# Statystyka dla studentów kierunków technicznych i przyrodniczych

Jacek Koronacki, Jan Mielniczuk

Opiniodawcy: *prof. dr hab. Stanisław Gnot*

*prof. dr hab. Aleksander Janicki*

Redaktor *Lilianna Szymańska*

Okładkę i strony tytułowe projektował *Wojciech J. Steifer*

Dzięki uprzejmości Muzeum Narodowego w Warszawie na okładce przedstawiono obraz Józefa Pankiewicza „Ulica w Madrycie"

Skład i korekta *Jan Ćwik, Jacek Koronacki, Jan Mielniczuk*

Wykonanie rysunków i łamanie *Jan Ćwik*

Książka jest nowoczesnym podręcznikiem statystyki. Jej pierwsza część (wstępna analiza danych, przejście od modelu probabilistycznego do wnioskowania statystycznego, podstawy wnioskowania statystycznego oraz analiza regresji) może być podstawą jednosemestralnego kursu wprowadzającego. W drugiej części przedstawiono wybrane, najbardziej istotne dla praktyka zagadnienia statystyki aplikacyjnej: analizę wariancji, analizę zależności cech jakościowych, metody próbkowania, zagadnienia symulacji komputerowej i metody rangowe. Całość stanowi podstawę dla rocznego wykładu ze statystyki. Przedmiot wykładu ujęty jest w sposób całościowy i w pełni wykorzystujący nowoczesne narzędzia statystyki obliczeniowej. Omawiane pojęcia i metody są motywowane i ilustrowane licznymi pochodzącymi z praktyki statystycznej przykładami. Po każdym rozdziale zamieszczono zadania do samodzielnego rozwiązania.
Książka jest przeznaczona przede wszystkim dla studentów kierunków technicznych, przyrodniczych, rolniczych i społecznych oraz dla wszystkich tych osób, które w swej pracy stykają się z problemami analizy danych.



Wydawnictwa Naukowo-Techniczne
00-048 Warszawa, ul. Mazowiecka 2/4
tel. 0-22 826 72 71, e-mail: wnt@wnt.pl
www.wnt.pl

ISBN 83-204-3242-1

J. Koronacki i J. Mielniczuk
Instytut Podstaw Informatyki PAN
01-248 Warszawa, ul. Jana Kazimierza 5
tel. 22 380 05 00, e-mail:
jacek.koronacki@ipipan.waw.pl
jan.mielniczuk@ipipan.waw.pl

Strona 3

# Statystyka dla studentów kierunków technicznych i przyrodniczych

Jacek Koronacki, Jan Mielniczuk

# Spis treści

# Przedmowa

Miniony wiek bywa nazywany wiekiem informacji. Moc obliczeniowa komputerów oraz pojemność ich pamięci rosły w ostatnich dziesięcioleciach nieomal z dnia na dzień. Doświadczyliśmy i nadal doświadczamy niebywałego rozwoju możliwości komunikacji w sieciach komputerowych. Oczywiście wiązał się z tym wszystkim ogromny wzrost możliwości gromadzenia informacji. W przypadku dużych baz danych mamy często do czynienia z megabajtami i nierzadko z terabajtami danych. Wielkiego znaczenia nabrała zatem potrzeba inteligentnego przetwarzania zebranych informacji.

Niejako w cieniu owej rewolucji informatycznej przez cały XX wiek trwał też niezwykły rozwój analizy danych i wnioskowania statystycznego, czyli – statystyki. Obydwa procesy nie były przy tym od siebie niezależne. Z jednej strony, bez statystyki nie ma możliwości pełnego zrozumienia i zinterpretowania wiedzy ukrytej w danych. Z drugiej zaś, techniczny rozwój komputerów umożliwił zalgorytmizowanie procedur statystycznych i rozwiązywanie zadań, z którymi człowiek sam nie mógłby sobie poradzić przez dziesiątki lat wytężonej pracy.

Jednym ze skutków rozwoju informatyki i statystyki jest upowszechnienie badań statystycznych w niemal wszystkich dziedzinach nauki i praktyki. We wszystkich sferach naszej działalności zbieramy dane, które – nie poddane odpowiedniej analizie – jawią się raczej jako niewiele mówiący chaos niż pewne uporządkowane *uniwersum*. Dzięki statystyce dostrzegamy ów ukryty w danych porządek oraz patrzymy na dane z właściwej perspektywy, tak jak autor obrazu, którego reprodukcję zamieszczamy na okładce, spojrzał na ulicę Madrytu – z bliska pełną południowego zamętu i chaotycznie poruszających się ludzi, a z perspektywy będącą częścią pięknego i zrozumiałego ładu.

Podręcznik ten jest wprowadzeniem w ład oferowany przez statystykę. Jest adresowany przede wszystkim do przyszłych techników i przyrodników, ale

uważamy, że będzie przydatny także dla studentów innych kierunków, zwłaszcza ekonomicznych, rolniczych, społecznych i medycznych. Powinien również zainteresować tych absolwentów wszystkich wymienionych kierunków, którzy uważają, że ich podstawowa wiedza statystyczna jest niedostateczna.

Oddawany do rąk Czytelnika podręcznik odbiega stylem od książek ze statystyki matematycznej już choćby dlatego, że jest adresowany do niematematyków. Swoją konstrukcją nawiązuje do anglosaskiej tradycji uczenia statystyki, którego celem jest danie dogłębnego i szerokiego, ale zarazem możliwie przystępnego wprowadzenia do przedmiotu.

Jesteśmy przekonani, że pojawienie się tego rodzaju podręcznika jest potrzebne, ponieważ przyczyni się do podniesienia poziomu zrozumienia i popularności statystyki wśród studentów. Niezbędne jest ukazanie się książki ujmującej wprowadzenie do statystyki w sposób całościowy, w pełni wykorzystującej nowoczesne techniki obliczeniowe. Jednocześnie Czytelnik musi nauczyć się wykorzystywania owych narzędzi w sposób odpowiedzialny i oparty na dobrym zrozumieniu przedmiotu. Nie trzeba nikogo przekonywać, że obecność na rynku licznych komputerowych pakietów statystycznych jest tyleż błogosławieństwem, co i przekleństwem, bowiem łatwo z nich korzystać bez żadnego zrozumienia oferowanych przez pakiet wyników.

Środkiem do lepszego zrozumienia przedmiotu nie może być nadmierna ścisłość formalna i przytaczanie wielu dowodów, lecz oparcie się na pouczających choć prostych przykładach i szerokiej argumentacji, odwołującej się do zdrowego rozsądku. Tak właśnie pisany jest nasz podręcznik. Jesteśmy przekonani, że w ten sposób można doskonale przekazać istotę rozumowania statystycznego. Za złożonym nawet matematycznym wywodem zawsze kryje się przejrzysta intuicja. To ją przede wszystkim powinen posiąść Czytelnik.

Zakres trzech pierwszych rozdziałów książki (rozdział zawierający wstępną analizę danych, rozdział poświęcony przejściu od modelu probabilistycznego do wnioskowania statystycznego oraz rozdział opisujący podstawy wnioskowania statystycznego) odpowiada typowym uczelnianym kursom ze statystyki; jednak różni się sposobem ujęcia materiału. W sposobie wykładu oraz wyborze tematów szczegółowych kierujemy się potrzebami praktyki oraz swymi doświadczeniami dydaktycznymi z uczelni w Warszawie (w ostatnich latach PJWSTK), The University of Michigan w Ann Arbor, Rice University w Houston i The University of New South Wales w Sydney. Niewątpliwy wpływ wywarły na nas najlepsze podręczniki anglosaskie, zwłaszcza książka Moore'a i McCabe'a „Introduction to the Practice of Statistics", Freeman & Co 1998, którą najczęściej sami wykorzystywaliśmy w nauczaniu.

Trzy pierwsze rozdziały tego podręcznika uzupełnione o omówioną w rozdziale 4 analizę regresji są pomyślane jako podstawa kursu semestralnego,

wprowadzającego słuchacza w zagadnienia statystyki i obejmującego tygodniowo przynajmniej dwugodzinny wykład oraz dwugodzinne laboratorium. W ramach kursu semestralnego udaje się omówić tylko zasadnicze kwestie analizy regresji. W pięciu następnych rozdziałach przedstawiono wybrane, najbardziej istotne dla praktyka zagadnienia statystyki: analizę wariancji i analizę zależności cech jakościowych, metody próbkowania, zagadnienia symulacji komputerowej i metod rangowych. Na podstawie tych rozdziałów oraz zaprezentowanej obszerniej analizy regresji, wykładowca może zaplanować drugi semestr wykładu ze statystyki.

Wśród zagadnień szczegółowych nie znalazło się miejsce dla niezwykle ważnych metod statystyki wielowymiarowej, które – mamy nadzieję – staną się treścią naszego następnego podręcznika. Nie mamy przy tym wątpliwości, że ze względu na występującą obecnie złożoność danych metody wielowymiarowe staną się już wkrótce elementem podstawowych wykładów ze statystyki.

Chcielibyśmy podkreślić fakt, że książka powstała w ramach działalności statutowej Instytutu Podstaw Informatyki Polskiej Akademii Nauk. W trakcie przygotowywania kolejnych wersji manuskryptu, korzystaliśmy z wnikliwych uwag Stanisława Gnota, Andrzeja Dąbrowskiego i Andrzeja Michalskiego oraz Elżbiety Ferenstein, którym serdecznie dziękujemy. Jesteśmy bardzo wdzięczni naszemu najbliższemu współpracownikowi, Janowi Ćwikowi, który sporządził wszystkie rysunki, przygotował ostateczny skład książki, przeliczył i sprawdził wiele przykładów oraz pomagał nam w trakcie kolejnych korekt. Składamy podziękowania Polsko-Japońskiej Wyższej Szkole Technik Komputerowych za finansowe wsparcie tego wydania naszego podręcznika. Dziękujemy Muzeum Narodowemu w Warszawie za wyrażenie zgody na reprodukcję obrazu Józefa Pankiewicza „Ulica w Madrycie". Z wdzięcznością myślimy o znakomitej pracy redakcyjnej Pani Lilianny Szymańskiej i o opiece Pani Redaktor Zofii Leszczyńskiej nad całością przedsięwzięcia. Bez ich wielkiego poświęcenia i zaangażowania szybkie wydanie tej książki byłoby niemożliwe.

*Jacek Koronacki i Jan Mielniczuk*

Warszawa, w lipcu 2001

W wydaniu trzecim powołujemy się na książkę Jacka Koronackiego i Jana Ćwika (2005): *Statystyczne systemy uczące się*, Warszawa, WNT. Książka ta jest poświęcona wybranym metodom statystyki wielowymiarowej i ogólnej analizie regresji. Odwołując się do tej monografii będziemy pisać krótko – p. (J. Koronacki, J. Ćwik (2005)).

ROZDZIAŁ 1

# Wstępna analiza danych

## 1.1. Wprowadzenie

W rozdziale tym opiszemy niezbędny zestaw działań podejmowanych w sytuacji, gdy spotykamy się po raz pierwszy z nowymi danymi. Naszym zadaniem wtedy jest opis *podstawowych* ich cech. Główne cechy danych mówią nam o zasadniczych własnościach zjawiska lub eksperymentu, który badamy. Ponadto, prawie zawsze potrzebny jest nam syntetyczny opis danych: bardzo trudno jest na przykład analizować „surowe" wyniki spisu powszechnego w Polsce. Konieczne jest dokonanie odpowiedniego ich przekształcenia i uproszczenia umożliwiającego analizę. Przede wszystkim musimy jednak ustalić, jaki jest typ danych. Jeśli mamy do czynienia z liczbami odpowiadającymi wartościom mierzonych wielkości, jak na przykład w przypadku pomiaru temperatury przy gruncie o godzinie ósmej rano na Śnieżce w kolejnych dniach listopada, to mówimy wtedy o **danych ilościowych**. W przypadku, gdy rejestrujemy cechę jakościową obiektów, na przykład płeć lub typ schorzenia pacjentów, mówimy o **danych jakościowych**. Oczywiście, jeśli dla jednego obiektu dokonujemy kilku pomiarów, to część z nich może być typu ilościowego, a część jakościowego. Możemy rejestrować jednocześnie wiek pacjenta (cecha ilościowa) i to, czy ma on lub nie problemy ze snem (cecha jakościowa). Określenie typu danych jest niezbędne przed przystąpieniem do ich wstępnej analizy.

## 1.2. Graficzne przedstawienie danych

Nie bez powodu rozpoczynamy rozdział o wstępnej analizie danych dyskusją dotyczącą konstrukcji i analizy wykresów. Wykres zawiera znacznie więcej informacji niż jeden, a nawet kilka wskaźników liczbowych obliczonych na podstawie danych. Często jest tak, że wartość pewnego wskaźnika

odpowiada dwóm zupełnie różnym wykresom i dlatego opieranie się wyłącznie na wartości tego wskaźnika może być mylące. Zarazem, wykres też jest pewną redukcją informacji w stosunku do oryginalnych danych, ale jest to redukcja bez porównania mniej drastyczna.

### 1.2.1. Wykresy dla danych jakościowych

Zacznijmy od sporządzenia wykresów dla danych jakościowych opisujących jedną cechę. Problem analizy danych dla kilku cech zostanie omówiony w rozdz. 6.

**Przykład 1.1.** Rozpatrzmy następujące dane dotyczące składu wyznaniowego ludności Warszawy w latach 1864 i 1917 (źródło: *400 lat stołeczności Warszawy*. Zakład Wydawnictw Statystycznych, Warszawa, 1997).

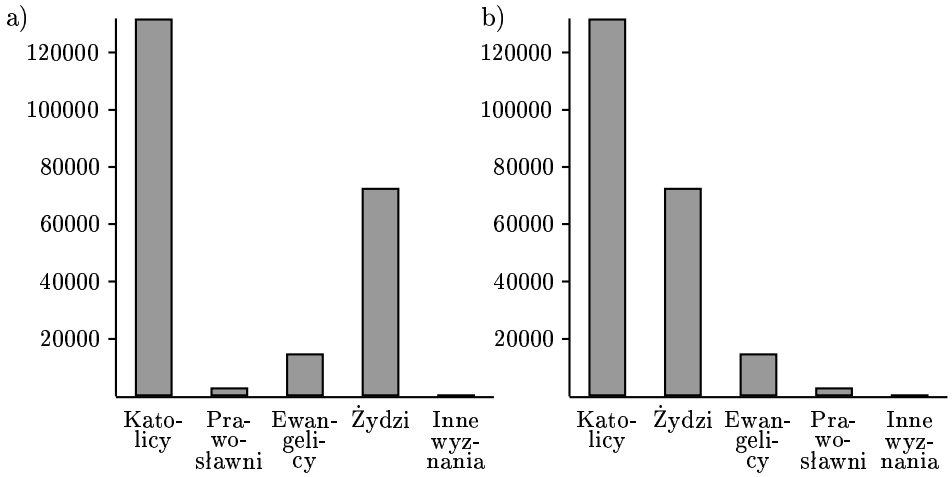
**Tabela 1.1. Skład wyznaniowy ludności Warszawy**

| Kategoria wyznaniowa | Rok 1864 | | Rok 1917 | |
|---|---|---|---|---|
| | Liczebność | % | Liczebność | % |
| Katolicy | 131808 | 59,1 | 387069 | 46,2 |
| Prawosławni | 3026 | 1,4 | 3961 | 0,5 |
| Ewangelicy | 15909 | 6,7 | 12147 | 1,5 |
| Żydzi | 72772 | 32,6 | 329535 | 39,3 |
| Inne wyznania | 287 | 0,2 | 104500 | 12,5 |

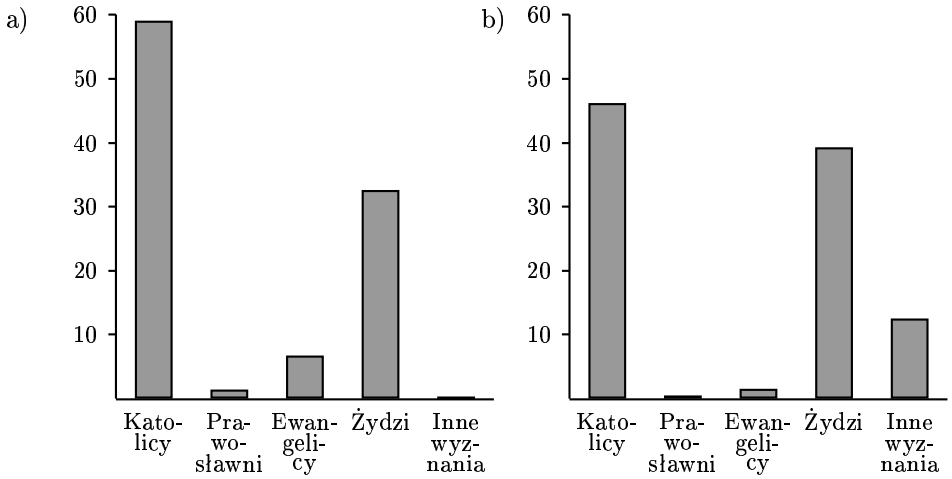
Liczebności poszczególnych grup wyznaniowych w 1864 roku (druga kolumna tab. 1.1) można przedstawić za pomocą wykresu słupkowego jak na rys. 1.1, na którym wysokości słupków są równe odpowiednim liczebnościom, a wspólna szerokość słupków jest dowolna.

Zauważmy, że na rys. 1.1 podstawy wszystkich słupków są takie same, a zatem porównanie liczności w dwóch różnych kategoriach wyznaniowych może polegać nie tylko na porównaniu wysokości, ale i pola słupków. Z wykresu natychmiast widać, że najliczniejsze kategorie wyznaniowe to katolicy i żydzi. Kolejność kategorii na wykresie jest nieistotna. Wykres 1.1b, na którym zamieniono miejsca kategorii „żydzi" i „prawosławni" zawiera dokładnie tyle samo informacji co wykres 1.1a. W przypadku danych jakościowych możemy w dowolny sposób ponumerować rozpatrywane kategorie na przykład liczbami od 1 do 5 i zastąpić nazwy na wykresie odpowiednią liczbą. W tym przypadku osoba, dla której wartość cechy wynosi 2, oznaczałaby osobę prawosławną.

Alternatywnie, zamiast liczności na wykresie możemy przedstawić częstości (frakcje) lub procentowe udziały odpowiednich wyznań. Sporządźmy na przykład wykres słupkowy procentowego składu wyznaniowego dla roku 1864 (rys. 1.2a).



