

PROPOZYCJA BADANIA SPOŁECZNEGO: JAKOŚCIOWA ANALIZA DANYCH

W rozdziale tym poznaliśmy niektóre dostępne badaczom społecznym metody jakościowej analizy danych. Ponieważ podczas pisania tej części propozycji badawczej nie przeanalizowaliśmy swoich danych, nie możemy też jeszcze nic powiedzieć na temat wniosków, które wyciągniemy. Możemy jednak opisać swoje wstępne plany dotyczące analizy. Mówię o „wstępnych” planach, ponieważ w miarę zbierania danych i pojawiania się wzorców możemy nieco zmienić kierunek. W niektórych przypadkach twoja analiza rozpocznie się w momencie dokonywania obserwacji i/lub zbierania innych danych albo też możesz zaplanować, że przed przystąpieniem do analizy najpierw zakończysz fazę zbierania danych.

W tym miejscu należy wskazać, czy planujesz zastosować określoną metodę analizy, taką jak teoria ugruntowana, semiotyka, analiza konwersacji itp. Jeśli zamierzasz użyć określonego programu QDA, wspomnij o tym również tutaj.

PYTANIA I ĆWICZENIA

1. Przypomnij sobie badania Goffmana dotyczące reklam związanych z płcią, a następnie znajdź i przeanalizuj kilka zdjęć reklamowych z czasopism lub internetu. Jaki zauważasz w nich związek między płcią a statusem społecznym?

2. Przyjrzyj się kwestii homoseksualizmu ujętej w Księdze Kapłańskiej, a następnie przedstaw sposoby możliwego ustrukturyzowania tych badań w przekrojowej analizie przypadków.
3. Wyobraź sobie, że prowadzisz przekrojącą analizę przypadków na podstawie dokumentów związanych z rewolucjami, takich jak Deklaracja Niepodległości czy Deklaracja Praw Człowieka i Obywatela. Określ kluczowe pojęcia, które można zakodować w następującym zdaniu:

Ilekoć wskutek biegu wypadków koniecznym się staje dla jakiegoś narodu, by zerwał więzy polityczne łączące go z innym narodem i zajął wśród potęg ziemskich oddzielne i równorzędne stanowisko, do którego upoważniają go prawa natury i jej Bóg, to właściwy respekt wymaga, aby naród ten podał powody, które zmusiły go do oderwania się⁷⁰.

4. Wejdź na stronę World Press Review (www.worldpress.org) i wybierz jakiś kontrowersyjny temat wiadomości omawiany w kilku gazetach. Sprawdź, czy zdołasz zidentyfikować cechy tych gazet (takie jak poglądy polityczne, region), które mogą wyjaśniać różne punkty widzenia wyrażane na konkretny temat.

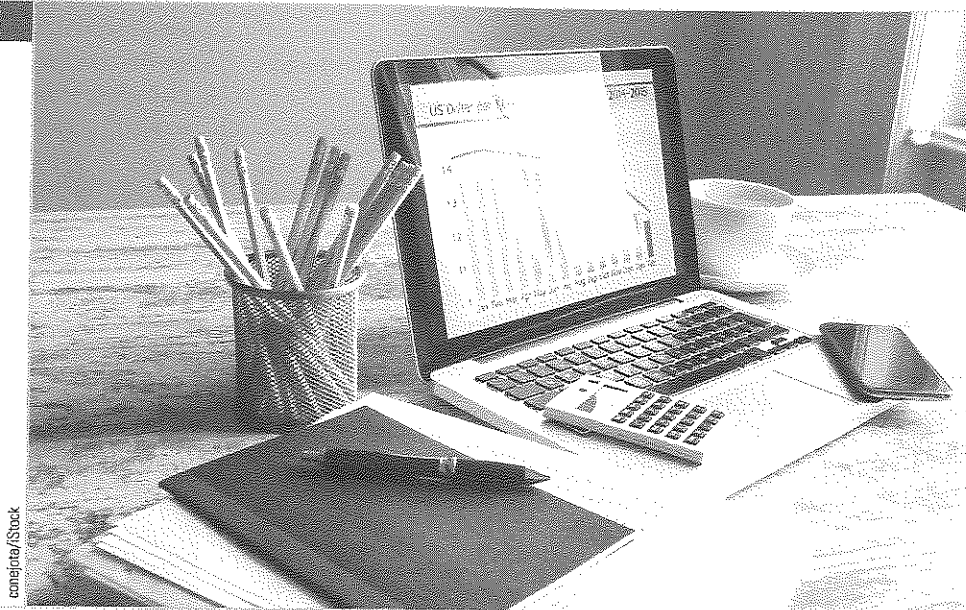
⁷⁰ Encyklopedia Historii Stanów Zjednoczonych Ameryki, red. A. Bartnicki, K. Michałek, I. Rusinowa, Warszawa: Egross Morex 1992, s. 67–68.

ROZDZIAŁ 14

Analiza danych ilościowych

PRZEGLĄD HOLOGRAFICZNY

Dane na temat życia społecznego na potrzeby statystyczne często są konwertowane do postaci liczbowej. W niniejszym rozdziale zaczniemy od procesu kwantyfikacji danych, a następnie przejdziemy do ich analizy. Analiza ilościowa może mieć charakter opisowy lub eksplanacyjny, a także obejmować jedną, dwie lub kilka zmiennych. Nasz opis sposobu przeprowadzania analiz ilościowych rozpoczniemy od kilku prostych, ale wyjątkowo przydatnych sposobów przetwarzania danych w celu wyciągnięcia wniosków badawczych.



Wprowadzenie

Kwantyfikacja danych

- Tworzenie kategorii kodowych
- Opracowanie książki kodowej
- Wprowadzanie danych

Analiza jednozmiennowa

- Rozkład wartości
- Tendencja centralna
- Rozproszenie
- Zmienne ciągłe i dyskretne
- Szczegółowość a czytelność

Porównywanie podgrup

- Łączenie kategorii odpowiedzi
- Sposoby postępowania z kategorią „nie wiem”
- Opisy liczbowe w badaniach jakościowych

Analiza dwuzmiennowa

- Procentowanie w tabeli
- Tworzenie i odczytywanie tabel dwuzmiennowych (krzyżowych)

Wprowadzenie do analizy wielozmiennowej

Diagnozy społeczne

Etyka i analiza danych ilościowych

Cele nauczania

Po przestudiowaniu tego rozdziału będziesz w stanie...

- podać przykład ilustrujący kwantyfikację danych,
- zidentyfikować i omówić kilka aspektów analizy jednozmiennowej,
- wyjaśnić korzyści wynikające z zastosowania porównania podgrup zamiast analiz jednozmiennowych,
- opisać i zilustrować różnicę między porównaniami podgrup a analizami dwuzmiennowymi,
- wymienić i wyjaśnić dodatkowe zalety analizy wielozmiennowej w porównaniu z analizą dwuzmiennową,
- podać przykład diagnozy socjologicznej i tego, jak można ją wykorzystać w odniesieniu do spraw społecznych,
- zdefiniować problemy etyczne, które mogą pojawić się w związku z analizami ilościowymi.

Wprowadzenie

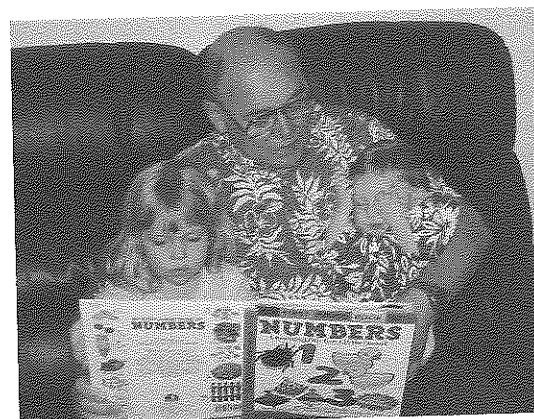
W rozdziale 13 poznaliśmy logikę i techniki, za pomocą których badacze społeczni analizują zebrane dane jakościowe. W tym rozdziale przyjrzymy się **analizie ilościowej**, czyli technikom, za pomocą których badacze przekształcają dane w liczby i poddają je analizom statystycznym.

Na początek przyjrzymy się **kwantyfikacji** – procesowi konwersji danych do formatu liczbowego. Wiąże się to z przekształceniem danych z nauk społecznych w formę nadającą się do odczytu maszynowego – formę, którą mogą odczytywać i przetwarzać komputery i podobne urządzenia używane w analizie ilościowej.

W dalszej części rozdziału przedstawimy logikę i niektóre techniki ilościowej analizy danych – zaczynając od najprostszego przypadku, analizy jednozmiennowej, która obejmuje jedną zmienną, a przez analizę dwuzmiennową, która obejmuje dwie zmienne, aż po krótkie wprowadzenie do analizy wielozmiennowej, czyli jednoczesnego badania kilku zmiennych, takich jak *wiek*, *wykształcenie* i *uprzedzenia*. Następnie omówimy diagnostykę socjologiczną. Na koniec przyjrzymy się etyce ilościowej analizy danych.

Zanim jednak będziemy mogli przeprowadzić jakąkolwiek analizę, musimy skwantyfikować nasze dane. Przejdźmy więc do podstawowych kroków związanych z przekształcaniem danych w formy nadające się do odczytu maszynowego – nadające się do przetwarzania i analizy komputerowej.

Analiza ilościowa – liczbowe przedstawienie i manipulacja obserwacjami w celu opisanie i wyjaśnienia zjawisk, które odzwierciedlają te obserwacje



Niektórzy uczniowie chętniej niż inni zajmują się statystyką

Kwantyfikacja danych

Obecnie analiza ilościowa niemal zawsze wykonywana jest przez programy komputerowe, takie jak Pakiet Statystyczny dla Nauk Społecznych (IBM SPSS Statistics⁷¹) oraz MicroCase. Aby jednak programy te mogły wykonać swoją magię, muszą być w stanie odczytać dane zebrane podczas badań. Jeśli na przykład skończył się proces zbierania ankiet w twoim badaniu, niektóre z danych, na przykład wiek lub dochód, są z natury liczbowe. Podczas gdy informacje i znaczniki wyboru na kwestionariuszu mają charakter jakościowy, nabazgrany wiek można łatwo przekształcić w dane ilościowe.

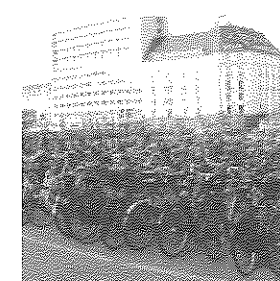
Inne dane również można łatwo określić ilościowo: przekształcenie mężczyzny i kobiety w „1” i „2” nie jest żadną filozofią. Równie łatwo badacze mogą przypisywać określenia liczbowe takim zmiennym, jak *przynależność religijna*, *partia polityczna* czy *region kraju*.

⁷¹ W skrócie nazywany po prostu SPSS.

Zastanów się

W rozdziale 13 dostrzeżliśmy kilka nieodłącznych braków związanych z danymi ilościowymi. Wady te związane były przede wszystkim z ich standaryzacją oraz powierzchownością w obliczu zróżnicowanej i głębokiej rzeczywistości społecznej. Czy można dowiedzieć się czegoś sensownego z danych, które rezygnują ze znaczących szczegółów, aby umożliwić manipulacje numeryczne?

Zob. ramkę *Zastanów się*... raz jeszcze pod koniec tego rozdziału.



Picture © Earl Babbie

Niektóre dane stanowią jednak większe wyzwanie. Jeśli respondent ankiety powie ci, że jego zdaniem największym problemem stojącym przed Woodbury w stanie Vermont jest „zanikająca warstwa ozonowa”, komputer nie jest w stanie przetworzyć tej odpowiedzi liczbowo. Musisz mu to przetłumaczyć, kodując odpowiedzi. Kodowanie omówiliśmy już w związku z analizą treści (rozdział 11) i ponownie w związku z jakościową analizą danych (rozdział 13). Teraz przyjrzymy się kodowaniu związanemu z analizą ilościową, która różni się od pozostałych dwóch przede wszystkim celem konwersji surowych danych na liczby.

Podobnie jak w przypadku analizy treści, zadaniem kodowania ilościowego jest zredukowanie szerokiej gamy jednostkowych elementów informacji do bardziej ograniczonego zestawu atrybutów składających się na zmienną. Załóżmy na przykład, że prowadzący sondaż badacz pyta respondentów: „Jaki jest Pana(-i) zawód?” Odpowiedzi na takie pytanie będą bardzo zróżnicowane. Chociaż możliwe będzie przypisanie każdemu zgłoszonemu zawodowi osobnego kodu numerycznego, procedura ta nie ułatwi analizy, która zwykle opiera się na fakcie, że wielu badanych ma te same cechy.

Zmienna *zawód* ma wiele z góry ustalonych schematów kodowania. Jeden z takich schematów rozróżnia zawody kierownicze i specjalistów, zawody biurowe, niewymagające wysokich kwalifikacji itd. Inny z kolei rozróżnia rozmaite sektory gospodarki: przemysł, ochronę zdrowia, edukację, handel itd. Jeszcze inne łączą te dwa schematy. Korzystanie z ustalonego schematu kodowania daje przewagę w postaci możliwości porównania wyników twoich badań z wynikami innych.

Wybrany przez ciebie schemat kodowania zawodów powinien odpowiadać teoretycznym pojęciom analizowanym w twoim badaniu. W niektórych przypadkach może wystarczyć zakodowanie wszystkich zawodów jako pracowników umysłowych lub fizycznych. W innych może okazać się użyteczne rozróżnienie na „samozatrudnionych” i „zatrudnionych przez inną osobę”. Z kolei badacz pacyfista

mógłby chcieć wiedzieć tylko, czy zawód związany jest z przemysłem zbrojeniowym.

Chociaż schemat kodowania powinien być dostosowany do konkretnych wymagań prowadzonej analizy, należy pamiętać o jednej ogólnej wytycznej. Jeśli dane są kodowane w sposób bardzo szczegółowy, ale sama analiza tego nie wymaga, kategorie kodów zawsze można łączyć. Jeśli jednak dane są zakodowane w stosunkowo niewielu ogólnych kategoriach, podczas analizy nie będziesz mieć możliwości odtworzenia oryginalnych szczegółów. Aby zostawić sobie możliwie szeroki wybór, dobrze jest kodować dane bardziej szczegółowo, niż wymaga tego planowana analiza.

