.2–7.4 oraz fragmenty opisujące zasady dekompozycji chi-kwadrat
model korespondencji	7.2–7.11

Większość z wymienionych wielkości wymaga znalezienia współrzędnych i korelacji kanonicznych. Dlatego ograniczę się do przykładów programów komputerowych, które pozwalają obliczyć te wielkości. Rozpocznę od programu LEM, który dostępny jest bezpłatnie w Internecie (część B.1). Program ten wykorzystany był w rozdziale 4 do oceny stopnia dopasowania do danych modelu zależności w tablicy. Współrzędne kanoniczne można również uzyskać za pomocą odpowiednich procedur komercyjnych pakietów statystycznych, co przedstawię na przykładzie pakietu STATA (część B.3). Czytelników zainteresowanych możliwościami skorzystania w tym zakresie z pakietu SPSS odsyłam do artykułu Jarosława Górniaka (2000). Mając współrzędne kanoniczne, obliczyć można wszystkie pozostałe parametry modeli wymienionych

w powyższym zestawieniu. Przy czym niektóre z tych parametrów dostępne są w ramach poszczególnych programów, niektóre zaś wymagają odrębnego obliczenia – co omawiam w części B.2.

Sposoby wykonania niezbędnych obliczeń zilustruję przykładem danych dotyczących związku między wykształceniem ojca a rodzajem szkoły ponadgimnazjalnej, w której uczy się dziecko. Dane te stanowią ilustrację największej liczby metod prezentowanych w książce, dlatego w tym miejscu również się nimi posłużę. Uzyskane w badaniu liczebności wykorzystanej tablicy przedstawione zostały w części [1] tabeli 5.1.

B.1 Program LEM

Program LEM autorstwa Jeroena K. Vermunta dostępny jest na portalu uniwersytetu w Tilburgu http://www.uvt.nl/faculteiten/fsw/organisatie/departementen/mto/ software2.html, skąd można go skopiować wraz z dokumentacją (Vermunt 1997). Sam program umieszczony jest w module LEM95.exe. Ponieważ pierwotnie przeznaczony był do uruchamiania w systemie operacyjnym DOS, został uzupełniony o nakładkę LEMWIN.EXE pozwalającą na pracę w środowisku Windows. Oba wymienione moduły najlepiej umieścić w folderze utworzonym bezpośrednio w katalogu głównym dysku (ang. *root*

Rycina B.1
Plik poleceń programu LEM wywołujących procedurę analizy korespondencji tablicy opisującej związek wykształcenia ojca z rodzajem szkoły ponadgimnazjalnej, do której uczęszcza dziecko

```
* Wykształcenie ojca a rodzaj szkoły ponadgimnazjalnej,
* w której uczy się dziecko
* PISA 2006

* Analiza korespondencji

man 2
dim 4 3
lab WO SZK
mod cor(1)
dat
[399    45    10
 627   454    92
 625  1036   465
 95    204   162]
```

directory), przy czym nazwa folderu nie powinna być dłuższa niż 8 znaków i powinna zawierać wyłącznie litery i cyfry – zgodnie z zasadami tworzenia nazw w systemie DOS. Nie zastosowanie się do tego wymogu spowodować może błędy w działaniu programu.

Program LEM przeznaczony jest przede wszystkim do testowania modeli log-liniowych. Możliwości jego wykorzystania w tym zakresie omawiają Domański i Przybysz (2007: 231–242) w pracy, która ukazała się jako tom 1 w ramach tej samej serii wydawniczej, co obecna książka. Program LEM ma również zaimplementowaną procedurę obliczania parametrów modelu analizy korespondencji (Vermunt 1997: 94). Jej uruchomienie wymaga przygotowania pliku poleceń (ang. *job*), którego zawartość przedstawiona została na rycinie B.1.

Plik poleceń najlepiej utworzyć za pomocą edytora plików w formacie tekstowym. Format ten nie zawiera żadnych dodatkowych znaków sterujących poza kombinacjami symboli oznaczających końce poszczególnych wierszy. Przykładem takiego edytora w systemie Windows jest program o nazwie *Notatnik*. Program nie interpretuje zawartości wierszy rozpoczynających się od symbolu „*" (gwiazdki), co daje możliwość umieszczenia komentarzy za tym znakiem. Warto w tym miejscu podać informacje o rodzaju przetwarzanych danych, gdyż wszystkie wiersze pliku poleceń przepisywane są do pliku wyników. Pozostałe wiersze są interpretowane przez program, toteż ich zawartość musi być ściśle zgodna z wymaganą składnią. W rozpatrywanym przykładzie znaczenie poszczególnych poleceń jest następujące

polecenie	*opis składni*
`man 2`	określa liczbę wymiarów tablicy. W wypadku tablic krzyżujących ze sobą dwie cechy wartość parametru polecenia zawsze jest równa 2;
`dim 4 3`	po identyfikatorze komendy 'dim' należy podać liczbę wierszy oraz liczbę kolumn tablicy;
`lab WO SZK`	polecenie opcjonalne, pozwalające nadać określenia obu cechom. Określenia muszą być krótkie (do 3 znaków) i nie mogą zawierać wewnętrznych spacji. Jeśli określenia nie zostaną podane, to program oznaczy cechę w wierszach literą A, zaś cechę w kolumnach literą B;
`mod cor(1,2)`	polecenie wywołuje procedurę estymacji modelu analizy korespondencji. Pierwszym parametrem w nawiasie jest zawsze 1, gdyż parametr ten oznacza rodzaj modelu. Drugi parametr określa zaś liczbę wyodrębnianych wymiarów kanonicznych. W modelu korespondencji liczbą tą jest na ogół 2, aczkolwiek w niektórych sytuacjach zasadne jest wyodrębnienie większej liczby wymiarów. Tę ostatnią sytuację ilustruje przykład analizy cytatów omawiany w podrozdziale 7.7;

`dat` polecenie określa, że począwszy od tego miejsca zostaną podane liczebności tablicy. Zestaw liczebności zawsze rozpoczyna się od symbolu [(otwierający nawias kwadratowy), zaś kończy symbolem] (zamykający nawias kwadratowy). Między tymi symbolami należy wyszczególnić wszystkie liczebności wnętrza tablicy w kolejności według wierszy. Poszczególne liczebności należy oddzielić od siebie jedną lub większą liczbą spacji. Liczebności dogodnie jest podzielić na wiersze w analogiczny sposób, jak występują w oryginalnej tablicy (por. tabela 5.1, część [1]), aczkolwiek nie jest to wymagane. Program czyta zawsze tyle liczb, ile wynosi iloczyn zadeklarowanej liczby wierszy i liczby kolumn. Podając liczebności w postaci liczb dziesiętnych (na przykład liczebności ważone) należy jako separator oddzielający część dziesiętną od całkowitej stosować kropkę, nie zaś przecinek.

Nazwie pliku poleceń warto nadać rozszerzenie „inp", gdyż takiego rozszerzenia oczekuje program dla plików poleceń. Po uruchomieniu programu LEM plik ten można wczytać wybierając z menu głównego komendy „File" oraz „Open". Polecenia można też wpisać bezpośrednio z klawiatury w oknie programu oznaczonym jako „Input".

Program uruchamia się wybierając z menu głównego komendy „File" oraz „Run", bądź też wciskając kombinację klawiszy <Ctrl>+<R>. Wyniki obliczeń wyświetlone zostają w oknie „Output", z którego mogą być przekopiowane do dowolnego pliku bądź zapisane jako plik tekstowy z rozszerzeniem „out". Wyniki dla rozpatrywanego przykładu przedstawione zostały na rycinie B.2.

Plik wynikowy rozpoczyna się od nagłówka, który zawiera informacje o programie LEM. W części oznaczonej jako ***INPUT*** wypisane zostają wszystkie wiersze pliku poleceń. Wyniki analizy zawiera kolejna część pliku, rozpoczynająca się od napisu ***CORRESPONDENCE OR CANONICAL CORRELATION ANALYSIS***. W pierwszej sekcji – oznaczonej jako *General* – podane są ogólne charakterystyki wyodrębnionych wymiarów kanonicznych. W rubryce opatrzonej nagłówkiem „rho" wypisane są wielkości korelacji kanonicznych, w rubryce „inertia" kwadraty korelacji kanonicznych – równoważne wartościom bezwładności dla pierwszego i kolejnych wymiarów. Kolejne rubryki przedstawiają udziały bezwładności obu wymiarów w bezwładności całkowitej (rubryka „prop.") oraz udziały w postaci skumulowanej (rubryka „cum.prop."). Dla rozpatrywanego przykładu interpretacje pierwszej i drugiej korelacji kanonicznej przedstawione zostały w rozdziale 6, przy okazji omawiania modeli kanonicznych, zaś ich wielkości podano w tabelach 6.3 i 6.8. Wielkości te omawiane były raz jeszcze w podrozdziale 7.10 jako wyniki analizy korespondencji (tabela 7.15).

Rycina B.2
Zawartość pliku wynikowego analizy korespondencji uzyskanego za pomocą programu LEM

```
LEM: log-linear and event history analysis with missing data.
Developed by Jeroen Vermunt (c), Tilburg University,
The Netherlands.
Version 1.0 (September 18, 1997).

*** INPUT ***
  * Wykształcenie ojca a rodzaj szkoły ponadgimnazjalnej,
  * w której uczy się dziecko
  * PISA 2006
  * Analiza korespondencji
  man 2
  dim 4 3
  lab WO SZK
  mod cor(1,2)
  dat
  [399    45    10
   627   454    92
   625  1036   465
    95   204   162]

*** CORRESPONDENCE OR CANONICAL CORRELATION ANALYSIS ***
 * General *
   dim     rho     inertia    prop.   cum.prop.   X-squared
    1    0.4124    0.1701     0.930     0.930      770.60
    2    0.1132    0.0128     0.070     1.000       54.01

 * Categories *
  Dimension 1
           weight      x     x*sqrt(rho)   x*rho   contribution
  WO  1    0.108   -2.249     -1.444      -0.928      0.545
  WO  2    0.278   -0.669     -0.430      -0.276      0.125
  WO  3    0.505    0.593      0.381       0.245      0.177
  WO  4    0.109    1.182      0.759       0.487      0.153
 SZK  1    0.414   -1.158     -0.744      -0.478      0.556
 SZK  2    0.413    0.628      0.404       0.259      0.163
 SZK  3    0.173    1.275      0.819       0.526      0.281
  Dimension 2
           weight      x     x*sqrt(rho)   x*rho   contribution
  WO  1    0.108   -1.541     -0.518      -0.174      0.256
  WO  2    0.278    0.969      0.326       0.110      0.262
  WO  3    0.505    0.236      0.079       0.027      0.028
  WO  4    0.109   -2.038     -0.686      -0.231      0.455
 SZK  1    0.414   -0.268     -0.090      -0.030      0.030
 SZK  2    0.413    1.014      0.341       0.115      0.424
 SZK  3    0.173   -1.776     -0.598      -0.201      0.546
```

Dane dotyczą zależności między wykształceniem ojca a rodzajem szkoły ponadgimnazjalnej, do której uczęszcza dziecko. Badanie PISA 2006.

Ostatnia rubryka – oznaczona jako „X-squared" – zawiera wielkości statystyki chi-kwadrat. Wartość dla pierwszego wymiaru odpowiada tablicy danych empirycznych, czyli jest równa wielkości, którą obliczalibyśmy, testując hipotezę, że cechy są niezależne w populacji, posługując się metodą opisaną w podrozdziale 3.10. Natomiast druga z wielkości – równa 54,01 – odpowiada modelowi kanonicznemu skonstruowanemu na bazie drugiej tablicy kanonicznej (podrozdział 6.6). Łatwo zauważyć, że autor programu dopuścił się tu pewnej niekonsekwencji, gdyż w wierszu dla pierwszego wymiaru należałoby raczej umieścić chi-kwadrat odpowiadający modelowi opartemu na pierwszej tablicy kanonicznej. Owa wielkość, której interpretację podaję w podrozdziale 6.5.5, wynosi 716,6. O niekonsekwencji tej należy pamiętać, gdy wyniki programu LEM mają służyć do określenia dekompozycji wskaźnika chi-kwadrat między poszczególne modele kanoniczne. Wielkość podaną w pierwszym wierszu należałoby wtedy przemnożyć przez proporcje podane w rubryce „prop.". W rozpatrywanym przykładzie wartość chi-kwadrat odpowiadająca modelowi opartemu na pierwszej tablicy kanonicznej wyniosłaby wtedy

$$770{,}60 * 0{,}930 = 716{,}66$$

zaś dla modelu opartego na drugiej tablicy kanonicznej

$$770{,}60 * 0{,}070 = 53{,}94$$

czyli w przybliżeniu tyle, ile wynosi wartość podana przez program.

Ostatnia sekcja pliku wynikowego, oznaczona *Categories*, zawiera informacje o wielkościach współrzędnych kanonicznych. Podane są one osobno dla pierwszego i drugiego wymiaru. W boczku poszczególnych wierszy podano oznaczenia kategorii, którym wiersze te odpowiadają. Program najpierw wypisuje wielkości odpowiadające kategoriom cechy w wierszach, a następnie, w jednym ciągu, wielkości dla kategorii cechy w kolumnach. Rubryka oznaczona napisem „weight" zawiera proporcje odpowiadające poszczególnym kategoriom obu cech, zwane w analizie korespondencji **masami** kategorii. Cztery pierwsze wielkości sumują się do jedności, podobnie jak trzy dalsze. Proporcje te odpowiadają profilom brzegowym obu cech przedstawionym w tabeli 5.1.