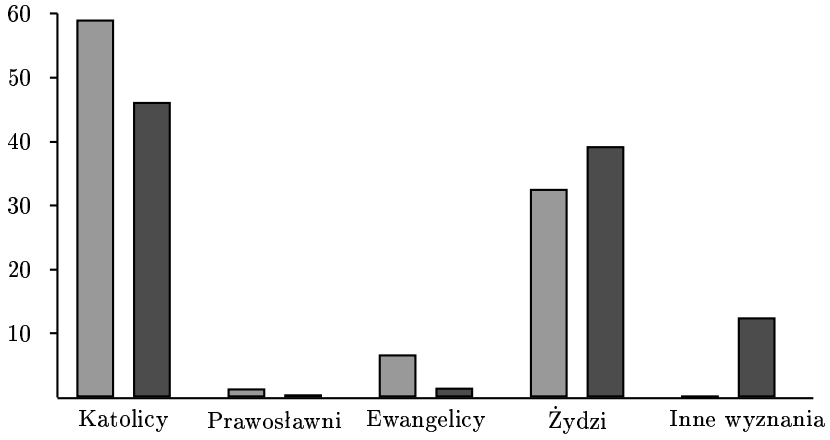

Rys. 1.1. Wykres słupkowy składu wyznaniowego ludności Warszawy w roku 1864




Rys. 1.2. Wykresy słupkowe procentowego składu wyznaniowego ludności Warszawy: a) rok 1864, b) rok 1917

Zauważmy, że jego kształt jest dokładnie taki sam jak wykresu na rys. 1.1a, mimo że wysokości słupków odpowiadają teraz udziałowi procentowemu, a nie liczebności danej kategorii. Możemy teraz łatwo znaleźć procentowy

udział ludności w połączonych kategoriach, na przykład katolików, prawosławnych i ewangelików było łącznie 59,1% + 1,4% + 6,7% = 67,2%. Procentowy wykres słupkowy jest bardziej użyteczny od opartego na liczebnościach, gdy chcemy porównać dane pogrupowane w tych samych kategoriach dla różnych lat. Skład wyznaniowy w Warszawie w latach 1864 i 1917 można przedstawić (rys. 1.3) także w trochę inny sposób, zestawiając obok siebie procentowe wykresy słupkowe dla kolumn 3 i 5 tab. 1.1. Pierwszy z przylegających dwu słupków odpowiada rokowi 1864.



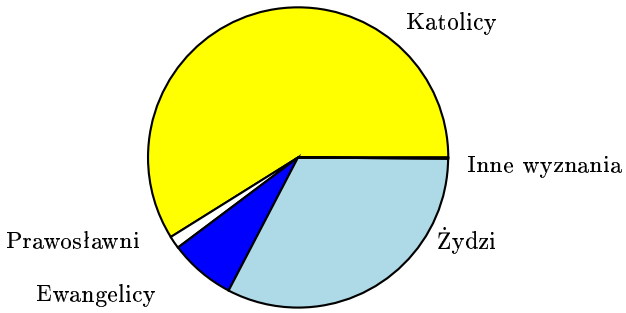

Rys. 1.3. Wykres słupkowy procentowego składu wyznaniowego ludności Warszawy z lat 1864 i 1917

Z powyższego wykresu można wyciągnąć ciekawe wnioski. W porównaniu z rokiem 1864, w roku 1917 nastąpił ponad 10-procentowy spadek udziału katolików w składzie wyznaniowym (przy jednoczesnym prawie trzykrotnym wzroście ich liczebności), ponad czterokrotny spadek udziału ewangelików i aż ponad sześćdziesięciokrotny wzrost udziału ludności innych wyznań (a raczej, jak należy przypuszczać, liczby ludzi deklarujących się jako niewierzących). Zauważmy, że połączenie wykresów słupkowych dla liczności nie dałoby możliwości porównania względnych (procentowych) zmian w poszczególnych kategoriach, a jedynie liczby ludzi w poszczególnych kategoriach.

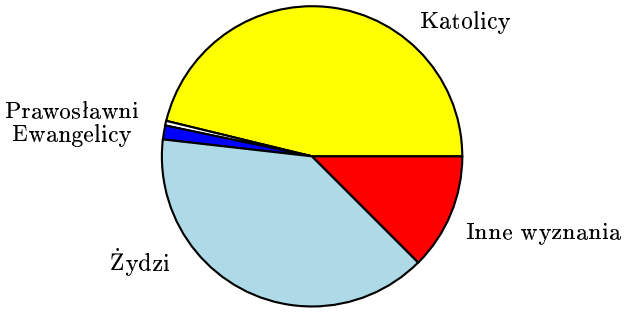
Wykresy słupkowe z rys. 1.2 można alternatywnie przedstawić za pomocą tak zwanych wykresów kołowych (rys. 1.4 i 1.5).

Na rysunku 1.4 kąt sektora odpowiadającego katolikom jest równy $0,59 \times$ $\times 360^\circ = 212{,}4^\circ$, ewangelikom $0{,}067 \times 360^\circ = 13{,}3^\circ$ itd. Zauważmy ograniczenia związane z wykresem kołowym: można za jego pomocą przedstawić tylko dane procentowe, wszystkie kategorie łącznie muszą dawać 100%, czyli każda obserwacja powinna być umieszczona w jednej z rozpatrywanych kategorii. W naszym przykładzie nie możemy jednoznacznie przedstawić udziału

jedynie czterech pierwszych kategorii wyznaniowych. Przy występowaniu wielu kategorii wykresy kołowe stają się mało czytelne, gdyż część sektorów będzie wąska i trudno porównywalna. Również wzajemna analiza dwóch wykresów kołowych jest bardziej kłopotliwa niż połączonego wykresu słupkowego.




Rys. 1.4. Skład wyznaniowy ludności Warszawy w 1864 r.




Rys. 1.5. Skład wyznaniowy ludności Warszawy w 1917 r.

### 1.2.2. Wykresy dla danych ilościowych

Rozpatrzmy następujący przykład.

**Przykład 1.2.** W stu kolejnych rzutach kostką otrzymaliśmy następujące wyniki:

5 2 2 6 3 2 5 3 1 2 5 3 6 2 5 4 4 6 1 6 4 5 5 2 4 6 1 4 4 3 4 2 4 2 4 4 1 1
4 5 3 1 5 6 5 6 1 5 6 2 4 5 5 2 5 4 5 5 1 1 2 2 5 5 2 6 3 5 5 4 1 4 5 5 1 4
3 2 1 2 6 1 2 1 6 5 1 3 6 1 5 6 6 2 2 3 5 5 2 4.

Oczywiście, mamy tu do czynienia z próbą wartości cechy ilościowej, będącą liczbą oczek w poszczególnych rzutach. Zauważmy, że na przykład liczba „2", oznaczająca wypadnięcie dwóch punktów na kostce nie podlega konwencji przypisania liczb kategoriom jak w przypadku danych jakościowych. Mając próbę wyników, chcielibyśmy ją w zwięzły sposób opisać. Najprostszym sposobem zrobienia tego jest podanie rozkładu cechy dla danej próby, będącego zapisem jakie wartości cecha przyjmuje w próbie i jak często. W naszym przykładzie obserwujemy wszystkie wartości od 1 do 6, a odpowiednie liczebności wystąpień wynoszą: 16, 19, 9, 17, 25, 14.

Zatem rozkład liczby oczek w próbie ma postać:

| Wartość (liczba oczek) | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| Liczność (liczba wystąpień) | 16 | 19 | 9 | 17 | 25 | 14 |
| Częstość | 0,16 | 0,19 | 0,09 | 0,17 | 0,25 | 0,14 |

Zauważmy, że jedyną informacją, którą tracimy, zastępując próbę przez jej rozkład, jest informacja o kolejności pojawiania się poszczególnych wartości. Często (ale jak dowiemy się z następnego punktu, nie zawsze) jest to informacja nieistotna. W rozpatrywanym przykładzie nieistotne jest dla nas, w jakich momentach pojawiała się na przykład liczba 6, tylko jak często się pojawiła.

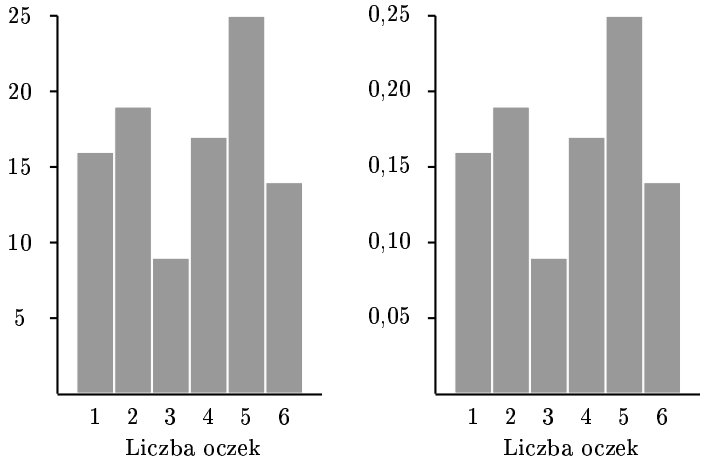
Ogólnie, gdy $x_1, \ldots, x_n$ są zaobserwowanymi wartościami cechy ilościowej, przez $y_1 < y_2 < \ldots < y_k$ oznaczymy różne, uporządkowane wartości spośród nich. Ponadto, niech $n_i$ będzie liczbą powtórzeń wartości $y_i$ w próbie, $i = 1, \ldots, k$. Wówczas **rozkładem cechy w próbie** $x_1, \ldots, x_n$ nazywamy ciąg $(y_1, n_1), \ldots, (y_k, n_k)$. Często w definicji rozkładu zamiast wartości $n_i$ podaje się częstość występowania wartości $y_i$, to jest $n_i/n$. Jeśli liczba wartości przyjmowanych przez cechę jest niewielka, jej rozkład w próbie można przedstawić za pomocą diagramu liczebności lub diagramu częstości. Diagramy liczebności i częstości przedstawiono na rys. 1.6.

W podobny sposób możemy zbudować diagram liczby przekroczeń przez sumy opadów w lipcu wartości 120 mm w ciągu dekady. Przedstawione dane dotyczą 15 dekad od roku 1811 do 1960 (Z. Kaczmarek (1970): *Metody statystyczne w hydrologii i meteoroologii.* Warszawa, WKiŁ).

| Liczba przekroczeń | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| Liczność | 2 | 5 | 4 | 3 | 0 | 1 |

Rozkłady takie są czasami przedstawiane również za pomocą modyfikowanego wykresu słupkowego, w którym słupki przylegają do siebie, a kategorie odpowiadają kolejnym liczbom przekroczeń. Z tak sporządzonego wy-

kresu zauważymy natychmiast, że najczęściej występująca liczba przekroczeń w dekadzie to 1, później 2, i że zdarzyła się jedna dekada, w której przekroczenie poziomu 120 mm nastąpiło aż 5 razy (były to lata 1851-1860, czego już z wykresu słupkowego nie odczytamy).




Rys. 1.6. Diagramy liczebności i częstości dla danych z przykł. 1.2

W przypadku dużej liczby wartości dokonujemy dalszej redukcji informacji, grupując obserwowane wartości w przedziały, co prowadzi do koncepcji histogramu.

**Przykład 1.3.** Rejestrujemy wiek 20 pracowników zgłaszających się na okresowe badania w pewnym zakładzie pracy. Zaobserwowane wielkości wynoszą (w latach):

36, 41, 33, 34, 38, 26, 33, 36, 30, 48, 39, 31, 35, 36, 38, 37, 22, 31, 25, 32.

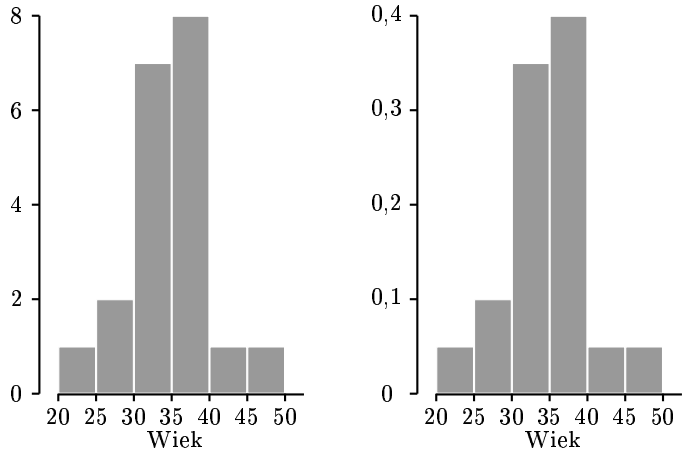
Liczba różnych wartości w próbie jest równa 16 i diagram rozkładu lat w próbie składający się z szesnastu słupków nie byłby specjalnie czytelny. Dlatego też dokonujemy agregacji danych, wybierając najpierw podział na pewne przedziały wiekowe, a następnie grupując obserwacje w klasy, w zależności od przedziału, do którego wpadają. Oczywiście, pierwszy przedział powinien być wybrany tak, aby najmniejsza obserwacja należała do odpowiadającej mu pierwszej klasy. Ponieważ najmłodszy z pracowników w próbie ma 22 lata, a najstarszy 48 lat, możemy na przykład rozpatrzeć następujące przedziały wiekowe:

$$[20,25),\ [25,30),\ [30,35),\ [35,40),\ [40,45),\ [45,50).$$

Odpowiedni podział próby na klasy wygląda następująco:

| Przedział | Klasa | Liczność klasy | Częstość |
|---|---|---|---|
| $[20, 25)$ | 22 | 1 | $1/20 = 0{,}05$ |
| $[25, 30)$ | 26, 25 | 2 | $2/20 = 0{,}1$ |
| $[30, 35)$ | 33, 34, 33, 30, 31, 31, 32 | 7 | $7/20 = 0{,}35$ |
| $[35, 40)$ | 36, 38, 36, 39, 35, 36, 38, 37 | 8 | $8/20 = 0{,}4$ |
| $[40, 45)$ | 41 | 1 | $1/20 = 0{,}05$ |
| $[45, 50)$ | 48 | 1 | $1/20 = 0{,}05$ |

Sporządzenie histogramu polega na naniesieniu na osi poziomej rozpatrywanych przedziałów i zbudowaniu nad nimi przylegających do siebie słupków, których wysokość jest równa liczebności lub częstości danej klasy. W naszym przykładzie histogramy liczebności i częstości wyglądają jak na rys. 1.7.




Rys. 1.7. Histogramy liczebności i częstości dla danych z przykł. 1.3

Wybór początku histogramu (początku pierwszego przedziału), jak i długości przedziału w dużej mierze zależy od nas; jednocześnie jak zobaczymy, ma on wpływ na wizualizację podstawowych cech danych. Problemem tym zajmiemy się dalej.

Zauważmy, że konstrukcja histogramu jest bardzo podobna do konstrukcji wykresu słupkowego. Poszczególne przedziały mają jednak teraz określoną długość odpowiadającą zakresowi wartości. Ponieważ długość przedziału jest stała, więc pola słupków są proporcjonalne do liczebności i częstości klas. Zmiana pola słupka odpowiada zatem zmianie częstości obserwacji w odpowiadającym przedziale. Zauważmy, że korzystając z histogramu częstości możemy natychmiast obliczyć częstość pracowników w próbie, mających co najmniej 30 lat. Wynosi ona $0{,}35 + 0{,}40 + 0{,}05 + 0{,}05 = 0{,}85$. Alternatywnie możemy obliczyć tę częstość, odejmując od 1 częstość pracowników mających mniej niż 30 lat: $1 - (0{,}05 + 0{,}1) = 0{,}85$.

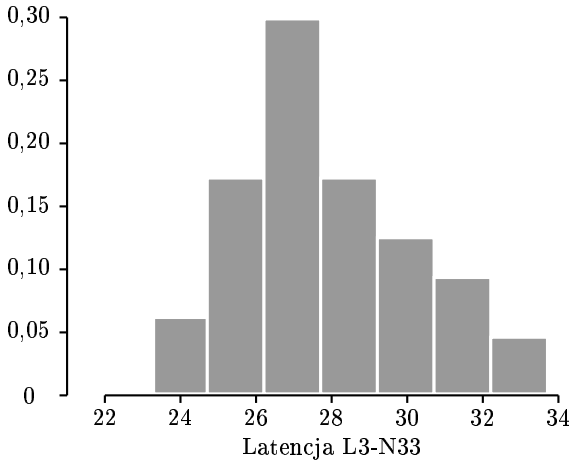
Kształt histogramu na rysunku jest w przybliżeniu symetryczny, ma on jedno maksimum, zwane często **modą**. Z tego powodu taki histogram jest nazywany **jednomodalnym**, w odróżnieniu od histogramów **wielomodalnych**, posiadających kilka maksimów lokalnych. Moda histogramu nie ma jednej wartości liczbowej, odpowiada jej cały przedział, do którego wpada najwięcej wartości w próbie, w naszym przykładzie przedział [35, 40). Zauważmy, że w tym przypadku modę można uznać za naturalny „środek" rozkładu wieku w próbie.

**Przykład 1.4.** Rozpatrzmy teraz inny przykład dotyczący dermatomalnych somatosensorycznych potencjałów wywołanych u zdrowych osobników (dane zebrane przez dr M. Rakowicz, Instytut Psychiatrii i Neurologii w Warszawie). Rozpatrywaną tu cechą jest jedna z charakterystyk tego potencjału zwana latencją L3-N33, jest to czas od momentu wzbudzenia potencjału w tzw. korzeniu L3 do osiągnięcia przez potencjał pierwszego maksimum lokalnego. W badaniu jest rejestrowany potencjał wzbudzony w kończynie lewej. Dane zebrane dla 62 pacjentów (w milisekundach) są następujące:

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 26,40 | 31,60 | 29,60 | 28,20 | 24,80 | 26,50 | 25,85 | 26,10 | 26,90 | 26,05 | 31,40 |
| 28,00 | 25,55 | 29,70 | 26,80 | 28,80 | 26,50 | 28,30 | 30,50 | 24,70 | 25,30 | 30,20 |
| 29,20 | 28,40 | 26,90 | 25,50 | 26,40 | 33,00 | 25,20 | 26,60 | 27,50 | 25,10 | 24,60 |
| 31,80 | 29,80 | 27,90 | 30,20 | 26,50 | 31,60 | 25,60 | 26,50 | 27,50 | 28,40 | 27,10 |
| 30,90 | 30,30 | 30,10 | 28,70 | 27,60 | 27,60 | 28,70 | 32,90 | 26,30 | 26,30 | 27,40 |
| 26,80 | 24,20 | 28,70 | 31,50 | 26,00 | 32,60 | 24,60 | | | | |

Zbudujmy histogram (rys. 1.8) dla powyższych danych oparty na 7 przedziałach długości 1,5 milisekundy, rozpoczynający się od punktu 23,25 milisekundy.