Tworzenie kategorii kodowych

Do procesu kodowania zasadniczo można podejść na dwa sposoby. Po pierwsze, można wyjść od dość rozbudowanego schematu kodowania zgodnego z twoim celem badawczym. W ten sposób wspomniany badacz pacyfista mógłby zakodować zawody wedle ich związku z przemysłem zbrojeniowym. Można również użyć gotowego klucza kodowego, aby umożliwić sobie porównanie wyników własnych badań z badaniami poprzedników.

Inna metoda polega na generowaniu kodów na podstawie danych, podobnie jak w rozdziale 13. Załóżmy, że zadaliśmy studentom w uniwersyteckiej ankiecie do samodzielnego wypełnienia pytanie o największy obecnie problem na ich uczelni. Oto niektóre z możliwych odpowiedzi:

Za wysokie czesne
Za mało miejsc parkingowych
Władze nie wiedzą, co robią
Opiekunowie roczników są wiecznie nieosiągalni
Zbyt mała oferta wykładów
Karałuchy w akademikach
Zbyt wysokie wymagania
Nieświeże jedzenie w stołówce
Za drogie książki
Za mało stypendiów

Poświęć chwilę na przyjrzenie się tym odpowiedziom, i sprawdź, czy reprezentują one jakieś szersze kategorie. Musisz mieć świadomość, że nie ma właściwej odpowiedzi: na podstawie tych odpowiedzi można wygenerować wiele schematów kodowania.

Zacznijmy od pierwszej odpowiedzi – „za wysokie czesne”. Na jaki szerszy zakres problemowy wskazuje ta odpowiedź? Jedną z oczywistych możliwości są „problemy finansowe”. Czy są jakieś inne odpowiedzi, które pasują do tej kategorii? W tabeli 14.1 zaznaczono te odpowiedzi, które mieszczą się w kategorii „problemy finansowe”.

W ogólniejszym ujęciu możemy także potraktować pierwszą kategorię jako odzwierciedlającą problemy pozaakademickie. Kategoryzacja taka byłaby trafna, gdyby twoje badania miały za przedmiot rozróżnienie problemów akademickich i pozaakademickich. Wówczas kody odpowiedzi wyglądałyby jak w tabeli 14.2.

Zwróć uwagę, że w tabeli 14.2 nie zakodowałem odpowiedzi „za drogie książki”, ponieważ odpowiedź tę można zakwalifikować do obu kategorii. Książki są częścią programu akademickiego, jednak ich ceny już nie. Wskazuje to na potrzebę uściślenia tworzonego przez nas schematu kodowania. Zależnie od celu nasze szczególne zainteresowania mogłyby koncentrować się na określeniu problemów związanych z nauczaniem akademickim – wówczas ten przypadek zakodowalibyśmy jako „akademickie”. Równie dobrze moglibyśmy jednak interesować się raczej określeniem problemów pozaakademickich i odpowiednio zakodować tę odpowiedź. W jeszcze innym przypadku moglibyśmy stworzyć odrębną kategorię dla odpowiedzi zawierających zarówno element akademicki, jak i pozaakademicki.

Jest jednak jeszcze inna możliwość: problemy pozaakademickie możemy podzielić na takie, które

TABELA 14.1

Odpowiedzi studentów, które mogą zostać zakodowane jako „problemy finansowe”

	problemy finansowe
Za wysokie czesne	X
Za mało miejsc parkingowych	
Władze wydziału nie wiedzą, co robią	
Kierownicy studiów są wiecznie nieosiągalni	
Zbyt mała oferta wykładów	
Karaluchy w akademikach	
Zbyt wysokie wymagania	
Nieświeże jedzenie w stołówce	
Za drogie książki	X
Za mało stypendiów	X

TABELA 14.2

Problemy studentów zakodowane jako „akademickie” i „pozaakademickie”

	akademickie	pozaakademickie
Za wysokie czesne		X
Za mało miejsc parkingowych		X
Władze wydziału nie wiedzą, co robią	X	
Kierownicy studiów są wiecznie nieosiągalni	X	
Zbyt mała oferta wykładów	X	
Karaluchy w akademikach		X
Zbyt wysokie wymagania	X	
Nieświeże jedzenie w stołówce		X
Za drogie książki		X
Za mało stypendiów		X

dotyczą kwestii administracyjnych, oraz takie, które wiążą się ze stanem infrastruktury uczelni. Sposób zakodowania 10 pierwszych odpowiedzi w takim przypadku przedstawia tabela 14.3.

Jak widać na tych kilku przykładach, istnieje wiele schematów kodowania zbioru danych. Twój wybór powinien korespondować z twoim celem badawczym i stanowić odbicie logiki wyłaniającej się z samych danych. Często, w miarę postępów procesu

TABELA 14.3

Problemy pozaakademickie kodowane jako „administracja” lub „infrastruktura”

	akademickie	administracja	infrastruktura
Za wysokie czesne		X	
Za mało miejsc parkingowych			X
Władze wydziału nie wiedzą, co robią	X		
Kierownicy studiów są wiecznie nieosiągalni	X		
Zbyt mała oferta wykładów	X		
Karaluchy w akademikach			X
Zbyt wysokie wymagania	X		
Nieświeże jedzenie w stołówce			X
Za drogie książki	X		
Za mało stypendiów		X	

kodowania, będzie trzeba modyfikować kategorie kodowe. Jednak zawsze gdy zmieniasz listę kategorii, musisz przejrzeć zakodowane już dane, by sprawdzić, czy zmiany zostały prawidłowo wprowadzone.

Podobnie jak zbiór wartości składających się na zmienną czy też kategorie odpowiedzi w zamkniętym pytaniu kwestionariusza, kategorie kodów powinny być zarazem wyczerpujące, jak i wzajemnie się wykluczające. Każda kodowana jednostka informacji powinna należeć do jednej i tylko jednej kategorii. Problem pojawia się, gdy wydaje się, że odpowiedź pasuje do więcej niż jednej lub nie pasuje do żadnej kategorii – oba te fakty sygnalizują rozbieżności między twoimi danymi a schematem kodowym.

Jeżeli masz szczęście i dysponujesz pomocą w procesie kodowania, wypada zapoznać swoich koderów z definicjami kategorii kodowych i nauczyć ich poprawnie używać tych kategorii. W tym celu wyjaśnij znaczenie każdej z kategorii i podaj kilka przykładów każdej z nich. Aby upewnić się, że koderzy dokładnie cię zrozumieli, zakoduj z góry kilka przypadków. Następnie poproś ich, aby zakodowali te same przypadki, nie uprzedzając jednak, że sam masz już to zrobione wcześniej. Na koniec porównaj ich pracę z własną. Jakiegokolwiek niezgodności będą wskazywać na nieefektywność komunikacji między tobą a koderami. Jednak nawet jeśli ty i koderzy zgadzacie się wręcz doskonale, najlepiej sprawdzić kodowanie przynajmniej części przypadków w toku procesu kodowania.

Jeżeli jednak nie masz takiego szczęścia i nikt nie pomaga ci w kodowaniu, i tak trzeba jakoś zweryfikować swoją wiarygodność jako koderki. Nikt nie jest doskonały, a już na pewno nie pełen entuzjazmu badacz na tropie wielkiego odkrycia. Załóżmy, że badasz nową grupę religijną i masz wrażenie, że istnieje duże prawdopodobieństwo, że osoby pozbawione normalnej rodziny będą postrzegały tę grupę jako jej substytut. Niebezpieczeństwo polega na tym, że zawsze, gdy natkniesz się na badanego, który zadeklaruje brak rodziny, będziesz podświadomie starać się znaleźć w jego wypowiedziach dowód potwierdzający fakt, że grupa religijna jest substytutem rodziny. Jeśli zatem jest to w ogóle możliwe, postaraj się, by ktoś inny zakodował część twoich przypadków, by sprawdzić, czy osoba ta dokonała takich samych wyborów jak ty.

Opracowanie książki kodowej

Produktem końcowym procesu kodowania jest przekształcenie danych w kody liczbowe. Kody te odpowiadają wartościom składającym się na zmienne, którym z kolei przypisane są pozycje w zbiorze danych. **Książka kodowa**⁷² to dokument, który opisuje

⁷² Wśród badaczy i często jest to nazywane w Polsce kluczem kodowym – przyp. M.J.L.

pozycje zmiennych i wymienia kody przyporządkowane wartościom składającym się na zmienne.

Książka kodowa spełnia dwie zasadnicze funkcje. Po pierwsze, jest to podstawowy przewodnik używany w procesie kodowania. Po drugie, w toku samej analizy służy pomocą w odszukiwaniu zmiennych i interpretacji kodów w zbiorze danych. Jeżeli w trakcie analizy danych zdecydujesz się skorelować dwie zmienne, książka kodowa podpowie ci, gdzie znajdziesz te zmienne i co oznaczają kody.

Ilustracja 14.1 przedstawia część książki kodowej, utworzoną z dwóch zmiennych z Generalnego Sondażu Społecznego. Przykład ten ukazuje kilka typowych elementów książki kodowej, chociaż oczywiście nie istnieje jeden prawidłowy format.

Zwróć przede wszystkim uwagę na to, że każda zmienna określona jest przez skrótową nazwę, jak POLVIEWS czy ATTEND. Na przykład fakt uczestniczenia respondentów w obrządkach religijnych możemy określić, posługując się zmienną ATTEND. Przykład ten stanowi zastosowanie formatu ustalonego przez Generalny Sondaż Społeczny, który został przeniesiony do SPSS. Inne zbiory danych i/lub programy do analizy mogą nadawać zmiennym inny format. Niektóre z nich jako nazwy zmiennych zamiast skrótów stosują na przykład kody liczbowe. Musisz jednak dysponować jakimś narzędziem identyfikacji, które pozwoli ci odnaleźć i wykorzystać konkretne zmienne.

Ponadto każda książka kodowa powinna zawierać pełną definicję każdej zmiennej. W przypadku kwestionariusza definicja składa się z literalnie przytoczonego pytania, ponieważ, jak się już przekonaliśmy, sformułowanie pytania wywiera duży wpływ na odpowiedź. W przypadku POLVIEWS wiesz, że respondent otrzymał stronę z kilkoma kategoriami politycznymi z prośbą, aby wybrał tę, która pasuje do niego najlepiej.

Książka kodowa wskazuje też wartości składające się na każdą ze zmiennych. Na przykład w POLVIEWS respondent mógł opisać swoje poglądy polityczne jako „zdecydowanie liberalne”, „liberalne”, „umiarkowanie liberalne” itd.

I na koniec, zwróć także uwagę, że każda ze zmiennych ma też oznaczenie liczbowe. W POLVIEWS „zdecydowanie liberalny” to kategoria kodowa 1. Tego rodzaju kody liczbowe wykorzystywane są w różnego rodzaju procesach przetwarzania danych. Gdybyś na przykład postanowił połączyć kategorie od 1 do 3 (wszyscy respondenci „liberalni”),

Książka kodowa – dokument używany do przetwarzania i analizy danych, który określa lokalizację różnych elementów danych w pliku danych. Zazwyczaj książka kodowa identyfikuje lokalizację elementów danych i znaczenie kodów używanych do reprezentowania różnych atrybutów zmiennych

POLVIEWS	ATTEND
Dużo się ostatnio słyszy o liberalach i konserwatystach. Przedstawie ci siedmiostopniową skalę, porządkującą poglądy ludzi od skrajnie liberalnych (punkt 1) do skrajnie konserwatywnych (punkt 7). Gdzie umieścisz(a)byś siebie na tej skali?	Jak często uczestniczysz w nabożeństwach?
1. Zdecydowanie liberalny	0. Nigdy
2. Liberalny	1. Rzadziej niż raz w roku
3. Umiarkowanie liberalny	2. Mniej więcej raz-dwa razy w roku
4. Zwolennik środka, centrum polityczne	3. Kilka razy w roku
5. Umiarkowanie konserwatywny	4. Mniej więcej raz w miesiącu
6. Konserwatywny	5. Dwa-trzy razy w miesiącu
7. Zdecydowanie konserwatywny	6. Prawie co tydzień
8. Nie wiem	7. Co tydzień
9. Brak odpowiedzi	8. Kilka razy w tygodniu lub częściej
	9. Nie wiem, brak odpowiedzi

ILUSTRACJA 14.1

Fragment książki kodowej

łatwiej to przeprowadzić na liczbach niż na długich nazwach⁷³.

Wprowadzanie danych

Oprócz przekształcania danych w formę liczbową badacze zainteresowani analizą ilościową muszą również konwertować dane do formatu nadającego się do odczytu maszynowego, co pozwoli komputerowi odczytywać dane i manipulować nimi. Istnieje wiele sposobów na wykonanie tego kroku, w zależności od pierwotnej postaci danych i programu komputerowego, którego użyjesz do analizy danych. Postaram się opisać to krok po kroku. Jeśli podejdziesz się tego zadania, musisz być w stanie dostosować swoją pracę do konkretnego źródła danych i używanego programu.

Jeśli twoje dane zostały zebrane za pomocą kwestionariusza, trzeba zakodować sam kwestionariusz. Następnie specjaliści od wprowadzania danych (w tym ty sam/sama) mogliby wprowadzić dane, powiedzmy, do macierzy danych SPSS lub do arkusza kalkulacyjnego Excel, który później zostałby zaimportowany do SPSS.

Czasami do zbierania danych badacze społeczni używają arkuszy skanowania optycznego. Arkusze te można skanować bezpośrednio do urządzeń przetwarzających czarne znaki na dane, które można zaimportować do programu analitycznego. Procedura

Analiza jednozmiennowa – analiza pojedynczej zmiennej, której celem jest opis. Jej przykładami są rozkłady częstości, średnie i miary dyspersji. Inne typy analiz to analizy dwuzmiennowa i wielozmiennowa

⁷³ Albo w sytuacji kiedy chcemy coś policzyć – np. średnią – przyp. M.J.L.

ta sprawdza się jednak tylko w przypadku osób, które w miarę swobodnie korzystają z takich arkuszy i zwykle ogranicza się do pytań zamkniętych.

Czasami wprowadzanie danych następuje w już w trakcie procesu ich zbierania. Na przykład podczas wywiadów telefonicznych wspomaganych komputerowo (CATI) ankietę wprowadza odpowiedzi bezpośrednio do komputera, gdzie dane są zestawiane do analizy (zob. rozdział 9). Jeszcze łatwiejsza sytuacja jest w przypadku sondażu przeprowadzanego online, w którym respondenci wprowadzają własne odpowiedzi bezpośrednio do bazy danych, bez potrzeby angażowania ankietera lub osoby wprowadzającej dane.