Rubryka „x" zawiera współrzędne kanoniczne w postaci standaryzowanej. Z współrzędnych tych korzystaliśmy w rozdziale 6, obliczając wielkości w polach tablic kanonicznych oraz dystanse chi-kwadrat między kategoriami. Wielkości współrzędnych dla pierwszego wymiaru kanonicznego prezentowane były w tabeli 6.3, zaś dla drugiego w tabeli 6.8. Porównanie tych wielkości z wynikami programu LEM pozwala zauważyć, że ich znaki w obu prezentacjach są odwrotne. Na przykład, zamieszczona w tabeli 6.3 wartość współrzęd-

nej kanonicznej dla ojców mających wykształcenie wyższe wynosi 2,249, zaś według programu LEM jest to –2,249. Wartości współrzędnych kanonicznych są jednak określane z dokładnością do przekształcenia liniowego, a więc zmiana ich znaku nie prowadzi do zmiany ich własności. Zmieniając ich znak (czyli mnożąc współrzędne przez –1), trzeba to konsekwentnie zrobić w wypadku obu cech. W przeciwnym wypadku nie oszacuje się poprawnie liczebności w polach tablicy kanonicznej (wzór 6.9). Natomiast odwrócenia skali wolno dokonywać niezależnie od siebie w każdym z rozpatrywanych wymiarów kanonicznych.

Wielkości podane w rubryce „x*sqrt(rho)" są tak zwanymi współrzędnymi w **normalizacji kanonicznej**, zwanej też **niesymetryczną**. Normalizację tę wykorzystać można do sporządzenia wykresu w analizie korespondencji, chociaż w książce podejścia tego nie omawiam (powody wyjaśniam w przypisie 6 w podrozdziale 7.3, podając odwołania do literatury). Bardziej przydatna jest kolejna kolumna „x*rho", przedstawiająca współrzędne główne, którymi posługiwaliśmy się przy sporządzaniu wykresów. Dla rozpatrywanego przykładu współrzędne te podane były uprzednio w części [2] tabeli 7.15. Podane tam współrzędne dla drugiego wymiaru mają odwrócony znak w stosunku do obliczonych przez program LEM. Wynika to stąd, że na wykresie kategoria ojców o wykształceniu wyższym miała się znaleźć jak najbliżej jego lewego górnego narożnika, czyli w najbardziej eksponowanym miejscu. Odpowiada to ujemnej wartości współrzędnej w pierwszym wymiarze oraz dodatniej w drugim.

Ostatnia rubryka nazwana „contribution" określa dekompozycję bezwładności związanej z danym wymiarem między kategorie obu cech. Tak jak w wypadku rozkładów brzegowych, wielkości te podane są w proporcjach. Suma pierwszych czterech wynosi 1, podobnie jak suma trzech kolejnych. Są one równe przedstawionym w tabeli 6.6 udziałom kwadratów średnich wskaźników Quételeta. Rozkłady brzegowe tej tabeli odpowiadają wartościom podanym w ostatniej rubryce wyników programu LEM dla pierwszego wymiaru. Wartości podane dla drugiego z wymiarów nie były natomiast w książce omawiane.

Wyniki programu LEM obejmują więc większość parametrów wymaganych do budowy modeli kanonicznych, czy równoważnych im modeli korespondencji. W szczególności wyniki te pozwalają sporządzić wykres korespondencji, uzupełnić jego opis o wielkości korelacji kanonicznych, a także dokonać dekompozycji bezwładności całkowitej między poszczególne wymiary. W analizie korespondencji pozwala to rozstrzygnąć, czy model ograniczony do dwóch wymiarów jest należycie dopasowany do danych.

B.2 Obliczanie parametrów niedostępnych w wynikach programu LEM

Wyniki programu LEM przydatne są również do obliczenia pozostałych parametrów, wymienionych w zestawieniu na początku aneksu.

B.2.1 Dopasowane średnie

Z podrozdziału 6.5.1 wynika, że współrzędne kanoniczne są równoważne wielkościom średnich uzyskanych metodą dopasowania (ang. *reciprocal averaging*). Jeśli więc badaczowi zależy wyłącznie na obliczeniu dopasowanych średnich, to jako ich wartości może przyjąć współrzędne pierwszego wymiaru kanonicznego, normalizując je ewentualnie do przedziału <0, 100> – tak jak to przedstawiono w tabeli 5.6. Na przykład, z tabeli tej odczytać można, że dopasowana średnia dla uczniów, których ojcowie mają wykształcenie zasadnicze zawodowe, wynosi 17. Korzystając z wyników programu LEM, wielkość tę uzyskać można podstawiając odpowiednie współrzędne kanoniczne pierwszego wymiaru do wzoru (5.3)

$$100 * \frac{-0{,}593 - (-1{,}182)}{2.249 - (-1.182)} = 100 * \frac{0{,}589}{3{,}431} = 17{,}2$$

Obliczając powyższą wartość skorzystaliśmy ze współrzędnych w postaci standaryzowanej, podanych w rubryce oznaczonej „x", chociaż skorzystanie ze współrzędnych w normalizacji niesymetrycznej „x*sqrt(rho)" bądź w postaci głównej „x*rho" doprowadziłoby do identycznego rezultatu. Warto odnotować, że przed wykonaniem obliczeń zmieniony został znak współrzędnych, tak aby kategorii ojców o wykształceniu wyższym odpowiadała najwyższa wartość skalowa.

B.2.2 Dekompozycja chi-kwadrat (bezwładności całkowitej) oraz wskaźniki Quételeta

W wypadku wielu zagadnień pomocne jest zdekomponowanie chi-kwadrat między wiersze, kolumny, a także poszczególne pola analizowanej tablicy. Dla rozpatrywanego przykładu dekompozycja chi-kwadrat przedstawiona została w tabeli 6.2.Pod względem koncepcyjnym dekompozycja chi-kwadrat jest równoważna dekompozycji bezwładności całkowitej. Korzystaliśmy z tego omawiając przykłady zastosowania analizy korespondencji w rozdziale 7. Dekompozycja bezwładności całkowitej odpowiada zarazem obliczeniu dla poszczególnych pól tablicy średnich kwadratów wskaźników Quételeta. Wszystkie wymienione parametry można więc traktować łącznie. Przy czym

obliczenie ich wielkości nie wymaga znajomości współrzędnych kanonicznych. Tym niemniej, warto poświęcić uwagę sposobowi ich liczenia, gdyż w analogiczny sposób obliczane są odpowiednie wielkości w modelach kanonicznych.

Punkt wyjścia stanowić może obliczenie wielkości kwadratów średnich wskaźników Quételeta dla pól tablicy, a następnie zsumowanie ich w wierszach i kolumnach. W tym celu wzór (4.9) na kwadrat średniego wskaźnika Quételeta warto sprowadzić do postaci

$$q_{ij}^2 = \frac{\left(n_{ij} - e_{ij}\right)^2}{n * e_{ij}} \tag{B.1}$$

która jest bardziej dogodna do wykonania obliczeń.

Obliczenia można przeprowadzić posługując się arkuszem kalkulacyjnym. W pierwszej kolejności w arkuszu tworzymy tablicę liczebności empirycznych, a następnie dodajemy do niej funkcję sumującą marginesy (rycina B.3). Następnie tablicę kopiujemy cztery razy, zmieniając jedynie napisy określające jej zawartość. W drugiej kolejnej tablicy umieszczamy liczebności modelu niezależności. Obliczamy je według wzoru (3.10), który w wypadku pola B13 przybierze postać następującej formuły

$$= B\$8 * \$E4/\$E\$8$$

Przypomnijmy, że w prezentowanej formule symbol $ (dolara) oznacza niezmienność umieszczonej bezpośrednio po nim współrzędnej podczas przenoszenia formuły do innych pól tablicy. Na przykład, składnik $*E*$8 odpowiada liczbie badanych uczniów i jest taki sam dla wszystkich pól modelu niezależności. Tak utworzoną formułę kopiujemy do pozostałych pól wnętrza tablicy. Otrzymane sumy brzegowe powinny być identyczne z sumami tablicy liczebności empirycznych zgodnie z wzorami (3.11)–(3.13).

W kolejnym kroku liczymy wartości kwadratów średnich wskaźników Quételeta według wzoru (B.1) i umieszczamy je w tabeli oznaczonej na rycinie B3 jako [3]. Formuła w polu B22 przybierze postać

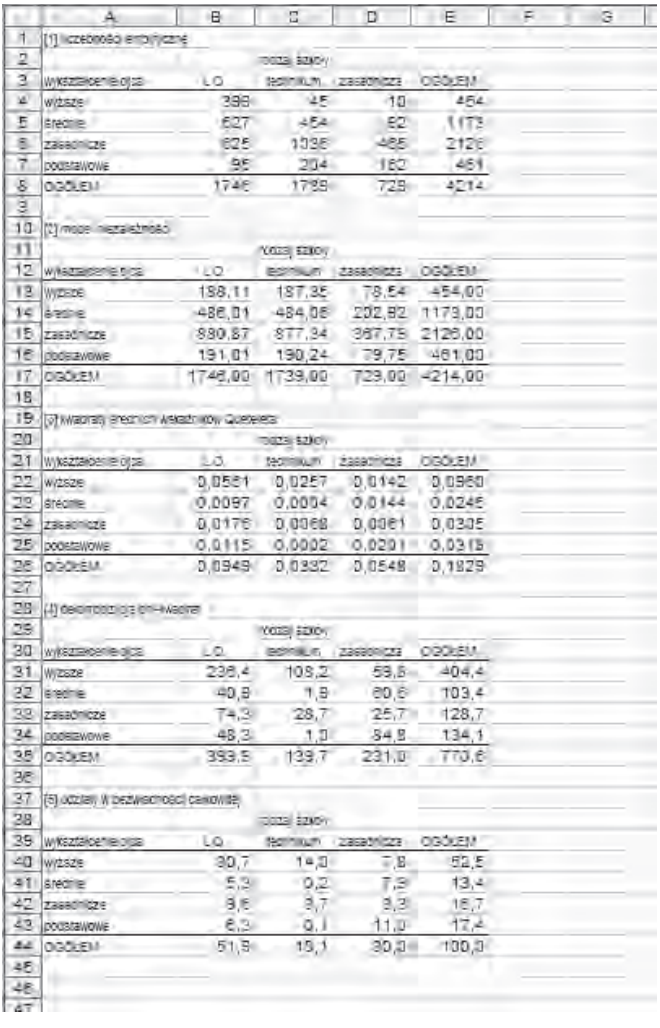
$$= (B4 - B13) * (B4 - B13)/B13/\$E\$8$$

Po rozszerzeniu formuły na wszystkie pola wnętrza tablicy uzyskamy prezentowane na rycinie B.3 wartości kwadratów średnich wskaźników Quételeta. Aby z wartości tych przejść do dekompozycji chi-kwadrat, każdą z nich należy przemnożyć przez liczbę badanych osób (wzór 4.10). Model dekompozycji chi-kwadrat przedstawiony został na rycinie B.3 jako tabela [4]. Wielkość w polu B31 określona jest jako

$$= \$E\$8 * B22$$

Suma ogółem równa jest 770,6, co odpowiada wartości chi-kwadrat dla całej tablicy. Jeśli przez wielkość tę podzielimy wielkości w poszczególnych polach, mnożąc je jednocześnie przez 100, to otrzymamy wyrażone w procentach

Rycina B.3
Obliczanie kwadratów średnich wskaźników Quételeta, dekompozycji chi-kwadrat oraz udziałów pól tablicy w bezwładności całkowitej za pomocą arkusza kalkulacyjnego



	A	B	C	D	E	F	G
1	(1) liczebności empiryczne						
2			rodzaj szkoły				
3	wykształcenie ojca	LO	technikum	zasadnicza	OGÓŁEM		
4	wyższe	398	45	10	454		
5	średnie	627	454	92	1173		
6	zasadnicze	625	1036	465	2126		
7	podstawowe	96	204	160	461		
8	OGÓŁEM	1746	1739	729	4214		
9							
10	(2) moce niezależności						
11			rodzaj szkoły				
12	wykształcenie ojca	LO	technikum	zasadnicza	OGÓŁEM		
13	wyższe	188,11	187,35	78,54	454,00		
14	średnie	486,01	484,06	202,92	1173,00		
15	zasadnicze	880,87	877,34	367,75	2126,00		
16	podstawowe	191,01	190,24	79,75	461,00		
17	OGÓŁEM	1746,00	1739,00	729,00	4214,00		
18							
19	(3) kwadraty średnich wskaźników Quételeta						
20			rodzaj szkoły				
21	wykształcenie ojca	LO	technikum	zasadnicza	OGÓŁEM		
22	wyższe	0,0581	0,0257	0,0142	0,0980		
23	średnie	0,0097	0,0004	0,0144	0,0246		
24	zasadnicze	0,0176	0,0068	0,0061	0,0305		
25	podstawowe	0,0115	0,0002	0,0291	0,0408		
26	OGÓŁEM	0,0949	0,0332	0,0648	0,1929		
27							
28	(4) dekompozycja chi-kwadrat						
29			rodzaj szkoły				
30	wykształcenie ojca	LO	technikum	zasadnicza	OGÓŁEM		
31	wyższe	236,4	108,2	59,8	404,4		
32	średnie	40,8	1,9	60,6	103,4		
33	zasadnicze	74,3	28,7	25,7	128,7		
34	podstawowe	48,3	1,0	84,8	134,1		
35	OGÓŁEM	393,5	139,7	231,0	770,6		
36							
37	(5) udziały w bezwładności całkowitej						
38			rodzaj szkoły				
39	wykształcenie ojca	LO	technikum	zasadnicza	OGÓŁEM		
40	wyższe	30,7	14,0	7,8	52,5		
41	średnie	5,3	0,2	7,9	13,4		
42	zasadnicze	9,6	3,7	3,3	16,7		
43	podstawowe	6,3	0,1	11,0	17,4		
44	OGÓŁEM	51,0	18,1	30,0	100,0		
45							
46							
47							

Dane dotyczą zależności między wykształceniem ojca a rodzajem szkoły ponadgimnazjalnej, do której uczęszcza dziecko. Badanie PISA 2006.

udziały poszczególnych pól wnętrza tabeli, a także udziały poszczególnych wierszy oraz udziały poszczególnych kolumn w bezwładności całkowitej. Formuła dla pola B40 będzie miała postać

$$= 100 * B31/\$E\$35$$

W ten sposób dokonaliśmy dekompozycji bezwładności całkowitej na poszczególne pola tablicy. Rezultat warto porównać z prezentowanym w części [3] tabeli 6.2. Tabela ta przedstawia dekompozycję chi-kwadrat, która zgodnie z wzorem (6.7) odpowiada dekompozycji bezwładności całkowitej. Podane udziały procentowe poszczególnych wierszy i kolumn odpowiadają analogicznym wielkościom podanym w części [5] ryciny 5.3.