Histogram ma wyraźną modę; jest nią przedział wartości [26,25, 27,75). Oznacza to, że dla największej liczby osobników ich czasy latencji L3-N33 były zawarte między 26,25 a 27,75 milisekundy. W odróżnieniu od histogramu z poprzedniego przykładu nie jest on w przybliżeniu symetryczny: wartości histogramu po prawej stronie mody maleją znacznie wolniej niż po jej lewej stronie. Czasami mówimy w tej sytuacji, że prawy ogon histogramu jest znacznie dłuższy i maleje wolniej niż jego lewy ogon. Taki histogram, a zarazem rozkład cechy w próbie, dla której jest on skonstruowany jest nazywany **prawostronnie skośnym** (**dodatnio skośnym** lub **prawostronnie asymetrycznym**). Gdy sytuacja po obu stronach mody jest odwrotna mówimy o **lewostronnej** (**ujemnej**) **skośności** lub **lewostronnej asymetrii**. Specjalnych komentarzy nie wymaga natomiast wyjaśnienie, co oznacza histogram wyostrzony lub spłaszczony.

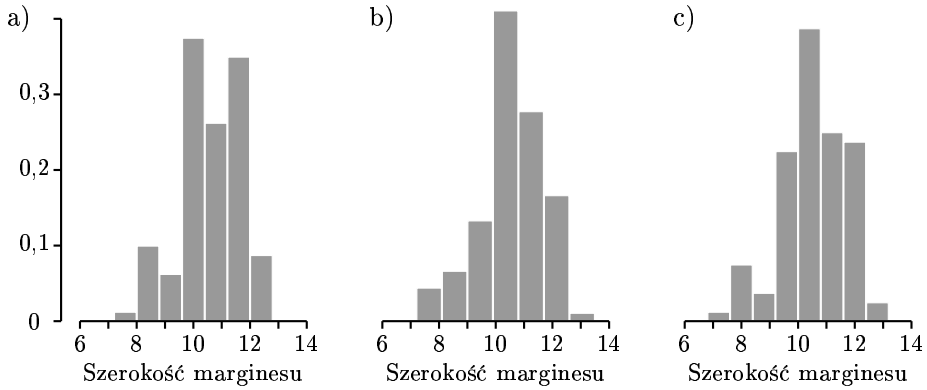



Rys. 1.8. Histogram częstości dla danych z przykł. 1.4

Rozpatrzmy jeszcze jeden przykład.

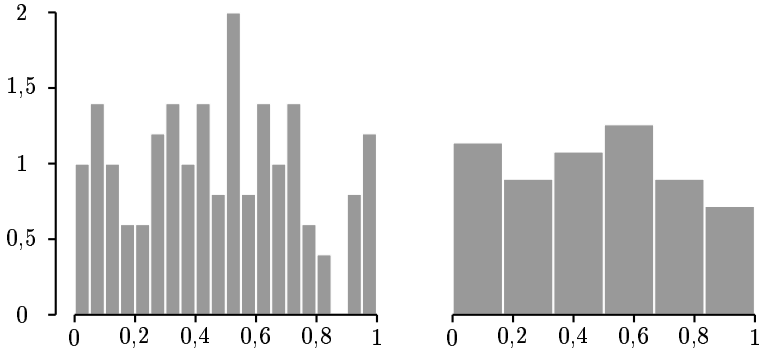
**Przykład 1.5.** Dane dotyczą szerokości (w milimetrach) dolnego marginesu 100 fałszywych banknotów dwudziestofrankowych (frank szwajcarski). Przy przyjęciu początku pierwszego przedziału jako 7,2 mm i jego długości $h = 0{,}8$ mm otrzymamy histogram, mający 3 mody (przedział drugi, czwarty i szósty na rys. 1.9a). Gdy zachowamy początek pierwszego przedziału i zmienimy długość na $h = 0{,}9$ mm histogram „straci" pierwszą i trzecią modę (rys. 1.9b). Z kolei zmiana początku histogramu na 6,8 mm przy zachowaniu pierwszej długości przedziału $h = 0{,}8$ mm prowadzi również do zmniejszenia liczby mód, ale tym razem tylko o jedną (rys. 1.9c).

Widzimy, że wybór początku histogramu i długości przedziału mogą mieć duży wpływ na jego kształt. Zanim przedstawimy pewne systematyczne podejście do rozwiązania tego problemu, zauważmy, że często dysponujemy dodatkową informacją pomagającą wybrać właściwy kształt spośród wielu zbudowanych dla różnych początków i długości przedziału. Na przykład trzy mody na rys. 1.9a mogą odpowiadać trzem różnym miejscom fałszowania banknotów. Jeśli wiemy, że banknoty pochodziły faktycznie od „producentów" z trzech źródeł, jest to istotny argument przemawiający za wyborem histogramu trójmodalnego. Ogólnie zauważmy, że histogram o kilku modach może wskazywać na to, że obserwacje pochodzą z kilku istotnie różnych populacji.




Rys. 1.9. Histogramy dla danych z przykł. 1.5

**Przykład 1.6.** Rozpatrzmy histogram zbudowany dla próby 100 losowo wybranych liczb z odcinka (0, 1). Za początek histogramu przyjęto 0, a długość przedziału jest równa 0,05 (rys. 1.10). Ponieważ duża zmienność wysokości słupków może być spowodowana stosunkowo małą wartością parametru $h$, zwiększamy jego wartość do $h = 1/6 = 0{,}167$.




Rys. 1.10. Histogramy dla danych z przykł. 1.6, dla długości przedziału $h = 0{,}05$ i $h = 0{,}167$

Zbliżone wysokości słupków sugerują, iż mniej więcej tyle samo obserwacji wpada do każdego przedziału o długości 0,167. Taki histogram nazywamy w przybliżeniu jednostajnym. Zauważmy, że mamy do czynienia z sytuacją bardzo podobną do sytuacji z przykł. 1.2. Tu rozpatrujemy sześć przedziałów takich, dla których częstość wpadania do każdego z nich wynosi 1/6. W przykładzie 1.2 częstość wypadnięcia każdej liczby oczek od 1 do 6 wynosiła tyle samo.

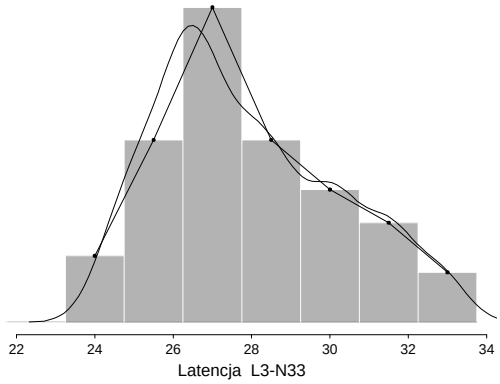
### Wybór długości przedziału i początku histogramu

Przedstawimy tylko jedną z metod wyboru długości przedziału. Reguła ta zwykle działa dobrze w praktyce. Opiera się ona na początkowym wyborze długości $h_0$, która jest adekwatna dla pewnego często występującego kształtu histogramu, tak zwanego kształtu normalnego (którym zajmiemy się w podrozdz. 1.4). Wielkość $h_0$ wynosi

$$h_0 = 2{,}64 \times IQR \times n^{-1/3}, \tag{1.1}$$

gdzie $IQR$ jest tak zwanym rozstępem międzykwartylowym, opisującym rozproszenie danych (def. 1.9), a $n$ oznacza liczebność próby. Podkreślmy, że stosowanie wzoru (1.1) ma sens tylko dla stosunkowo licznych prób ($n \geq 50$). Dla małych prób ($30 < n < 50$) stosuje się z reguły nie więcej niż 4–5 przedziałów. Co jednak zrobić, gdy podejrzewamy, że kształt histogramu adekwatnie opisującego dane może znacznie odbiegać od kształtu normalnego? Sensowne wydaje się wtedy stopniowe zmniejszanie lub zwiększanie długości przedziału i obserwowanie, jaki wpływ będzie miała ta zmiana na kształt histogramu. Pamiętamy z przykł. 1.5 i 1.6, że zmniejszanie długości przedziału powoduje zwiększenie stopnia zmienności histogramu i odwrotnie, zwiększanie $h$ prowadzi do coraz większego jego wygładzenia. Jeśli histogram dla początkowej długości $h_0$ wydaje nam się za bardzo nieregularny, staramy się go wygładzić, zastępując $h_0$ kolejno przez coraz większe wartości $ah_0, a^2h_0$ itd, gdzie $a$ przyjmuje się na przykład równe 1,2 lub 1,5. Zwiększanie długości przedziału powinniśmy przerwać w momencie, gdy stwierdzamy, że histogram staje się zbyt wygładzony. Pamiętajmy, że zwiększanie $h$ jest związane z coraz większą redukcją informacji: wartości cechy są zastępowane przez zliczanie ich wartości w coraz dłuższych przedziałach. Zwiększając długość coraz bardziej, otrzymamy w końcu histogram składający się tylko z jednego słupka! Odwrotnie, gdy początkowy histogram wydaje się nam zbytnio wygładzony, zastępujemy długość przedziału $h_0$ przez coraz mniejsze wartości $a^{-1}h_0, a^{-2}h_0$ itd. i przerywamy proces w momencie wystąpienia zbyt dużych nieregularności. Oczywiście, pojęcia zbytniego wygładzenia i nieregularności mogą się w praktyce okazać bardzo subiektywne, dlatego są tu pomocne wszelkie informacje dodatkowe, na przykład dotyczące liczby mód dla „właściwego" histogramu. Pamiętajmy również, że zmiana długości przedziału (jak i początku histogramu) powoduje zawsze duże zmiany kształtu w przypadku małych prób.

Problem wyboru początku histogramu nie ma również jednego rozwiązania. Godny polecenia wydaje się wybór początku tak, aby najmniejsza wartość była środkiem pierwszego przedziału histogramu. Skuteczną metodą uniezależnienia się od wpływu początku histogramu na otrzymany kształt jest uśrednienie pewnej liczby histogramów, których początki są nieznacznie




Rys. 1.11. Łamana częstości i krzywa estymatora jądrowego dla danych z przykł. 1.4

przesunięte względem siebie (metoda ASH; D. Scott (1992): *Multivariate density estimation*. Wiley, New York).

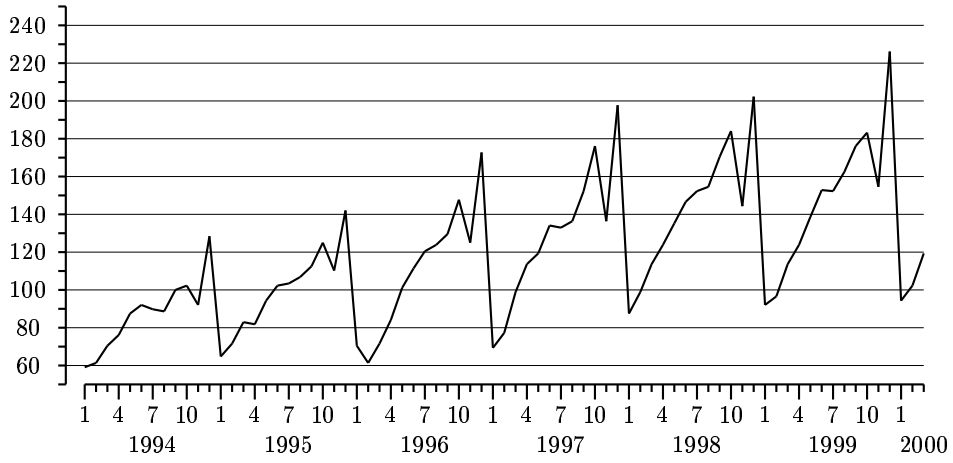
Na koniec zauważmy, że problem braku ciągłości histogramu możemy rozwiązać, łącząc środki górnych odcinków jego słupków i otrzymując tzw. łamaną częstości. W przypadku, gdy interesuje nas kształt bardziej gładki, możemy narysować krzywą tak zwanego estymatora jądrowego (rys. 1.11) lub opartego na funkcjach sklejanych. Estymatory takie są dostępne w pakietach i nie będziemy ich tu omawiać. Nie unikniemy jednak w ten sposób problemu wyboru pewnego parametru tego estymatora, będącego odpowiednikiem długości przedziału dla histogramu. Nie zatrzymując się dłużej nad tymi problemami, zwróćmy uwagę na to, że wybór długości przedziału i jego początku w każdym pakiecie jest wynikiem pewnego zautomatyzowanego procesu, zazwyczaj podobnego do opartego na równaniu (1.1), który nie musi dawać najlepszego wyniku w przypadku aktualnie rozpatrywanych przez nas danych. Dlatego bezpiecznie jest stwierdzić jak wygląda histogram przy kilku alternatywnych wyborach tych wielkości.

### 1.2.3. Wykresy przebiegu

Jeśli dane ilościowe są zbierane w następujących po sobie momentach czasowych, dobrym pomysłem na ich wizualizację jest sporządzenie ich wykresu w funkcji czasu. Dane tego typu noszą nazwę **szeregu czasowego**, a odpowiedni wykres będziemy nazywać **wykresem przebiegu**. Na jego podstawie można się przekonać, czy wartości zebrane w różnych odcinkach czasowych zachowują się podobnie i czy istnieje zależność między wartościami obserwowanymi w sąsiednich momentach czasowych. Tego typu

informacji nie można uzyskać po przeanalizowaniu histogramu, który rejestruje tylko zagregowane w przedziały wartości cechy, pomijając momenty czasowe, w których się one pojawiły.

Popatrzmy na wykres przebiegu produkcji sprzedanej budownictwa od stycznia 1994 do stycznia 2000 roku (rys. 1.12, na podstawie danych GUS-u). Wartości są rejestrowane co miesiąc przy przyjęciu średniej produkcji miesięcznej w 1995 roku jako 100. Obserwacje dla kolejnych momentów czasowych połączono odcinkami i otrzymano wykres w postaci linii łamanej. Dwie cechy wykresu są łatwo zauważalne: powolna, ale wyraźna ogólna tendencja wzrostu oraz powtarzający się cyklicznie kształt wykresu w poszczególnych latach. Produkcja sprzedana jest najniższa w styczniu i lutym każdego roku, później rośnie do października, po czym następuje późnojesienny zwrot powodujący spadek w listopadzie, a następnie pojawia się zwrot w przeciwnym kierunku, którego rezultatem jest największa (w skali roku!) produkcja sprzedana w grudniu (na co niepośledni wpływ ma tak zwana ulga podatkowa na budowę oraz remont i modernizację mieszkań).




Rys. 1.12. Wykres przebiegu produkcji sprzedanej budownictwa przy średniej miesięcznej produkcji w 1995 r. przyjętej jako 100 (1, 4, 7, 10 oznaczają początki kwartałów, czyli odpowiednio miesiące: styczeń, kwiecień, lipiec i październik)

Ogólną, stałą tendencję wzrostową lub spadkową nazywamy **trendem**, a kształt wycinka wykresu pojawiający się cyklicznie w kolejnych przedziałach czasowych, **zmiennością sezonową**. Ważnym zadaniem statystycznym jest wyodrębnienie trendu i zmienności sezonowej oraz analiza szeregu czasowego po odjęciu tych składników. Często opisane składniki szeregu czasowego nie są tak ewidentne jak na rys. 1.12. W szczególności trend

może zacząć być widoczny dopiero przy analizie danych dla bardzo długiego odcinka czasowego.

## 1.3. Wskaźniki sumaryczne

Poprzedni podrozdział pozwolił nam docenić pierwsze zalety histogramu, który w dogodny sposób opisuje rozkład cechy ilościowej w próbie. Histogram jest sugestywnym środkiem syntezy informacji zawartej w próbie, ponieważ jest opisem graficznym, a taki opis przemawia najłatwiej do wyobraźni. Naturalne jest także pokuszenie się o skonstruowanie niewielu liczbowych miar, opisujących przynajmniej podstawowe własności rozkładu cechy. Miary takie, zwane **wskaźnikami**, nie mogą zastąpić samego histogramu, ale mogą stanowić rozsądny, kolejny etap agregacji informacji o rozkładzie. Ich omówieniu poświęcony jest ten podrozdział.

Pierwsze dwa pytania, na jakie chciałoby się mieć odpowiedź liczbową, to pytanie gdzie leży „centrum" lub „środek" próby oraz jak duże jest rozproszenie cechy w próbie wokół owego „centrum". Wskaźniki określające „centrum" lub „środek" próby nazywamy **wskaźnikami położenia**. Wskaźniki określające rozproszenie cechy wokół wskaźnika położenia nazywamy **wskaźnikami rozproszenia**. W przypadku wskaźników obydwu typów użyliśmy liczby mnogiej, ponieważ ze względów, które przedstawimy w dalszym ciągu tego podrozdziału, warto mieć więcej niż jedną miarę położenia i więcej niż jedną miarę rozproszenia. Krótko mówiąc, wybór najbardziej właściwych wskaźników często zależy od typu rozkładu, z jakim mamy do czynienia.

### 1.3.1. Wskaźniki położenia

Niech $x_1, x_2, \ldots, x_n$ oznacza próbę o liczności $n$.

**Definicja 1.1.** ***Wartością średnią w próbie*** *(lub prościej, wartością średnią próby), oznaczaną $\bar{x}$, nazywamy średnią arytmetyczną wartości cechy w próbie*

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{1.2}$$

W przykładzie 1.3, w którym $n = 20$, wartość średnia próby wynosi

$$\bar{x} = \frac{1}{20}(36 + 41 + 33 + \ldots + 25 + 32) = 34{,}05.$$

W tym przypadku otrzymaną wartość średnią można uznać za bliską modzie histogramu, czyli przedziałowi [35,40). Zauważmy, że moda zależy od

przyjętego podziału próby na klasy. Na przykład podział, który prowadziłby do wystąpienia w nim przedziału [34,39), dałby taką właśnie modę. Zwróćmy też uwagę, że obliczona średnia ma dokładność do setnych części roku, gdy tymczasem dane były podawane w pełnych latach. Zwykle otrzymaną wartość średnią zaokrągla się do wartości o jedno miejsce dziesiętne dokładniejszej niż wynosi precyzja zapisu danych. W naszym przykładzie zdecydowaliśmy się zachować większą dokładność zapisu, taki bowiem zapis daje więcej infomacji niż zaokrąglenie do wartości 34 lub 34,1. Trzeba jednak pamiętać, że dane były mierzone w jednostkach całkowitych.