Gdy dane zostaną już w pełni skwantyfikowane i wprowadzone do komputera, naukowcy mogą rozpocząć analizę ilościową. Przyjrzyjmy się zatem trzem przypadkom wspomnianym już na początku tego rozdziału: analizie jednozmiennowej, dwuzmiennowej i wielozmiennowej.

Analiza jednozmiennowa

Najprostsza forma analizy ilościowej, **analiza jednozmiennowa**, polega na opisywaniu przypadku w kategoriach pojedynczej zmiennej, ze szczególnym uwzględnieniem rozkładu jej wartości. Gdyby przedmiotem pomiaru była na przykład *pleć*, sprawdzilibyśmy, ilu badanych było mężczyznami, a ile kobiet.

Rozkład wartości

Podstawowy format prezentacji danych jednozmiennowych wymaga uwzględnienia wszystkich przypadków indywidualnych, co oznacza konieczność wyszczególnienia wartości badanej zmiennej

dla każdego przypadku. Weźmy na przykład dane Generalnego Sondażu Społecznego (GSS) dotyczące uczestnictwa w nabożeństwach, ATTEND.

Ilustracja 14.2 pokazuje, w jaki sposób można uzyskać te dane, korzystając z omówionego wcześniej internetowego programu analitycznego Berkeley SDA. (Dostęp do tego programu można uzyskać pod adresem <http://sda.berkeley.edu/sdaweb>).

Na ilustracji zobaczysz, że jako zmienną Row wprowadzono ATTEND, a filtr wyboru ograniczyłem do analizy danych zebranych w GSS w 2014 roku. Szybko zauważysz skutki tego działania.

Ilustracja 14.3 ukazuje tabelaryczną odpowiedź na naszą prośbę o dostęp do danych. Możemy na przykład stwierdzić, że 26,2% badanych twierdzi, że nigdy nie uczestniczyło w nabożeństwach. Gdy przejdziemy z kolei w dół tabeli, zauważymy, że 17,6% odpowiedziało, że uczęszcza na nie co tydzień. Aby uprościć wyniki, możemy połączyć ostatnie trzy kategorie i stwierdzić, że 28,8% uczęszcza na nabożeństwa „mniej więcej raz w tygodniu”.

Opis tego, ile razy odpowiedź pojawia się w próbie, nazywamy **rozkładem częstości**. Czasami rozkład ten najłatwiej przedstawić na wykresie. Ilustracja 14.4 została stworzona przez program SDA po wybraniu odpowiednich pól w opcjach wykresów jak na ilustracji 14.2. Pionowa skala po lewej stronie wykresu wskazuje procent respondentów wybierających każdą z odpowiedzi wymienionych wzdłuż poziomej osi wykresu. Przyjrzyj się przez chwilę, w jaki sposób procenty na ilustracji 14.3 odpowiadają wysokości słupków z ilustracji 14.4.

Program ten oferuje również inne możliwości graficzne. Na ilustracji 14.2 mogliśmy wybrać typ wykresu i zamiast „wykres słupkowy” wskazać „wykres kołowy”.

Ponieważ wykresy kołowe są szczególnie odpowiednie do zmiennych nominalnych, ilustracja 14.5 pochodzi z innego badania. Prowadząc badania nad religiami w mieście Navi Mumbai w Indiach, demograf Prem Saxena (2012: 20) uznał, że wykres kołowy najlepiej zobrazuje dominację hinduizmu nad innymi wyznaniami.

SDA Frequencies/Crosstabulation Program
Help: General / Recoding Variables

Row: ATTEND (Required)

Column:

Control:

Selection Filter(s): YEAR(2014)

Weight: COMPTWT - Composite weight: WTSSALL * OVERSAMP * FOR

Output Options

Chart Options

Decimal Options

Run the Table Clear Fields

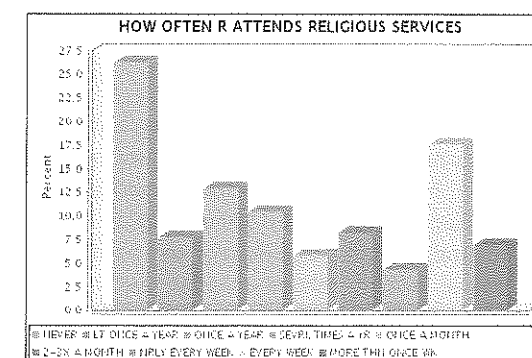
ILUSTRACJA 14.2

Okno wniosku o analizę jednozmiennową ATTEND, 2014

Frequency Distribution	
Cells contain: -Column percent -Weighted N	Distribution
0: nigdy	26.2 660.6
1: rzadziej niż raz do roku	7.7 193.4
2: raz do roku	12.9 326.9
3: kilka razy w roku	10.4 263.7
4: raz w miesiącu	5.8 146.8
5: 2-3 razy w miesiącu	8.2 207.0
6: prawie co tydzień	4.3 109.1
7: co tydzień	17.6 443.4
8: kilka razy w tygodniu	6.9 175.1
łącznie	100.0 2,526.0

ILUSTRACJA 14.3

Częstotliwość uczęszczania w nabożeństwach, 2014



ILUSTRACJA 14.4

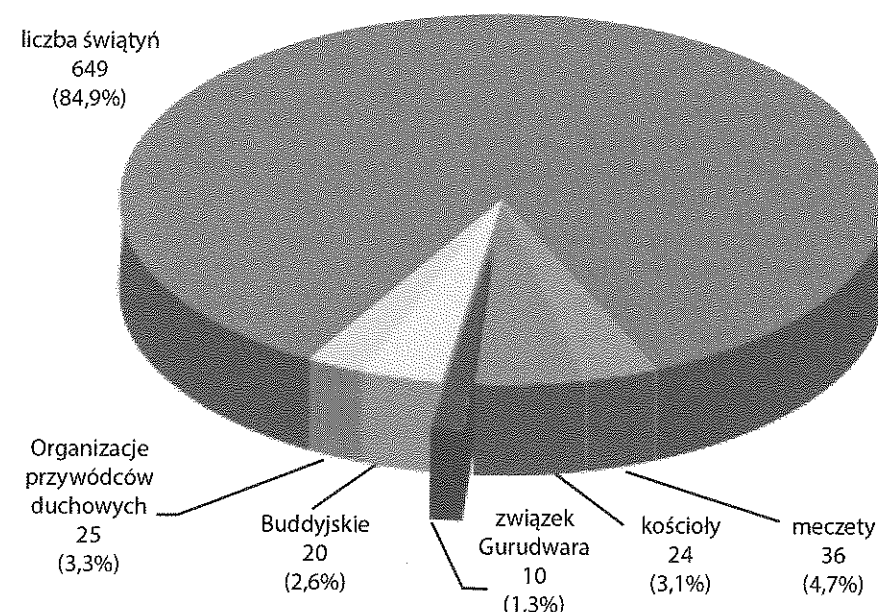
Wykres słupkowy GSS ATTEND, 2014

Tendencja centralna

Poza prostym opisem ogólnego rozkładu wartości, zwanym niekiedy *częstościami brzegowymi* lub po prostu *marginami*, być może zechcesz przedstawić swoje dane w formie **wartości** średniej lub tendencji centralnej. Pojęcie tendencji centralnej jest ci już

Rozkład częstości – opis, ile razy różne atrybuty zmiennej są obserwowane w próbie. Jego prostym przykładem byłby raport, że 53% próby stanowili mężczyźni, a 47% kobiety. Innym przykładem może być raport, zgodnie z którym 15 badanych miało populację poniżej 10 000, 23 miało populację od 10 000 do 25 000 itd.

Wartość przeciętna – mało precyzyjne pojęcie sugerujące typową lub normalną tendencję centralną. Przykładami konkretnych przeciętnych wartości matematycznych są *średnia*, *dominanta* i *mediana*



ILUSTRACJA 14.5

Wykres kołowy ukazujący liczbę i odsetek instytucji religijnych według typu

Źródło: P. Saxena (2012), *Social, Religious and Regional Institutions in Navi Mumbai* (Navi Mumbai, India: City and Industrial Development Corporation of Maharashtra), Ilustracja 2.3, s. 20

znane z rozmaitych przeciętnych, których używasz w życiu codziennym, by wyrazić „typową” wartość zmiennej. Na przykład w baseballu przeciętna celność pałkarza na poziomie 0,300 oznacza, że zawodnik ten trafia w piłkę średnio trzy razy na dziesięć. W trakcie sezonu pałkarz mógłby mieć dłuższe okresy bez żadnych trafień oraz takie, kiedy trafiał za każdym razem. Z upływem czasu jednak tendencja centralna celności pałkarza może być opisana jako trzy trafienia w każdych dziesięciu próbach. Podobnie średnia twoich ocen wyraża „typową” wartość wszystkich twoich ocen wziętych, mimo że niektóre z nich to piątki, inne – czwórki, a może masz nawet jedną lub dwie trójki (dobrze wiem, że nigdy nie dostajesz mniej niż trójkę).

Tego rodzaju uśrednienie poprawniej nazywa się średnimi arytmetycznymi (wynik dzielenia sumy

Średnia – uśredniona wartość obliczana przez zsumowanie wartości kilku obserwacji i podzielenie przez liczbę obserwacji. Jeśli masz teraz średnią ocen 4,0 na podstawie 10 przedmiotów i uzyskasz ocenę 2 z jednego przedmiotu, twoja nowa średnia ocen wyniesie 3,6

Dominanta – uśredniona wartość reprezentująca najczęściej obserwowaną wartość lub atrybut. Jeśli próbka obejmuje 1000 protestantów, 275 katolików i 33 Żydów, to kategorią modalną jest „protestant”

Mediana – średnia reprezentująca wartość „środkowego” przypadku w zbiorze obserwacji uporządkowanym według rang. Jeśli wiek pięciu mężczyzn wynosi 16, 17, 20, 54 i 88 lat, mediana wyniosłaby 20 (z kolei średnia wyniosłaby 39)

wartości przez ogólną liczbę przypadków). Jednak **średnia** to tylko jeden ze sposobów pomiaru tendencji centralnej lub wartości „typowych”. Dwie inne możliwości to **dominanta**, zwana czasem modalną (najczęściej występująca wartość), oraz **mediana** (środkowa wartość w uporządkowanym rozkładzie wszystkich zaobserwowanych wartości). A oto w jaki sposób na podstawie zbioru danych wyliczamy te trzy uśrednione wartości.

Załóżmy, że prowadzisz eksperyment, w którym badanymi są nastolatki. Ich wiek wynosi od 13 do 19 lat, zgodnie z poniższą tabelą:

wiek	liczba
13	3
14	4
15	6
16	8
17	4
18	3
19	3

Teraz, kiedy znasz już faktyczny wiek każdego z 31 badanych, jak określiłbyś ogólnie ich „przeciętny” wiek? Spójrzmy na trzy możliwe odpowiedzi na to pytanie.

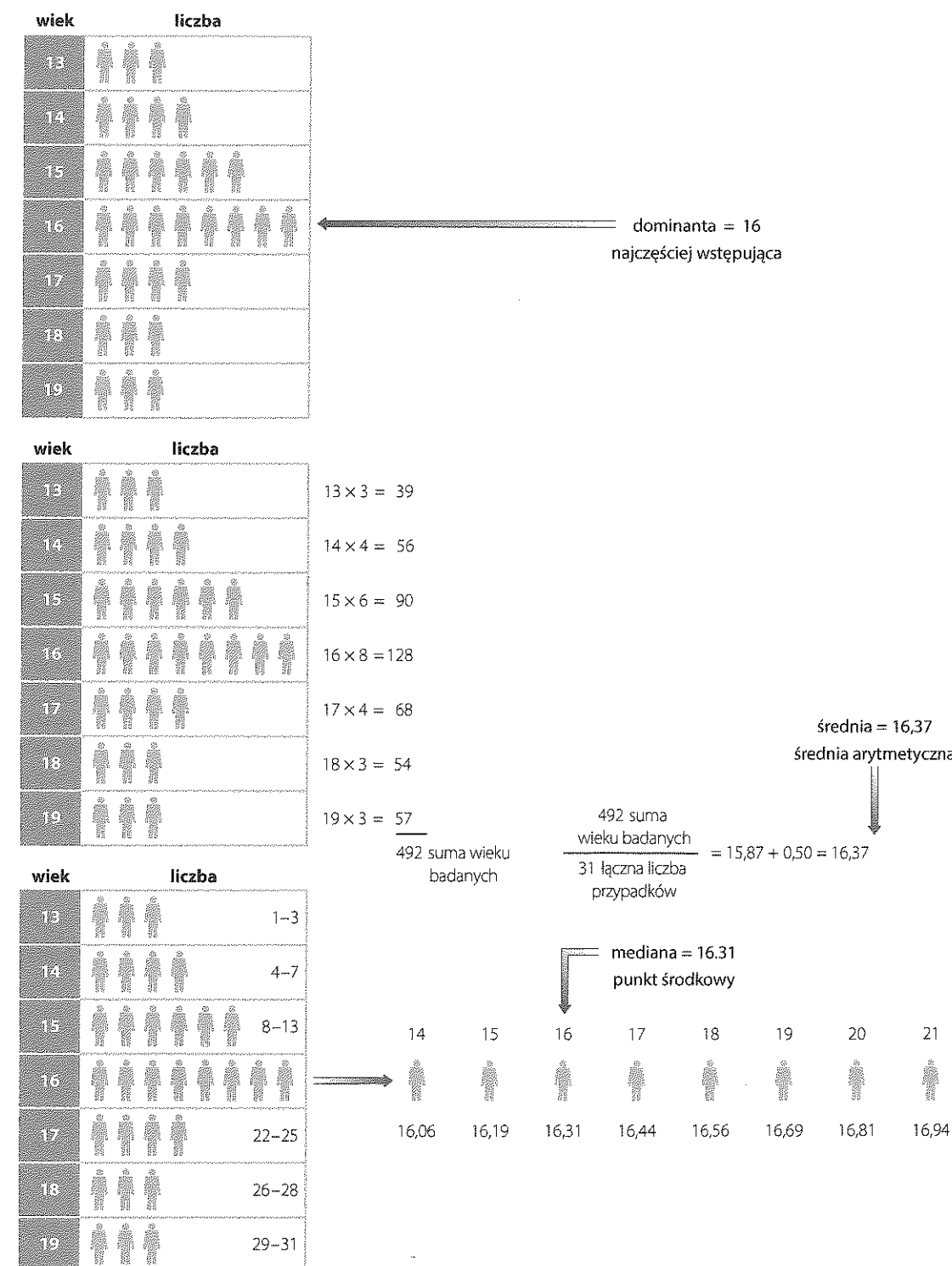
Wartość przeciętna, którą najłatwiej obliczyć, to dominanta, czyli najczęściej występująca wartość. Jak widzisz, szesnastolatków było więcej niż osób w jakimkolwiek innym wieku, więc dominanta

wieku wynosi 16, co ukazano na ilustracji 14.6. Technicznie rzecz ujmując, dominanta wieku obejmuje kategorię „16”, do której mogą należeć także osoby, które są bliżej 17. niż 16. roku życia, ale nie miały jeszcze 17. urodzin.