B.2.3 Dystanse chi-kwadrat

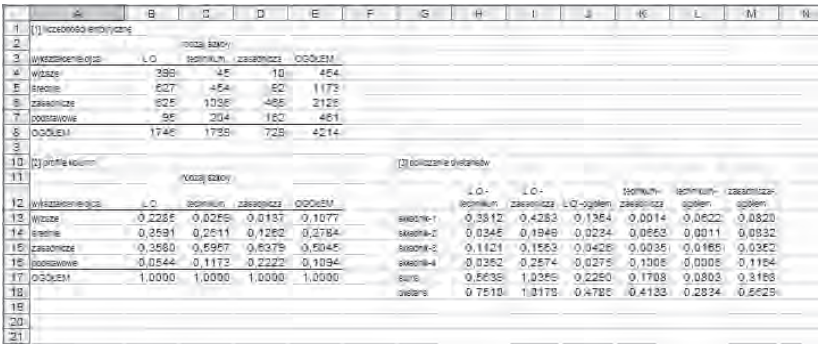
Obliczenie dystansów chi-kwadrat dla tablicy liczebności empirycznych również nie wymaga korzystania ze współrzędnych kanonicznych. Rycina B.4 obrazuje sposób obliczenia dystansów chi-kwadrat między profilami w kolumnach tablicy – w tym wypadku dystansów między rodzajami szkół. Punkt wyjścia stanowi tablica liczebności empirycznych [1]. Tablicę tę należy skopiować poniżej w celu umieszczenia w niej profili. Wielkości w polach profili najlepiej jest wyrazić w proporcjach. W sytuacji tej suma każdej kolumny tablicy jest równa 1. Profile obliczamy dla wszystkich szkół oraz dla kolumny „ogółem".

Wartość dystansu chi-kwadrat między profilami w kolumnach określa wzór (6.2). Wykonując obliczenia, sumę pod pierwiastkiem najlepiej rozłożyć na składniki odpowiadające wierszom tablicy. Wiąże się to z koniecznością sporządzenia tabeli pomocniczej w sposób przedstawiony w części [3] ryciny B.4, której kolumny odpowiadają parom porównywanych profili. Przykładowo, funkcja w polu H13, odpowiadająca składnikowi-1 przy porównywaniu profili w liceach ogólnokształcących (LO) i technikach, przedstawia się następująco

$$= (\$B13 - C13) * (\$B13 - C13)/\$E13$$

Zadeklarowanie za pomocą symbolu $ kolumny B jako stałej pozwala skopiować formułę do wszystkich pól bloku H13:J16, odpowiadających porównywaniu profili w liceach z pozostałymi profilami. W analogiczny sposób należy utworzyć formuły dla par profili obejmujących technika i szkoły zasadnicze zawodowe. Pola w wierszu „suma" zawierają sumy czterech składników, zaś pola w wierszu „dystans" pierwiastki z owych sum – czyli szukane wielkości dystansów. Z zestawu tego w części [2] tabeli 6.2 prezentowane są dystanse poszczególnych rodzajów szkół od profilu brzegowego. Wielkości te odpowiadają podanym na rycinie B.4.

Rycina B.4
Obliczanie dystansów chi-kwadrat między profilami w kolumnach dla tablicy liczebności empirycznych

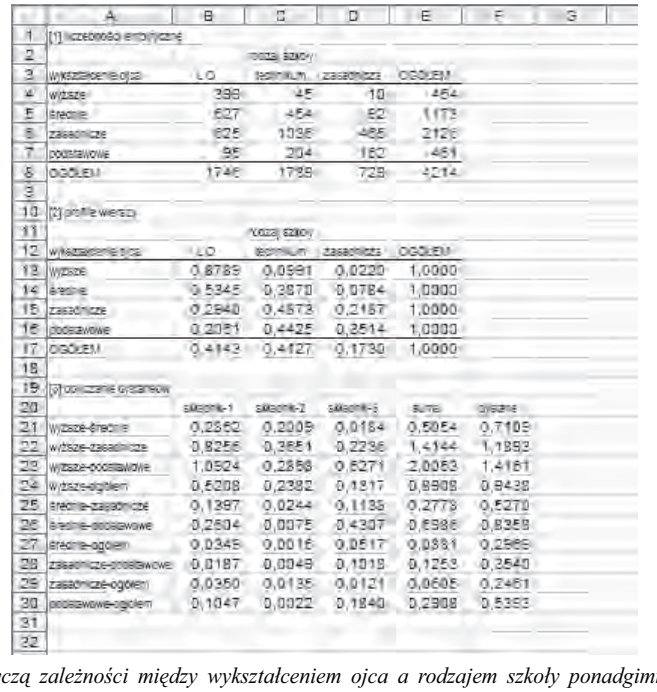


	A	B	C	D	E
1	[1] liczebności empiryczne				
2			rodzaj szkoły		
3	wykształcenie ojca	LO	technikum	zasadnicza	OGÓŁEM
4	wyższe	398	45	10	454
5	średnie	627	454	92	1173
6	zasadnicze	625	1036	465	2126
7	podstawowe	96	204	160	461
8	OGÓŁEM	1746	1739	729	4214
9					
10	[2] profile kolumn				
11			rodzaj szkoły		
12	wykształcenie ojca	LO	technikum	zasadnicza	OGÓŁEM
13	wyższe	0,2265	0,0259	0,0137	0,1077
14	średnie	0,3591	0,2611	0,1262	0,2784
15	zasadnicze	0,3580	0,5957	0,6379	0,5045
16	podstawowe	0,0544	0,1173	0,2222	0,1094
17	OGÓŁEM	1,0000	1,0000	1,0000	1,0000
18					
19					
20					
21					

	G	H	I	J	K	L	M
10	[3] obliczanie dystansów						
11							
12		LO - technikum	LO - zasadnicza	LO - ogółem	technikum - zasadnicza	technikum - ogółem	zasadnicza - ogółem
13	składnik-1	0,3812	0,4263	0,1364	0,0014	0,0622	0,0820
14	składnik-2	0,0345	0,1849	0,0234	0,0653	0,0011	0,0932
15	składnik-3	0,1121	0,1553	0,0426	0,0035	0,0165	0,0362
16	składnik-4	0,0352	0,2574	0,0275	0,1305	0,0005	0,1154
17	suma	0,5630	1,0239	0,2299	0,1708	0,0803	0,3168
18	odległość	0,7503	1,0119	0,4795	0,4133	0,2834	0,5629

Dane dotyczą zależności między wykształceniem ojca a rodzajem szkoły ponadgimnazjalnej, do której uczęszcza dziecko. Badanie PISA 2006.

Rycina B.5
Obliczanie dystansów chi-kwadrat między profilami w wierszach dla tablicy liczebności empirycznych



	A	B	C	D	E
1	[1] liczebności empiryczne				
2			rodzaj szkoły		
3	wykształcenie ojca	LO	technikum	zasadnicza	OGÓŁEM
4	wyższe	398	45	10	454
5	średnie	627	454	92	1173
6	zasadnicze	625	1036	465	2126
7	podstawowe	96	204	160	461
8	OGÓŁEM	1746	1739	729	4214
9					
10	[2] profile wierszy				
11			rodzaj szkoły		
12	wykształcenie ojca	LO	technikum	zasadnicza	OGÓŁEM
13	wyższe	0,8789	0,0991	0,0220	1,0000
14	średnie	0,5345	0,3870	0,0784	1,0000
15	zasadnicze	0,2940	0,4873	0,2187	1,0000
16	podstawowe	0,2081	0,4425	0,3514	1,0000
17	OGÓŁEM	0,4143	0,4127	0,1730	1,0000
18					
19	[3] obliczanie dystansów				
20		składnik-1	składnik-2	składnik-3	suma	dystans
21	wyższe-średnie	0,2852	0,2009	0,0184	0,5054	0,7109
22	wyższe-zasadnicze	0,8256	0,3851	0,2236	1,4344	1,1982
23	wyższe-podstawowe	1,0524	0,2858	0,6271	2,0653	1,4161
24	wyższe-ogółem	0,5208	0,2382	0,1217	0,8808	0,9438
25	średnie-zasadnicze	0,1397	0,0244	0,1135	0,2778	0,5270
26	średnie-podstawowe	0,2604	0,0075	0,4307	0,6986	0,8358
27	średnie-ogółem	0,0345	0,0016	0,0517	0,0881	0,2966
28	zasadnicze-podstawowe	0,0187	0,0049	0,1015	0,1253	0,3540
29	zasadnicze-ogółem	0,0350	0,0135	0,0121	0,0605	0,2461
30	podstawowe-ogółem	0,1047	0,0022	0,1840	0,2908	0,5393
31						
32						

Dane dotyczą zależności między wykształceniem ojca a rodzajem szkoły ponadgimnazjalnej, do której uczęszcza dziecko. Badanie PISA 2006.

Sposób obliczania dystansów chi-kwadrat między profilami w wierszach jest podobny (rycina B.5). Tym razem pomocniczą tablicę wygodniej jest umieścić nie obok, a pod tablicą zawierającą profile dla wierszy. Postać przykładowej funkcji pozwalającej obliczyć wartość pierwszego składnika dla porównania profili szkoły, do których uczęszczają dzieci mające ojców o wykształceniu wyższym i średnim, przedstawia się następująco

$$= (B\$13 - B14) * (B\$13 - B14)/B\$17$$

Tym razem jako stałą warto przyjąć wiersz 13, co pozwoli skopiować formułę do wszystkich pól odpowiadających porównywaniu profili dla ojców z wykształceniem wyższym. Obliczone w kolumnie F dystanse prezentowane uprzednio były jako część [2] tabeli 6.1.

B.2.4 Modele kanoniczne

Rozważmy obecnie sposób obliczenia liczebności w polach tablicy odtworzonych przez model kanoniczny, który odpowiada pierwszej tablicy kanonicznej. W rozdziale 6 szukane wielkości prezentowane są w części [4] tabeli 6.4. Początek obliczeń jest analogiczny jak w wypadku obliczania udziałów bezwładności (B.2.2). W arkuszu należy utworzyć tablicę liczebności empirycznych oraz obliczyć dla niej liczebności modelu niezależności. Jeśli dysponujemy poprzednio utworzonym arkuszem, to różnica pojawi się dopiero w wypadku trzeciej kolejnej tablicy (rycina B.6). Liczebności modelu kanonicznego określone są przez wzór (6.16), który w wypadku rozważanego modelu jest równoważny złożeniu wzorów (6.11) i (6.8)

$$m_{ij} = e_{ij} + r^{(1)} e_{ij} x_i^{(1)} y_j^{(1)} \tag{B.2}$$

Aby obliczyć te wielkości, dogodnie jest dopisać w każdym z wierszy, a także w każdej z kolumn, wielkości współrzędnych kanonicznych w postaci standaryzowanej. W wynikach programu LEM podane są one w kolumnie oznaczonej „x". W książce wartości te prezentowane są w tabeli 6.3. Dopisując dodatkowo w jednym z pól arkusza wartość pierwszej korelacji kanonicznej, która również potrzebna jest do oszacowania liczebności modelu, otrzymamy konfigurację wielkości przedstawioną jako trzecia tabela ryciny B.6. Formuła na liczebność $m_{11}^{(1)}$ w polu B22 arkusza wyraża się następująco

$$= B13 + \$G\$28 * B13 * \$G22 * B\$28$$

Po jej rozszerzeniu na wszystkie wewnętrzne pola tablicy uzyskujemy liczebności pierwszego modelu kanonicznego, równe co do wielkości podanym w części [3] tabeli 6.4.