Może się zdarzyć, że nie dysponujemy oryginalnym zbiorem danych, dysponujemy zaś wyłącznie powstałym na ich podstawie histogramem. Istnieje wówczas możliwość przybliżonego obliczenia średniej. Wystarczy w tym celu liczność każdej klasy pomnożyć przez środkową wartość przedziału określającego tę klasę, następnie obliczyć sumę tych iloczynów i otrzymany wynik podzielić przez liczność próby. Przybliżenie polega zatem na zastąpieniu dokładnych wartości elementów próby środkową wartością przedziału, do którego należy dany element. Na przykład, jeśli tak obliczymy przybliżoną wartość średnią w próbie z przykł. 1.3, to otrzymamy (środki kolejnych klas wypadają w punktach: 22,5, 27,5, ... , 47,5):

$$\frac{1}{20}(1 \times 22{,}5 + 2 \times 27{,}5 + 7 \times 32{,}5 + 8 \times 37{,}5 + 1 \times 42{,}5 + 1 \times 47{,}5) = 34{,}75.$$

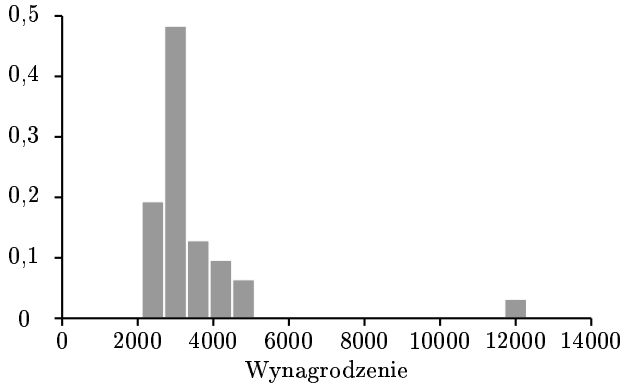
Wartość średnia nie budzi kontrowersji jako wskaźnik położenia, gdy rozkład cechy w próbie jest przynajmniej w przybliżeniu symetryczny, czyli gdy wartości cechy rozkładają się w przybliżeniu symetrycznie wokół średniej. Tak właśnie jest w przykł. 1.3, gdzie wartość średnia leży rzeczywiście w „środku" próby, czyli w punkcie bliskim modzie histogramu. Inaczej jest jednak, gdy rozkład cechy w próbie jest prawostronnie skośny.

**Przykład 1.7.** Rozważmy rozkład miesięcznych zasadniczych wynagrodzeń pracowników z wyższym wykształceniem, zatrudnionych w pewnej firmie. Sześciu spośród pracowników ma wynagrodzenie 2500 zł, ośmiu ma 3000 zł, siedmiu 3100 zł, czterech 3500 zł, trzech 4000 zł, dwóch 5000 zł i jeden zarabia 12 000 zł. Średnie wynagrodzenie pracownika z wyższym wykształceniem wynosi (gdyż $n = 31$)

$$\bar{x} = \frac{1}{31}(6 \times 2500 + 8 \times 3000 + 7 \times 3100 + 4 \times 3500 + 3 \times 4000 + \\ + 2 \times 5000 + 12000) = 3506 \text{ zł}.$$

Z histogramu na rys. 1.13 wynika, że rozkład jest prawostronnie skośny i ma jedną wartość drastycznie przewyższającą inne zarobki. W rezultacie wartość średnia rozkładu jest wyraźnie przesunięta na prawo od mody

histogramu (gdyby rozkład był lewostronnie skośny, co w przypadku rzeczywistych organizacji gospodarczych jest raczej trudno wyobrażalne, średnia byłaby przesunięta na lewo od mody histogramu). Nie ma w tym nic złego, ale warto zastanowić się nad konsekwencjami opisanego faktu. Wyobraźmy sobie, że młody absolwent wyższej uczelni zgłasza się na rozmowę kwalifikującą do pracy w tej właśnie firmie. Kandydat dowiaduje się, że wprawdzie będzie zarabiał miesięcznie 2500 zł, ale że średnie miesięczne wynagrodzenie w firmie wynosi 3500 zł. Zatem, powiada wiceprezes firmy, ma pan przed sobą wspaniałe możliwości awansu i znacznie wyższego wynagrodzenia. Czego jednak kandydat nie słyszy, to tego, że około 2/3 pracowników firmy, mających wyższe wykształcenie, nie zarabia miesięcznie więcej niż 3100 zł. Wysoka średnia jest konsekwencją bardzo wysokich zarobków kierownictwa, do którego nasz kandydat nie trafi jeszcze przez długie lata. Przeciwnie, przez lata jego wynagrodzenie będzie najprawdopodobniej typowe dla firmy, czyli będzie bliskie środkowemu poziomowi zarobków w tym sensie, że zarobki połowy kadry techniczno-administracyjnej leżą poniżej owego poziomu środkowego, zarobki zaś drugiej połowy są od tego poziomu wyższe. Kandydat dobrze uczyniłby, pytając o wskaźnik położenia zwany **medianą**, która stanowi środkową wartość próby uporządkowanej niemalejąco, od wartości najmniejszej w próbie do wartości największej. Na osi liczbowej, na lewo i na prawo od mediany jest położona taka sama liczba danych z próby.




Rys. 1.13. Histogram częstości dla danych z przykł. 1.7

Aby ściśle zdefiniować medianę, oznaczmy niemalejąco uporządkowane elementy próby w następujący sposob:

$$x_{(1)}, x_{(2)}, \ldots, x_{(n-1)}, x_{(n)},$$

gdzie $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n-1)} \leq x_{(n)}$. W naszym przykładzie

$$x_{(1)} = x_{(2)} = \ldots = x_{(6)} = 2500, \qquad x_{(7)} = x_{(8)} = \ldots = x_{(14)} = 3000,$$
$$x_{(15)} = x_{(16)} = \ldots = x_{(21)} = 3100, \qquad x_{(22)} = x_{(23)} = x_{(24)} = x_{(25)} = 3500,$$
$$x_{(26)} = x_{(27)} = x_{(28)} = 4000, \qquad x_{(29)} = x_{(30)} = 5000, \quad x_{(31)} = 12000.$$

**Definicja 1.2.** ***Medianą w próbie*** *(lub medianą próby), oznaczaną $x_{med}$, nazywamy następującą wielkość*

$$x_{med} = \begin{cases} x_{((n+1)/2)}, & \text{gdy } n \text{ nieparzyste} \\ \frac{1}{2}(x_{(n/2)} + x_{(n/2+1)}), & \text{gdy } n \text{ parzyste.} \end{cases} \tag{1.3}$$

Ścisła definicja mediany wymaga rozróżnienia dwóch przypadków: gdy liczność próby jest nieparzysta oraz gdy jest parzysta. Rzecz w tym, że gdy $n$ jest nieparzyste, to $x_{((n+1)/2)}$ jest środkowym co do wartości elementem próby (na lewo i na prawo od niej leży tyle samo elementów próby). Gdy zaś $n$ jest parzyste, takiego elementu leżącego w środku próby nie ma. Można natomiast umówić się, że mediana leży w środku odcinka łączącego elementy $x_{(n/2)}$ oraz $x_{(n/2+1)}$. Na lewo i na prawo od tak określonej mediany leży dokładnie połowa elementów próby.

W naszym przykładzie mediana wynosi $x_{((31+1)/2)} = x_{(16)} = 3100$ zł i znacznie lepiej oddaje zarobkowe perspektywy nowo zatrudnianego kandydata niż wartość średnia $\bar{x} = 3506$ zł. W przykładzie 1.4, gdzie histogram był prawostronnie skośny, $x_{med} = 27{,}50$, podczas gdy $\bar{x} = 27{,}91$.

Pożądaną cechą mediany jest jej brak wrażliwości na **wartości odstające**, czyli wartości bardzo wyraźnie oddalone od innych wartości w próbie i w tym sensie zdecydowanie nietypowe dla zaobserwowanego rozkładu pozostałych wartości cechy w próbie. Przez brak wrażliwości, zwany dalej **odpornością na obserwacje odstające**, rozumiemy to, że obserwacje takie wcale lub tylko nieznacznie wpływają na wartość danego wskaźnika (w tym przypadku mediany).

**Przykład 1.8.** Jak w przykładzie 1.3 rejestrujemy wiek 20 pracowników zgłaszających się na okresowe badania lekarskie. Zaobserwowane wielkości wynoszą (w latach):

36, 41, 33, 34, 38, 26, 33, 36, 30, **84**, 39, 31, 35, 36, 38, 37, 22, 31, 25, 32.

Zaobserwowana próba jest w rzeczywistości powtórzeniem danych z przykł. 1.3 z jednym tylko wyjątkiem: zamiast wieku 48 lat, wpisaliśmy „przez pomyłkę" wiek 84 lata (takie błędy rzeczywiście się zdarzają i są nierzadko przyczyną wystąpienia obserwacji odstających). Pojawienie się takiej wartości odstającej nie ma w naszym przypadku żadnego wpływu na wartość mediany. Skutkiem błędu przy zapisywaniu danych największą wartość w prawdziwym zbiorze obserwacji (liczbę 48) zastąpiliśmy wartością jeszcze większą (84), co nie mogło zmienić faktu, że liczba 34 jest nadal dziesiątym, a liczba 35 jedenastym, elementem próby uporządkowanej od elementu

najmniejszego do największego. Oczywiście, w nowej próbie inna musi być wartość średnia (nowa wartość średnia wynosi 35,85 lat, podczas gdy „stara" wartość średnia wynosi 34,05 lat).

W tym miejscu niezbędne jest ostrzeżenie Czytelnika przed nieprzemyślanym usunięciem z próby wartości odstających. Wartości te możemy usunąć wtedy tylko, gdy potrafimy wykazać, iż znalazły się w próbie skutkiem błędu, faktycznie zaś do próby nie należą. Niekiedy pojawienie się wartości odstających nie jest wynikiem błędu, lecz niejednorodności badanego zjawiska. W pewnym sensie kontrola jakości produkcji jest metodą wykrywania wystąpienia takiej niejednorodności. Bada się na przykład średnice toczonych wałków i w jakimś momencie obserwuje pojawienie się średnic nietypowych, ponieważ wyraźnie odstających co do swych wartości, od średnic otrzymywanych, gdy proces produkcji przebiega bez zakłóceń. Takie wartości odstające niosą bardzo istotną informację o procesie produkcji, a ściślej o wystąpieniu jakiejś zmiany w tym procesie (np. uszkodzeniu noża tokarskiego).

**Przykład 1.9.** W roku 1985 wiele rozgłosu zyskał sobie spór o to, czy nad biegunem południowym pojawiła się dziura ozonowa. Jej wystąpienie stwierdzili pracownicy naziemnego obserwatorium w Wielkiej Brytanii. Wynikom tym nie dano początkowo wiary. Wszak pomiary satelitarne, prowadzone regularnie i automatycznie od roku 1979 żadnej dziury nie wykazały. A wiadomo, że automaty się nie mylą, natomiast ludziom pomyłki zdarzają się nierzadko. Uznano, że rzekomo zaobserwowana przez Brytyjczyków dziura ozonowa jest wynikiem obserwacji odstających i fałszywych. Nie zrażeni Brytyjczycy kontynuowali swoje obserwacje i nadal otrzymywali zerową wartość grubości warstwy ozonu nad biegunem południowym. W ten sposób zmusili zespół, prowadzący pomiary satelitarne do przeanalizowania komputerowego programu zbierania i analizy danych. Okazało się, że wartości bliskie zera były automatycznie traktowane przez system jako odstające i fałszywe. Szczęśliwie software sterujący systemem, choć o otrzymaniu takich danych nie informował, pozostawiał je w bazie danych. Dopiero zatem przestudiowanie archiwizowanych danych umożliwiło potwierdzenie słuszności brytyjskiej tezy – od pewnego czasu nad biegunem południowym tworzyła się dziura ozonowa!

Jeżeli liczność próby jest mała i kolejne, uporządkowane elementy próby są od siebie dość odległe, mediana jest niestabilnym wskaźnikiem położenia. Jeżeli na przykład do próby z przykł. 1.3 dołączymy jedną obserwację o wartości 32 i jedną obserwację o wartości 33, to mediana nowej próby wyniesie 33,5 (w próbie oryginalnej $x_{med} = 34{,}5$). Tymczasem średnia w próbie bez

wartości dodatkowych wynosi, jak pamiętamy, 34,05, w próbie zaś z tymi wartościami jest równa 33,9. Zważywszy, że dane były mierzone w jednostkach całkowitych, mediana zmieniła się istotnie, bo o cały rok, natomiast średnia prawie nie uległa żadnej zmianie. Przyczyna podanej niestabilności mediany jest dość oczywista i zwykle znika wraz ze wzrostem liczności próby. Rzecz w tym, że medianę definiujemy jako środkowy element próby, a w próbie mało licznej kolejne co do wielkości elementy próby mogą się od siebie znacznie różnić. Jednakże dołączenie dodatkowej obserwacji do próby zmienia wartość średnią proporcjonalnie do odległości obserwacji dodatkowej od średniej otrzymanej bez tej obserwacji.

W uniknięciu niestabilności podanego typu może pomóc wprowadzenie innego jeszcze wskaźnika położenia, mianowicie **średniej ucinanej**, która powstaje przez obliczenie średniej próby powstałej przez usunięcie z próby oryginalnej $k$ wartości najmniejszych i $k$ wartości największych (parametr $2k$ jest zawsze znacznie mniejszy od $n$).

**Definicja 1.3.** ***Średnią ucinaną*** *(z parametrem $k$), oznaczaną $\bar{x}_{tk}$, nazywamy wielkość*[1]

$$\bar{x}_{tk} = \frac{1}{n-2k} \sum_{i=k+1}^{n-k} x_{(i)}. \tag{1.4}$$

Zauważmy, że jeżeli rozkład cechy w próbie jest w przybliżeniu symetryczny oraz gdy nie występują w niej obserwacje odstające, średnia $\bar{x}$ i średnia ucinana powinny mieć bliskie wartości. W przykładzie 1.3 średnia ucinana z $k = 2$, $\bar{x}_{t2}$, ma wartość 34,1. Gdyby do tej próby dołączyć jeszcze obserwację o wartości 32 oraz obserwację o wartości 33, to otrzymalibyśmy $\bar{x}_{t2} = 33{,}9$.

Ucinanie wartości skrajnych ma oczywiście na celu pozbycie się wpływu ewentualnych wartości odstających na wartość wskaźnika położenia. Zauważmy, że w przypadku średniej ucinanej musimy zdecydować jaką wartość $k$ zastosować. Wartość ta powinna być nie mniejsza niż liczba wartości odstających na każdym z dwóch krańców rozkładu próby. W przypadku mediany, która co prawda może być w opisanym sensie niestabilna, wpływ wartości odstających jest usuwany automatycznie.

Punkt ten zakończymy podaniem jeszcze jednego wskaźnika położenia, mianowicie średniej winsorowskiej, którą warto się posłużyć na przykład wtedy, gdy skrajne, na przykład największe, obserwowane wyniki charakteryzują się dużą niepewnością co do ich rzeczywistych wartości. W takiej sytuacji rozsądne może być posłużenie się wskaźnikiem, który wprawdzie nie zależy

[1] Literka $t$ w symbolu średniej ucinanej pochodzi od angielskiego słowa *trimmed*, czyli *ucinana*.

od rzeczywistych wartości obserwacji skrajnych, ale na którego wartość ma wpływ informacja o obecności w próbie oryginalnej owych wartości skrajnych. Własność taką ma tzw. **średnia winsorowska**[2]. Jak w przypadku średniej ucinanej, obliczenie średniej winsorowskiej opiera się na $n-2k$ „środkowych" elementach próby, otrzymanych w wyniku pominięcia $k$ najmniejszych i $k$ największych jej elementów. Chcąc jednak uwzględnić fakt pojawienia się w próbie oryginalnej $k$ wartości nie większych niż $x_{(k+1)}$ oraz $k$ wartości nie mniejszych niż $x_{(n-k)}$, przy obliczaniu średniej postępuje się tak, jakby $x_{(k+1)}$ i $x_{(n-k)}$ wystąpiły dodatkowo $k$ razy (te dodatkowe wystąpienia wymienionych dwóch statystyk niejako zastępują wartości $x_{(1)}, \ldots, x_{(k)}$ oraz $x_{(n-k+1)}, \ldots, x_{(n)}$).

**Definicja 1.4.** ***Średnią winsorowską*** *(o parametrze k), oznaczaną $\bar{x}_{wk}$, nazywamy wielkość*

$$\bar{x}_{wk} = \frac{1}{n}[(k+1)x_{(k+1)} + \sum_{i=k+2}^{n-k-1} x_{(i)} + (k+1)x_{(n-k)}]. \tag{1.5}$$

Innym przykładem zastosowania średniej winsorowskiej jest przypadek utraty z bazy danych obserwacji skrajnych. Szczególnie ciekawym przykładem takiej sytuacji jest występowanie w próbie obserwacji „uciętych". W przypadku „ucinania prawostronnego" o każdej takiej obserwacji wiadomo tylko, że jej nieznana wartość jest nie mniejsza niż pewna znana liczba. Sytuacja taka może się zdarzyć w przypadku zastosowania niewłaściwego miernika, gdy wartości cechy mierzonej pochodzą z szerszego przedziału niż założono przed przystąpieniem do pomiarów. Często badania dotyczące czasu działania (tzw. czasu życia) urządzenia prowadzi się w ustalonym z góry przedziale czasu, notując jedynie fakt, że część urządzeń w badanej próbie dotrwała do końca owego przedziału.

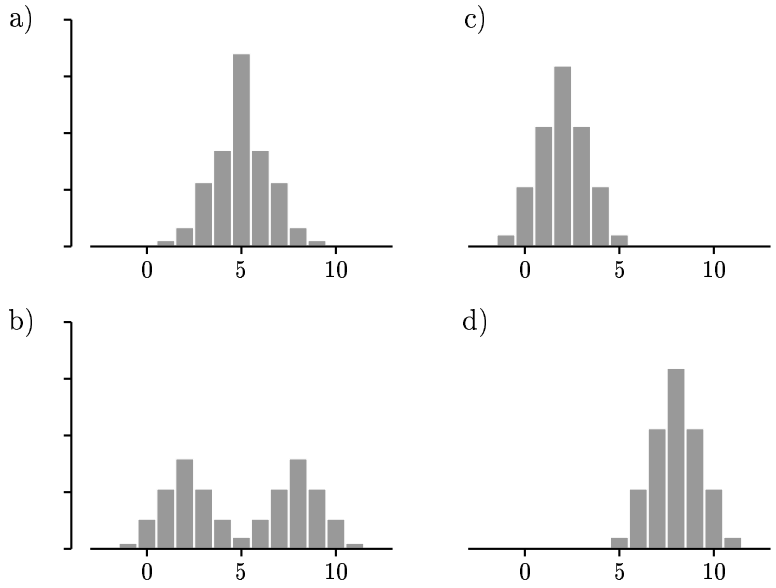
**Przykład 1.10.** Wyobraźmy sobie, że dane z przykł. 1.3 zostały zapisane w bazie danych, poczynając od pomiaru najmniejszego do największego. Omyłkowo, z bazy usunięto dwa ostatnie wiersze, czyli dwa największe wyniki. Dla $k = 2$, $\bar{x}_{w2} = 33{,}75$, a zatem prawie 34 lata. Gdyby natomiast pogodzić się z faktem usunięcia dwóch obserwacji i obliczyć zwykłą średnią pozostałych 18 obserwacji, otrzymalibyśmy $\bar{x} = 32{,}9$, czyli – zważywszy, że obserwacje były prowadzone w jednostkach całkowitych – wartość istotnie mniejszą niż 33,75.