Ilustracja 14.6 także przedstawia wyliczenie średniej. Obejmuje ono trzy kroki: (1) pomnóż

każdy wiek przez liczbę badanych w tym wieku, (2) dodaj wszystkie uzyskane wyniki i (3) podziel sumę przez liczbę badanych.

W przypadku wieku potrzebne jest dodatkowe założenie. Omawiając dominantę zwróciłem uwagę, że osoby określające siebie „trzynastolatkami” to zarówno ci, którzy mają dokładnie 13 lat, jak i ci, którzy



ILUSTRACJA 14.6

Trzy „przeciętne”

są tuż przed 14 urodzinami. Co więcej, można zakładać, że w całym kraju grupa trzynastolatków rozkłada się równomiernie na przestrzeni jednego roku, co daje przeciętny wiek 13,5 roku. Prawdopodobieństwo ta zachodzi w każdej z grup wiekowych. Z tego też powodu wydaje się, że do ostatecznego wyniku należy dodać 0,5 roku, co, jak pokazano na ilustracji 14.6, daje średni wiek 16,37.

Trzeci rodzaj wartości środkowej, mediana, reprezentuje wartość „środkową”: połowa jest powyżej, połowa poniżej. Gdybyśmy znali dokładny wiek każdego badanego (na przykład 17 lat i 124 dni), moglibyśmy uporządkować wszystkich 31 badanych według wieku, a medianą całej grupy byłby wiek środkowego badanego.

Jak jednak widzisz, nie znamy dokładnego wieku – nasze dane są pod tym względem „danymi grupowymi”. Na przykład trzy osoby, które nie są dokładnie w tym samym wieku, zostały umieszczone w kategorii „trzynastolatki”.

Ilustracja 14.6 obrazuje zasadę obliczenia mediany danych grupowych. Jako że jest 31 badanych, osobą środkową byłby numer 16, gdyby uszeregowano ich według wieku – 15 nastolatków byłoby starszych, a 15 młodszych. Spójrz na dolny fragment rysunku 14.6, a zobaczysz, że środkową osobą jest jedno z ośmiorga szesnastolatków. Przyglądając się dokładniej tej konkretnej grupie, można zauważyć, że numer 16 to trzecia osoba od lewej.

Ponieważ nie znamy dokładnego wieku badanych w tej grupie, statystyczna konwencja nakazuje nam zakładać, że wiek rozkłada się równomiernie w całej tej grupie. W tym przypadku możliwy wiek badanego wynosi od 16 lat i 0 dni do 16 lat i 364 dni. Dokładniej rzecz ujmując, skala możliwości wynosi 364/365 dni. Do celów praktycznych spokojnie możemy to nazwać jednym rokiem.

Gdyby ośmiu badanych w tej grupie zostało równomiernie rozłożonych od jednej granicy do drugiej, byłoby od siebie oddaleni o jedną ósmą roku – interwał wynosiłby 0,125 roku. Spójrz na ilustrację, a przekonasz się, że jeżeli umieścimy pierwszego badanego o pół interwału od dolnej granicy, po czym dodamy pełną wartość interwału, aby obliczyć wiek każdego kolejnego badanego, ostatni wypada pół interwału od górnej granicy.

Właśnie obliczyliśmy hipotetycznie dokładny wiek każdego z ośmiu badanych – zakładając równy rozkład ich wieku. W tym momencie możemy zauważyć, że wiek środkowego badanego wynosi 16,31 i jest to mediana wieku całej grupy.

Rozproszenie – rozkład wartości wokół jakiejś centralnej wartości, takiej jak średnia. Jej zakres jest prostym przykładem miary rozproszenia. Można więc podać, że średni wiek grupy wynosi 37,9 lat, a zakres od 12 do 89 lat

Kiedy z kolei całkowita liczba badanych jest parzysta, oznacza to oczywiście, że nie ma przypadku środkowego. By zatem otrzymać medianę, należy wyliczyć średnią z dwóch wartości po obu stronach punktu środkowego w uszeregowanym zbiorze danych. Załóżmy na przykład, że w naszej grupie znalazłby się jeszcze jeden dziewiętnastolatek, co dałoby w sumie liczbę 32 przypadków. Punkt środkowy wypadłby wtedy między badanymi numer 16 i 17. W tej sytuacji medianę możemy policzyć w następujący sposób: $(16,31 + 16,44) \div 2 = 16,38$.

Jak widać na ilustracji 14.6, trzy mierniki tendencji centralnej dają dla naszego zbioru danych trzy różne wartości. Jest to częsta sytuacja, aczkolwiek nie oznacza to, że zawsze mamy z nią do czynienia. Zatem który z tych mierników najlepiej oddaje wartość „przeciętną”? Ogólniej: który miernik tendencji centralnej powinniśmy przedkładać nad inne? Odpowiedź zależy od typu danych i celu analizy. Na przykład prezentując średnią, trzeba mieć świadomość, że jest ona podatna na oddziaływanie wartości skrajnych – kilku bardzo małych lub bardzo dużych liczb. By podać tylko jeden przykład: średnia wartość majątku mieszkańca Mediny w stanie Waszyngton wynosi ponad milion dolarów. Gdybyś jednak odwiedził Medinę, stwierdziłbyś zapewne, że „przeciętny” mieszkaniec tej miejscowości raczej nie pasuje do twoich wyobrażeń o milionerach. Wysokość tej średniej stanowi odbicie wpływu jednego przypadku ekstremalnego wśród liczącego 3000 mieszkańców Mediny – jest nim Bill Gates, właściciel Microsoftu, którego majątek szacuje się na kilkadziesiąt miliardów dolarów. Zdecydowanie lepszy wgląd w sytuację ogółu mieszkańców Mediny dałaby ci mediana.

Przykład ten powinien stanowić dobrą ilustrację potrzeby starannego wyboru pomiędzy różnymi miarami tendencji centralnej. Więcej informacji na temat różnych sytuacji, w których można zastosować każdą z nich, znajdziesz w podręczniku do statystyki.

Rozproszenie

Wartości średnie pozwalają badaczowi zredukować surowe dane do poręczniejszej formy: pojedyncza liczba (lub cecha) może reprezentować wszystkie szczegółowe dane zebrane na temat konkretnej zmiennej. Ułatwienie to uzyskujemy jednak pewnym kosztem, ponieważ na podstawie wartości średniej czytelnik nie jest w stanie odtworzyć oryginalnych danych. Wpływ tej wady może jednak zmniejszyć sumaryczne ujęcie rozproszenia odpowiedzi.

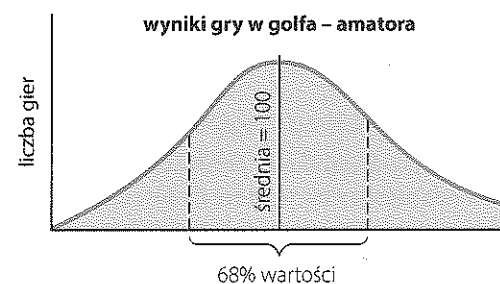
Rozproszenie odnosi się do rozkładu wartości wokół pewnej wartości centralnej, takiej jak średnia. Najprostszym miernikiem rozproszenia jest zakres: odległość między najwyższą a najniższą wartością. W ten sposób, poza stwierdzeniem, że średnia wieku

naszych badanych wynosi 15,87, moglibyśmy też wskazać, że ich wiek mieści się w zakresie od 13 do 19 lat.

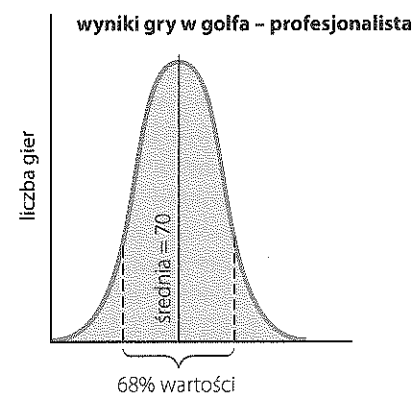
Bardziej wyrafinowanym miernikiem rozproszenia jest **odchylenie standardowe**. Miernik ten wspomnieliśmy już w rozdziale 7 jako błąd standardowy rozkładu z prób. Najprościej rzecz ujmując, odchylenie standardowe jest wskaźnikiem poziomu zmienności w zbiorze danych. Wyższe odchylenie standardowe oznacza, że dane są bardziej rozproszone, a niższe odchylenie standardowe, że są bardziej skupione. Tę zasadniczą cechę ukazuje ilustracja 14.7. Zwróć uwagę, że profesjonalny golfista ma nie tylko niższą średnią wyników – wyniki te są też spójniejsze, co odzwierciedla niższe odchylenie standardowe. Z kolei amator ma wyższą przeciętną, ale jego wyniki są mniej spójne – czasem idzie mu całkiem dobrze, a czasem dużo gorzej.

Istnieje jeszcze wiele innych mierników rozproszenia. Opisując wyniki testów na inteligencję, badacze mogliby określić zakres międzykwartylowy (*interquartile range*), czyli zakres wyników dla środkowych 50% badanych. Gdyby w górnej ćwiartce (górnym kwartyle) wystąpiły wyniki od 120 do 150, a w dolnej ćwiartce (dolnym kwartyle) wyniki od 60 do 90 punktów, raport mógłby stwierdzić, że zakres międzykwartylowy wynosi od 90 do 120 punktów (czyli 30 punktów), ze średnim wynikiem 102.

a. wysokie odchylenie standardowe = wartości rozproszone



b. niskie odchylenie standardowe = wartości skupione



ILUSTRACJA 14.7

Wysokie i niskie odchylenie standardowe

Zmienne ciągłe i dyskretne

Omówione wyżej wyliczenia nie stosują się do wszystkich zmiennych. By to zrozumieć, musimy rozróżnić dwa typy zmiennych: ciągłe i dyskretne. **Zmienna ciągła** (zwana inaczej zmienną ilorazową) nieustannie wzrasta w minimalnych przedziałach. Przykładem jest **wiek**, który stale wzrasta w każdej jednostce czasu. Z kolei **zmienna dyskretna** skacze od kategorii do kategorii bez kroków pośrednich. Przykładami mogą być: **pleć**, **stopień wojskowy**, **rok studiów** (przechodzisz z pierwszego na drugi w jednej chwili).

Analizując zmienną dyskretną – na przykład nominalną albo porządkową – nie możesz skorzystać z niektórych omówionych powyżej technik. Ściśle rzecz biorąc, dominantę powinno się obliczać dla danych nominalnych, medianę – dla interwałowych, a średnią dla danych ilorazowych, nie dla nominalnych (zob. rozdział 5). Jeżeli analizowaną zmienną jest na przykład **pleć**, same liczby (23 osoby w naszej próbie motocyklistów transwestytów to kobiety) lub procenty (7% to kobiety) mogą być użytecznym i stosownym narzędziem analizy, ale ani mediana, ani średnia nie miałyby tu sensu. Chociaż wyliczenie dominanty mogłoby być uzasadnione, nie wyjaśniłoby zbyt wiele, dowiedzielibyśmy się bowiem tylko tyle, że „większość to mężczyźni”. Dominanta z danych na temat przynależności religijnej mogłaby jednak okazać się bardziej interesująca, jak na przykład w przypadku zdania „większość obywateli Stanów Zjednoczonych to protestanci”.

Szczegółowość a czytelność

Przedstawiając zarówno dane jednozmiennowe, jak i inne, będziesz się starać osiągnąć dwa cele. Z jednej strony, chcesz dążyć do tego, aby zapewnić czytelnikowi możliwie najbardziej szczegółowe informacje

Odchylenie standardowe – miara rozproszenia wokół średniej, obliczona w taki sposób, że około 68% przypadków mieści się w granicach plus minus jednego odchylenia standardowego od średniej, 95% mieści się w granicach plus minus dwa odchylenia standardowych, a 99,9% mieści się w granicach trzech odchylenia standardowych. Na przykład, jeśli średni wiek w grupie wynosi 30 lat, a odchylenie standardowe 10, to 68% osób jest w wieku od 20 do 40 lat. Im mniejsze odchylenie standardowe, tym wartości są ściślej skupione wokół średniej, a jeśli odchylenie standardowe jest wysokie, to wartości są szeroko rozproszone

Zmienna ciągła – zmienna, której cechy tworzą stałą progresję, na przykład wiek lub dochód. Na przykład wiek grupy ludzi może obejmować osoby liczące 21, 22, 23, 24 lata itd., a nawet można go podzielić na ułamki lat

Zmienna dyskretna – zmienna, której cechy są od siebie oddzielone lub ściśle rozgraniczone, jak w przypadku płci lub przynależności religijnej. Innymi słowy, w przypadku płci nie ma progresji od mężczyzny do kobiety

dotyczące tych danych. Z drugiej jednak strony, dane powinny zostać przedstawione w zrozumiałej formie. Ponieważ jednak często oba te cele są ze sobą sprzeczne, trzeba będzie nieustannie poszukiwać najlepszego kompromisu między nimi. Praktycznym rozwiązaniem jest przedstawienie konkretnego zbioru danych w więcej niż jednej formie. W przypadku wieku można na przykład przedstawić rozkład wieku bez podziału na grupy oraz średnią wieku i odchylenie standardowe.

Jak widać na podstawie tak krótkiego wstępu na temat analizy jednozmiennej, ta pozornie prosta materia bywa dosyć złożona. Wiedza zawarta w niniejszym podrozdziale stanowi jednak pomost do porównań podgrup i analizy dwuzmiennej.