Dla pól otrzymanego modelu kanonicznego obliczyć można kwadraty średnich wskaźników Quételeta korzystając z wzoru (B.1). Tym razem jednak z liczebnościami e_{ij} pól w modelu niezależności nie zestawia się liczebności obserwowanych n_{ij}, lecz estymowane liczebności $m_{ij}^{(1)}$. Formuła na wielkość wskaźnika Quételeta w polu B33 przybierze więc postać

Rycina B.6
Estymacja liczebności pierwszego modelu kanonicznego

	A	B	C	D	E	F	G
1	(1) liczebności empiryczne						
2			rodzaj szkoły				
3	wykształcenie ojca	LO	technikum	zasadnicza	OGÓŁEM		
4	wyższe	398	45	10	454		
5	średnie	627	454	92	1173		
6	zasadnicze	625	1036	465	2126		
7	podstawowe	96	204	160	461		
8	OGÓŁEM	1746	1739	729	4214		
9							
10	(2) model niezależności						
11			rodzaj szkoły				
12	wykształcenie ojca	LO	technikum	zasadnicza	OGÓŁEM		
13	wyższe	188,11	187,35	78,54	454,00		
14	średnie	486,01	484,06	202,92	1173,00		
15	zasadnicze	880,57	877,34	367,75	2126,00		
16	podstawowe	191,01	190,24	79,75	461,00		
17	OGÓŁEM	1746,00	1739,00	729,00	4214,00		
18							
19	(3) pierwszy model kanoniczny						
20			rodzaj szkoły				współrzędne
21	wykształcenie ojca	LO	technikum	zasadnicza	OGÓŁEM		kanoniczne
22	wyższe	380	79	-14	454		-2,245
23	średnie	641	460	132	1173		-0,669
24	zasadnicze	631	1012	482	2126		0,593
25	podstawowe	93	248	129	461		1,162
26	OGÓŁEM	1746	1739	729	4214		
27							
28	współrzędne kanoniczne	-1,158	0,623	1,275		korelacja	0,412
29							
30	(4) kwadraty średnich wskaźników Quételeta						
31			rodzaj szkoły				
32	wykształcenie ojca	LO	technikum	zasadnicza	OGÓŁEM		
33	wyższe	0,0515	0,0151	0,0251	0,0525		
34	średnie	0,0118	0,0034	0,0060	0,0212		
35	zasadnicze	0,0169	0,0045	0,0036	0,0302		
36	podstawowe	0,0144	0,0042	0,0073	0,0260		
37	OGÓŁEM	0,0945	0,0277	0,0478	0,1700		
38							
39	(5) udziały w bezwskaźnikości całkowitej, w procentach						
40			rodzaj szkoły				
41	wykształcenie ojca	LO	technikum	zasadnicza	OGÓŁEM		
42	wyższe	30,3	8,9	15,3	54,5		
43	średnie	6,5	2,0	3,5	12,5		
44	zasadnicze	9,9	2,5	5,0	17,7		
45	podstawowe	8,5	2,5	4,3	15,3		
46	OGÓŁEM	55,6	15,9	28,1	100,0		
47							

Dane dotyczą zależności między wykształceniem ojca a rodzajem szkoły ponadgimnazjalnej, do której uczęszcza dziecko. Badanie PISA 2006.

$$= (B22 - B13) * (B22 - B13)/B13/\$E\$8$$

Obliczone tą drogą wielkości przedstawione są na rycinie 5.6 w części [4]. Ich suma wynosi 0,1700, co odpowiada wielkości chi-kwadrat obliczonej dla

Rycina B.7
Estymacja parametrów drugiego modelu kanonicznego drogą zmiany współrzędnych i korelacji kanonicznych w arkuszu użytym do estymacji pierwszego modelu kanonicznego

	A	B	C	D	E	F	G
1	(1) liczebności empiryczne						
2			rodzaj szkoły				
3	wykształcenie ojca	LO	technikum	zasadnicza	OGÓŁEM		
4	wyższe	398	45	10	454		
5	średnie	627	454	92	1173		
6	zasadnicze	625	1036	465	2126		
7	podstawowe	96	204	160	461		
8	OGÓŁEM	1746	1739	729	4214		
9							
10	(2) model niezależności						
11			rodzaj szkoły				
12	wykształcenie ojca	LO	technikum	zasadnicza	OGÓŁEM		
13	wyższe	188,11	187,35	78,54	454,00		
14	średnie	486,01	484,06	202,92	1173,00		
15	zasadnicze	880,57	877,34	367,75	2126,00		
16	podstawowe	191,01	190,24	79,75	461,00		
17	OGÓŁEM	1746,00	1739,00	729,00	4214,00		
18							
19	(3) drugi model kanoniczny						
20			rodzaj szkoły				współrzędne
21	wykształcenie ojca	LO	technikum	zasadnicza	OGÓŁEM		kanoniczne
22	wyższe	197	154	103	454		-1,541
23	średnie	472	538	163	1173		0,965
24	zasadnicze	875	901	350	2126		0,286
25	podstawowe	203	146	112	461		-2,085
26	OGÓŁEM	1746	1739	729	4214		
27							
28	współrzędne kanoniczne	-0,268	1,314	-1,775		korelacja	0,113
29							
30	(4) kwadraty średnich wartości wiersz						
31			rodzaj szkoły				
32	wykształcenie ojca	LO	technikum	zasadnicza	OGÓŁEM		
33	wyższe	0,0001	0,0014	0,0015	0,0033		
34	średnie	0,0001	0,0014	0,0015	0,0033		
35	zasadnicze	0,0000	0,0002	0,0002	0,0004		
36	podstawowe	0,0002	0,0025	0,0032	0,0058		
37	OGÓŁEM	0,0004	0,0054	0,0070	0,0128		
38							
39	(5) udziały w bezwładności całkowitej	w procentach					
40			rodzaj szkoły				
41	wykształcenie ojca	LO	technikum	zasadnicza	OGÓŁEM		
42	wyższe	0,8	10,8	14,0	25,6		
43	średnie	0,8	11,1	14,3	26,1		
44	zasadnicze	0,1	1,2	1,5	2,8		
45	podstawowe	1,4	19,3	24,8	45,5		
46	OGÓŁEM	3,0	42,4	54,6	100,0		
47							

Dane dotyczą zależności między wykształceniem ojca a rodzajem szkoły ponadgimnazjalnej, do której uczęszcza dziecko. Badanie PISA 2006.

tablicy liczebności modelu kanonicznego i podzielonej przez liczbę badanych uczniów (wzór 6.29), a także jest równa kwadratowi pierwszej korelacji kanonicznej (wzór 6.15). Obliczenie odsetków wielkości w polach względem ich sumy da nam dekompozycję chi-kwadrat dla modelu kanonicznego między pola tablicy. Dekompozycję tę przedstawia tabela [5] na rycinie 5.4. Zarówno obliczane obecnie kwadraty średnich wskaźników Quételeta, jak też dekompozycja bezwładności całkowitej między pola tabeli, omawiane były w podrozdziale 6.5.4 i przedstawione w tabeli 6.6.

Dysponując sformatowanym w ten sposób arkuszem kalkulacyjnym łatwo jest obliczyć liczebności modelu odpowiadającego drugiej tablicy kanonicznej, który to model był przedmiotem dyskusji w podrozdziale 6.6. W tym celu w arkuszu należy zastąpić zestaw współrzędnych pierwszego wymiaru przez zestaw współrzędnych dla wymiaru drugiego, zaś zamiast pierwszej korelacji

Rycina B.8
Obliczanie dystansów chi-kwadrat między profilami w wierszach dla liczebności pierwszego modelu kanonicznego

	A	B	C	D	E	F	G
1	[1] liczebności pierwszego modelu kanonicznego						
2			rodzaj szkoły				
3	wykształcenie ojca	LO	technikum	zasadnicza	OGÓŁEM		
4	wyższe	390	78	-14	454		
5	średnie	541	490	132	1173		
6	zasadnicze	831	1012	482	2126		
7	podstawowe	83	246	125	454		
8	OGÓŁEM	1745	1799	729	4214		
9							
10	[2] profile wierszy						
11			rodzaj szkoły				
12	wykształcenie ojca	LO	technikum	zasadnicza	OGÓŁEM		
13	wyższe	0,8593	0,1723	-0,0316	1,0000		
14	średnie	0,5467	0,3412	0,1121	1,0000		
15	zasadnicze	0,2870	0,4761	0,2269	1,0000		
16	podstawowe	0,1805	0,5380	0,2805	1,0000		
17	OGÓŁEM	0,4143	0,4127	0,1730	1,0000		
18							
19	[3] obliczanie dystansów						
20		odległość-1	odległość-2	odległość-3	suma	dystans	
21	wyższe-średnie	0,2358	0,0681	0,1194	0,4243	0,6514	
22	wyższe-zasadnicze	0,7530	0,2236	0,3952	1,3718	1,1717	
23	wyższe-podstawowe	1,1121	0,3259	0,5621	2,0011	1,4146	
24	wyższe-ogółem	0,4778	0,1400	0,2419	0,8597	0,9272	
25	średnie-zasadnicze	0,1505	0,0441	0,0762	0,2708	0,5203	
26	średnie-podstawowe	0,3237	0,0948	0,1639	0,5825	0,7632	
27	średnie-ogółem	0,0423	0,0124	0,0214	0,0761	0,2758	
28	zasadnicze-podstawowe	0,0328	0,0096	0,0166	0,0590	0,2429	
29	zasadnicze-ogółem	0,0332	0,0097	0,0169	0,0598	0,2445	
30	podstawowe-ogółem	0,1320	0,0387	0,0668	0,2375	0,4874	
31							
32							

Dane dotyczą zależności między wykształceniem ojca a rodzajem szkoły ponadgimnazjalnej, do której uczęszcza dziecko. Badanie PISA 2006.

kanonicznej wpisać wielkość drugiej korelacji kanonicznej. Tak zmodyfikowany arkusz przedstawiony został na rycinie B.7. Otrzymane wielkości są zgodne z prezentowanymi w tabeli 6.7.

Znając współrzędne oraz korelacje kanoniczne jesteśmy też w stanie określić wielkości dystansów chi-kwadrat między kategoriami. Można w tym celu skorzystać z wzorów (6.12) i (6.13). Jeśli jednak dysponujemy arkuszem użytym w B.2.3 do obliczania dystansów chi-kwadrat, to w arkuszu tym wystarczy zastąpić tablicę liczebności empirycznych przez liczebności odpowiedniego modelu kanonicznego. Rozwiązanie to przedstawione zostało na rycinie B.8. Jest to arkusz kalkulacyjny z ryciny B.5, do którego wklejono liczebności pierwszego modelu kanonicznego (rycina B.7). Podane w kolumnie „dystans" wartości dystansów chi-kwadrat są równe wielkościom prezentowanym w części [2] tabeli 6.5.

B.3 Wykonanie obliczeń za pomocą pakietu STATA

Pakiet STATA jest coraz szerzej stosowany przez badaczy i studentów nauk społecznych (Treiman 2009). Powodem jest zapewne stosunkowo niska cena, bezawaryjność oraz niewielkie wymagania sprzętowe. Zalety te wynikają stąd, że organizacja pakietu wywodzi się z rozwiązań specyficznych dla systemu DOS. Ograniczone możliwości interaktywnej pracy są z kolei wadą tego pakietu. Dlatego, podobnie jak w wypadku programu LEM, dobrze jest sformułować zadanie w postaci pliku poleceń. Dla rozpatrywanego przykładu plik ten przedstawiony został na rycinie B.9.

Przykładowy plik instrukcji rozpoczynają wiersze komentarza. Analogicznie, jak pakiecie LEM, wiersze te rozpoczynają się symbolem * (gwiazdki). Po nich warto umieścić instrukcję

```
#delimit ;
```

która oznacza, że każde następne polecenie może być zapisane w więcej niż jednym wierszu. Jest to wygodne w wypadku instrukcji zawierających deklaracje danych, gdyż w przeciwnym wypadku całość danych musiałaby zostać umieszczona w jednym wierszu pliku. Podana instrukcja powoduje, że program będzie interpretował każde kolejne polecenie aż do napotkania znacznika jego końca, którą to funkcję pełni symbol średnika.

Polecenie rozpoczynające się od słów „matrix input" zawiera definicję rozpatrywanej tablicy[1]. Opatrzono ją identyfikatorem „woxszk". Po znaku rów-

[1] Omawiam wyłącznie sytuację, gdy użytkownik wykonuje obliczenia na danych w formacie tabelarycznym. Pakiet STATA umożliwia również pracę z plikami danych, w których informacje zorganizowane są w postaci zmiennych. Prezentowany zestaw poleceń miałby wtedy nieco inną zawartość.

Rycina B.9
Plik poleceń programu STATA wywołujących procedurę analizy korespondencji

```
* Wykształcenie ojca a rodzaj szkoły ponadgimnazjalnej,
* w której uczy się dziecko
* PISA 2006

* Analiza korespondencji

#delimit ;

matrix input woxszk = (
  399   45  10\
  627  454  92\
  625 1036 465\
   95  204 162);

matrix colnames woxszk = LO technikum zsz;

matrix rownames woxszk = wyższe średnie zasadnicze podstawowe;

camat woxszk, dim(2) norm(principal)
   rowname(wykszt_ojca) colname(rodz_szkoły);
```

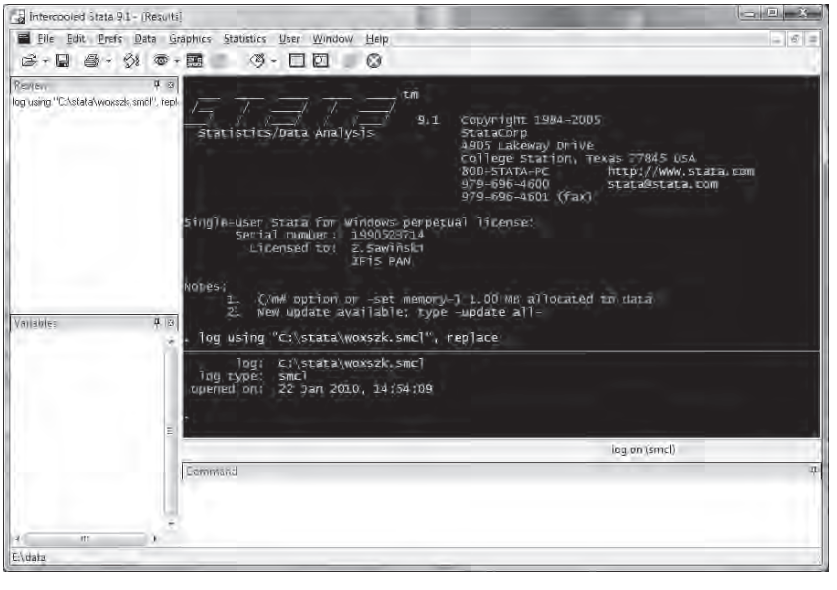
ności w nawiasach okrągłych wyszczególnione zostały wszystkie liczebności jej wnętrza, przy czym poszczególne wiersze rozdzielone są symbolem \ (tzw. *backslash*). Jak widać, definicja tablicy nie zawiera osobnych deklaracji liczby wierszy i kolumn, toteż separatory oddzielające od siebie wiersze tablicy muszą być umieszczone we właściwych miejscach.