Przy analizie danych warto badać różne wskaźniki położenia, by w ten spo-

[2]Wskaźnik ten został zaproponowany przez C. P. Winsora.

sób móc uchwycić różne możliwe aspekty badanej cechy. Powszechna dziś dostępność komputerów oraz pakietów statystycznych umożliwia to bez żadnego kłopotu. Z tych samych powodów także w następnym punkcie podamy nie jeden tylko, lecz kilka wskaźników rozproszenia.

Podkreślmy jeszcze, że interpretując wartości otrzymanych wskaźników nie wolno zapomnieć o możliwości, czy raczej potrzebie wykorzystania takich graficznych środków analizy jak histogram. Nietrudno na przykład wyobrazić sobie dwa symetryczne rozkłady cechy w próbie o takiej samej wartości średniej, równej ze względu na symetrię wspólnej medianie tych rozkładów, z których jeden rozkład jest jednomodalny, drugi zaś ma dwie mody (patrz rys. 1.14a i 1.14b - drugi z rozkładów jest idealnie symetryczny, dwumodalny). W drugim przypadku możemy mieć do czynienia z „mieszaniną" dwóch symetrycznych rozkładów jednomodalnych przedstawionych na rys. 1.14c i 1.14d.




Rys. 1.14. Histogramy częstości. a) rozkład symetryczny jednomodalny (średnia = 5), b) rozkład symetryczny dwumodalny (średnia = 5), c) rozkład symetryczny (średnia = 2), d) rozkład symetryczny (średnia = 8)

W obydwu przypadkach sens wartości średniej (a zarazem mediany) jest zupełnie inny. W przypadku rozkładu jednomodalnego parametr położenia wskazuje obszar, w którym skupiają się obserwacje. Innymi słowy, parametr ten wskazuje, gdzie leży centralna część próby. Z kolei, w przedstawionym przykładzie rozkładu dwumodalnego parametr położenia niejako rozdziela dwa skupienia obserwacji. Mówienie o centralnej części próby, rozumianej

analogicznie jak w przypadku jednomodalnym, nie ma tym razem sensu. Interpretacja rozkładów dwumodalnych (i w ogóle wielomodalnych) wymaga jeszcze dodatkowej ostrożności. Mianowicie, bywa że przyczyną dwumodalności jest niewłaściwe połączenie w jeden zbiór danych dwóch różnych rodzajów. Na przykład, w przypadku mierzenia średnic i wysokości partii walców nie ma sensu potraktowanie otrzymanych wyników pomiarów jako jednego zbioru obserwacji liczbowych – jakkolwiek średnicę i wysokość każdego walca możemy mierzyć w tych samych jednostkach, mamy tu do czynienia z dwoma jakościowo odmiennymi i w efekcie nieporównywalnymi pomiarami. Dwumodalność takiej „łącznej" próby oznacza, że w rzeczywistości mamy do czynienia z dwiema próbami niebacznie połączonymi w jedną.

### 1.3.2. Wskaźniki rozproszenia

W końcu lat osiemdziesiątych sensację wśród specjalistów przemysłu samochodowego wzbudziła dokładność toczenia wałów głównych skrzyni biegów przez pewnego producenta japońskiego. Znany koncern stosował w swoich samochodach te i inne, „takie same" wały niejapońskie. Po jakimś czasie okazało się, że skrzynie biegów pierwszego typu sprawują się lepiej niż skrzynie drugiego typu. Przeprowadzono zatem stosowne analizy i stwierdzono, że rozkład średnicy wybranego przekroju wałów w próbie wałów japońskich ma ten sam kształt co rozkład odpowiedniej średnicy w próbie wałów niejapońskich. Wartości średnie w obydwu próbach były identyczne. Średnice wszystkich wałów pozostawały w granicach tolerancji. Ale odchyłki średnic wałów japońskich od wartości średniej były przynajmniej o rząd wielkości mniejsze od odchyłek wałów niejapońskich, tak jak to symbolicznie pokazano na rys. 1.15. Próbka wałów japońskich charakteryzowała się mniejszym rozproszeniem niż próbka innych wałów.

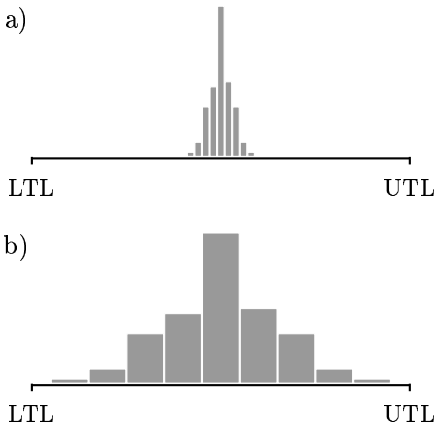
Z powyższego przykładu wynika, że zredukowanie informacji o rozkładzie cechy w próbie do wartości średniej jest zdecydowanie zbyt drastyczne. Potrzebujemy jeszcze przynajmniej informacji o stopniu **rozproszenia** próby wokół wartości średniej.

Najprostszym wskaźnikiem rozproszenia cechy w próbie jest jej **rozstęp**, czyli różnica między największą i najmniejszą wartością cechy w próbie.

**Definicja 1.5.** ***Rozstępem próby*** *o liczności $n$, oznaczanym $R$, nazywamy wielkość*

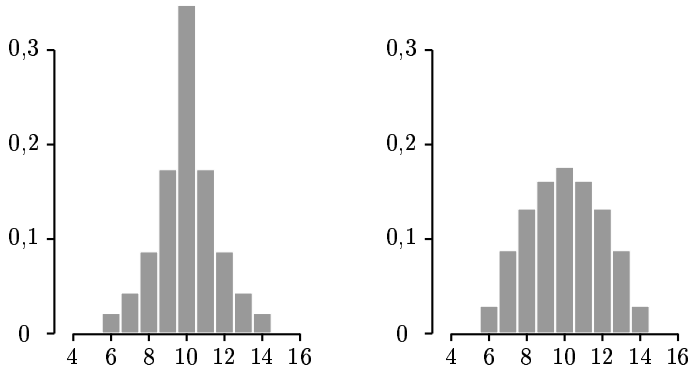
$$R = x_{(n)} - x_{(1)}, \tag{1.6}$$

*gdzie $x_{(1)}$ i $x_{(n)}$ są, odpowiednio, najmniejszym i największym elementem w próbie.*




Rys. 1.15. Rozkład średnicy wałów. a) Wały japońskie, b) wały niejapońskie (LTL = dolna granica tolerancji, UTL = górna granica tolerancji)

W przykładzie 1.3 $R = 26$. Pomijając oczywisty brak odporności rozstępu $R$ na obserwacje odstające, wskaźnik ten nie odróżnia na przykład takich dwóch rozkładów jak na rys. 1.16: **jeden jest wyostrzony, drugi spłaszczony, obydwa są symetryczne, o tym samym nośniku**. W przypadku pierwszego rozkładu większy procent obserwacji jest skupiony w bezpośrednim otoczeniu środka rozkładu niż to występuje w drugim rozkładzie, gdzie stosunkowo dużo obserwacji leży blisko jego krańców. Ta niewrażliwość rozstępu na kształt rozkładu wynika stąd, iż wartość rozstępu zależy od zaledwie dwóch skrajnych wartości cechy w próbie.




Rys. 1.16. Rozkłady symetryczne o tym samym nośniku: wyostrzony i spłaszczony

Kolejnym wskaźnikiem rozproszenia jest **wariancja**.

**Definicja 1.6.** ***Wariancją w próbie****, oznaczaną $s^2$, nazywamy wielkość*

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2, \tag{1.7}$$

*gdzie, jak zwykle, $\bar{x}$ oznacza średnią w próbie. Pierwiastek z wariancji jest nazywany* ***odchyleniem standardowym cechy w próbie*** *i jest oznaczany przez $s$*

$$s = \sqrt{s^2}.$$

W przykładzie 1.3 $s^2 = 34{,}37$, skąd $s = 5{,}9$ (po zaokrągleniu do jednego miejsca po przecinku).

Jeśliby w definicji wariancji ułamek $1/(n-1)$ zastąpić ułamkiem $1/n$, wariancja mierzyłaby średnią wartość kwadratów odchyłek wartości cechy w próbie od wartości średniej $\bar{x}$. Przyczyna zastąpienia mianownika $n$ mianownikiem $n-1$ wyjaśni się w podrozdz. 2.4. Dla dostatecznie dużych $n$ jest to oczywiście różnica nieistotna. Ponieważ wariancja bazuje na kwadratach odchyleń, więc jest mierzona w jednostkach, które stanowią kwadrat jednostki pomiarowej cechy w próbie. Naturalne jest wprowadzenie jako wskaźnika rozproszenia próby również pierwiastka z wariancji (jeżeli średnice wałów skrzyni biegów mierzone byłyby w mikronach, to mianem wariancji jest kwadrat mikrona). Zgodnie z podaną definicją, taki wskaźnik nazywamy odchyleniem standardowym.

Zauważmy, że wprowadzenie kwadratów odchyleń powiększa wpływ obserwacji znacznie odbiegających od wartości średniej $\bar{x}$ na wartość sumy $\sum_{i=1}^{n}(x_i-\bar{x})^2$ we wzorze na wariancję. Jednocześnie, z tego samego powodu, jest pomniejszany wpływ na tę samą sumę obserwacji bliskich wartości średniej $\bar{x}$. Jeżeli na przykład odległość wartości cechy w próbie od średniej wynosi 2, to jej udział w sumie definiującej wariancję wynosi 4; jeśli natomiast zaobserwowaliśmy wartość, której odległość od średniej wynosi 0,1, to w sumie $\sum_{i=1}^{n}(x_i - \bar{x})^2$ pojawia się składnik 0,01.

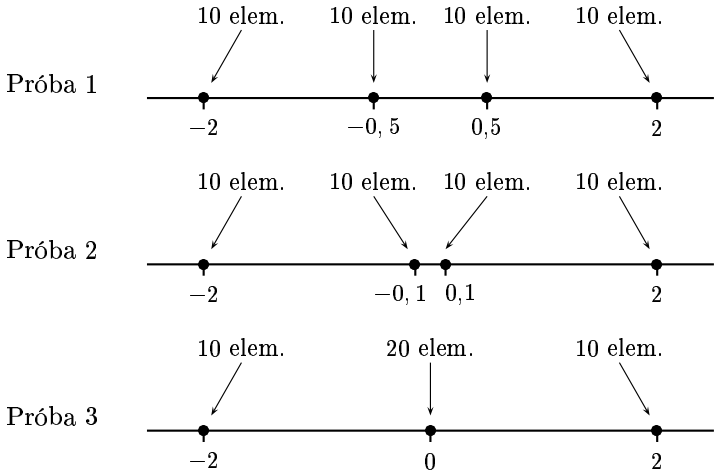
Niekiedy pożądane jest by wszystkie obserwacje miały udział w ogólnej wartości wskaźnika rozproszenia proporcjonalny do wielkości odchyłek tych obserwacji od wartości średniej. Wówczas stosuje się wskaźnik rozproszenia zwany **odchyleniem przeciętnym** od wartości średniej.

**Definicja 1.7.** ***Odchyleniem przeciętnym*** *od wartości średniej, oznaczanym $d_1$, nazywamy wielkość*

$$d_1 = \frac{1}{n}\sum_{i=1}^{n} |x_i - \bar{x}|. \tag{1.8}$$

Aby lepiej uzmysłowić sobie różnice między podanymi wskaźnikami rozproszenia, rozważmy sztuczny wprawdzie, ale pouczający przykład.

**Przykład 1.11.** Wyobraźmy sobie trzy próby, symbolicznie przedstawione na rys. 1.17, każda o liczności 40 i zerowej wartości średniej. W pierwszej próbie mamy 10 elementów o wartości 0,5, 10 elementów o wartości $-0{,}5$, 10 elementów o wartości 2 i 10 elementów o wartości $-2$. W drugiej próbie zamiast elementów o wartości 0,5 mamy elementy o wartości 0,1 i zamiast elementów o wartości $-0{,}5$ mamy elementy $-0{,}1$. W trzeciej próbie mamy 20 elementów o wartości 0, poza tym, jak w poprzednich próbach, 10 elementów o wartości 2 i 10 elementów o wartości $-2$.

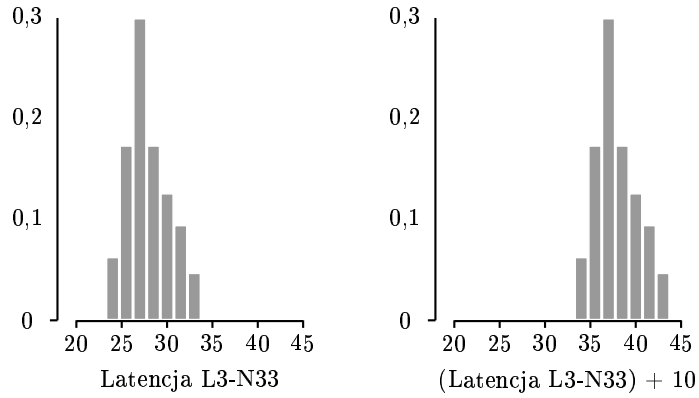



Rys. 1.17. Trzy próby z przykładu 1.11

W pierwszej próbie $s^2 = 2{,}18$, $s = 1{,}48$, $d_1 = 1{,}25$. W drugiej próbie $s^2 = 2{,}06$, $s = 1{,}43$, $d_1 = 1{,}05$. W trzeciej próbie $s^2 = 2{,}05$, $s = 1{,}43$, $d_1 = 1$. Jak widać, udział odchyłek od średniej w otrzymanej wartości odchylenia przeciętnego $d_1$ jest proporcjonalny do wartości tych odchyłek (w przypadku pierwszej próby, 20 odchyłek o wartości 2 daje po podzieleniu przez 40 wartość 1, do której dodaje się 20 odchyłek o wartości 0,5, podzielonych znowu przez 40; analogicznie jest w przypadku dwóch pozostałych

prób). Inaczej powstaje wartość wariancji i odchylenia standardowego – maleniu odchyłek towarzyszy nieproporcjonalnie szybki spadek ich wpływu na ostateczną wartość wskaźnika rozproszenia.

W tym miejscu warto poczynić jeszcze dwie uwagi ogólnej natury. Należy mianowicie podkreślić, że żaden wskaźnik rozproszenia nie powinien zmieniać swej wartości, gdy do wszystkich elementów próby zostanie dodana ta sama liczba (dodatnia lub ujemna) – mówimy wówczas o przesunięciu elementów próby o zadaną wartość. Postulat ten wynika stąd, że takie przesunięcie w ogóle nie zmienia kształtu rozkładu i w szczególności nie zmienia stopnia rozproszenia próby (por. rys. 1.18, gdzie przedstawiono histogram próby z przykł. 1.4 oraz histogram tej samej próby po przesunięciu jej elementów o 10). Wszystkie znane wskaźniki rozproszenia mają tę własność. Na przykład, podane już wskaźniki opierają się na obliczeniu kolejnych różnic między elementem próby i średnią. Ale przesunięcie elementów próby o zadaną wartość przesuwa o tę samą wartość także średnią i w rezultacie wspomniane różnice nie zmieniają swych wartości.




Rys. 1.18. Histogram próby (latencja L3-N33) z przykł. 1.4 oraz histogram tej samej próby po przesunięciu jej elementów o 10 milisekund

Wszystkie wskaźniki rozproszenia mierzone w tych samych jednostkach co elementy próby, np. odchylenie standardowe i odchylenie przeciętne, powinny spełniać, i spełniają, także następujący postulat. Pomnożenie każdego elementu próby przez tę samą liczbę (dodatnią lub ujemną) powinno prowadzić do pomnożenia wskaźnika przez wartość bezwzględną tej liczby. Rzecz w tym, że taki mnożnik mówi ile razy zmienia się stopień rozproszenia próby. Jeżeli najmniejszy element próby jest równy $-3$, natomiast największy 4, czyli, jeżeli rozstęp wynosi 7, to pomnożenie każdego elementu próby przez 6 zmienia wartość rozstępu na $24-(-18)=42$, a zatem sześciokrotnie, jak należy sobie tego życzyć. Pomnożenie przez 6 każdego elementu próby

prowadzi do pomnożenia przez ten sam mnożnik średniej. W efekcie przez 6 zostaje pomnożony każdy składnik sumy określającej wskaźnik $d_1$, i przeto, sam wskaźnik $d_1$.

Sytuacja jest tylko nieco bardziej złożona w przypadku odchylenia standardowego $s$, które jest określone za pomocą wariancji $s^2$. Wariancja opiera się na kwadratach odchyłek, w rezultacie czego pomnożenie elementów próby przez liczbę, powiedzmy $c$, prowadzi do pomnożenia wariancji przez $c^2$. Ale ponieważ $s$ jest pierwiastkiem z wariancji $s^2$, znowu mamy pożądaną $c$-krotną zmianę $s$.

Podane dotąd wskaźniki rozproszenia, zwłaszcza wariancja, nie są odporne na wartości odstające w próbie. Przyczyna tego jest podobna do przyczyny braku odporności wartości średniej. Dlatego nie będziemy się nad tym faktem dłużej zatrzymywać. Wspomnimy jedynie, że wskaźnik rozproszenia, który obecnie wprowadzimy, jest z oczywistych względów odporny na wartości odstające. Zasadnicza przyczyna jego wprowadzenia jest jednak nieco inna.

Analizując wskaźniki położenia, zauważyliśmy, że mediana może być uważana za lepszy wskaźnik niż średnia w próbie, gdy rozkład cechy w próbie jest asymetryczny. W przypadku takiego rozkładu na wartość podanych wskaźników rozproszenia (zwłaszcza wariancji) zbyt duży wpływ mogą mieć wartości skrajne, pochodzące z długiego ogona rozkładu. Wartości takich nie jest zwykle zbyt wiele w próbie, ale są to wartości bardzo odległe od średniej i stąd mające istotny wpływ na sumę $\sum_{i=1}^{n}(x_i - \bar{x})^2$. Dlatego, gdy mamy do czynienia z rozkładami asymetrycznymi, rozproszenie cechy w próbie warto określać na podstawie elementów położonych w centralnej części tej próby, nie uwzględniając zachowania się cechy w ogonach jej rozkładu.