Porównywanie podgrup

Analiza jednozmienowa opisuje jednostki analizy zastosowane w badaniach i pozwala, jeżeli jednostki te stanowią próbę większej populacji, wyciągnąć wnioski opisowe dotyczące całej populacji. Z kolei analizy dwuzmiennej i wielozmiennej służą przede wszystkim wyjaśnianiu. Zanim jednak przejdziemy do wyjaśniania, powinniśmy przyjrzeć się bliżej przypadkom opisu podgrup.

Często zasadne jest opisywanie podzbiorów przypadków, osób badanych lub respondentów. Weźmy przykład zaczerpnięty z GSS. W 2014 roku przeprowadzono sondaż, podczas którego respondentom zadano zapytanie: „Czy marihuana powinna być legalna?”. W odpowiedzi 55,5% stwierdziło, że powinna, a 44,5%, że nie powinna. W tabeli 14.4 przedstawiono odpowiedzi udzielone na to pytanie przez respondentów w różnych kategoriach wiekowych.

Jak widzimy, porównania podgrup mówią nam, jak różne grupy w populacji odpowiedziały na to pytanie. Nie ma żadnej wątpliwości, że w wynikach można dostrzec pewien wzorec, choć być może nie do końca taki, jakiego oczekiwaliśmy – za chwilę jeszcze do tego wrócimy. Najpierw przyjrzyjmy się, jak na to pytanie odpowiedział inny zestaw podgrup.

TABELA 14.4

Poparcie legalizacji marihuany według wieku respondentów, 2012

	poniżej 25	25–39	40–54	55 lub starsi
powinna być zalegalizowana	71%	60%	55%	47%
nie powinna być zalegalizowana	29	40	45	53
100% =	(161)	(441)	(429)	(533)

Źródło: Generalny Sondaż Społeczny, 2014, Narodowe Centrum Badania Opinii Społecznej

TABELA 14.5

Legalizacja marihuany, według orientacji politycznej, 2014

	powinna być zalega- lizowana	nie powinna być zalegali- zowana	100% =
zdecydowanie liberalny	84%	16	(58)
liberalny	76%	24	(172)
umiarkowanie liberalny	69%	31	(168)
centrysta	54%	46	(597)
umiarkowanie konserwatywny	54%	46	(219)
konserwatywny	33%	67	(222)
zdecydowanie konserwatywny	34%	66	(72)

Źródło: Generalny Sondaż Społeczny, 2014, Narodowe Centrum Badania Opinii Społecznej

Tabela 14.5 przedstawia postawy wobec legalizacji marihuany różnych podgrup politycznych w zależności od tego, czy respondenci określili się jako konserwatyści czy liberałowie. Zanim przeanalizujesz tabelę, możesz spróbować postawić hipotezę, jakie mogą być wyniki i dlaczego. Zwróć uwagę, że zmieniłem kierunek procentowania tej tabeli, aby była łatwiejsza do odczytania. W tym przypadku, aby porównać podgrupy, należy czytać kolumny, a nie wiersze.

Zanim przejdziemy do rozważań na temat logiki analizy przyczynowej, rozważmy inny przykład porównań podgrup, który zarazem pozwoli nam poruszyć kilka zagadnień z zakresu formatu tabel.

Łączenie kategorii odpowiedzi

Na ogół „podręcznikowe przykłady” tabel są zdecydowanie prostsze niż te, które możesz znaleźć w opublikowanych raportach z badań albo we własnej analizie danych. Dlatego też podrozdział ten, podobnie jak następny, przybliży ci dwa często spotykane problemy oraz zasugeruje, jak sobie z nimi poradzić.

Zacznijmy od analizy tabeli 14.6, która przedstawia dane z hipotetycznego sondażu politycznego. Jak widać, respondentów zapytano, na którego z kandydatów oddaliby głos, gdyby wybory odbyły dzisiaj: Smitha czy Jonesa. Następnie poproszono ich o określenie stopnia ich poparcia wybranego kandydata. Wyniki zostały przedstawione w odniesieniu do sympatii partyjnych respondentów.

Chociaż prawdopodobnie nie mielibyśmy problemu z odgadnięciem, który kandydat był republikaninem, a który demokratą, możemy jednak odnieść wrażenie, że w tabeli było o kilka liczb za dużo, aby można je było łatwo zinterpretować. Możemy nieco uprościć sprawę, łącząc skrajne i umiarkowane kategorie dla każdego kandydata. Zrobiłem tak w tabeli 14.7.

TABELA 14.6

Hipotetyczny sondaż polityczny

Gdyby wybory odbyły się dzisiaj, jak myślisz, na kogo byś zagłosował(a): na Smitha czy Jonesa?

	Sympatie polityczne respondentów		
	demokrata	republikanin	inny
zdecydowanie Smith	40%	10%	10%
prawdopodobnie Smith	30	10	20
prawdopodobnie Jones	10	25	20
zdecydowanie Jones	5	35	15
nie wiem	15	20	35
ogółem	100%	100%	100%

TABELA 14.7

Hipotetyczny sondaż polityczny z połączonymi kategoriami odpowiedzi

Gdyby wybory odbyły się dzisiaj, jak myślisz, na kogo byś zagłosował(a): na Smitha czy Jonesa?

	sympatie polityczne respondentów		
	demokrata	republikanin	inny
zdecydowanie lub prawdopodobnie			
Smith	70%	20%	30%
Jones	15	60	35
nie wiem	15	20	35
ogółem	100%	100%	100%

Sposoby postępowania z kategorią „nie wiem”

Chociaż w trakcie kampanii wyborczej warto wiedzieć, jaka część elektoratu jest niezdecydowana, oczywiście też jest, że w dniu wyborów nie można zagłosować „niezdecydowany”. Z tego powodu warto zbadać względną siłę dwóch kandydatów, wyłączając ze swoich obliczeń kategorię „nie wiem”. Wyniki te przedstawiono w tabeli 14.8.

Te obliczenia są już dość proste. Ponieważ demokraci zadeklarowali w sumie 85% planowanych głosów na dwóch kandydatów, wystarczy podzielić każdy z pierwotnych odsetków przez 0,85. Tym samym, jeżeli wykluczmy odpowiedzi „nie wiem”, 70% poparcia Smitha wśród demokratów staje się 82%. Republikanie oddali 80% swoich głosów na dwóch kandydatów, więc dzielimy każdy z ich procentów w tabeli 14.8 przez 0,80. Z kolei odsetek osób

TABELA 14.7

Hipotetyczny sondaż polityczny z połączonymi kategoriami odpowiedzi

Gdyby wybory odbyły się dzisiaj, jak myślisz, na kogo byś zagłosował(a): na Smitha czy Jonesa?

	sympatie polityczne respondentów		
	demokrata	republikanin	inny
zdecydowanie lub prawdopodobnie			
Smith	82%	25%	46%
Jones	18	75	54
ogółem	100%	100%	100%

deklarujących inne sympatie polityczne zostanie podzielony przez 0,65.

W tym momencie, mając przed sobą trzy wersje danych, możesz zadać sobie pytanie: „Która jest właściwa?”. Odpowiedź zależy od celu analizy i interpretacji danych. Jeżeli na przykład nie jest konieczne odróżnienie „zdecydowanie” od „prawdopodobnie”, warto je połączyć, ponieważ łatwiej jest odczytać tabelę.

Dużo trudniejsze jest podjęcie poza kontekstem konkretnych badań decyzji, czy uwzględnić, czy też wyłączyć odpowiedzi „nie wiem”. Odkrycie, że tak duży odsetek wyborców nie jest zdecydowany, może być naprawdę ważne. Z drugiej strony, przy założeniu, że osoby deklarujące się jako niezdecydowane nie zagłosują lub w ostatecznym rozrachunku ich głosy najprawdopodobniej zostaną podzielone równo między dwóch kandydatów, odpowiedniejsze może być ich wykluczenie z obliczeń⁷⁴.

W każdym razie na podstawie swoich danych możesz stwierdzić, że pewien odsetek osób powiedział, że nie wie, na kogo zagłosuje, a odpowiedzi pozostałych w jakiś sposób się podzieliły. Często właściwym rozwiązaniem jest opublikowanie danych w obu formach – z i bez „nie wiem” – aby również twoi czytelnicy mogli wyciągnąć własne wnioski. Z drugiej strony ty również będziesz czytelnikiem podobnych tabel, sporządzonych przez innych badaczy, a znajomość logiki ich tworzenia pomoże ci być bardziej świadomym konsumentem danych ilościowych.

⁷⁴ W badaniach preferencji politycznych i sondażach przedwyborczych często odpowiedzi „nie wiem” nie są wykluczane z analiz, ale na podstawie zestawu dodatkowych pytań – o głosowanie w poprzednich wyborach, wybierane partie polityczne i ogólne poglądy – dokonuje się prognozy na kogo takie osoby mogły zagłosować. W odróżnieniu od wyników sondażu przedwyborczego mówimy wtedy o prognozie wyborczej – przyp. M.J.L.

Opisy liczbowe w badaniach jakościowych

Chociaż w rozdziale tym zajmujemy się przede wszystkim badaniami ilościowymi, rozważania te znajdują również zastosowanie w analizach jakościowych. Wyniki pogłębionych studiów jakościowych często dają się sprawdzić za pomocą testów liczbowych. Na przykład David Silverman, kiedy postanowił porównać leczenie pacjentów chorych na raka w klinikach prywatnych i w placówkach brytyjskiej Narodowej Służby Zdrowia, zdecydował się przede wszystkim na pogłębioną analizę interakcji między pacjentami a lekarzami:

Moja metoda analizy była w znacznym stopniu jakościowa [...]. Wykorzystałem fragmenty wypowiedzi lekarzy i pacjentów, podając jednocześnie zwięzły opis etnograficzny otoczenia oraz pewnych danych behawioralnych. Dodatkowo jednak skonstruowałem też formę kodowania, która pozwoliła mi ustalić kilka ogólnych mierników interakcji między pacjentami a lekarzami.

(Silverman 1993: 163)

Dane liczbowe doprecyzowały intuicyjne wnioski Silvermana oparte na jego obserwacjach jakościowych, zarazem jednak pogłębione zrozumienie sytuacji pozwoliło mu ma wypracowanie tym trafniejszych metod analizy ilościowej. Czytając poniższy obszerny wywód, zwróć uwagę na relacje między podejściem ilościowym a jakościowym:

Moje ogólne wrażenie było takie, że konsultacje prywatne trwają znacznie dłużej niż w klinikach publicznych. Po dokładnym zbadaniu danych okazało się, że te pierwsze trwały praktycznie niemal dwa razy dłużej (20 minut w stosunku do 11) i że różnica ta była statystycznie bardzo istotna. Przypomniałem sobie jednak, że jedna z klinik publicznych miała, ze szczególnych względów, niezwykle krótki czas konsultacji. Uznałem, że chcąc uczciwie porównać czas konsultacji w tych dwóch sektorach, należy wykluczyć tę klinikę oraz porównywać tylko czas konsultacji z udziałem konkretnego lekarza pracującego w obu sektorach. Badanie tej ograniczonej próbki wykazało, że różnica czasu trwania konsultacji prywatnych i publicznych spadła do poziomu poniżej 3 minut. Wynik ten był nadal istotny statystycznie, ale w mniejszym stopniu. W końcu jednak, gdyby porównać czas poświęcony tylko nowym pacjentom tego samego lekarza, to pacjentom w publicznej

Analiza dwuzmiennowa – jednoczesna analiza dwóch zmiennych w celu określenia empirycznego związku między nimi. Jej przykładem jest konstruowanie prostej tabeli procentowej lub obliczanie prostego współczynnika korelacji

służbie zdrowia poświęcano średnio o 4 minuty więcej – 34 minuty w porównaniu do 30 w klinikach prywatnych.

(Silverman 1993: 163–164)

Powyższy przykład ujawnia korzyści związane z łączeniem różnych metod w badaniach społecznych. Szczególnie wartościowe może być połączenie metod ilościowych i jakościowych.

Analiza dwuzmiennowa

W przeciwieństwie do analizy jednozmiennowej porównania podgrup wymagają dwóch zmiennych. Pod tym względem analiza podgrup stanowi rodzaj **analizy dwuzmiennowej** – jednoczesnej analizy dwóch zmiennych. Jednakże, podobnie jak w przypadku analizy jednozmiennowej, celem porównań podgrup jest przede wszystkim opis. Analiza dwuzmiennowa w badaniach społecznych na ogół zawiera jeszcze jeden element: określenie zależności między zmiennymi jako takimi. Tak więc, o ile analiza jednozmiennowa i porównania podgrup mają na celu opis ludzi (czy innych jednostek analizy), analiza dwuzmiennowa koncentruje się na zmiennych i empirycznych zależnościach między nimi.

Przykład porównania podgrup zawiera tabela 14.9: ukazuje ona niezależnie od siebie stopnień uczestnictwa w nabożeństwach kobiet i mężczyzn, zgodnie z wynikami GSS z 2014 roku. Tabela pokazuje – porównawczo i opisowo – dane dowodzące, że badane kobiety uczestniczyły w nabożeństwach dużo częściej niż mężczyźni. Ta sama tabela jednak, gdy potraktujemy ją jako materiał do służącej wyjaśnianiu analizy dwuzmiennowej, opowiada o czymś zupełnie innym. Dowiadujemy się z niej, że zmienna płeć ma wpływ na zmienną uczestnictwo w nabożeństwach. Oznacza to, że ten typ zachowania możemy postrzegać jako zmienną zależną, która jest po części zdeterminowana przez zmienną niezależną, czyli płeć.

Eksplanacyjna analiza dwuzmiennowa wymaga zatem zastosowania wprowadzonego już w rozdziale 1 „języka zmiennych”. Przesuwając lekko akcenty, nie

TABELA 14.9

Mężczyźni i kobiety deklarujący udział w nabożeństwach, 2012

	mężczyźni	kobiety
co tydzień	24%	33%
rzadziej	76	67
100% =	(1148)	(1378)

Źródło: Generalny Sondaż Społeczny, 2014, Narodowe Centrum Badania Opinii Społecznej

mówimy już o mężczyznach i kobietach jako różnych podgrupach, lecz o *płci* jako zmiennej, która ma wpływ na inne zmienne. Teoretyczna interpretacja tabeli 14.9 mogłaby się opierać na omówionej w rozdziale 2 hipotezie pocieszenia Charlesa Glocka:

1. Kobiety w Stanach Zjednoczonych nadal traktowane są jako obywatele drugiej kategorii.
2. W społeczeństwie świeckim ludzie, którym odmawia się nagród w postaci statusu społecznego, mogą zwrócić się w stronę religii jako alternatywnego źródła statusu.
3. Oznacza to, że kobiety powinny być bardziej religijne niż mężczyźni.