Kolejne dwa polecenia mają charakter pomocniczy. Za pomocą polecenia „matrix colnames" po symbolu = (równości) zadeklarować można nazwy kategorii cechy umieszczonej w kolumnach tablicy. Należy dążyć do tego, aby długości nazw nie przekraczały 8 znaków, gdyż w przeciwnym wypadku w niektórych z wypisywanych wyników nazwy te będą skracane. Polecenie „matrix rownames" służy z kolei do zadeklarowania nazw kategorii cechy w wierszach. W nazwach można stosować polskie znaki diakrytyczne, natomiast nie jest dopuszczalne użycie symbolu spacji – gdyż służy ona rozdzieleniu kolejnych nazw od siebie.

Polecenie rozpoczynające się od identyfikatora „camat" wywołuje procedurę analizy korespondencji. Po identyfikatorze należy podać nazwę tablicy, a następnie, po przecinku, parametry analizy. Parametr „dim" określa liczbę wyodrębnianych wymiarów rozwiązania, którą należy wpisać w nawiasach. Pominięcie tego parametru spowoduje, że procedura domyślnie przyjmie dwa

wymiary. Parametr „norm(principal)" określa, że wyniki przedstawiać mają współrzędne kanoniczne w postaci współrzędnych głównych. Przypomnijmy, że z tej postaci współrzędnych korzystaliśmy w rozdziale 7 przy sporządzaniu wykresów. Program pozwala otrzymać współrzędne znormalizowane również na inne sposoby. Parametr „rowname" określa nazwę cechy w wierszach tablicy. Długość nazwy, umieszczonej w nawiasach, nie powinna przekraczać 12 znaków. Analogiczną funkcję pełni parametr „colname". Polecenie kończy, tak jak każde inne, symbol średnika.

Rycina B.10
Interface programu STATA po wykonaniu komendy inicjalizacji pliku wyników



Pierwszą operacją, której warto dokonać po uruchomieniu programu STATA jest inicjalizacja pliku wyników. Domyślnie program nie zapamiętuje bowiem wyników na osobnym pliku, pozwalając obejrzeć je jedynie na ekranie. Inicjalizacji dokonujemy wybierając z menu kolejno opcje „File" – „Log" – „Begin". Rycina B.10 przedstawia *interface* programu STATA po wykonaniu tej operacji. Plik wyników otwarty został pod nazwą „woxszk.smcl".

Następnie można wczytać plik poleceń. Przyłącza się go poprzez opcje „File" – „Do" , po czym w okienku dialogowym należy wybrać żądany plik. Plik poleceń wykonuje się automatycznie po wczytaniu, zaś wyniki wyświetlane są sukcesywnie na ekranie. Po ich przejrzeniu i zaakceptowaniu należy

zamknąć plik wydruku za pomocą komend „File" – „Log" – „Close". W tym momencie plik ten zapisany jest w formacie systemowym programu STATA. Aby móc z niego niezależnie korzystać, należy dokonać jego konwersji do formatu tekstowego. W tym celu przyłączamy go ponownie za pomocą sekwencji „File" – „Log" – „Translate", a następnie w okienku dialogowym podajemy nazwę pliku tekstowego. Przy czym tę ostatnią nazwę należy podać z pełną ścieżką dostępu, na przykład „c:\stata\woxszk.txt". Po wykonaniu konwersji plik wyników zapisany jest w formacie tekstowym i może być przeglądany pod dowolnym edytorem.

Na rycinie B.11 przedstawiono fragment pliku wyników zawierający rezultaty analizy korespondencji. W prawym górnym rogu podane są podstawowe informacje dotyczące wykonanej analizy, jak łączna liczba przypadków, wielkość chi-kwadrat dla tablicy liczebności obserwowanych, czy bezwładność całkowita. Poniżej przedstawione są korelacje i bezwładności odpowiadające kolejnym wymiarom. W rubryce „singular values" podane są wartości korelacji kanonicznych, zaś w rubryce „principal inertia" ich kwadraty. Warto zwrócić uwagę, że, w odróżnieniu od programu LEM, STATA dekomponuje chi-kwadrat na poszczególne wymiary, co odpowiada przedstawionej w rozdziale 6 metodzie dekompozycji tablicy na modele kanoniczne.

Kategorie obu cech opisane są w kolejnym fragmencie wyników. W rubryce „mass" podane są proporcje badanych jednostek należące do poszczególnych kategorii, zaś w rubryce „inertia" bezwładności odpowiadające tym kategoriom. Bezwładności te sumują się do bezwładności całkowitej w obrębie każdej z cech. Rubryka „overall quality" zawiera stopień odtworzenia bezwładności kategorii przez zastosowany model. Ponieważ w rozważanym przykładzie mniejszy z rozmiarów tablicy wyjściowej jest równy 3, to dwuwymiarowy model w pełni wyjaśnia odchylenia obserwowanych liczebności od modelu niezależności. Wszystkie podane wielkości są przez to równe 1.

Wyniki odpowiadające każdemu z wymiarów obejmują trzy rubryki. W rubryce „coord" podane są współrzędne kanoniczne. W tym wypadku są to wartości współrzędnych głównych, prezentowane w książce w tabeli 7.15. Uzyskanie współrzędnych w postaci standaryzowanej – potrzebnych na przykład do oszacowania liczebności w modelu kanonicznym w sposób opisany w części B.3 – wymaga podzielenia tych wielkości przez wartość korelacji kanonicznej (podrozdział 7.6). Na przykład, wartość współrzędnej standaryzowanej dla ojców mających wyższe wykształcenie jest równa

$$0{,}928 \,/\, 0{,}4123719 = 2{,}250$$

Rycina B.8
Fragment pliku wynikowego programu STATA zawierający rezultaty analizy korespondencji

```
Correspondence analysis                         Number of obs  =    4214
                                              Pearson chi2(6)  =  770.60
                                                 Prob > chi2   =  0.0000
                                               Total inertia   =  0.1829
  4 active rows                                Number of dim.  =       2
  3 active columns                            Expl.inertia(%)  =  100.00

                | singular   principal                    cumul
    Dimensions  |   values     inertia    chi2   percent  percent
----------------+-------------------------------------------------
          dim 1 | .4123719   .1700506  716.59    92.99    92.99
          dim 2 |  .113211   .0128167   54.01     7.01   100.00
----------------+-------------------------------------------------
          total |            .1828673  770.60   100

Statistics for row and column categories in principal normalization

                |       overall          |      dimension_1
     Categories | mass  quality  inertia | coord  sqcorr  contrib
----------------+------------------------+----------------------
   wykszt_ojca  |                        |
         wyższe |0.108   1.000   0.096  | 0.928  0.966   0.545
       średnie |0.278   1.000   0.025  | 0.276  0.863   0.125
     zasadnicze |0.505   1.000   0.031  |-0.245  0.988   0.177
     podstawowe |0.109   1.000   0.032  |-0.487  0.817   0.153
----------------+------------------------+----------------------
   rodz_szkoły  |                        |
              LO |0.414   1.000   0.095  | 0.478  0.996   0.556
       technikum |0.413   1.000   0.033  |-0.259  0.836   0.163
             zsz |0.173   1.000   0.055  |-0.526  0.872   0.281
----------------+------------------------+----------------------

                |        dimension_2
     Categories |   coord   sqcorr  contrib
   -------------+--------------------------
   wykszt_ojca  |
         wyższe |   0.174    0.034    0.256
       średnie |  -0.110    0.137    0.262
     zasadnicze |  -0.027    0.012    0.028
     podstawowe |   0.231    0.183    0.455
   -------------+--------------------------
   rodz_szkoły  |
              LO |   0.030    0.004    0.030
       technikum |  -0.115    0.164    0.424
             zsz |   0.201    0.128    0.546
   ---------------------------------------
```

Dane dotyczą zależności między wykształceniem ojca a rodzajem szkoły ponadgimnazjalnej, do której uczęszcza dziecko. Badanie PISA 2006.

co odpowiada wielkości podanej w tabeli 6.3. Współrzędne standaryzowane można też otrzymać zmieniając w poleceniu „camat" normalizację „principal" na „standard"[2].

W rubryce „sqcorr" podane są kwadraty tak zwanych korelacji kategorii z wymiarami kanonicznymi. Z wielkości tych w książce nie korzystaliśmy. Ich interpretację podają Blasius i Greenacre (1994: 67). Ostatnia rubryka zatytułowana „contrib" zawiera proporcje bezwładności, związane z poszczególnymi kategoriami, wyjaśnione przez dany wymiar. Wartości dla pierwszego wymiaru odpowiadają wielkościom prezentowanym w części [2] tabeli 6.6.

Zakres wyników uzyskanych do tego momentu odpowiada w zasadzie temu, co można uzyskać za pomocą programu LEM. Program STATA oferuje jednak kilka dalszych możliwości. Gdy tablica została wczytana do programu, to wtedy dalsze analizy wywołać można wpisując z klawiatury odpowiednie polecenia w oknie oznaczonym „Command" (zob. rycina B.10).

Rycina B.12
Fragment pliku wyników programu STATA zawierający wielkości dystansów chi-kwadrat między profilami

```
. estat distances

Chi2 distances between the row profiles

    wykszt_ojca |  średnie  zasadni~e  podstaw~e |    center
   -------------+-------------------------------+----------
         wyższe |   0.7109     1.1893     1.4161 |    0.9438
        średnie |              0.5270     0.8358 |    0.2969
     zasadnicze |                        0.3540 |    0.2461
     podstawowe |                               |    0.5393
   ------------------------------------------------------------

Chi2 distances between the column profiles

    rodz_szkoły | technikum        zsz |    center
   -------------+----------------------+-----------
             LO |    0.7510     1.0178 |    0.4786
       technikum |               0.4133 |    0.2834
            zsz |                      |    0.5629
   ---------------------------------------------------
```

Dane dotyczą zależności między wykształceniem ojca a rodzajem szkoły ponadgimnazjalnej, do której uczęszcza dziecko. Badanie PISA 2006.

[2] Opcja dostępna począwszy od 10 wersji programu STATA. W starszych wersjach należy procedurę wywołać dwukrotnie z normalizacjami „row" oraz „column".

Wpisanie polecenia „estat distances" spowoduje obliczenie i wypisanie dystansów chi-kwadrat między kategoriami cechy w wierszach, między kategoriami cechy w kolumnach oraz pomiędzy kategoriami każdej z tych cech a profilami brzegowymi. Fragment wyników prezentujący te dystanse przedstawiony jest na rycinie B.12. Odpowiadają one wielkościom prezentowanym uprzednio na rycinach B.4 i B.5, których sposób obliczania omówiony został w B.2.3. Należy zaznaczyć, że chodzi tu o dystanse obliczone dla tablicy liczebności obserwowanych. Wersja programu STATA, z której korzystałem przy przygotowywaniu tego aneksu, nie dawała możliwości obliczenia dystansów chi-kwadrat w modelach kanonicznych.

Polecenie „estat inertia" spowoduje wypisanie bezwładności dla poszczególnych pól wnętrza tablicy (rycina B.13). Są to wielkości, które w części B.2.2 obliczaliśmy za pomocą arkusza kalkulacyjnego (zob. rycina B.3).

Rycina B.13
Fragment pliku wyników programu STATA zawierający bezwładności w poszczególnych polach tablicy

```
. estat inertia

Inertia (=Pearson-Chi2/N) contributions (with N = 4214)

                |        LO  technikum        zsz
    ------------+---------------------------------
          wyższe |    0.0561     0.0257     0.0142
         średnie |    0.0097     0.0004     0.0144
       zasadnicze |    0.0176     0.0068     0.0061
       podstawowe |    0.0115     0.0002     0.0201
    ------------------------------------------------
```

Dane dotyczą zależności między wykształceniem ojca a rodzajem szkoły ponadgimnazjalnej, do której uczęszcza dziecko. Badanie PISA 2006.

Z kolei polecenie „estat table" spowoduje wypisanie wielkości wyjaśnionych przez model kanoniczny oparty na zadanej liczbie wymiarów. Wielkości te podawane są jednak w postaci proporcji, nie zaś liczebności, toteż aby obliczyć liczebności należy przemnożyć je przez liczbę badanych osób. Na rycinie B.14 przedstawiono omawiane proporcje oszacowane dla modelu, w którym dopuszczono nie dwa, lecz jeden kanoniczny wymiar. Odpowiada to prezentowanemu w podrozdziałach 6.4 i 6.5 modelowi opartemu na pierwszej tablicy kanonicznej. Jeśli przykładowo, wielkość równą 0,0926 a odpowiadającą ojcom, którzy mają wykształcenie wyższe, zaś ich dzieci wybrały licea ogólnokształcące, przemnożymy przez liczbę badanych uczniów (4214), to otrzyma-

Rycina B.14
Fragment pliku wyników programu STATA określający proporcje pól modelu kanonicznego utworzonego na podstawie pierwszej tablicy kanonicznej

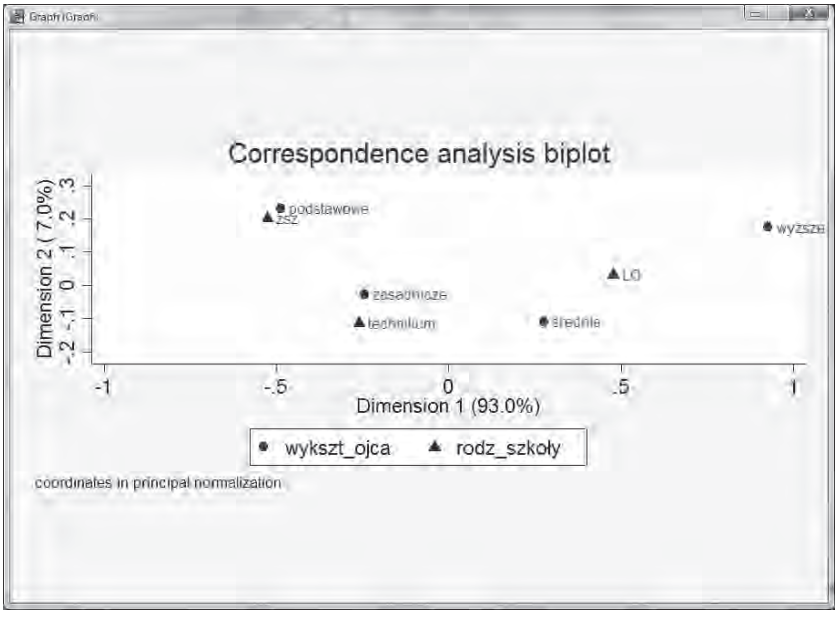
```
. estat table

Approximation for dim = 1 (normalized to overall sum = 1)

                |        LO  technikum        zsz
----------------+----------------------------------
      wyższe    |    0.0926     0.0185    -0.0034
      średnie   |    0.1522     0.0950     0.0312
    zasadnicze  |    0.1498     0.2402     0.1145
    podstawowe  |    0.0197     0.0590     0.0307
-------------------------------------------------
```

Dane dotyczą zależności między wykształceniem ojca a rodzajem szkoły ponadgimnazjalnej, do której uczęszcza dziecko. Badanie PISA 2006.