Wskaźnikiem opartym na pomiarze rozproszenia centralnej części próby jest **rozstęp międzykwartylowy**. Aby podać jego ścisłą definicję, wprowadźmy najpierw pojęcia **dolnego kwartyla** (inaczej pierwszego kwartyla) i **górnego kwartyla** (inaczej trzeciego kwartyla) próby.

**Definicja 1.8.** ***Dolnym (pierwszym) kwartylem*** *próby nazywamy medianę podpróby, składającej się ze wszystkich elementów próby o wartościach mniejszych od mediany całej próby.* ***Górnym (trzecim) kwartylem*** *próby nazywamy medianę podpróby, składającej się ze wszystkich elementów próby o wartościach większych od mediany całej próby. Medianę całej próby nazywamy również* ***drugim kwartylem*** *całej próby. Dolny kwartyl oznaczamy symbolem $Q_1$, górny symbolem $Q_3$; medianę oznaczamy niekiedy symbolem $Q_2$.*

Kwartyle dzielą próbę uporządkowaną od wartości najmniejszej do wartości największej na cztery części: do pierwszej części należą elementy próby od najmniejszego do dolnego kwartyla, do drugiej części elementy od dolnego kwartyla do mediany, do trzeciej od mediany do górnego kwartyla i do ostatniej części należą elementy próby od górnego kwartyla do elementu największego. Można powiedzieć, że kwartyle dzielą próbę na cztery części o równej liczności (nie jest to stwierdzenie całkowicie ścisłe, nie rozstrzygamy tu bowiem, do których części należą kwartyle, ani nie przejmujemy się, czy liczność próby jest np. podzielna przez cztery).

**Definicja 1.9.** ***Rozstępem międzykwartylowym****, oznaczanym* $IQR$*, nazywamy wielkość*[3]

$$IQR = Q_3 - Q_1. \tag{1.9}$$

Rozstęp międzykwartylowy jest zatem rozstępem w sensie def. 1.5, ale odniesionym do centralnej połowy wartości cechy w próbie.

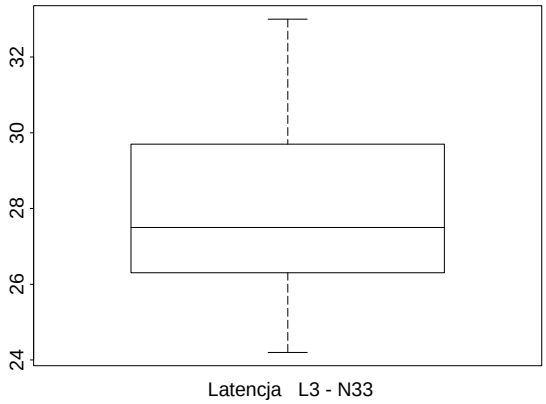
Kończąc ten punkt, zwróćmy uwagę na to, że w przypadku pomiarów dodatniej wielkości fizycznej ważne jest zbadanie stosunku wskaźnika położenia do wartości wskaźnika rozproszenia (mierzonego w tych samych jednostkach co wskaźnik położenia). Fakt, że wskaźnik rozproszenia wynosi 0,1 w sytuacji, gdy wskaźnik położenia wynosi 0,1 ma zupełnie inne znaczenie niż w sytuacji, gdy wskaźnik położenia wynosi 15.

### 1.3.3. Wykres ramkowy

Bardzo pożytecznym, graficznym środkiem wstępnej analizy danych jest **wykres ramkowy**, zwany też wykresem pudełkowym. Wykres ramkowy dla danych z przykł. 1.4 jest pokazany na rys. 1.19. Skala na osi pionowej odpowiada wartościom obserwacji (wartości te można by oczywiście odłożyć na osi poziomej, co wymagałoby obrócenia wykresu o 90°, ale poza tym jego interpretacja przebiegałaby zupełnie analogicznie). Na wykresie, współrzędna $y$ dolnej podstawy ramki jest równa pierwszemu kwartylowi $Q_1$. Współrzędna $y$ górnej podstawy ramki jest równa trzeciemu kwartylowi $Q_3$. Długość boku (wysokość) ramki jest zatem równa rozstępowi międzykwartylowemu $IQR$. Poziomy odcinek wewnątrz ramki, niekiedy zastępowany przez mały kwadracik, wyznacza medianę cechy w próbie. Odcinek wychodzący z górnej podstawy ramki kończy się poziomą linią, wyznaczającą największą obserwację (w próbie), spełniającą dodatkowy warunek, iż jest nie większa niż

$$Q_3 + 1{,}5 \times IQR.$$

[3] Skrót $IQR$ pochodzi od terminu angielskiego *interquartile range*.




Rys. 1.19. Wykres ramkowy dla danych z przykł. 1.4

Innymi słowy, odcinek ten, który za Anglosasami nazywamy żartobliwie **wąsem**, nie może być dłuższy niż półtora rozstępu międzykwartylowego. Obserwacje o wartościach większych niż $Q_3 + 1{,}5 \times IQR$, o ile występują w próbie, są nanoszone na wykres indywidualnie, tak jak to pokazano na rys. 1.22, do którego dokładniejszego omówienia wrócimy później.

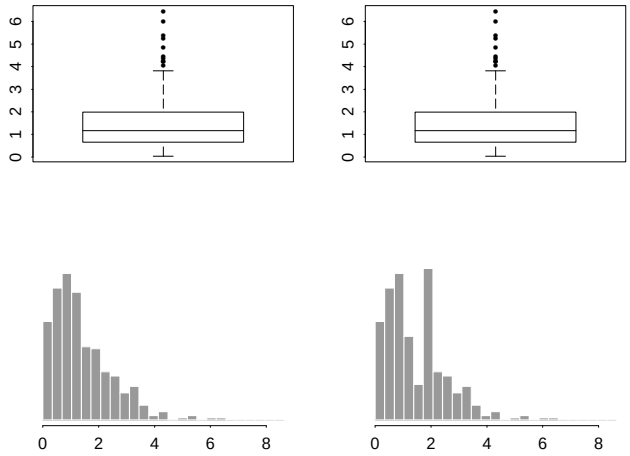
Podobnie do górnego wąsa tworzy się dolny wąs, sięgający od dolnej podstawy ramki do najmniejszej zaobserwowanej wartości, spełniającej dodatkowy warunek, iż jest nie mniejsza niż

$$Q_1 - 1{,}5 \times IQR.$$

Jak poprzednio, wąs nie może być dłuższy niż półtora rozstępu międzykwartylowego, obserwacje zaś o wartościach mniejszych niż $Q_1 - 1{,}5 \times IQR$ (o ile występują w próbie) są nanoszone na wykres indywidualnie.

Przyczyna, dla której za maksymalną możliwą długość wąsów uznano wartość $1,5 \times IQR$ zostanie wyjaśniona w p. 1.4.2. W tym miejscu nadmienimy jedynie, że występowanie obserwacji poza przedziałem wyznaczonym przez rozstęp międzykwartylowy z dołączonymi doń wąsami świadczy, iż badany rozkład cechy w próbie daleko odbiega od typowych rozkładów symetrycznych. Jak zobaczymy w p. 1.4.2, maksymalna długość wąsów mogłaby zresztą być inna niż akurat tu przyjęta krotność wartości $IQR$, ale ta właśnie krotność jest najczęściej stosowana w praktyce.

Z wykresu ramkowego dla danych z przykł. 1.4 wyraźnie widać dodatnią (inaczej prawostronną) skośność rozkładu. Górny (dla wykresu ułożonego poziomo – prawy) wąs jest wyraźnie dłuższy od wąsa dolnego (lewego dla wykresu poziomego) i, co więcej, mediana położona jest bliżej pierwszego kwartyla niż trzeciego kwartyla.



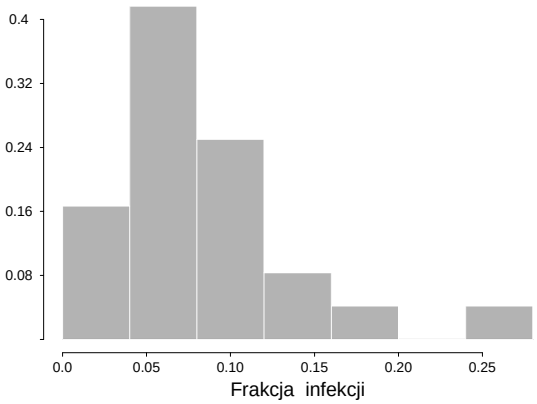

Rys. 1.20. Histogramy i wykresy ramkowe dla przykł. 1.12

Wskazanie za pomocą czytelnych i prostych środków graficznych dokładnych wartości mediany, pozostałych kwartyli oraz najmniejszej i największej obserwacji daje dobrą sumaryczną informację o badanym rozkładzie. Zarazem informacja ta nie zastępuje histogramu – wykres ramkowy nie mówi na przykład, czy rozkład cechy jest jednomodalny czy wielomodalny. Łatwo sobie wyobrazić rozkłady o różnych liczbach mód i takim samym wykresie ramkowym.

**Przykład 1.12.** Na rysunku 1.20 przedstawiono histogramy i wykresy ramkowe dla dwóch prób: jednej ze skośnego rozkładu dwumodalnego, drugiej ze skośnego rozkładu jednomodalnego. Wykresy ramkowe są takie same, natomiast histogramy są różne.

W wielu pakietach statystycznych obserwacje, które są nanoszone na wykres ramkowy indywidualnie, nazywa się automatycznie wartościami odstającymi, choć nie zawsze takie ich określenie jest uzasadnione.

**Przykład 1.13.** W pewnym szpitalu kierownictwo poleciło zbadanie frakcji infekcji pooperacyjnych po wymianie stawu biodrowego na endoprotezę. Frakcje te notowano od 24 miesięcy. Zebrane dane są podane w tab. 1.2 (J.R. Thompson i J. Koronacki (1994): *Statystyczne sterowanie procesem.* Akademicka Oficyna Wydawnicza PLJ, Warszawa). Rozkład frakcji pokazano na rys. 1.21.



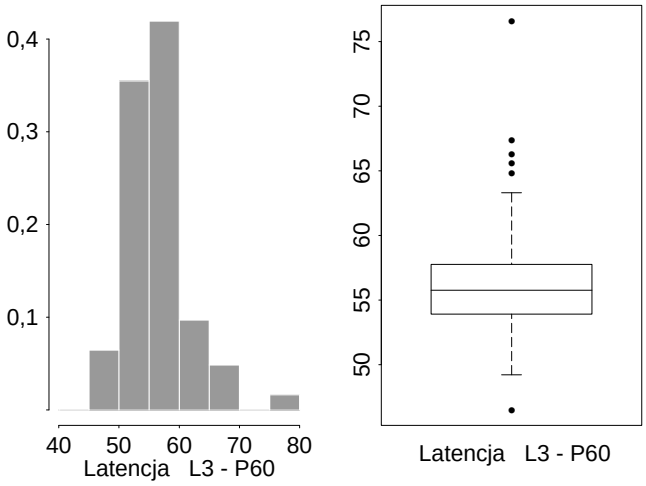

Rys. 1.21. Histogram częstości dla danych z przykł. 1.13

**Tabela 1.2. Infekcje pooperacyjne po wymianie stawu**

| Miesiąc | Pacjenci | Infekcje | Frakcja | Miesiąc | Pacjenci | Infekcje | Frakcja |
|---|---|---|---|---|---|---|---|
| 1 | 50 | 3 | 0,060 | 13 | 66 | 8 | 0,121 |
| 2 | 42 | 2 | 0,048 | 14 | 49 | 5 | 0,102 |
| 3 | 37 | 6 | 0,162 | 15 | 55 | 4 | 0,073 |
| 4 | 71 | 5 | 0,070 | 16 | 41 | 2 | 0,049 |
| 5 | 55 | 6 | 0,109 | 17 | 29 | 0 | 0 |
| 6 | 44 | 6 | 0,136 | 18 | 40 | 3 | 0,075 |
| 7 | 38 | 10 | 0,263 | 19 | 41 | 2 | 0,049 |
| 8 | 33 | 2 | 0,061 | 20 | 48 | 5 | 0,104 |
| 9 | 41 | 4 | 0,098 | 21 | 52 | 4 | 0,077 |
| 10 | 27 | 1 | 0,037 | 22 | 55 | 6 | 0,109 |
| 11 | 33 | 1 | 0,030 | 23 | 49 | 5 | 0,102 |
| 12 | 49 | 3 | 0,061 | 24 | 60 | 2 | 0,033 |

W tym przypadku jedna obserwacja jest rzeczywiście obserwacją odstającą. Wyjątkowo dużą frakcję infekcji pooperacyjnych zaobserwowano w siódmym miesiącu zbierania danych. Był to czwarty, wyjątkowo pechowy miesiąc pracy nowo utworzonego zespołu operującego: na 13 operacji przeprowadzonych przez ten zespół, infekcje wystąpiły aż po sześciu. Do przykładu tego wrócimy po omówieniu przykładu poniższego.

**Przykład 1.14.** Badanie, o którym była już mowa w przykł. 1.4, dotyczyło wielu charakterystyk potencjałów wywołanych u zdrowych osobników, w tym także tzw. latencji L3-P60 (czasu od momentu wzbudzenia potencjału w korzeniu L3 do chwili osiągnięcia przez potencjał drugiego minimum lokalnego). Na rysunku 1.22 przedstawiono histogram i wykres ramkowy rozkładu tej latencji.




Rys. 1.22. Histogram częstości i wykres ramkowy dla danych z przykł. 1.14

Tym razem, pięć obserwacji leży poza górnym wąsem, ale tylko jedna z nich może być niewątpliwie uznana za odstającą. Jest to latencja uzyskana w przypadku osoby wyjątkowo wysokiej. Osiąganie kolejnych minimów dla tej osoby następuje z coraz większym opóźnieniem w stosunku do czasów potrzebnych w przypadku osobników o przeciętnym wzroście. W rezultacie, wzrost ma duży wpływ na czas osiągnięcia drugiego minimum lokalnego. Obserwacje bliskie górnemu wąsowi należy prawdopodobnie przypisać dużej skośności rozkładu cechy w próbie (skośność rozkładu może pochodzić stąd, że rozkład wzrostu w próbie nie jest symetryczny, lecz także jest prawostronnie skośny). Indywidualnie odnotowana obserwacja w zakresie małych wartości latencji jest nietypowa dla skośnego rozkładu cechy w próbie, ale nie pochodzi od osobnika szczególnie niskiego. Przyczyna jej otrzymania powinna zostać przeanalizowana przez zbierających i zapisujących dane oraz, jeśli nie popełniono żadnego błędu, także przez osobę odpowiedzialną za merytoryczny aspekt prowadzonych badań (dopiero głębsza analiza przypadku może odpowiedzieć na pytanie czy badany osobnik był pod jakimś względem nietypowy i jego przypadek może być usunięty z analiz, czy też rozkład latencji L3-P60, jakkolwiek prawostronnie skośny ma też wyraźnie zaznaczający się lewy ogon).

Bardzo ważną zaletą wykresów ramkowych jest możliwość ich łatwego wykorzystania do porównania rozkładów pewnej cechy w różnych próbach. Może nas np. interesować porównanie rozkładów zarobków w różnych grupach zawodowych albo w populacjach ludzi o różnym wykształceniu (np. popu-

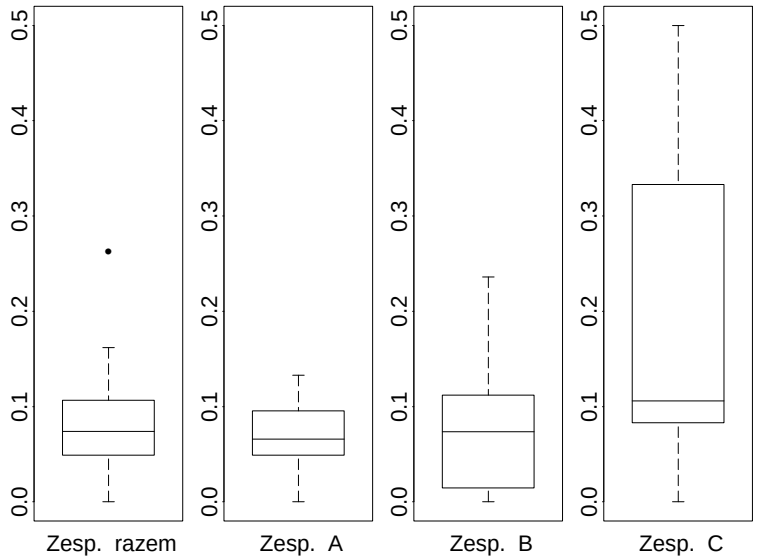
lacji osób z wykształceniem podstawowym, z wykształceniem średnim oraz z wykształceniem wyższym technicznym). Z kolei w przykł. 1.13 konieczne jest głębsze wejrzenie w przyczyny dużego rozproszenia interesującego nas rozkładu frakcji infekcji pooperacyjnych.

Mianowicie, rozstęp międzykwartylowy, równy 0,057, jest duży w porównaniu z medianą, równą 0,074. Rozstęp $R$ jest ogromny, jeśli uwzględnić obserwację odstającą ($R = 0{,}263$), i pozostaje bardzo duży po jej usunięciu (jego wartość maleje do wartości 0,162). Duże rozproszenie rozkładu może wynikać z różnej jakości pracy zespołów operujących. Jeżeli okaże się, że jakiś zespół osiągał lepsze wyniki niż pozostałe, wskazane będzie poznanie tego przyczyny i wykorzystanie zdobytej wiedzy do poprawy pracy gorszych zespołów. W ten sposób powinno uzyskać się zmniejszenie rozproszenia rozkładu infekcji pooperacyjnych w przyszłości. Słowem, pogłębioną analizę problemu należy rozpocząć od porównania rozkładów dla różnych zespołów.

**Przykład 1.13 cd.** W interesującym nas szpitalu, endoprotezy są zakładane przez trzy zespoły operacyjne (patrz tab. 1.3). Na rysunku 1.23 przedstawiono wykresy ramkowe frakcji infekcji pooperacyjnych uzyskanych przez te zespoły w ciągu dwóch lat. Widać wyraźnie, że najlepszy jest zespół A, natomiast C jest zespołem zdecydowanie najgorszym.

Okazało się, że szczególnie słabe wyniki zespołu C wynikały z braku rutyny i zbyt wczesnego skierowania do samodzielnej pracy części chirurgów z zespołu.