Rozumowanie to potwierdzają dane przedstawione w tabeli 14.9: co tydzień w nabożeństwach uczestniczy 33% kobiet i 24% mężczyzn.

Istotne implikacje dla tworzenia i odczytywania tabel procentowych ma dodatek w postaci logiki zależności przyczynowych. Jednym z zasadniczych problemów początkujących analityków jest ustalenie „kierunku procentowania” tabeli. Na przykład w tabeli 14.9 podzieliłem grupę badanych na dwie podgrupy – mężczyzn i kobiety – po czym opisałem zachowanie każdej z podgrup. Jest to jak najbardziej poprawna metoda tworzenia tego typu tabel. Zwróć jednak uwagę, że byłibyśmy w stanie – ale w sposób niepoprawny – skonstruować tę tabelę inaczej. Moglibyśmy najpierw podzielić badanych wedle częstości ich uczestnictwa w nabożeństwach, a następnie ustalić odsetek kobiet i mężczyzn w każdej z podgrup. Metoda taka byłaby jednak bezwartościowa z punktu widzenia wyjaśniania. Tabela 14.9 sugeruje, że płeć ma wpływ na częstość uczestnictwa w nabożeństwach. Gdybyśmy użyli innej metody konstrukcji, tabela sugerowałaby, że uczestnictwo wybranej osoby w nabożeństwach ma wpływ na to, czy jest ona mężczyzną, czy kobietą, co nie ma sensu. Zachowanie nie może określać płci.

Z podobnym problemem zmagają się często początkujący analitycy. Jak odczytać tabelę procentową? Istnieje pokusa, aby tabelę 14.9 odczytać w następujący sposób: „spośród kobiet, tylko 33% uczestniczyło co tydzień w nabożeństwach, a 67% uczestniczyło w nich rzadziej: bycie kobietą powoduje zatem spadek prawdopodobieństwa, że dana osoba będzie często uczestniczyła w nabożeństwach”. Taki sposób odczytania tabeli jest oczywiście niepoprawny. Jakikolwiek wniosek co do wpływu zmiennej *płeć* na uczestnictwo w nabożeństwach musi opierać się na porównaniu kobiet i mężczyzn. W szczególności możemy porównać 33% z 24% i stwierdzić, że jest prawdopodobniejsze, że kobieta będzie co tydzień uczestniczyła w nabożeństwach. Tym samym do odczytania tabeli dwuzmiennowej służącej wyjaśnianiu konieczne jest porównanie podgrup.

Tworząc i przedstawiając tabelę 14.9, zastosowałem konwencję zwaną *procentowaniem w dół*

(w kolumnach). Termin ten oznacza, że w każdej kolumnie odsetki sumują się do pełnych 100%. Tęgo typu tabelę czyta się rzędami. W rzędzie oznaczonym „co tydzień” sprawdzasz, jaki odsetek mężczyzn co tydzień uczestniczył w nabożeństwach oraz jaki odsetek kobiet co tydzień uczestniczył w nabożeństwach.

Kierunek procentowania w tabeli jest czymś arbitralnym – niektórzy badacze wolą procentowanie poziome. Skonstruowaliby oni tabelę 14.9 w taki sposób, aby „mężczyźni” i „kobiety” znaleźli się po lewej stronie tabeli, tworząc dwa rzędy, natomiast „co tydzień” i „rzadziej” pojawiłyby się na górze jako oznaczenie kolumn. Liczby w polach tabeli przesunęłyby się odpowiednio, a odsetki w każdym rzędzie sumowałyby się do 100%. W takim przypadku tabelę czytałoby się w dół kolumny, szukając odpowiedzi na pytanie, jaki odsetek kobiet i mężczyzn uczestniczył w nabożeństwach z taką częstotliwością. Logika i wnioski byłyby w obu przypadkach takie same, zmieniłaby się jedynie forma.

Odczytując tabelę skonstruowaną przez kogoś innego, musisz ustalić, w jakim kierunku przebiegało procentowanie. Na ogół będzie to oznaczone lub w oczywisty sposób będzie wynikać z logiki analizowanych zmiennych. W ostateczności możesz dodać odsetki w każdej kolumnie i każdym rzędzie. Jeżeli każda z kolumn daje 100%, procentowano w dół (w kolumnach). Jeżeli w każdym rzędzie otrzymasz 100%, procentowanie jest poziome (w wierszach). Zasada jest następująca:

1. Jeżeli procentowano w kolumnach, czytaj wierszami.
2. Jeżeli procentowano w wierszach, czytaj kolumnami.

Procentowanie w tabeli

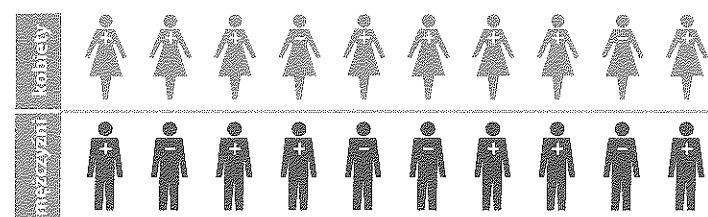
Logikę tworzenia tabeli procentowej dla dwóch zmiennych ukazuje ilustracja 14.8. Wykorzystałem tu zmienne *płeć* oraz *postawy wobec równouprawnienia mężczyzn i kobiet*.

A to inny przykład. Załóżmy, że interesujemy się polityką redakcyjną gazet w kwestii legalizacji marihuany. Prowadzimy analizę treści artykułów redakcyjnych na ten temat, które pojawiły się w ciągu danego roku w próbie dzienników w całym kraju. Każdy artykuł klasyfikujemy jako sprzyjający, neutralny lub wrogi względem legalizacji marihuany. Moglibyśmy na przykład zbadać zależność między poglądami wyrażonymi w artykule a typem społeczności, w której dany dziennik jest wydawany, zakładając, że społeczności wiejskie powinny być pod tym względem bardziej konserwatywne niż miejskie. Tym samym każdy dziennik – a więc i każdy artykuł redakcyjny – został zaklasyfikowany również w kategoriach liczby ludności w miejscowości, w której jest wydawany.

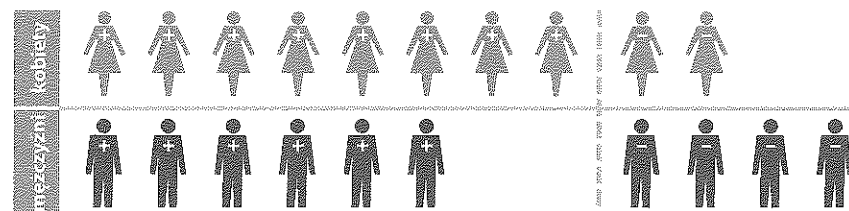
a. mężczyźni i kobiety popierający (+) lub niepopierający (-) równouprawnienia płci



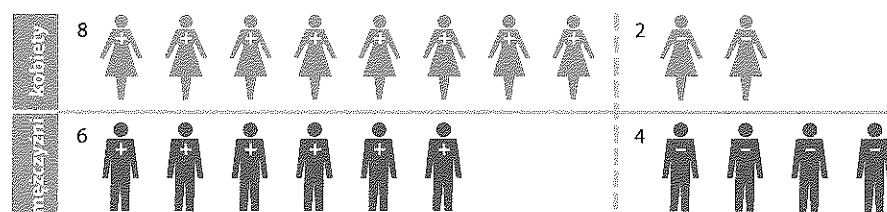
b. oddzielamy mężczyzn od kobiet (zmienna niezależna)



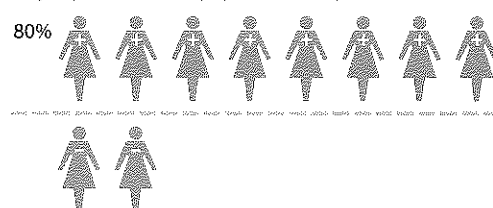
c. w obrębie każdej z płci oddzielamy tych, którzy popierają równouprawnienie, od tych, którzy go nie popierają (zmienna zależna)



d. liczymy, ile przypadków jest w każdej komórce tabeli



e. jaki procent kobiet popiera równouprawnienie?



g. wnioski

Prawdopodobieństwo, że kobieta popiera równouprawnienie, jest większe niż w przypadku mężczyzn.

Wydaje się, że płeć jest jedną z przyczyn postaw wobec równouprawnienia płci.

ILUSTRACJA 14.8

Procentowanie w tabeli

Tabela 14.10 przedstawia hipotetyczne dane opisujące poglądy wyrażane w artykułach redakcyjnych dzienników wydawanych w miastach i na wsi. Warto zauważyć, że jednostką analizy jest w tym przypadku pojedynczy artykuł. Tabela 14.10 zawiera informację, że 127 artykułów redakcyjnych na temat marihuany w naszej próbie dzienników zostało opublikowanych w społecznościach o populacji poniżej

100 tysięcy mieszkańców. (Uwaga: ten punkt graniczny został wybrany dla uproszczenia przykładu i nie oznacza, że słowo „wieś” w jakimkolwiek obiektywnym sensie odnosi się do społeczności poniżej 100 tysięcy mieszkańców). Spośród nich 11% (14 artykułów) sprzyjało legalizacji marihuany, 29% było neutralnych, a 60% wrogich. Spośród 438 artykułów, które pojawiły się w próbie dzienników wydawanych

	kobiety	mężczyźni
popierają równouprawnienie	80%	60%
nie popierają równouprawnienia	20	40
ogółem	100%	100%

TABELA 14.10

Hipotetyczne dane dotyczące treści artykułów redakcyjnych na temat legalizacji marihuany

polityka redakcyjna wobec legalizacji marihuany	wielkość społeczności	
	poniżej 100 000	powyżej 100 000
sprzyjająca	11%	32%
neutralna	29	40
nieprzychylna	60	28
100% =	(127)	(438)

w miejscowościach powyżej 100 tysięcy mieszkańców, 32% (140 artykułów) zaliczono do sprzyjających legalizacji marihuany, 40% do obojętnych, a 28% do wrogich.

Porównując politykę redakcyjną gazet wydawanych w miastach i na wsi w naszych hipotetycznych badaniach, stwierdzamy – zgodnie z oczekiwaniami – że dzienniki wiejskie w dużo mniejszym stopniu niż miejskie sprzyjają legalizacji marihuany. Określamy to na podstawie faktu, że więcej (32%) artykułów redakcji miejskich niż wiejskich (11%) sprzyjało legalizacji marihuany. Moglibyśmy również zauważyć, że więcej artykułów redakcji wiejskich niż miejskich było nieprzychylnie nastawionych do tego pomysłu (60% do 28%). Zauważalne jest również, że tabela ta zakłada, że wielkość społeczności może wpływać na poglądy wyrażane przez redakcję dziennika na ten temat, a nie, że poglądy redakcji wpływają na wielkość miejscowości.

Tworzenie i odczytywanie tabel dwuzmiennowych (krzyżowych)

Przyjrzyjmy się teraz kolejnym etapom tworzenia tabeli dwuzmiennowej (krzyżowej) służącej wyjaśnianiu:

1. Przypadki dzielimy na grupy zgodnie z wartościami zmiennej niezależnej.
2. Następnie opisujemy każdą z podgrup w kategoriach wartości zmiennej zależnej.
3. Na koniec odczytujemy tabelę, porównując między sobą podgrupy zmiennej niezależnej w kategoriach wartości zmiennej zależnej.

Powtórzmy naszą analizę płci i postaw wobec równouprawnienia, postępując zgodnie z powyższym opisem. Z przedstawionych wcześniej względów płeć jest zmienną niezależną, a postawy wobec równouprawnienia płci zmienną zależną. Postępujemy zatem w sposób następujący:

1. Dzielimy przypadki na mężczyzn i kobiety.
2. Każdą podgrupę płci zostaje opisana w kategoriach poparcia lub braku poparcia równouprawnienia płci.

3. Porównujemy mężczyzn i kobiety pod kątem odsetka osób popierających równouprawnienie.

W przykładzie dotyczącym polityki redakcyjnej wobec legalizacji marihuany zmienną niezależną była wielkość społeczności, a zmienną zależną polityka redakcyjna gazety. Tabelę należałoby wówczas stworzyć w następujący sposób:

1. Podziel artykuły na podgrupy według wielkości społeczności, w których wydawane są badane dzienniki.
2. Opisz każdą podgrupę artykułów według odsetka artykułów przyjaznych, obojętnych lub wrogich wobec legalizacji marihuany.
3. Porównaj obie podgrupy w kategoriach odsetków artykułów przyjaznych legalizacji marihuany.

Analizy dwuzmiennowe mają na ogół na celu wyjaśnianie przyczynowe. Powyższe hipotetyczne przykłady dają pewien wgląd w naturę związku przyczynowego, jakim posługują się badacze społeczni.

Tabele podobne do przed chwilą analizowanych na ogół nazywamy **tabelami krzyżowymi (tabelami kontyngencji)**. Wartości zmiennej zależnej uzależnione są od wartości zmiennej niezależnej. Pomimo popularności tego typu tabel w naukach społecznych ich format nigdy nie został zestandaryzowany, wskutek czego w literaturze badawczej znajdziesz wiele rozmaitych formatów. Z drugiej strony prawdopodobnie nie ma większego sensu dążyć do standaryzacji, póki tabele są łatwe do odczytania i interpretacji. Jest jednak kilka wskazówek, którymi powinniśmy się kierować podczas prezentowania większości danych tabelarycznych:

1. Tabela powinna mieć nagłówek lub tytuł związane opisujący jej zawartość.
2. Jeśli tylko to możliwe, należy jasno przedstawić oryginalną treść zmiennej – w samej tabeli lub w tekście wraz z parafrazą w tabeli. Informacja ta jest szczególnie istotna, gdy zmienna została wyprowadzona z odpowiedzi na pytania o postawy, ponieważ w tej sytuacji znaczenie odpowiedzi w znacznym stopniu zależy od sformułowania pytania.
3. Należy wyraźnie wskazać wartości każdej zmiennej. Mimo że konieczne okaże się skrócenie rozbudowanych kategorii, ich znaczenie powinno zostać objaśnione w tabeli, a oprócz tego w tekście powinien oczywiście znaleźć się pełny opis.