Rycina B.15
Wykres przedstawiający wyniki analizy korespondencji otrzymany za pomocą programu STATA



my liczebność tego pola w omawianym modelu kanonicznym równą 390 (por. tabela 6.4, część [4]). Rozwiązanie uzyskane w programie STATA pozwala więc łatwo oszacować liczebności pierwszego modelu kanonicznego. Przypomnijmy, że w wypadku rezultatów programu LEM wielkości te musieliśmy obliczać osobno, korzystając ze współrzędnych kanonicznych.

Jedną z zalet programu STATA stanowi możliwość przedstawienia rozwiązania w postaci graficznej. W tym celu do polecenia wywołującego procedurę analizy korespondencji należy dopisać parametr „plot". Polecenie przybierze wtedy następującą postać

```
camat woxszk, dim(2) norm(principal)
   rowname(wykszt_ojca) colname(rodz_szkoły) plot
```

W efekcie wykonania polecenia w podanej formie otworzy się dodatkowe okno z wykresem, który przedstawiony został na rycinie B.15. Porównanie tego wykresu z wykresem prezentowanym na rycinie 7.12 [b] pozwala zauważyć, że wykres z programu STATA ma odwróconą orientację. W lewym górnym narożniku – czyli w miejscu najbardziej eksponowanym – znaleźli się ojcowie o wykształceniu podstawowym, których dzieci kształcą się głównie w szkołach zasadniczych. Również estetyka wykresu nie jest przekonująca. Dlatego tworzone przez program wykresy traktować raczej należy jako wstępny ogląd struktury analizowanych tablic. W wypadku tablic przeznaczonych do prezentacji wykresy warto wykonać osobno, ustalając odpowiednio ich orientację, wielkość obszaru zajętego przez punkty, a także treść objaśnień przypisanych kategoriom. Kwestie te mają kluczowe znaczenie dla czytelności obrazu prezentowanego zjawiska.

Literatura cytowana

Agresti, Alan (1984). *Analysis of Ordinal Categorical Data*. New York: Wiley.

(1990). *Categorical Data Analysis*. New York: Wiley.

(2002). *Categorical Data Analysis*. Wydanie 2. Hoboken, New Jersey: Wiley.

(2007). *An Introduction to Categorical Data Analysis*. Wydanie 2. Hoboken, New Jersey: Wiley.

Agresti, Alan i Barbara Finlay (2008). *Statistical Methods for the Social Sciences*. Wydanie 4. Uppet Saddle River, New Jersey: Pearson Education.

Andrews, Frank M., Laura Klem, Terrence N. Davidson, Patrick M. O'Malley i Willard L. Rodgers (1981). *A Guide for Selecting Statistical Techniques for Analyzing Social Science Data*. Ann Arbor: Institute for Social Research, University of Michigan.

Arum, Richard, Adam Gamoran i Yossi Shavit (2007). „More inclusion than diversion: Expansion, differentiation, and market structure in higher education", [w:] Yossi Shavit, Richard Arum, Adam Gamoran i Gila Menahem (red.), *Stratification in Higher Education. A Comparative Study*. Stanford: Stanford University Press, s. 1–35.

Babbie, Earl (2008). *Podstawy badań społecznych*. Warszawa: Wydawnictwo Naukowe PWN.

Beh, Eric J. (2004). „Simple correspondence analysis: A bibliographic review". *International Statistical Review* 72: 257–284.

Benzécri, Jean-Paul (1969). „Statistical analysis as a tool to make patterns emerge from data", [w:] Satose Watanabe (red.), *Methodologies of Pattern Recognition*. New York: Academic Press, s. 35–60.

(1992). *Correspondence Analysis Handbook*. New York: Marcel Dekker.

Benzécri, Jean-Paul wraz z współpracownikami (1973). *L'Analyse des données. 1. La taxonomie, 2. L'Analyse des correspondances*. Paris: Dunod.

Bielby, William T., Robert M. Hauser i David L. Featherman (1977), „Response errors of Black and Nonblack males in models of the intergenerational transmission of social status", *American Journal of Sociology* 82: 1242–1288.

Billiet, Jaak (2007). „Non-response bias in cross-national surveys: designs for detection and adjustment in the European Social Survey". Referat na

konferencji: *Europejski Sondaż Społeczny. Procesy społeczne w początkach XXI wieku*. Warszawa: Instytut Filozofii i Socjologii PAN.

Bishop, Yvonne M., Stephen E. Fienberg i Paul W. Holland (1975). *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, Massachusetts: The MIT Press.

Blasius, Jörg (1994). „Correspondence analysis in social science research", [w:] Michael Greenacre i Jörg Blasius (red.), *Correspondence Analysis in the Social Sciences*. San Diego: Academic Press, s. 23–52.

Blasius, Jörg i Michael Greenacre (1994). „Computation of correspondence analysis", [w:] Michael Greenacre i Jörg Blasius (red.), *Correspondence Analysis in the Social Sciences*. San Diego: Academic Press, s. 53–78.

(2006a). *Multiple Correspondence Analysis and Related Methods* (red.), Boca Raton, Fl.: Chapman & Hall/CRC.

(2006b). „Correspondence analysis and related methods in practice", [w:] Michael Greenacre i Jörg Blasius (red.), *Multiple Correspondence Analysis and Related Methods*. Boca Raton, Fl.: Chapman & Hall/CRC, s. 3–40.

Bojko, Agnieszka (2006). „Using eye tracking to compare web page designs: A case study", *Journal of Usability Studies* 1, Issue 3: 112–120.

Bourdieu, Pierre (2006 [1979]). *Dystynkcja. Społeczna krytyka władzy sądzenia*. Warszawa: Wydawnictwo Scholar.

Box, George E. P. i Norman R. Draper (1987). *Empirical Model-Building and Response Surfaces*. New York: Wiley.

Böckenholt, Ulf i Yoshio Takane (1994). „Linear constraints in correspondence analysis", [w:] Michael Greenacre i Jörg Blasius (red.), *Correspondence Analysis in the Social Sciences*. San Diego: Academic Press, s. 112–127.

Brzeziński, Jerzy (2007). *Metodologia badań psychologicznych*. Wydanie V. Warszawa: Wydawnictwo Naukowe PWN.

Buck, R. Creighton (1980). „Sherlock Holmes in Babylon", *American Mathematical Monthly* 87: 335–345.

Buławski, Rajmund (1932). „Warstwy społeczne". *Kwartalnik Statystyczny* 9, zeszyt 2.

Charszewski, Adam (1931) [piszący jako A. Charczyński]. „Struktura społeczna młodzieży szkół średnich i wyższych". *Miesięcznik Literacki*, luty 1931, nr 15. Wybór obszernych fragmentów [w:] Dariusz Poliński (opr.), *Warunki pracy i życia w Polsce międzywojennej*. Warszawa 1980: Książka i Wiedza, s. 429–438.

Clausen, Sten-Erik (1998), *Applied Correspondence Analysis: An Introduction*. Series: Quantitative Applications in the Social Sciences, Vol. 121. Sage Publications.

Cohen, Jack K., R. Frank Falk i Charles F. Cortese (1976). „Reply to Tauber and Tauber". *American Sociological Review* 41: 889–893.

Cortese, Charles F., R. Frank Falk i Jack K. Cohen (1976). „Further considerations on the methodological analysis of segregation indices". *American Sociological Review* 41: 630–637.

Croxton, Frederick E. i Dudley J. Cowden (1955). *Applied General Statistics*. Englewood Cliffs, N. J.: Prentice-Hall.

Deming, W. Edwards (1984). *Statistical adjustment of data*. Dover Publications.

Deming, W. Edwards i Frederick F. Stephan (1940). „On a least squares adjustment of a sampled frequency table when the expected marginal totals are known". *The Annals of Mathematical Statistics* 11: 427–444.

Denzin, Norman K. i Yvonna S. Lincoln. (2005). "Introduction: The discipline and practice of qualitative research", [w:] Norman K. Denzin i Yvonna S. Lincoln (eds.), *The Sage Handbook of Qualitative Research*. Wydanie 3. Thousand Oaks, CA: Sage, s. 1–32.

Dietz, E. Jacquelin (2004). „Interview with Alan Agresti", *STATS: The Magazine for Students of Statistics* 39: 10–14.

Domański, Henryk (2000). „Wpływ kategorii »trudno powiedzieć« na wyniki analiz". *Ask* 9: 77–93.

(2007). *Struktura społeczna*. Warszawa: Wydawnictwo Naukowe Scholar.

(2009). *Społeczeństwa europejskie. Stratyfikacja i systemy wartości*. Warszawa: Wydawnictwo Naukowe Scholar.

Domański, Henryk i Dariusz Przybysz (2007). *Homogamia małżeńska a hierarchie społeczne.* Studia z Socjologii Ilościowej 1. Warszawa: Wydawnictwo IFiS PAN.

(2009). „Bariery zawierania małżeństw w Polsce w latach 1977–2007", *Studia Socjologiczne* nr 1(192): 53–85.

Domański, Henryk, Zbigniew Sawiński i Kazimierz M. Słomczyński (2007). *Nowa klasyfikacja i skale zawodów. Socjologiczne wskaźniki pozycji społecznej w Polsce*. Warszawa: Wydawnictwo IFiS PAN.

(2009). *Sociological Tools Measuring Occupations. New Classification and Scales*. Warsaw: IFiS Publishers.

Duncan, Otis Dudley (1979). „How destination depends on origin in the occupational mobility table". *American Journal of Sociology* 84: 793–803.

Duncan, Otis Dudley i Beverly Duncan (1955). „A methodological analysis of segregation indexes". *American Sociological Review* 20: 210–217.

Eckart, Carl i Gale Young (1936). „The approximation of one matrix by another of lower rank". *Psychometrika* 1: 211–218.

Ehrenberg, Andrew S. C. (1981). „The problem of numeracy". *The American Statistician* 35, No. 2: 67–71.

(1986). „Reading a table: An example". *Applied Statistics* 35: 237–244.
Falguerolles, Antoine de (2008). „L'analyse des données : before and around", *Electronic Journ@l for History of Probability and Statistics* 4, no. 2 (December). Dostępny na: www.emis.de/journals/JEHPS/
Featherman, David L. i Robert M. Hauser (1978). *Opportunity and Change*. New York: Academic Press.
Ferguson, George A. i Yoshio Takane (2007). *Analiza statystyczna w psychologii i pedagogice*. Warszawa: Wydawnictwo Naukowe PWN.
Fienberg, Stephen E. (1970). „An iterative procedure for estimation in contingency tables", *The Annals of Mathematical Statistics* 41: 907–917.
(1977). *The Analysis of Cross-Classified Categorical Data*. Cambridge: The MIT Press.
Firebaugh, Glenn (2008). *Seven Rules for Social Research*. Princeton i Oxford: Princenton University Press.
Fisher, R. A. (1940). „The precision of discriminant functions". *Annals of Eugenics* 10: 422–429.
Garcia-Álvares, Ercilia i Jordi López-Sintas (2002). „Contingency table: A two-way bridge between qualitative and quantitative methods". *Field Methods* 14: 270–287.
Garcia-Pérez, Miguel A. i Vicente Núñez-Antón (2003). „Cellwise residual analysis in two-way contingency tables". *Educational and Psychological Measurement* 63: 825–839.
Gautam, Shiva i George Kimeldorf (1999). „Some Results on the Maximal Correlation in 2 x k Contingency Tables". *The American Statistician* 53, No. 4: 336–341.
Gernert, Dieter (2009). „Ockham's Razor and its improper use". *Cognitive Systems* 7: 133–138.
Gilula, Zvi i Shelby J. Haberman (1986). „Canonical analysis of contingency tables by maximum likelihood". *Journal of the American Statistical Association* 81: 780–788.
(1988). „The analysis of multivariate contingency tables by restricted canonical and restricted association models". *Journal of the American Statistical Association* 83: 760–771.
Goodman, Leo A. (1965). „On the statistical analysis of mobility tables". *American Journal of Sociology* 70: 564–585.
(1978). *Analyzing Qualitative/Categorical Data. Log-Linear Models and Latent-Structure Analysis*. Cambridge, Massachusetts: Abt Books.
(1979). „Multiplicative models for the analysis of occupational mobility tables and other kinds of cross-classification tables". *American Journal of Sociology* 84: 804–819.

(1984). *The Analysis of Cross-Classified Data Having Ordered Categories*. Cambridge, Massachusetts: Harvard University Press.

(1986). „Some useful extensions of the usual correspondence analysis approach and the usual log-linear models approach in the analysis of contingency tables". *International Statistical Review* 54, No. 3 (December): 243–270.