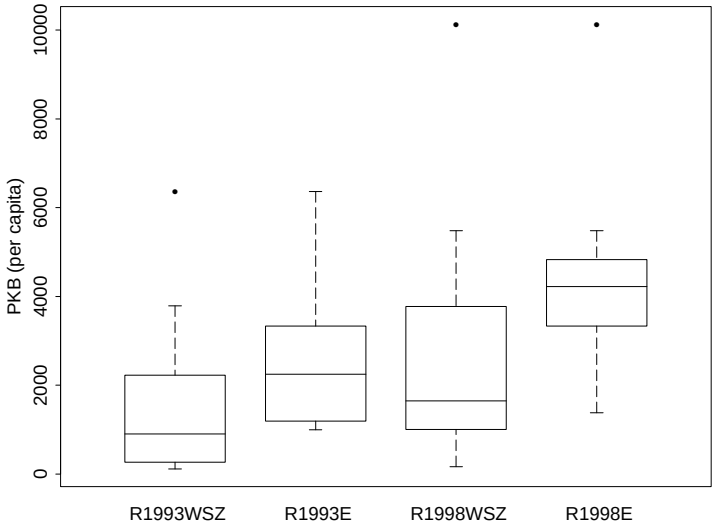
Przykład 1.13 wymaga dodatkowego komentarza. Szpital, o którym mowa w przykładzie, nie analizował danych we właściwy sposób. Dane z tab. 1.3 były archiwizowane i dopiero po dwóch latach kierownictwo zainteresowało się ogólną średnią frakcji infekcji pooperacyjnych. Wartość średnia, równa 0,084, choć wyższa od mediany (jak pamiętamy, równej 0,074), nie wzbudziła niepokoju kierownictwa, ponieważ odpowiadała średniej wykazywanej w owym czasie przez inne szpitale. Dopiero zatrudniony w szpitalu statystyk zauważył niepokojące rozproszenie rozkładu i faktycznie zainicjował bardziej wnikliwą analizę. A przecież dane były zbierane miesiąc po miesiącu, w pewnym momencie stworzono nowy zespół operujący, i nie tylko można było, ale trzeba było wyniki uzyskiwane przez każdy zespół nanosić na bieżąco, na trzy indywidualne wykresy przebiegu. Na szczęście, na podstawie danych dla zespołu C (patrz tab. 1.3) można sądzić, że zespół ten dokładnie śledził wyniki swej pracy, doskonalił procedury postępowania i dzięki temu, po około piętnastu miesiącach dogonił wprawniejszych kolegów.




Rys. 1.23. Wykresy ramkowe dla danych z przykł. 1.13

**Tabela 1.3. Infekcje pooperacyjne - dane dla zespołów A, B i C ($n$ – liczba operacji, $m$ – liczba infekcji)**

| | Zespół A | | | Zespół B | | | Zespół C | | |
|---|---|---|---|---|---|---|---|---|---|
| Miesiąc | $n$ | $m$ | Frakcja | $n$ | $m$ | Frakcja | $n$ | $m$ | Frakcja |
| 1 | 20 | 1 | 0,050 | 30 | 2 | 0,067 | 0 | 0 | - |
| 2 | 22 | 2 | 0,091 | 20 | 0 | 0 | 0 | 0 | - |
| 3 | 20 | 2 | 0,100 | 17 | 4 | 0,236 | 0 | 0 | - |
| 4 | 30 | 2 | 0,067 | 35 | 1 | 0,029 | 6 | 2 | 0,333 |
| 5 | 17 | 2 | 0,118 | 25 | 2 | 0,080 | 13 | 2 | 0,154 |
| 6 | 20 | 1 | 0,050 | 15 | 2 | 0,133 | 9 | 3 | 0,333 |
| 7 | 15 | 2 | 0,133 | 10 | 2 | 0,200 | 13 | 6 | 0,462 |
| 8 | 21 | 1 | 0,048 | 9 | 0 | 0 | 3 | 1 | 0,333 |
| 9 | 19 | 1 | 0,053 | 19 | 2 | 0,106 | 3 | 1 | 0,333 |
| 10 | 10 | 0 | 0 | 15 | 0 | 0 | 2 | 1 | 0,500 |
| 11 | 15 | 1 | 0,067 | 15 | 0 | 0 | 3 | 0 | 0 |
| 12 | 25 | 1 | 0,040 | 20 | 1 | 0,050 | 4 | 1 | 0,250 |
| 13 | 31 | 2 | 0,065 | 20 | 2 | 0,100 | 15 | 4 | 0,267 |
| 14 | 19 | 1 | 0,053 | 20 | 1 | 0,050 | 10 | 3 | 0,300 |
| 15 | 25 | 1 | 0,025 | 20 | 2 | 0,100 | 10 | 1 | 0,100 |
| 16 | 19 | 2 | 0,106 | 15 | 0 | 0 | 7 | 0 | 0 |
| 17 | 10 | 0 | 0 | 9 | 0 | 0 | 10 | 0 | 0 |
| 18 | 14 | 1 | 0,071 | 16 | 1 | 0,063 | 10 | 1 | 0,100 |
| 19 | 10 | 1 | 0,100 | 10 | 1 | 0,100 | 21 | 0 | 0 |
| 20 | 15 | 1 | 0,067 | 10 | 2 | 0,200 | 23 | 2 | 0,087 |
| 21 | 20 | 1 | 0,020 | 20 | 2 | 0,100 | 12 | 1 | 0,083 |
| 22 | 19 | 2 | 0,106 | 17 | 2 | 0,118 | 19 | 2 | 0,106 |
| 23 | 14 | 1 | 0,071 | 15 | 2 | 0,133 | 20 | 2 | 0,100 |
| 24 | 20 | 1 | 0,050 | 20 | 1 | 0,050 | 20 | 0 | 0 |




Rys. 1.24. Wykresy ramkowe dla danych z przykł. 1.15

**Przykład 1.15.** Międzynarodowy Fundusz Walutowy regularnie przygląda się grupie państw takich jak Polska, przeżywających okres ekonomicznej transformacji. W niektórych analizach państwa te są traktowane jako jeden blok lub wyróżnia się wśród nich pewną ich podgrupę. Na rysunku 1.24 są pokazane wykresy ramkowe czterech rozkładów dochodu krajowego brutto per capita (na głowę mieszkańca) w dolarach USA, w latach 1993 i 1998. Pierwsze dwa wykresy dotyczą roku 1993: najpierw jest podany rozkład dochodu w całej grupie państw w trakcie transformacji i następnie w tej samej grupie, ale z wyłączeniem – poza Estonią – byłych republik Związku Radzieckiego (Estonia została zachowana, jako kraj, którego wskaźniki ekonomiczne są podobne do innych europejskich państw przeżywających transformację). Dwa kolejne wykresy odnoszą się do tych samych grup w roku 1998. We wszystkich przypadkach obserwacją odstającą jest Słowenia (w przypadku bez byłych republik Związku Radzieckiego, w roku 1993 Słowenia znajdowała się u szczytu górnego wąsa). Za długi dolny ogon na ostatnim wykresie ramkowym odpowiada Bułgaria, której produkt krajowy per capita był w roku 1998 mniejszy niż Białorusi. Interpretację wykresów pozostawiamy Czytelnikowi.

Jak to wynika z podanych przykładów, wykresy ramkowe rzeczywiście umożliwiają łatwą i przekonującą analizę zjawisk, które najlepiej opisywać, po-

dając rozkład interesującej cechy, a nie jej pojedynczą, liczbową charakterystykę.

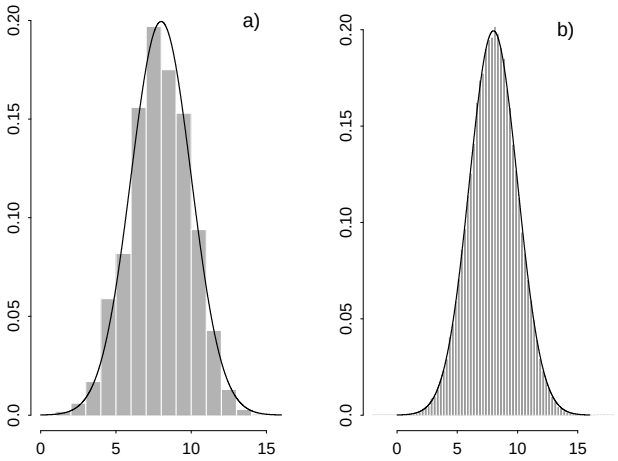
Nietrudno się domyślić, że każdy z pakietów statystycznych oferuje różne odmiany wykresów ramkowych. Na przykład, pakiety stwarzają możliwość zastąpienia nanoszonej przez nas wartości mediany wartością średnią oraz umożliwiają dobieranie maksymalnej długości wąsów wedle własnego uznania. Niektóre pakiety graficznie rozróżniają obserwacje odstające oraz obserwacje ekstremalnie odstające. Obserwacjami ekstremalnie odstającymi nazywa się obserwacje oddalone od przedziału $[Q_1, Q_3]$ o więcej niż $3 \times IQR$. Obserwacjami odstającymi nazywa się wówczas obserwacje leżące w przedziale $(Q_3 + 1{,}5 \times IQR, Q_3 + 3 \times IQR]$, gdy chodzi o wartości duże oraz w przedziale $[Q_1 - 3 \times IQR, Q_1 - 1{,}5 \times IQR)$, gdy chodzi o wartości małe. Ogólna idea wszystkich wykresów ramkowych proponowanych przez pakiety jest jednak taka sama.

## 1.4. Gęstości rozkładów – wprowadzenie

### 1.4.1. Podstawowe pojęcia

Najbardziej istotną dla nas cechą histogramu jest jego kształt, który często daje się zwięźle opisać za pomocą bliskiej niemu, ciągłej i regularnej krzywej. Rozpatrzmy histogram częstości dla 1000 wyników pewnego pomiaru, przy czym skalę na osi poziomej wybierzmy tak, aby przedział histogramu miał jednostkową długość (rys. 1.25a). W tym przypadku pola słupków są nie tylko proporcjonalne do częstości wpadnięcia do odpowiednich przedziałów wartości, ale są po prostu im równe i całkowita suma pól słupków jest równa 1.

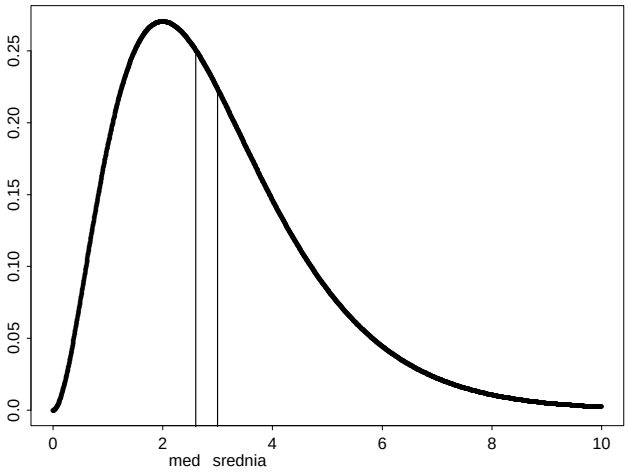
Zauważmy, że dla małej długości przedziału $h = 0{,}25$ (dla 100000 wyników pomiaru), słupki histogramu (rys. 1.25b) są bardzo wąskie, a zmiany wysokości przylegających do siebie słupków nieznaczne. Na skutek tego, po pominięciu pionowych odcinków łączących podstawy słupków, histogram daje się bardzo dobrze opisać przez wyrysowaną krzywą ciągłą. Zachowuje ona zasadnicze cechy histogramu: jej maksimum jest bliskie wartości modalnej histogramu, a szybkość malenia przy oddalaniu się od mody jest zbliżona do szybkości malenia częstości histogramu. Oczywiście, pole pod krzywą nad każdym przedziałem histogramu powinno być bliskie odpowiedniej częstości. Tak jest w istocie, na przykład dla przedziału (10, 11) częstość pomiarów w tym przedziale jest równa 0,094, podczas gdy odpowiadające pole pod krzywą wynosi 0,091848. Pole pod całą krzywą wynosi 1 i jest ono równe sumie częstości odpowiadających wszystkim przedziałom histogramu. Krzywą ciągłą przybliżającą histogram i taką, że pole pod nią wynosi 1 nazywamy **krzywą (funkcją) gęstości** lub po prostu **gęstością**. Stanowi ona bardzo




Rys. 1.25. Histogramy dla a) 1000, b) 100000 wyników pomiaru

użyteczny punkt odniesienia, który syntetyzuje podstawowe cechy kształtu histogramu. Często określa się ją jako wyidealizowany histogram, odpowiadający bardzo dużej liczności próby i bardzo małej długości przedziału. Im większa liczność próby tym mniejsza może być długość przedziału histogramu i tym bliższy staje się histogram histogramowi idealnemu. Taki idealny histogram odpowiada rozkładowi pewnej ciągłej cechy $X$ w próbie. Ponieważ długość przedziału histogramu jest bardzo mała, nie tracimy informacji, przechodząc od rozkładu cechy w próbie do jej histogramu idealnego. Idealny histogram (gęstość) i rozkład cechy $X$ są w tej sytuacji równoważne. Dlatego o gęstości mówimy wtedy jako o **gęstości rozkładu cechy** $X$.

Oczywiście, ogromnej różnorodności kształtów histogramów odpowiada wielka różnorodność kształtów gęstości. Terminologia służąca do opisu kształtów gęstości pozostaje na szczęście taka sama jak dla histogramów. Możemy mieć zatem gęstość na przykład prawostronnie skośną. Podobnie jest z definicją mody i mediany: **modą** nazywamy każdy punkt na osi poziomej, w którym występuje maksimum lokalne gęstości, a **mediana** oznacza taki punkt, że pole pod krzywą na lewo od niego jest równe polu na prawo (i oczywiście równe 1/2). Nieco bardziej skomplikowana jest definicja wartości średniej gęstości. Precyzyjnie zostanie ona przedstawiona w p. 2.2.5. Tu omówimy tylko intuicyjnie to pojęcie. Wyobraźmy sobie, że kształt gęstości został wycięty z blachy i staramy się znaleźć na jego poziomej podstawie punkt, w którym moglibyśmy utrzymać wycięty kształt w równowadze. Środek ciężkości wykonanej figury jest dokładnie **wartością średnią gęstości**. Oznaczać ją będziemy w książce przez $\mu$. W przypadku gęstości symetrycznej i jednomodalnej wszystkie definicje środka się pokrywają: wartość średnia jest




Rys. 1.26. Prawostronnie skośna gęstość

równa medianie i modzie. Na rysunku 1.26 przedstawiono prawostronnie skośną gęstość i usytuowanie jej wartości średniej i mediany: podobnie jak w przypadku histogramu, siła grawitacji działająca na długi prawy ogon „przeciąga" wartość średnią na prawą stronę mediany. Natomiast znacznie trudniej intuicyjnie jest przedstawić **odchylenie standardowe gęstości**, które będziemy oznaczać symbolem $\sigma$. Poprzestańmy na razie na stwierdzeniu, że jest ono odpowiednikiem odchylenia standardowego rozkładu cechy w próbie.

Potrzebny nam będzie jeszcze odpowiednik kwartyli dla gęstości. Podamy najpierw bardziej ogólną definicję. **Kwantylem rzędu** *$p$ nazywamy taki punkt $q_p$ na osi poziomej, że pole pod gęstością na lewo od niego jest równe dokładnie $p$, a pole na prawo równe $1-p$.* Ponieważ całkowite pole pod krzywą wynosi 1, więc definicja ta ma oczywiście sens tylko dla $p$ takiego, że $0 < p < 1$. Tak więc mediana to nic innego niż kwantyl rzędu 0,5, a pierwszy (trzeci) kwartyl to kwantyl rzędu 0,25 (0,75). Pożyteczne jest uzmysłowienie sobie, że drugim kwartylem jest mediana. W podobny sposób możemy zdefiniować kwantyle w próbie. Rozpatrzmy **rozstęp międzykwartylowy gęstości** zdefiniowany jako $q_{0,75} - q_{0,25}$. Jest on oczywiście odpowiednikiem rozstępu międzykwartylowego cechy w próbie i stanowi konkurencyjną wobec odchylenia standardowego $\sigma$ miarę rozproszenia.

Podkreślmy raz jeszcze, że choć wszystkie wskaźniki liczbowe rozkładu cechy w próbie mają swoje odpowiedniki wśród wskaźników gęstości opisanych w tab. 1.4, jest między nimi zasadnicza różnica, wynikająca z różnicy między histogramem dla skończonej próby a gęstością. Gęstość odpowiada histogramowi zbudowanemu dla nieskończenie wielkiej próby i nie ma znaczenia, jaką

konkretnie nieskończenie wielką próbę rozpatrywanej cechy weźmiemy: gęstość pozostanie taka sama. Jednakże histogramy, na przykład wzrostu dla dwóch prób 100 dorosłych Polaków, będą się z reguły różnić. Zatem również wskaźniki liczbowe rozkładu w próbie zmieniają się od próby do próby, w przeciwieństwie do wskaźników gęstości. Wartość wskaźnika w próbie wartości pewnej cechy służy z reguły do oszacowania odpowiedniego wskaźnika gęstości tej cechy. Na ile dobre jest to oszacowanie przekonujemy się, badając rozkład wskaźnika, traktowanego jako nowa cecha o wartościach losowych. Zajmiemy się tym w rozdz. 3.

**Tabela 1.4. Odpowiedniość między wskaźnikami liczbowymi rozkładu cechy w próbie i wskaźnikami gęstości rozkładu cechy**

| Rozkład cechy w próbie | Gęstość rozkładu cechy |
|---|---|
| Wartość średnia $\bar{x}$ | Wartość średnia $\mu$ |
| Odchylenie standardowe $s$ | Odchylenie standardowe $\sigma$ |
| Pierwszy kwartyl $Q_1$ | Pierwszy kwartyl $q_{0,25}$ |
| Mediana $x_{med}$ | Mediana $q_{0,5}$ |
| Trzeci kwartyl $Q_3$ | Trzeci kwartyl $q_{0,75}$ |

### 1.4.2. Gęstości normalne

Jednej klasie gęstości chcielibyśmy poświęcić więcej uwagi. Jest to klasa gęstości normalnych. Cechę, której rozkład jest zadany przez gęstość normalną nazywamy **cechą o rozkładzie normalnym** lub prościej **cechą normalną**. Gęstość normalna jest określona przez dwa parametry $\mu$ i $\sigma$, gdzie $\mu$ jest dowolną, a $\sigma$ dodatnią liczbą rzeczywistą i jest opisana przez funkcję

$$\phi_{\mu,\sigma}(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/(2\sigma^2)}. \tag{1.10}$$

Funkcja ta dla każdej wartości $x$ zadaje wartość krzywej gęstości $\phi_{\mu,\sigma}(x)$. Rozkład cechy normalnej jest symbolicznie zapisywany jako $N(\mu, \sigma)$. Z gęstością tą zetknęliśmy się już w poprzednim punkcie, gdzie przybliżała ona histogram 1000 wartości pewnego pomiaru (dla $\mu = 8$ i $\sigma = 2$, rys. 1.25).