Tabela krzyżowa (kontyngencji) – format prezentacji relacji między zmiennymi w postaci rozkładów procentowych. Zwykle używana do ujawnienia wpływu zmiennej niezależnej na zmienną zależną

4. Jeżeli w tabeli podano wartości procentowe, należy wskazać podstawę, na której zostały obliczone. Nie ma sensu podawać wszystkich nieprzetworzonych liczb dla każdej kategorii, ponieważ można je odtworzyć na podstawie wartości i podstaw procentowych. Co więcej, prezentacja zarówno liczb, jak i odsetków często komplikuje tabelę i utrudnia jej odczytanie.
5. Jeżeli w tabeli ze względu na brakujące dane (na przykład brak odpowiedzi) pominięto jakieś przypadki, ich liczba powinna zostać podana w tabeli.

Chociaż przedstawiłem tu logikę dwuzmiennowej analizy przyczynowej w postaci tabel procentowych, można do tego wykorzystać również wiele innych formatów. Jedną z możliwości jest wykres rozrzutu, który zapewnia wizualne ukazanie zależności między dwiema zmiennymi. Poszukując interesujących przykładów takich wykresów, możesz skorzystać z dostępnego w internecie programu GapMinder. Używając krajów jako jednostek analizy, możesz na przykład zbadać związek między współczynnikiem urodzeń a śmiertelnością niemowląt. Co więcej, możesz nawet obserwować rozwój tej zależności w czasie.

Wprowadzenie do analizy wielozmiennowej

Logika **analizy wielozmiennowej**, czyli jednoczesna analiza więcej niż dwóch zmiennych, może być postrzegana jako rozszerzenie analizy dwuzmiennowej. Możemy w szczególności konstruować tabele wielozmiennowe na podstawie bardziej skomplikowanego opisu podgrupy, wykonując zasadniczo te same kroki, które opisano dla tabel dwuzmiennowych (krzyżowych). Zamiast jednej zmiennej niezależnej i jednej zmiennej zależnej będziemy mieli jednak więcej niż jedną zmienną niezależną. Zamiast wyjaśniania zmiennej zależnej za pomocą jednej zmiennej niezależnej, poszukujemy wyjaśnienia, wykorzystując więcej niż jedną zmienną niezależną.

Wróćmy do naszego przykładu z udziałem w nabożeństwach. Załóżmy, że jesteśmy przekonani, że na tego rodzaju zachowania wpływa również wiek (według hipotezy pocieszenia Glocka ludzie starsi są bardziej religijni niż młodsi). W pierwszym kroku powinniśmy zatem podzielić próbę na podgrupy na podstawie wartości obu zmiennych niezależnych jednocześnie: młodzi mężczyźni, starsi mężczyźni,

młode kobiety i starsze kobiety. Następnie należy opisać kilka podgrup w kategoriach zmiennej zależnej *uczestniczenia w nabożeństwach*, po czym przesłizbyśmy do etapu porównywania. Efekt przedstawiono w tabeli 14.11 sporządzonej na podstawie analizy wyników danych Generalnego Sondażu Społecznego z 2006 roku.

W tabeli 14.11 procentowania dokonano w kolumnach, dlatego też należy ją czytać poziomo. Interpretacja tej tabeli prowadzi do kilku wniosków:

1. Zarówno wśród mężczyzn, jak i kobiet osoby starsze częściej uczestniczą w praktykach religijnych niż młode. Co tydzień w nabożeństwach uczestniczyło 27% kobiet poniżej 40. roku życia i 41% kobiet w wieku 40 lat lub więcej. W przypadku mężczyzn liczby te wynoszą odpowiednio 19% i 31%.
2. W obrębie każdej z grup wiekowych kobiety uczestniczyły w nabożeństwach częściej niż mężczyźni. Wśród respondentów poniżej 40. roku życia co tydzień uczestniczyło w nich 27% kobiet i tylko 19% mężczyzn. Z kolei wśród osób w wieku 40 lat i więcej co tydzień w nabożeństwach uczestniczyło 41% kobiet i 31% mężczyzn.
3. Dane w tabeli dowodzą, że wiek może mieć większy wpływ na zaangażowanie w życie religijne niż płeć.
4. Płeć i wiek mają niezależny wpływ na uczestnictwo w nabożeństwach. Przy danej wartości jednej ze zmiennych niezależnych zmiana wartości drugiej nadal wywiera wpływ na zachowanie.
5. Dwie zmienne niezależne wywierają również łączny wpływ na zachowanie. Najczęściej w nabożeństwach uczestniczą starsze kobiety (41%), a najrzadziej młodzi mężczyźni (19%).
6. Należy również zauważyć, że sensowne byłoby też stwierdzenie, że istnieje *pozytywna* zależność między wiekiem a uczestnictwem w nabo-

TABELA 14.11

Zależność wielozmiennowa: uczestnictwo w nabożeństwach, wiek i płeć, 2006

Jak często uczestniczysz w nabożeństwach?	poniżej 40 lat		40 lat i więcej	
	mężczyźni	kobiety	mężczyźni	kobiety
mniej więcej co tydzień*	19%	27%	31%	41%
rzadziej	81	73	69	59
100% =	(832)	(958)	(1211)	(1477)

* Mniej więcej co tydzień = „częściej niż raz na tydzień”, „co tydzień” i „prawie co tydzień”

Źródło: Generalny Sondaż Społeczny, 2006, Narodowe Centrum Badania Opinii Społecznej

żeństwach. Wzrost wieku wiąże się ze wzrostem zaangażowania w życie religijne. Gdyby starsi ludzie rzadziej uczestniczyli w nabożeństwach, moglibyśmy stwierdzić, że istnieje zależność *negatywna*. Podobnych stwierdzeń nie możemy jednak sformułować w przypadku płci, ponieważ jest to zmienna nominalna, bez wewnętrznego porządku „wyższego” lub „niższego” różniącego kobiety i mężczyzn.

Zanim zakończę tę część, warto zwrócić uwagę na alternatywny format prezentacji takich danych. Kilka tabel przedstawionych w tym rozdziale jest dość mało efektywnych. Kiedy zmienna zależna *zaangażowanie religijne* jest dychotomiczna (posiada dokładnie dwa atrybuty), znajomość jednego atrybutu pozwala czytelnikowi łatwo zrekonstruować drugi. Na przykład, jeśli wiemy, że 27% kobiet poniżej 40. roku życia co tydzień uczestniczy w nabożeństwach, to automatycznie wiemy, że 73% czyni to rzadziej. Tym samym podawanie odsetka osób, które rzadziej uczestniczą w nabożeństwach jest niepotrzebne.

Zgodnie z powyższym rozumowaniem, tabela 14.11 może być traktowana jako alternatywny format tabeli 14.12. W tej tabeli odsetki osób, które twierdzą, że uczęszczają na nabożeństwa mniej więcej co tydzień, podane zostały w komórkach reprezentujących punkty przecięcia dwóch zmiennych niezależnych. Liczby przedstawione w nawiasach poniżej każdego odsetka odpowiadają liczbie przypadków, na podstawie której obliczono dany odsetek. W ten sposób czytelnik wie na przykład, że próba obejmuje 958 kobiety poniżej 40. roku życia, spośród których 27% co tydzień uczestniczy w praktykach religijnych. Na tej podstawie możemy obliczyć, że co tydzień w nabożeństwach uczestniczy 259 spośród tych 958 kobiet, a pozostałe 699 młodszych kobiet (lub 73%) uczestniczy w nich rzadziej. Nowa tabela jest łatwiejsza do odczytania niż poprzednia, a przy tym nie tracimy żadnego szczegółu.

TABELA 14.12

Uproszczona wersja tabeli 14.11

	odsetek osób, które w nabożeństwach uczestniczą mniej więcej co tydzień	
	mężczyźni	kobiety
poniżej 40 lat	19% (832)	27% (958)
40 lub więcej	31% (1,211)	41% (1,477)

Źródło: Generalny Sondaż Społeczny, 2006, Narodowe Centrum Badania Opinii Społecznej

Diagnozy społeczne

Rozważane przez nas techniki wielozmiennowe mogą służyć jako wspaniałe narzędzia do diagnozowania problemów społecznych. Można je wykorzystać, aby zastąpić opinie faktami oraz do rozstrzygania ideologicznych debat dzięki analizie danych.

Wróćmy na przykład do rozważanej już wcześniej kwestii płci i dochodów. Przedstawiono wiele argumentów, które miały wyjaśniać obserwowany od dawna wzorec kobiet na rynku pracy jako zarabiających mniej niż mężczyźni. Jednym z wyjaśnień jest to, że ze względu na tradycyjne schematy rodzinne kobiety jako grupa mają mniejszy udział w sile roboczej, a wiele z nich rozpoczyna pracę poza domem dopiero po ukończeniu przynajmniej podstawowych zadań związanych z wychowywaniem dzieci. Tym samym kobiety jako grupa prawdopodobnie będą pracowały krócej niż mężczyźni, a dochody rosną wraz ze stażem pracy. Przełomowe w tej kwestii było badanie przeprowadzone w 1984 roku przez amerykański Urząd Statystyczny, które wykazało, że to rozumowanie jest częściowo prawdziwe, co pokazano w tabeli 14.13.

Tabela 14.13 pokazuje przede wszystkim, że okres zatrudnienia rzeczywiście miał wpływ na dochody. Zarówno w przypadku mężczyzn, jak i kobiet, osoby z dłuższym stażem pracy zarabiały więcej. Widać to dobrze po odczytaniu pierwszych dwóch kolumn tabeli.

Z tabeli wynika jednak również, że kobiety zarabiały mniej niż mężczyźni, niezależnie od stażu pracy. Można to zauważyć, porównując średnie płace w wierszach tabeli, a w trzeciej kolumnie pokazano stosunek wynagrodzeń kobiet do mężczyzn. Jak widać, kobiety ze stażem pracy wynoszącym ponad 10 lat zarabiają wyraźnie mniej (7,91 USD

TABELA 14.13

Płeć i staż pracy a zarobki (pracownicy pełnoetatowi w wieku od 21 do 64 lat)

staż pracy u aktualnego pracodawcy	średnia stawka za godzinę pracy		stosunek zarobków kobiet do mężczyzn
	mężczyźni	kobiety	
mniej niż 2 lata	8,46 \$	6,03 \$	0,71
2–4 lata	9,38 \$	6,78 \$	0,72
5–9 lat	10,42 \$	7,56 \$	0,73
10 lat i więcej	12,38 \$	7,91 \$	0,64

Źródło: Biuro Spisowe, Current Population Reports, Series P-70, No. 10, Male–Female Differences in Work Experience, Occupation, and Earning, 1984 (Washington, DC: U.S. Government Printing Office, 1987), 4

Analiza wielozmiennowa – jednoczesna analiza zależności między kilkoma zmiennymi. Jej przykładem mogłaby być analiza jednoczesnego wpływu wieku, płci i klasy społecznej na religijność

za godzinę) niż mężczyźni ze stażem nieprzekraczającym 2 lat (8,46 USD za godzinę).

Skoro czas trwania stosunku pracy nie wyjaśnia różnicy między płacami mężczyzn a kobiet, możliwe są również inne wyjaśnienia: poziom wykształcenia, obowiązki związane z opieką nad dziećmi itp. Badacze, którzy stworzyli tabelę 14.13, przeanalizowali także inne zmienne, które mogłyby wyjaśnić różnice zarobków, nie będąc przy tym wyrazem dyskryminacji ze względu na płeć. Oprócz stażu pracy u obecnego pracodawcy, wzięli oni pod uwagę następujące zmienne:

- * liczba lat pracy w obecnym zawodzie,
- * całkowity staż pracy (niezależnie od zawodu),
- * fakt, czy respondent(ka) pracował(a) zazwyczaj na pełny etat,
- * stan cywilny,
- * wielkość miejscowości zamieszkania,
- * fakt bycia objętym zbiorowym układem pracy,
- * rodzaj wykonywanej pracy,
- * liczba pracowników w zakładzie pracy,
- * pracodawca prywatny lub państwowy,
- * przyczyny zwolnienia z poprzedniego miejsca pracy,
- * luka w zatrudnieniu między ostatnią a obecną pracą,
- * rasa,
- * niepełnosprawność,
- * stan zdrowia,
- * wiek dzieci,
- * uczestnictwo w zajęciach akademickich w szkole średniej,
- * liczba zajęć z matematyki, nauk ścisłych i przyrodniczych oraz języków obcych w szkole średniej,
- * fakt uczęszczania do prywatnej lub publicznej szkoły średniej,
- * osiągnięty poziom wykształcenia,
- * odsetek kobiet zatrudnionych w danym zawodzie,
- * główny kierunek studiów.

Każda z tych zmiennych mogłaby wpływać na zarobki i służyć jako wyjaśnienie różnic zarobków kobiet i mężczyzn, gdyby okazało się, że różnią się pod któryś z powyższych względów. Wziąwszy pod uwagę te wszystkie zmienne, badacze zdołali wyjaśnić 60% nierówności zarobków kobiet i mężczyzn. Pozostałe 40% jest więc funkcją innych „rozsądnych” zmiennych oraz/lub uprzedzeń. Do tego typu wniosków można dojść tylko drogą badania wpływu wielu zmiennych jednocześnie – to znaczy w drodze analizy wielozmiennej.

Mam nadzieję, że ten przykład pokazuje, w jaki sposób logika ukryta w codziennych rozmowach może być ukazana i testowana w tego rodzaju ilościowej analizie danych. Przyjrzyjmy się niektórym danym przedstawionym w rozdziale 11.

W 2016 roku średni tygodniowy dochód pracownika płci męskiej zatrudnionego w pełnym wymiarze

godzin przez cały rok wynosił 969 USD. Z kolei przeciętna pracownica zatrudniona w pełnym wymiarze godzin przez cały rok zarabiała 784 USD, czyli około 81% kwoty jej męskiego odpowiednika (Biuro Statystyki Pracy 2017, tabela 9). Pytanie brzmi, czy ta różnica oznacza dyskryminację ze względu na płeć, czy też odzwierciedla uzasadnione czynniki?