(1987). „New methods for analyzing the intrinsic character of qualitative variables using cross-classified data". *American Journal of Sociology* 93: 529–583.

(1996). „A single general method for the analysis of cross-classified data: Reconciliation and synthesis of some methods of Pearson, Yule, and Fisher, and also some methods of correspondence analysis and association analysis". *Journal of the American Statistical Association* 91: 408–428.

(2000). „The analysis of cross-classified data: Notes on a century of progress in contingency table analysis, and some comments on its prehistory and its future", [w:] C. R. Rao i Gábor Székely (red.), *Statistics for the 21st Century. Methodologies for Applications of the Future.* New York, Basel: Marcel Dekker, s. 189–231.

Goodman, Leo A. i William H. Kruskal (1954). „Measures of Association for Cross Classifications". *Journal of the American Statistical Association* 49: 732–764.

(1959). „Measures of Association for Cross Classifications, II: Further Discussion and References". *Journal of the American Statistical Association* 54: 123–163.

(1963). „Measures of Association for Cross Classifications, III: Approximate Sampling Theory". *Journal of the American Statistical Association* 58: 310–364.

(1972). „Measures of Association for Cross Classifications, IV: Simplifications of Asymptotic Variances". *Journal of the American Statistical Association* 67: 415–421.

Góralski, Andrzej (1979). *Algorytmy i programy statystyki kwalitytatywnej*. Warszawa: Wydawnictwa Naukowo-Techniczne.

Górniak, Jarosław (2000). „Zastosowanie analizy korespondencji w badaniach społecznych i marketingowych". *Ask* 9: 115–134.

Górniak, Jarosław i Janusz Wachnicki (2008). *Pierwsze kroki w analizie danych. SPSS for Windows*. Wydanie piąte. Kraków: SPSS Polska.

Greenacre, Michael (1984). *Theory and Applications of Correspondence Analysis*. London: Academic Press.

(1993). *Correspondence Analysis in Practice*. London: Academic Press.

(1994). „Correspondence analysis and its interpretation", [w:] Michael Greenacre i Jörg Blasius (red.), *Correspondence Analysis in the Social Sciences*. San Diego: Academic Press, s. 3–22.
(2006). „From simple to multiple correspondence analysis", [w:] Jörg Blasius i Michael Greenacre (red.), *Multiple Correspondence Analysis and Related Methods*. Boca Raton, Fl.: Chapman & Hall/CRC, s. 41–76.

Greenacre, Michael i Jörg Blasius (1994). *Correspondence Analysis in the Social Sciences* (red.). San Diego: Academic Press.

Greenacre, Michael i Trevor Hastie (1987). „The geometric interpretation of correspondence analysis". *Journal of the American Statistical Association* 82: 437–447.

Grether, David M. (1976). „On the use of ordinal data in correlation analysis". *American Sociological Review* 41: 908–912.

GUS (2007). *Rocznik demograficzny 2007*. Warszawa: Główny Urząd Statystyczny (wydanie elektroniczne: www.stat.gov.pl).

Hand, David, Heikki Mannila i Padhraic Smyth (2005). *Eksploracja danych*. Warszawa: Wydawnictwa Naukowo-Techniczne.

Haberman, Shelby J. (1978). *Analysis of Qualitative Data*. Vol. 1. *Introductory Topics*. Orlando: Academic Press Inc.

Hauser, Robert M. (1978). „A structural model of the mobility table". *Social Forces* 56: 919–953.

Hill, M. O. (1974). „Correspondence analysis: A neglected multivariate method". *Applied Statistics* 23: 340–354.

Hirschfeld, H. O. (1935). „A connection between correlation and contingency". *Mathematical Proceedings of the Cambridge Philosophical Society* 31: 520–524.

Hout, Michael (1983). *Mobility Tables*. Sage University Paper series on Quantitative Application in the Social Sciences, 07–031. Beverly Hills and London: Sage.

Huff, Darrell (1954). *How to Lie with Statistics*. New York, London: W. W. Norton & Company.

Ireland, C. T. i S. Kullback (1968). „Contingency tables with given marginals", *Biometrika* 55(1): 179–188.

ISCED (1997). *International Standard Classification of Education*. UNESCO: Institute for Statistics.

Jahn, Julius A., Calvin F. Schmid i Clarence Schrag (1941). „The measurement of ecological segregation". *American Sociological Review* 12: 293–303.

Jencks, Christopher, Marshall Smith, Henry Acland, Mary Jo Bane, David Cohen, Herbert Gintis, Barbara Heyns i Stephan Michelson (1972). *Inequality: A Reassessment of the Effect of Family and Schooling in America*. New York: Basic Books.

Jencks, Christopher, Susan Bartlett, Mary Corcoran, James Crouse, David Eaglesfield, Gregory Jackson, Kent McClelland, Peter Mueser, Michael Olneck, Joseph Schwartz, Sherry Ward i Jill Williams (1979). *Who Gets Ahead? The Determinants of Economic Success in America*. New York: Basic Books.

Jones, F. Lancaster (1985). „New and (very) old mobility ratios: is there life after Benini". *Social Forces* 63: 838–850.

Karplus, Elizabeth, Robert Karplus i Warren Wollman (1974). „Intellectual development beyond elementary school IV: Ratio, the influence of cognitive style". *School Science and Mathematics* 74: 476–482.

Kendall, Maurice G. i William R. Buckland (1975). *Słownik terminów statystycznych*. Warszawa: Państwowe Wydawnictwo Ekonomiczne.

Kendall, Maurice G. i Alan Stuart (1979). *The Advanced Theory of Statistics. Volume 2. Inference and Relationship*. Fourth Edition. New York: Macmillan.

Keuzenkamp, Hugo A., Michael McAleer i Arnold Zellner (2001). „The enigma of simplicity", [w:] Arnold Zellner, Hugo A. Keuzenkamp i Michael McAleer (eds.), *Simplicity, Inference and Modelling*. Cambridge: University Press, s. 1–12.

Klatzky, Sheila R. i Robert W. Hodge (1971). „A canonical correlation analysis of occupational mobility". *Journal of the American Statistical Association* 66: 16–22.

Kopaliński, Władysław (2007). Słownik wyrazów obcych i zwrotów obcojęzycznych z almanachem. Warszawa: Rytm Oficyna Wydawnicza.

Koseła, Krzysztof i Krystyna Utzig (1980). „Skalowanie wielowymiarowe – zastosowania uprawnione i nieuprawnione". *Studia Socjologiczne* nr 2: 251–277.

Krauze, Tadeusz i Kazimierz M. Słomczyński (1985). „How far to meritocracy? Empirical tests of a controversial thesis". *Social Forces* 63: 623–642.

Kruskal, Joseph B. i Myron Wish (1978). *Multidimensional Scaling*. Sage University Paper series on Quantitative Application in the Social Sciences, 07–011. Beverly Hills and London: Sage Publications.

Kunovich, Sheri i Kazimierz M. Słomczyński (2007). „Systems of distribution and a sense of equity: A multilevel analysis of meritocratic attitudes in post-industrial societies". *European Sociological Review* 23: 649–663.

Labovitz, Sanford (1970). „The assignment of numbers to rank order categories". *American Sociological Review* 35: 515–524.
(1971). „In defense of assigning numbers to ranks". *American Sociological Review* 36 (June): 521–522.

Lancaster, H. O. (1969). *The Chi-squared Distribution*. New York: Wiley.

Lang, Thomas A. i Michelle Secic (2006). *How to Report Statistics in Medicine*. ACP Press.

Larose, Daniel T. (2006). *Odkrywanie wiedzy z danych. Wprowadzenie do eksploracji danych*. Warszawa: Wydawnictwo Naukowe PWN.

Lebart, Ludovic (2006). „Validation techniques in multiple correspondence analysis", [w:] Michael Greenacre i Jörg Blasius (red.), *Multiple Correspondence Analysis and Related Methods*. Boca Raton, Fl.: Chapman & Hall/CRC, s. 179–196.

(2008). „Exploratory multivariate data analysis from its origins to 1980: Nine contributions", *Electronic Journ@l for History of Probability and Statistics* 4, no. 2 (December). Dostępny na: www.emis.de/journals/JEHPS/

Levine, Joel H. (1993). *Exceptions are the Rule. An Inquiry into Methods in the Social Sciences*. Boulder, Colorado: Westview Press.

(2005). „Extended correlation: Not necessarily quadratic or quantitative". *Sociological Methods & Research* 34, No. 1: 31–75.

Lissowski, Grzegorz (1984). „Zastosowanie modeli logarytmiczno-liniowych do analizy związków między wieloma zmiennymi jakościowymi", *Studia Socjologiczne*, nr 2 (93): 239–263.

Lissowski, Grzegorz, Jacek Haman i Mikołaj Jasiński (2008). *Podstawy statystyki dla socjologów*. Warszawa: Wydawnictwo Naukowe Scholar.

Lutyńska, Krystyna (1999). *Odpowiedzi „trudno powiedzieć" w badaniach CBOS. Wybrane Problemy.* Warszawa: Centrum Badania Opinii Społecznej.

Maddala, G. S. (2006). *Ekonometria*. Warszawa: Wydawnictwo Naukowe PWN.

Marsh, Catherine (1988). *Exploring Data. An Introduction to Data Analysis for Social Scientists*. Polity Press and Basil Blackwell Inc.

Maung, K. (1941). „Measurement of association in a contingency table with special reference to the pigmentation of hair and eye colours of Scottish school children". *Annals of Eugenics* 11: 189–223.

Mayer, Lawrence S. (1970). „Comment on »The assignment of numbers to rank order categories«". *American Sociological Review* 35: 916–917.

(1971). „A note on treating ordinal data as interval data". *American Sociological Review* 36: 519–520.

McGarigal, Kevin, Sam Cushman i Susan Stafford (2000). *Multivariate Statistics for Wildlife Ecology Research*. New York: Springer-Verlag.

McKnight, Patric E., Katherine M. McKnight, Souraya Sidani i Aurelio José Figueredo (2007). *Missing Data. A Gentle Introduction*. New York: Guilford Press.

Merton, Robert W. (2002). *Teoria socjologiczna i struktura społeczna*. Warszawa: Wydawnictwo Naukowe PWN.

Mirkin, Borys (2001). „Eleven ways to look at the Chi-squared coefficient for contingency tables". *The American Statistician* 55, No. 2: 111–120.

Moore, David S., George P. McCabe i Bruce Craig (2007). *Introduction to the Practice of Statistics*. Wydanie 6. New York: W. H. Freeman.

Mosteller, Frederick (1968). „Association and estimation in contingency tables". *Journal of the American Statistical Association* 63: 1–28.

Murtagh, Fionn (2007). „Review of »Multiple Correspondence Analysis and Related Methods«, by M. Greenacre and J. Blasius". *Psychometrika* 72: 275–277.

(2008). „Origins of modern data analysis linked to the beginning and early development of computer science and information engineering", *Electronic Journ@l for History of Probability and Statistics* 4, No 2 (December). Dostępny na: www.emis.de/journals/JEHPS/.

Nenadić, Oleg i Michael Greenacre (2006). „Computation of multiple correspondence analysis with Code in R", [w:] Michael Greenacre i Jörg Blasius (red.), *Multiple Correspondence Analysis and Related Methods*. Boca Raton, Fl.: Chapman & Hall/CRC, s. 523–551.

Niepokojczycki, Wojciech i Zbigniew Sawiński (1984). *Przygotowanie danych socjologicznych do analiz na maszynie cyfrowej*. Zespół Badań Socjologicznych nad Problemami Oświaty, Zeszyt 35. Warszawa: Instytut Socjologii Uniwersytetu Warszawskiego.

Noelting, Gerald (1980). „The development of proportional reasoning and the ratio concept". *Educational Studies in Mathematics* 11: 217–253.

Nosal, Czesław S. (1987). „Interpretacja zależności między zbiorami zmiennych w ramach modelu analizy kanonicznej", [w:] Jerzy Brzeziński (red.), *Wielozmiennowe modele statystyczne w badaniach psychologicznych*. Warszawa–Poznań: PWN, s.152–170.

O'Brien, Robert M. (1979). „The use of Pearson's R with ordinal data". *American Sociological Review* 44: 851–857.

Perek-Białas, Jolanta i Urszula Korzeniecka (2006). „Wykorzystanie metod ilościowych w badaniach marketingowych w Polsce". *Ask* 15: 51–73.

Piaget, Jean (1972). *The Psychology of the Child*. New York: Basic Books.

PKW (2005). *Wyniki wyborów do Sejmu Rzeczypospolitej Polskiej w dniu 25 września 2005*. Państwowa Komisja Wyborcza (wydanie elektroniczne: www.wybory2005.pkw.gov.pl).

Quételet, Adolphe (1832). „Sur la Possibilité de Mesurer l'Influence des Causes qui Modifient les Éléments Sociaux". *Lettre à M. Willermé de l'Institut de France*. Bruxelles.

(1849). „Letters addressed to H. R. H. the Grand Duke of Saxe Coburg and Gotha on the Theory of Probabilities as Applied to the Moral and Political Sciences", (tłumaczył z francuskiego Olinthus Gregory Downs). London: Charles and Edwin Layton.

Radelet, Michael L. i Glenn L. Pierce (1991). „Choosing those who will die: Race and the death penalty in Florida". *Florida Law Review* 43: 1–34.

Richardson, M. i G. F. Kuder (1933). „Making a rating scale that measures", *Personnel Journal* 12: 36–40.

Ritov, Ya'acov i Zvi Gilula (1993). „Analysis of contingency tables by correspondence models subject to order constraints". *Journal of the American Statistical Association* 88: 1380–1387.

Rodgers, Joseph Lee i W. Alan Nicewander (1988). „Thirteen ways to look at the correlation coefficient", *The American Statistician* 42: 59–66.