Niektórzy twierdzą, że tym, co wpływa na dochody, jest na przykład wykształcenie i że w przeszłości kobiety zdobywały niższe wykształcenie niż mężczyźni. Można by więc zacząć od sprawdzenia, czy różnice w wykształceniu wyjaśniają, dlaczego kobiety zarabiają dziś średnio mniej niż mężczyźni. Dane pozwalające sprawdzić tę hipotezę zawiera tabela 14.14.

Jak widać na podstawie danych w tabeli, na każdym poziomie porównywalnego wykształcenia kobiety zarabiają znacznie mniej niż mężczyźni. Najwyraźniej wykształcenie nie wyjaśnia tych różnic.

Płeć i gender są na ogół uważane za prosty i wyraźny podział na mężczyzn i kobiety. Badacze społeczni przyznają jednak, że sprawa jest bardziej złożona. Na przykład osoby transseksualne, które decydują się operację i suplementację hormonalną mające uzgodnić ich płeć biologiczną. Oczywiście taka radykalna zmiana niesie za sobą wiele zmian i wyzwań, które mogłyby stanowić podstawę interesujących badań. Podchodząc do tego problemu, Kristen Schilt (2006) przyjęła niezwykłą taktykę.

O ile wiele rodzajów badań wskazuje na gorszy status kobiet w miejscu pracy, badania Schilt dotyczące osób transseksualnych ujawniają wpływ płci na poziomie osobistym. W wielu przypadkach osoby badane uzgodniły płeć, zachowując to samo miejsce pracy. Po korekcie płci w przypadku osób transseksualnych, którzy urodzili się jako kobiety, istniało

TABELA 14.14

Średnie tygodniowe zarobki pracowników pełnoetatowych, 2016

	mężczyźni	kobiety	stosunek zarobków kobiet do mężczyzn
wszyscy pracujący	969 \$	784 \$	0,81
podstawowe	551 \$	423 \$	0,77
średnie	769 \$	599 \$	0,81
wyższe niepełne	896 \$	688 \$	0,77
licencjat	1348 \$	994 \$	0,74
wyższe pełne	1707 \$	1257 \$	0,74

Uwaga: Dane te wskazują na utrzymującą się różnicę między dochodami kobiet i mężczyzn, nawet jeśli obie grupy osiągnęły ten sam poziom wykształcenia.

Źródło: Biuro Statystyki Pracy, Usual Weekly Earnings of Wage and Salary Workers, Fourth Quarter 2016. 27 stycznia 2017, tabela 9. Quartiles and selected deciles of usual weekly earnings of full-time wage and salary workers by selected characteristics, 2016 annual averages.

Zastanów się... raz jeszcze

Rozdział ten rozpoczęliśmy się od pytania, czy możliwe jest dowiedzenie się czegoś sensownego lub użytecznego z analizy danych, które zostały pozbawione wielu szczegółów, by umożliwić manipulację statystyczną. Odpowiedź, jak widzieliśmy, to bezwarunkowe „tak”.

Analiza ilościowa może być narzędziem służącym zmianom społecznym. Na przykład obliczenie średnich dochodów mężczyzn i kobiet lub białych i kolorowych mniejszości może wykazać nierówne traktowanie osób wykonujących dokładnie tę samą pracę. Takie analizy ilościowe mogą mieć większą wartość niż niepotwierdzone dowody

duże prawdopodobieństwo, że będą mogli cieszyć się z podwyżki płac i większego autorytetu. W innych badaniach osoby transseksualne, które korygowały płeć z męskiej na żeńską, zgłaszały dokładnie przeciwnie doświadczenia. Tego rodzaju osobiste relacje stanowią uzupełnienie badań statystycznych, które konsekwentnie pokazują, że kobiety zarabiają mniej niż mężczyźni, nawet jeśli wykonują tę samą pracę.

Poszukując kolejnego przykładu wielozmiennej analizy danych w prawdziwym życiu, warto przyjrzeć się powszechnej opinii, że członkom grup mniejszościowych częściej odmawia się kredytów bankowych niż białym. Innym wyjaśnieniem tego zjawiska może być to, że należący do mniejszości wnioskodawcy, o których mowa, byli bardziej narażeni na bankructwo lub mieli słabsze zabezpieczenie w celu zagwarantowania wnioskowanej pożyczki – z punktu widzenia banków są to rozsądne podstawy do udzielenia lub odmowy pożyczki. Spór ten może z łatwością rozwiązać ten sam rodzaj analizy wielozmiennej, którą właśnie omówiliśmy.

Załóżmy, że bierzemy pod uwagę tylko tych, którzy nie ogłaszali wcześniej bankructwa i mają określony poziom zabezpieczeń. Czy biali i osoby należące do mniejszości mają takie same szanse na uzyskanie żądanej pożyczki? Taką samą analizę powinniśmy przeprowadzić w podgrupach określonych przez poziom zabezpieczenia. Gdyby obie te grupy miały równe szanse na uzyskanie pożyczek w każdej z podgrup, musieliśmyby stwierdzić, że nie doszło do żadnej dyskryminacji etnicznej. Gdyby jednak członkowie mniejszości nadal mieli mniejsze szanse na uzyskanie pożyczek, oznaczałoby to, że bankructwo i różnice w zabezpieczeniach nie były wyjaśnieniem – co wzmocniłoby argument, że mamy do czynienia z dyskryminacją.

Wszystko to powinno jasno pokazać, że badania społeczne mogą odgrywać potężną rolę w służbie społeczności ludzkiej. Mogą nam pomagać określić aktualny stan rzeczy i wskazać, dokąd chcemy podążać.

Witamy w świecie diagnoz społecznych!

mówiące o konkretnych kobietach lub członkach mniejszości, którzy zarabiają duże pieniądze. Widzieliśmy również, że ilościową analizę zjawisk jakościowych, takich jak badania preferencji wyborczych, można przeprowadzić z dużą precyzją i pożytkiem.

Kluczowym wnioskiem naszych rozważań jest to, że zarówno badania jakościowe, jak i ilościowe są uzasadnionymi i ważnymi metodami pozwalającymi na zrozumienie życia społecznego. Co więcej, są one szczególnie przydatne, gdy są używane razem.

Etyka i analiza danych ilościowych

W rozdziale 13 zwróciłem uwagę na to, że towarzyszący analizie danych jakościowych subiektywizm zwiększa ryzyko stronniczych analiz, których doświadczeni badacze nauczyli się unikać. Niektórzy z kolei uważają, że analizy ilościowe nie są podatne na tego rodzaju błędy. Niestety nie jest to prawda. Nawet najdokładniejsza matematycznie analiza daje dużo miejsca na definiowanie i mierzenie zmiennych w sposób, który zachęca do jednego odkrycia zamiast drugiego, czego muszą wystrzegać się analitycy ilościowi. Czasami wcześniejsze staranne określenie hipotez może zapewnić jakąś ochronę, lecz z drugiej strony może to także uniemożliwić pełną eksplorację tego, co mogą nam powiedzieć dane.

Analitik ilościowy ma obowiązek zgłosić wszelkie formalne hipotezy i inne założenia, których nie udało się potwierdzić. Założmy, że uważasz, że jakaś zmienna okaże się istotną przyczyną uprzedzeń związanych z płcią, ale analiza danych temu zaprzecza. Brak korelacji należy zgłosić, ponieważ taka informacja przyda się innym, którzy prowadzą badania na ten temat. Chociaż oczywiście bardziej satysfakcjonujące byłoby odkrycie, co powoduje uprzedzenia, dość istotne jest również to, aby wiedzieć, co ich nie powoduje.

Ochrona prywatności podmiotów badań jest równie ważna w analizie ilościowej, jak i jakościowej. Dzięki metodom ilościowym często łatwiej jest zbierać i rejestrować dane w sposób, który utrudnia identyfikację podmiotu. Nabierze to tym większego znaczenia, kiedy urzędnicy państwowi zażądają ujawnienia nazwisk studentów, którzy w ankiecie zgłosili, że zażywają nielegalne narkotyki. (Nawiasem mówiąc, nie ujawniaj nazwisk. W razie potrzeby spal kwestionariusze – oczywiście „przypadkowo”).

GŁÓWNE ZAGADNIENIA

Wprowadzenie

- Większość danych ma początkowo charakter jakościowy, więc aby umożliwić analizę statystyczną, należy je określić liczbowo.
- Analiza ilościowa obejmuje techniki, za pomocą których badacze przekształcają dane w postać liczbową i poddają je analizom statystycznym.

Kwantyfikacja danych

- Niektóre dane, takie jak wiek i dochód, są z samej swojej natury liczbowe.
- Kwantyfikacja często obejmuje kodowanie danych w kategorie, którym następnie przypisuje się wartości liczbowe.
- Badacze mogą korzystać z istniejących schematów kodowania, takich jak kategoryzacja zawodów według amerykańskiego Biura Statystycznego, lub opracowywać własne kategorie kodowania. W obu przypadkach schemat kodowania musi być odpowiedni do charakteru i celów badania⁷⁵.
- Książka kodowa to dokument opisujący identyfikatory przypisane różnym zmiennym oraz kody przypisane do atrybutów określających te zmienne.

Analiza jednozmiennowa

- Analiza jednozmiennowa to analiza pojedynczej zmiennej. Ponieważ nie obejmuje ona związków między dwiema i więcej zmiennymi, służy raczej celom opisowym niż wyjaśnianiu.
- Istnieje wiele technik, które pozwalają badaczom podsumować ich pierwotne dane, by uczynić je bardziej zrozumiałymi, przy jednoczesnym zachowaniu tak wielu szczegółów, jak to tylko możliwe. Rozkład częstości, wartości przeciętne, grupowanie danych, miary rozproszenia – to sposoby podsumowania danych dotyczących jednej zmiennej.

Porównywanie podgrup

- Porównania podgrup mogą być wykorzystywane do opisu podobieństw i różnic między podgrupami w zakresie tej samej zmiennej.
- Łączenie kategorii odpowiedzi i radzenie sobie z kategorią „nie wiem” to dwie techniki prezentacji i interpretacji danych.

Analiza dwuzmiennowa

- Analiza dwuzmiennowa koncentruje się raczej na zależnościach między zmiennymi niż porównywaniu podgrup. Analiza dwuzmiennowa analizuje statystyczne powiązania między zmienną zależną a niezależną. Na ogół służy raczej wyjaśnianiu niż tylko opisowi.
- Wyniki analizy dwuzmiennowej często przedstawiane są w formie tabel krzyżowych (tabel kontyngencji), które tworzone są w celu wyjaśniania wpływu zmiennej niezależnej na zmienną zależną.

⁷⁵ W Polsce można np. skorzystać z Klasyfikacji Zawodów i Specjalności przygotowanej przez Główny Urząd Statystyczny i dostępnej na ich stronie – przyp. M.J.L.

Wprowadzenie do analizy wielozmiennowej

- Analiza wielozmiennowa jest metodą analizy jednoczesnych zależności między wieloma zmiennymi. Może być także wykorzystywana do pełniejszego zrozumienia zależności między dwiema zmiennymi.
- Logika i techniki badań ilościowych mogą być przydatne także dla badaczy jakościowych.

Diagnozy społeczne

- Diagnoza społeczna to technika analizy ilościowej służąca do określania natury problemów społecznych, takich jak dyskryminacja etniczna czy płciowa.

Etyka i analiza danych ilościowych

- Bezstronna analiza i referowanie wyników to kwestia etyczna w takim samym stopniu w analizie ilościowej, jak w analizie jakościowej.
- I w trakcie ilościowej analizy danych, i referowania wyników należy chronić prywatność badanych podmiotów.

KLUCZOWE POJĘCIA

Analiza dwuzmiennowa	Rozproszenie
Analiza jednozmiennowa	Średnia
Analiza wielozmiennowa	Tabela krzyżowa (kontyngencji)
Książka kodowa	Wartość przeciętna
Mediana	Zmienna ciągła
Dominanta	Zmienna dyskretna
Odchylenie standardowe	Zmienna dyskretna
Rozkład częstości	

PROPOZYCJA BADANIA SPOŁECZNEGO: ANALIZA DANYCH IŁOŚCIOWYCH

W tym ćwiczeniu powinniśmy nakreślić swoje plany analizy. We wcześniejszych ćwiczeniach określiliśmy zmienne do analizy, a także to, w jaki sposób będziemy je mierzyć.

Teraz czas na przedstawienie raportu na temat tego, jak planujesz przeprowadzić analizę. Czy twoje cele mają głównie charakter opisowy czy eksplanacyjny? Jeśli wyjaśniający, to planujesz prostą analizę dwuzmiennową czy wielozmiennową? Trzeba też określić, czy planujesz analizę tabelaryczną, czy coś bardziej złożonego niż to, co zostało omówione w tym rozdziale. Tak naprawdę nie ma większego znaczenia, jakiego programu komputerowego używasz [SPSS, system Analizy Statystycznej (SAS) itd.], chyba że jest to program specjalistyczny lub taki, który nie jest powszechnie używany.

Jeśli udało ci się sformułować precyzyjną hipotezę, możesz chcieć ustalić poziomy istotności statystycznej, które określą znaczenie wyników. Nie zawsze jest to jednak konieczne.

Pytania i ćwiczenia

- W jaki sposób różne kierunki na twojej uczelni można podzielić na kategorie? Stwórz system kodowania, który pozwoli ci sklasyfikować je według jakiejś znaczącej zmiennej. Następnie utwórz inny system kodowania, używając innej zmiennej.
- Na ile sposobów możesz opisać to w kategoriach liczbowych? Jakie są niektóre z twoich najważniejszych atrybutów liczbowych? Czy potrafisz wyrazić niektóre ze swoich cech jakościowych w kategoriach ilościowych?
- Jak możesz stworzyć i zinterpretować tabelę krzyżową na podstawie następujących informacji: 150 demokratów popiera podniesienie płacy minimalnej, a 50 jest temu przeciwnych; 100 republikanów popiera podniesienie płacy minimalnej, a 300 jest temu przeciwnych.
- Używając hipotetycznych danych z poniższej tabeli, stwórz i zinterpretuj tabele ilustrujące trzy poniższe zależności:

wiek	orientacja polityczna	postawa wobec aborcji	częstość
młody	liberalny	zwolennik	90
młody	liberalny	przeciwnik	10
młody	konserwatywny	zwolennik	60
młody	konserwatywny	przeciwnik	40
stary	liberalny	zwolennik	60
stary	liberalny	przeciwnik	40
stary	konserwatywny	zwolennik	20
stary	konserwatywny	przeciwnik	80