Sakoda, James M. (1981). „A generalized index of dissimilarity". *Demography* 18: 245–50.

Santos, J. Reynaldo A. (2000). „Getting the most out of multiple response questions". *Journal of Extention* 38 (http://www.joe.org/).

Sarmanov, O. V. (1958). „Maksimalnyj koefficient korrelacji (Niesimmietricznyj sluczaj)". *Doklady AN SSSR* 121: 52–55.

Sawiński, Zbigniew (1979). *Koncepcja maksymalizacji współczynników korelacji*. Praca magisterska. Warszawa: Instytut Socjologii, Uniwersytet Warszawski.

(1981). „Mierniki ruchliwości społeczno-zawodowej", *Studia Socjologiczne* nr 2(81): 171–187.

(1984). *Koncepcja alokacji merytokratycznej. Część I. Prezentacja koncepcji*. Zespół Badań Socjologicznych nad Problemami Oświaty, Zeszyt 34. Warszawa: Instytut Socjologii, Uniwersytet Warszawski.

(1985). *Analiza maksymalnej korelacji*. Zespół Badań Socjologicznych nad Problemami Oświaty, Zeszyt 39. Warszawa: Instytut Socjologii, Uniwersytet Warszawski.

(1986). *Pomiar i skalowanie wykształcenia w badaniach socjologicznych*. Warszawa: Uniwersytet Warszawski.

(1988). „Błędy pomiaru w badaniach osiągnięć społeczno-zawodowych". *Studia Socjologiczne* nr 2 (109): 45–62.

(1994). „Postrzeganie roli szkół niepaństwowych w systemie nierówności edukacyjnych". *Kwartalnik Pedagogiczny* nr 1/2 (151–152): 115–123.

(1996). „Sondaże telefoniczne". *Ask* nr 1/1996:7–36.

(2004). „Źródła rozwoju metodologii badań marketingowych", [w:] Paweł B. Sztabiński, Franciszek Sztabiński, Zbigniew Sawiński (red.), *Nowe metody, nowe podejścia badawcze w naukach społecznych*. Warszawa: Wydawnictwo IFiS PAN, s. 23–28.

(2007a). „Badania trackingowe", [w:] Dominika Maison i Artur Noga-Bogomilski (red.), *Badania marketingowe. Od teorii do praktyki*. Gdańsk: Gdańskie Wydawnictwo Psychologiczne, s. 103–118.

(2007b). „Metody doboru prób w badaniach marketingowych", [w:] Dominika Maison i Artur Noga-Bogomilski (red.), *Badania marketingowe. Od teorii do praktyki*. Gdańsk: Gdańskie Wydawnictwo Psychologiczne, s. 119–134.

(2007c). *Europejski Sondaż Społeczny 2006. Opis schematu doboru próby oraz dyskusja jej realizacji* (tekst niepublikowany). Warszawa: Instytut Filozofii i Socjologii, Polska Akademia Nauk.

(2008). „Zmiany systemowe a nierówności w dostępie do wykształcenia", [w:] Henryk Domański (red.), *Zmiany stratyfikacji społecznej w Polsce*. Warszawa: Wydawnictwo IFiS PAN, s. 13–43.

(2009). „Nierówności edukacyjne w teoriach struktury społecznej". *Studia Socjologiczne*, nr 1(192): 89–114.

Sawiński, Zbigniew i Henryk Domański (1986). *Wymiary struktury społecznej. Analiza porównawcza*. Wrocław: Ossolineum.

(1989). „Dimensions of Social Stratification: a Comparative Analysis". *International Journal of Sociology* 19: 1–102.

Sawiński, Zbigniew i Marta Zahorska (1991). „Wiesław Wiśniewski (28 IV 1921—2 XI 1991)", *Studia Socjologiczne* nr 3/4.

Schweitzer, Sybil i Donald G. Schweitzer (1971). „Comment on the Person R in random number and precise functional scale transformations". *American Sociological Review* 36: 518–519.

Simpson, Edward H. (1951). „The interpretation of interaction in contingency tables". *Journal of the Royal Statistical Society*, Ser. B 13: 238–241.

Słomczyński, Kazimierz M. (1983). *Pozycja zawodowa i jej związki z wykształceniem*. Warszawa: Instytut Filozofii i Socjologii PAN.

(1989). *Social Structure and Mobility: Poland, Japan and the United States*. Warszawa: Polish Academy of Sciences, Institute of Philosophy and Sociology.

(2002). „Introduction: Social structure, its changes and linkages", [w:] Kazimierz M. Słomczyński (ed.), Social Structure: Changes and Linkages. The Advanced Phase of the Post-Communist Transition in Poland. Warsaw: IFiS Publishers, s. 11–28.

Słomczyński, Kazimierz M., Ireneusz Białecki, Henryk Domański, Krystyna Janicka, Bogdan W. Mach, Zbigniew Sawiński, Joanna Sikorska i Wojciech Zaborowski (1989). *Struktura społeczna: schemat teoretyczny i warsztat badawczy*. Warszawa: Polska Akademia Nauk, Instytut Filozofii i Socjologii.

Słomczyński, Kazimierz M., Tadeusz K. Krauze i Zbigniew Peradzyński (1988) „The dynamics of status trajectory: a model and its empirical assessment", *European Sociological Review* 4: 46–64.

Sober, Elliott (2001). „What is the problem of simplicity?", [w:] Arnold Zellner, Hugo A. Keuzenkamp i Michael McAleer (eds.), *Simplicity, Inference and Modelling*. Cambridge: University Press, s. 13–31.

Steczkowski, Jan i Aleksander Zeliaś (1981). *Statystyczne metody analizy cech jakościowych*. Warszawa: Państwowe Wydawnictwo Ekonomiczne.

Stigler, Stephen M. (1986). *The History of Statistics. Measurement of Uncertainty before 1900*. Cambridge, Massachusetts: Harvard University Press.
(1999). *Statistics on the Table. The History of Statistical Concepts and Methods*. Cambridge, Massachusetts: Harvard University Press.

Stoop, Ineke A. L. (2005). *The Hunt for the Last Respondent. Nonresponse in Sample Surveys*. The Hague: Social and Cultural Planning Office.

Szacki, Jerzy (1981). *Historia myśli socjologicznej.* Tom 1 i 2. Warszawa: Wydawnictwo Naukowe PWN.

Sztabiński, Paweł B. (2004). „Metodologia badania Europejski Sondaż Społeczny". *Ask* 13: 27–37.

Sztabiński, Paweł B. i Franciszek Sztabiński (2006). „Europejski Sondaż Społeczny: integracja w dziedzinie badań", [w:] Henryk Domański, Antonina Ostrowska i Paweł B. Sztabiński (red.), *W środku Europy? Wyniki Europejskiego Sondażu Społecznego*. Warszawa: Wydawnictwo IFiS PAN, s. 13–26.

Sztabiński, Paweł B., Zbigniew Sawiński i Franciszek Sztabiński [red.] (2005). *Fieldwork jest sztuką*. Warszawa: Wydawnictwo IFiS PAN.

Szymczak, Mieczysław [red.] (1978). *Słownik języka polskiego*. Warszawa: Państwowe Wydawnictwo Naukowe.

Tauber, Karl E. i Alma F. Tauber (1976). „A practitioner's perspective on the index of dissimilarity". *American Sociological Review* 41: 884–889.

Tenenhaus, Michel i Forrest W. Young (1985). „An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis and other methods for quantifying categorical multivariate data", *Psychometrika* 50: 91–119.

Ter Braak, Cajo J. F. (1987). „The analysis of vegetation-environment relationships by canonical correspondence analysis". *Vegetatio* 69: 69–77.

Thompson, Bruce (1984). *Canonical Correlation Analysis. Uses and Interpretation*. Sage University Paper series on Quantitative Applications in the Social Sciences, 47. Beverly Hills i London: Sage.

Timofiejuk, Igor, Mirosława Lasek i Marek Pęczkowski (1997). *Miary statystyczne*. Warszawa: Główny Urząd Statystyczny.

Toth, Jeffrey P. (2000) „Nonconscious Forms of Human Memory", [w:] Endel Tulving i Fergus Craik (red.), *The Oxford Handbook of Memory*. Oxford University Press, s. 245–261.

Tovee, M. J., S. Reinhardt, J. L. Emery i P. L. Cornelissen (1998). „Optimum body-mass index and maximum sexual attractiveness". *Lancet* 352 (9127): 548.

Treiman, Donald J. (2009). *Quantitative Data Analysis: Doing Social Research to Test Ideas*. San Francisco: Jossey-Bass.

Umesh, U. N. (1995). „Predicting nominal variable relationships with multiple response". *Journal of Forecasting* 14, Issue 7: 585–596.

Umesh, U. N., Martin Tan i Donald E. Stem Jr. (1992). „Analysis of multiple response in marketing research: Estimating the degree of association". *Marketing Letters* 3, No. 2 (April): 107–114.

Van Koppen, Frans (2006). „Miscellaneous old babylonian period documents", [w:] Mark W. Chavalas (red.), *The ancient Near East: historical sources in translation*. Blackwell Publishing, s. 107–133.

Van Meter, Karl M., Marie-Ange Schiltz, Philippe Cibois i Lise Mounier (1994). „Correspondence analysis: A history and French sociological perspective", [w:] Michael Greenacre i Jörg Blasius (ed.), *Correspondence Analysis in the Social Sciences*. San Diego: Academic Press, s. 128–137.

Vapnik, Vladimir (2006). *Estimation of Dependencies Based on Empirical Data*. New York: Springer.

Vermunt, Jeroen K. (1997). *LEM 1.0: A general program for the analysis of categorical data*. Tilburg: Tilburg University.

Wainer, Howard (1992). „Understanding graphs and tables". *Ed Researcher* 21: 14–23.

Wang, Chunlei (2002). „Maritocratic allocation of persons to jobs", [w:] Kazimierz M. Słomczyński (ed.), *Social Structure: Changes and Linkages. The Advanced Phase of the Post-Communist Transition in Poland.* Warsaw: IFiS Publishers, s. 57–78.

Wansbeek, Tom i Michel Wedel (1999). „Marketing and econometrics: Editors' introduction". *Journal of Econometrics* 89: 1–14.

Wesołowski, Włodzimierz i Tadeusz Krauze (1981). „Socialist society and the meritocratic principle of remuneration", [w:] Gerald D. Berreman (red.), *Social Inequality. Comparative and Developmental Approaches*. New York: Academic Press, s. 337–349.

White, Stuart (2008). *Równość*. Warszawa: Wydawnictwo Sic!

Whittaker, Robert H. (1967). „Gradient analysis of vegetation". *Biological Review* 42: 207–264.

Williams, E. J. (1952). „Use of scores for the analysis of association in contingency tables". *Biometrika* 39: 274–289.

Wolfram, Stephen (2002). *A New Kind of Science*. Champaign Il.: Wolfram Media.

Yancy, D. Edwards i Greg M. Allenby (2003). „Multivariate analysis of multiple response data". *Journal of Marketing Research*, Vol. XL, No 3 (August): 321–334.

Yule, G. Udny (1912). „On the methods of measuring association between two attributes", *Journal of the Royal Statistical Society* 75: 579–652.
(1921 [1910]). *Wstęp do teorji statystyki*. Tłumaczenie Z. Limanowskiego. Warszawa: Gebethner i Wolff.

Zieliński, Ryszard (1972). *Tablice statystyczne*. Warszawa: Państwowe Wydawnictwo Naukowe.

Indeks

E

F

G

H

I

J

K

N

O

S

Y

Z

Zbigniew Sawiński, adiunkt w Instytucie Filozofii i Socjologii PAN; twórca wielu rozwiązań w zakresie doboru prób, implementacji technik badawczych oraz metod analizy wyników badań, prezentowanych między innymi w artykułach: *Analiza maksymalnej korelacji* (1985), *Sondaże telefoniczne* (1996), *Źródła rozwoju metodologii badań marketingowych* (2004), *Badania trackingowe* (2007), a także w książce *Sociological Tools Measuring Occupations* (2009, wspólnie z Henrykiem Domańskim i Kazimierzem M. Słomczyńskim). Jest współautorem podręcznika *Fieldwork jest sztuką* (2005), stanowiącego kompendium wiedzy na temat standardów realizacji badań w Polsce. Osobny obszar jego zainteresowań obejmuje problematykę formowania się nierówności społecznych. Publikacje z tego zakresu to między innymi: *Dimensions of Social Stratification* (1989; wspólnie z Henrykiem Domańskim), *Zmiany systemowe a nierówności w dostępie do wykształcenia* (2008) czy *Nierówności edukacyjne w teoriach struktury społecznej* (2009).

(...) Właściwości prezentowanych metod znane są specjalistom od analiz ilościowych, a na przykład analiza korespondencji znajduje coraz szersze zastosowanie w praktyce badawczej. Nikt jednak przed Sawińskim nie powiązał ich w całość, nie pokazał kryjących się za nimi założeń i możliwości aplikacyjnych (...) Z oryginalnością poznawczą łączą się walory dydaktyczne książki. Jej tematyczna struktura ma logikę wykładu. Autor zaczynając od spraw elementarnych przechodzi do bardziej złożonych, omawiając krok po kroku właściwości rozpatrywanych metod (...), ilustruje posługiwanie się nimi na konkretnych przykładach, mówi jak należy je interpretować, a jakich wniosków unikać (...) Książka ta może być również podręcznikowym *exemplum* pisania o rzeczach niełatwych w odbiorze. Autor posługuje się językiem przyjaznym dla czytelnika – logicznym, precyzyjnymi zdaniami, z robieniem przerywników (...). A równocześnie napisane jest to z dystansem, dzięki komentarzom odautorskim i ogólniejszej refleksji (...)

z recenzji
prof. dr. hab. Henryka Domańskiego

ISBN 978-83-7683-013-1