Przede wszystkim, w krzywej normalnej nie ma nic normalnego! Chcemy przez to powiedzieć, że nie stanowi ona jedynego wzorca, do którego będziemy odnosili wszystkie inne gęstości i że nazwa gęstość normalna nie powinna sugerować, że gęstość znacznie odbiegająca od niej kształtem jest w jakiś sposób nieprawidłowa. Sama nazwa jest swoistym wybiegiem i pozwala uniknąć drażliwej kwestii, czy gęstość ta powinna być nazywana krzywą Gaussa czy krzywą de Moivre'a, którzy to badacze niezależnie wprowadzili tę

krzywą w naukach przyrodniczych. Swoją popularność i częstą stosowalność (nie zawsze uzasadnioną!) gęstość normalna zawdzięcza kilku przyczynom.

Po pierwsze, łatwo można ją przedstawić graficznie. Jest to krzywa symetryczna, jednomodalna, o szybko malejących ogonach, mająca charakterystyczny kształt dzwonu. Po drugie, histogramy zbliżone do gęstości normalnych pojawiają się często przy analizie różnych rozkładów dla danych eksperymentalnych i zjawisk losowych. O przyczynach tego powiemy w podrozdz. 2.4. Po trzecie, parametry $\mu$ i $\sigma$ opisujące gęstość normalną mają bardzo prostą interpretację. Mianowicie, $\mu$ jest modą krzywej gęstości i, ponieważ gęstość jest symetryczna, również jej wartością średnią i medianą. Zmiana $\mu$ przy ustalonym $\sigma$ powoduje jedynie przesunięcie wykresu gęstości bez jakiejkolwiek zmiany jego kształtu. Liczba $\sigma$ okazuje się być równa odchyleniu standardowemu gęstości normalnej. Nie jest to bynajmniej oczywiste ze wzoru (1.10) i wymaga precyzyjnego uzasadnienia. Wartość $\sigma$ określa zatem wartość rozproszenia rozkładu cechy normalnej, jej zwiększenie powoduje wypłaszczenie centralnej części wykresu i pogrubienie ogonów. Ze wzoru (1.10) wynika natomiast prosto, że punkty $\mu \pm \sigma$ są punktami zerowania się drugiej pochodnej funkcji $\phi_{\mu,\sigma}$. Zatem punkt $\mu + \sigma$ może być zlokalizowany na wykresie jako ten punkt, w którym coraz szybszy spadek krzywej gęstości przy poruszaniu się na prawo od $\mu$ przechodzi w spadek coraz wolniejszy.

Jak wspomnieliśmy powyżej, gęstości normalne dla różnych $\mu$ i $\sigma$ mają podobny kształt. Umożliwia to opisanie kształtu gęstości $N(\mu, \sigma)$ za pomocą kształtu ustalonej gęstości normalnej, na przykład $N(0, 1)$ i liczb $\mu$ i $\sigma$. Służy temu tak zwana standaryzacja.

**Standaryzacja.** Jeśli cecha $X$ ma rozkład normalny $N(\mu, \sigma)$, to cecha $Z = (X - \mu)/\sigma$ ma rozkład $N(0, 1)$.

Rozkład $N(0, 1)$ nazywamy standardowym rozkładem normalnym, a $\phi_{0,1}$ standardową gęstością normalną. W dalszym ciągu zamiast $\phi_{0,1}$ będziemy pisali po prostu $\phi$. Zasada standaryzacji orzeka, że własności wszystkich krzywych normalnych są dokładnie takie same, gdy ze zwykłych jednostek przejdziemy na jednostki standaryzowane, mówiące o ile odchyleń standardowych $\sigma$ znajdujemy się od wartości średniej $\mu$. Rozważmy na przykład punkt $x$, znajdujący się o dwa odchylenia standardowe na prawo od wartości średniej $\mu$, czyli punkt $x = \mu + 2\sigma$. W jednostkach standaryzowanych jego wartość wynosi $z = (x - \mu)/\sigma = 2$. Z zasady standaryzacji wynika, że pole pod krzywą normalną na lewo od $x = \mu + 2\sigma$ jest równe polu pod standardową krzywą normalną na lewo od odpowiadającej jej wartości $z = 2$. Tę ostatnią wartość otrzymujemy z tabl. I zamieszczonej na końcu książki. Wynosi ona w przybliżeniu 0,975. Podobnie, pole na lewo od punktu $x = \mu - 2\sigma$ wynosi około 0,025. Otrzymujemy stąd łatwą do zapamiętania regułę.

**Reguła pięciu procent**. *Pole zawarte pod gęstością normalną $\phi_{\mu,\sigma}$ między punktami $\mu-2\sigma$ i $\mu+2\sigma$ jest równe 0,95. Pole pod tą krzywą* **na zewnątrz** *odcinka $(\mu-2\sigma, \mu+2\sigma)$ wynosi 0,05, czyli 5% całego pola.*

Pamiętając, że dla cechy normalnej $X$ gęstość $N(\mu,\sigma)$ opisuje jej idealny histogram częstości, możemy powyższą zasadę sformułować następująco: w przybliżeniu 95% warości cechy $X$ w próbie jest zawartych między $\mu-2\sigma$ i $\mu+2\sigma$. Oznacza to, że przedział $(\mu-2\sigma, \mu+2\sigma)$ reprezentuje typowe wartości cechy normalnej i tylko w około 5% przypadków otrzymamy nietypową wartość spoza tego przedziału.

W podobny sposób możemy interpretować pole pod krzywą normalną nad odcinkiem $(\mu-c\sigma, \mu+c\sigma)$ dla dowolnego dodatniego $c$. Zauważmy, że standaryzowane wartości końców odcinka wynoszą $(\mu \pm c\sigma)/\sigma = \pm c$. Zatem żeby obliczyć wartość pola, zgodnie z zasadą standaryzacji, wystarczy obliczyć pole pod standardową gęstością normalną nad odcinkiem $(-c, c)$, czyli wyrażając pole pod krzywą przez odpowiednią całkę, otrzymamy

$$\int_{-c}^{c} \phi(z)\,dz = \int_{-\infty}^{c} \phi(z)\,dz - \int_{-\infty}^{-c} \phi(z)\,dz = \Phi(c) - \Phi(-c)\,, \tag{1.11}$$

gdzie $\Phi(x) = \int_{-\infty}^{x} \phi(z)\,dz$ jest tak zwaną **dystrybuantą** standardowej gęstości normalnej, równą polu na lewo od punktu $x$. Zauważmy, że wzór (1.11) można jeszcze uprościć, pamiętając o tym, że standardowa gęstość normalna jest symetryczna względem zera, a zatem pole pod nią na lewo od $-c$ jest równe polu na prawo od $c$. Tak więc $\Phi(c) = 1 - \Phi(-c)$ i wartość pola w (1.11) jest równa $2\Phi(c) - 1$.

Zatem, korzystając ze standaryzacji, obliczenie wartości pola pod krzywą normalną nad dowolnym odcinkiem możemy zawsze sprowadzić do obliczenia odpowiednich wartości dystrybuanty! Jedyny problem polega na tym, że $\Phi(x)$ nie daje wyrazić się jawnie jako prosta funkcja $x$ i dlatego musiano się uciec do stablicowania jej wartości obliczonych numerycznie. Przedstawiono je na końcu książki w formie tabl. I – jest to niewątpliwie najczęściej używana tablica statystyczna. Wiersze tablicy odpowiadają częściom dziesiętnym wartości $x$, a kolumny częściom setnym.

**Przykład 1.16. Obliczenie pola pod krzywą normalną.** Stwierdzono, że wzrost dorosłych Polaków jest cechą o rozkładzie normalnym o wartości średniej $\mu = 176$ cm i odchyleniu standardowym $\sigma = 6{,}5$ cm. Znajdźmy centralny przedział wokół $\mu$ reprezentujący 95% warto-

ści wzrostu wszystkich dorosłych Polaków. Ponieważ $\mu + 2\sigma = 189$ cm i $\mu - 2\sigma = 163$ cm, zgodnie z regułą 5% typowe, czyli obejmujące 95% częstości wartości są zawarte między 163 i 189 cm.

Oczywiście, co uważamy za typowe, jest kwestią umowy. Załóżmy, że za nietypowych uznano wszystkich mających powyżej 195 cm wzrostu i obliczmy, jaki procent dorosłych Polaków zostanie tak sklasyfikowanych. Tak więc chodzi nam o obliczenie pola pod gęstością normalną $\phi_{176,\,6{,}5}$ na prawo od punktu $x = 195$ cm. Standaryzowana wartość $x$ wynosi

$$z = \frac{195 - 176}{6{,}5} = 2{,}92\,.$$

Z tablicy I otrzymujemy $\Phi(2{,}92) = 0{,}9982$, zatem częstość przekroczenia 195 cm wzrostu wynosi $1 - 0{,}9982 = 0{,}0018$. Oznacza to, że wśród 10 000 Polaków przeciętnie osiemnastu będzie wzrostu powyżej 195 cm, a wśród miliona będzie ich przeciętnie 1800.

**Obliczenie kwantyla rozkładu normalnego**. Rozważmy teraz problem następujący: ile trzeba mieć wzrostu, aby znaleźć się wśród 10% najniższych Polaków? Innymi słowy pytamy, ile wynosi kwantyl rzędu 0,1 dla gęstości $\phi_{176,\,6{,}5}$, gdyż 10% populacji ma wzrost poniżej tego właśnie kwantyla. Oznaczmy wartość takiego kwantyla przez $x$. Po standaryzacji $x$ zmieni się w wartość kwantyla tego samego rzędu, ale dla standardowej gęstości normalnej, którą łatwo znaleźć w tabl. I. W tym celu szukamy wartości 0,9 **wewnątrz** tablicy, a następnie odczytujemy odpowiadający jej argument, będący poszukiwaną wartością kwantyla. Wynosi ona $-1{,}29$. Zatem

$$\frac{x - 176}{6{,}5} = -1{,}29$$

i $x = 176$ cm $- 1{,}29 \times 6{,}5$ cm $= 167{,}6$ cm. Tak więc każdy dorosły Polak mający mniej niż 167,6 cm wzrostu należy do 10% najniższych Polaków. Zauważmy, że rozwiązaliśmy problem w pewnym sensie **odwrotny** do problemu poprzedniego, gdzie dla danej wartości wzrostu szukaliśmy częstości z jaką ta wartość jest przekraczana. Tutaj, dla danej częstości przekroczenia 0,9 znaleźliśmy odpowiadającą jej graniczną wartość wzrostu 167,6 cm.

W tym miejscu wyjaśnimy skąd bierze się podane w p. 1.3.3 ograniczenie na długość wąsów. Wartości, które są nanoszone na wykres indywidulanie, są obserwacjami w pewnym sensie bardzo odległymi od centralnej połowy próby. Są to albo obserwacje odstające, albo wskazujące, że rozkład cechy w próbie ma bardzo długi jeden lub obydwa ogony. Pytaniem, na które trzeba zatem jeszcze odpowiedzieć, jest: na jakiej podstawie uznać obserwacje za „bardzo odległe" od centralnej połowy próby. Po pierwsze zauważmy, że rozstępowi międzykwartylowemu $IQR$ na wykresie ramkowym odpowiada rozstęp międzykwartylowy gęstości $q_{0,75} - q_{0,25}$.

Po drugie zgódźmy się uznać rozkład normalny za swego rodzaju rozkład odniesienia, jako mający własność symetrii oraz niezbyt grube i długie ogony. Można obliczyć (patrz zad. 1.10), że jeżeli mamy do czynienia z rozkładem normalnym, to poza przedziałem

$$[q_{0,25} - 1{,}5 \times (q_{0,75} - q_{0,25}), q_{0,75} + 1{,}5 \times (q_{0,75} - q_{0,25})]$$

znajduje się średnio 7 obserwacji na 1000. Zatem, jeżeli wykres ramkowy wykazuje istnienie obserwacji leżących poza przedziałem

$$[Q_1 - 1{,}5 \times IQR, Q_3 + 1{,}5 \times IQR],$$

to możemy uznać, że rozkład cechy w próbie bardzo daleko odbiega od rozkładu normalnego. I stąd, na podstawie powszechnie przyjętej umowy, długość wąsów ograniczono do wartości $1{,}5 \times IQR$. Jest jasne, że z podobnym skutkiem można by zastosować nieco inną krotność rozstępu międzykwartylowego. Ponadto, dla rozkładu normalnego poza przedziałem

$$[q_{0,25} - 3 \times (q_{0,75} - q_{0,25}), q_{0,75} + 3 \times (q_{0,75} - q_{0,25})]$$

znajdują się średnio 24 obserwacje na 10 milionów. I stąd wzięło się nazywanie tych obserwacji ekstremalnie odstającymi.

## 1.5. Zadania

**1.1.** Przedstawione poniżej dane są wynikami ankiety dotyczącej problemów ze snem przeprowadzonej w Polsce i kilku krajach Europy Zachodniej (oznaczonej EZ i obejmującej Niemcy, Włochy, Francję i Wielką Brytanię). Ankietowani wybierali jedną z następujących odpowiedzi: nigdy nie mam problemów ze snem; mam je każdej nocy; mam je kilka razy w miesiącu; mam je kilka razy w tygodniu. Wyniki są podane w procentach.

| | EZ | Polska |
|---|---|---|
| nigdy | 64% | 61% |
| co noc | 14% | 8% |
| kilka razy w miesiącu | 7% | 11% |
| kilka razy w tygodniu | 15% | 20% |

(źródło: Gazeta Wyborcza za Pentorem z 11 kwietnia 2000). Czy podana kolejność kategorii jest logiczna? Jeśli nie, dokonać innego wyboru kolejności kategorii dla wykresu słupkowego i uzasadnić go. Sporządzić wykresy słupkowe dla wyników z Polski i Europy Zachodniej i porównać je. Zakładając, że z reguły bezsenność pogłębia się z czasem i leczenie jej jest trudne, przedstawić roboczą hipotezę, jak sytuacja będzie wyglądać za 5 lat.

**1.2.** Przykład 4.3 zawiera dane dotyczące parametrów 24 modeli samochodów znajdujących się w roku 2000 na rynku polskim.

**a)** Sporządzić histogram i wykres ramkowy zmiennej zużycia paliwa na 100 km (zmienna $Y$). Przeanalizować, jak długość przedziału histogramu wpływa na jego charakter. Opisać charakter otrzymanych wykresów, znaleźć medianę i wartość średnią w próbie. Czy zależność między medianą a wartością średnią zgadza się z charakterem wykresów? Zidentyfikować samochód odpowiadający najmniejszej i największej obserwacji. Czy któraś z nich jest obserwacją odstającą?

**b)** Powtórzyć czynności z punktu a dla szerokości pojazdów (zmienna $x_6$). Czy najmniejszym i największym obserwacjom w a i b odpowiadają te same samochody?

**1.3.** W tabeli 1.5 podano wartości sumy opadów w Warszawie w lipcu w kolejnych 150 latach poczynając od roku 1811 (Z. Kaczmarek (1970): *Metody statystyczne w hydrologii i meteorologii.* Warszawa, WKiŁ, tab. 1.3.2). W punkcie 1.2.2 omówiliśmy diagram przekroczeń 120 mm w ciągu dekady dla tych danych.

**a)** Sporządzić histogram i wykres ramkowy sumy opadów. Przeanalizować charakter wykresów, obliczyć podstawowe wkaźniki położenia i rozproszenia. Wymienić przynajmniej jedną informację zawartą w danych, którą można uzyskać na podstawie histogramu, a nie można na podstawie diagramu przekroczeń i odwrotnie.

**b)** Zidentyfikować obserwacje odstające. Usunąć je ze zbioru danych i porównać, o ile zmieniła się średnia i odchylenie standardowe. Uzasadnić wyniki.

**1.4.** Dane w tabeli 3.5.2 w Andrews i Herzberg (D. Andrews, P. Herzberg (1985): *Data.* Springer, New York) i udostępnione do ogólnego użytku pod adresem `http://lib.stat.cmu.edu/datasets/Andrews` zawierają dane dotyczące poziomu cukru we krwi na czczo (cecha $x_4$) i w godzinę po podaniu dawki 100 g glukozy (cecha $x_7$) dla grupy 52 kobiet w trzecim trymestrze ciąży.

**a)** Sporządzić wykresy histogramów oraz (na jednym rysunku) wykresy ramkowe wartości obu cech. Opisać charakter obu histogramów, ich wzajemne położenie i porównać wartości wskaźników położenia i rozproszenia.

**b)** Porównać kwartyl $Q_1$ cechy $x_4$ z kwartylem $Q_3$ cechy $x_7$ i opisać słowami. Co oznacza ta zależność?

**c)** Obliczyć średnie ucinane dla cech $x_4$ i $x_7$ dla obu cech przy kilku wartościach parametru $k$; porównać wyniki z wartościami zwykłych średnich.

**Tabela 1.5. Suma opadów (mm) i temperatura (°C) w Warszawie w lipcu w latach 1811-1960**

| Lata | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1811- | opady | 35 | 82 | 48 | 75 | 77 | 123 | 117 | 75 | 92 | 101 |
| 1820 | temp. | 22,5 | 18,5 | 18,5 | 20,1 | 15,7 | 16,9 | 17,5 | 18,2 | 18,3 | 15,9 |
| 1821- | opady | 116 | 113 | 42 | 44 | 36 | 71 | 9 | 74 | 114 | 49 |
| 1830 | temp. | 16 | 19,8 | 18,9 | 17 | 17,3 | 21,4 | 18,6 | 19,4 | 18,3 | 17,3 |
| 1831- | opady | 83 | 94 | 223 | 28 | 57 | 46 | 33 | 86 | 85 | 74 |
| 1840 | temp. | 18,6 | 14,1 | 17,3 | 21,8 | 19,7 | 16,8 | 16,8 | 17,5 | 20,7 | 18,6 |
| 1841- | opady | 72 | 104 | 37 | 229 | 41 | 50 | 73 | 40 | 76 | 100 |
| 1850 | temp. | 18,2 | 16,4 | 18,5 | 15,3 | 20,3 | 19,8 | 17 | 18,9 | 16,7 | 18,8 |
| 1851- | opady | 171 | 41 | 160 | 120 | 144 | 46 | 143 | 105 | 29 | 92 |
| 1860 | temp. | 17,6 | 19,1 | 18,8 | 19,7 | 19 | 17 | 18,2 | 19,9 | 20,4 | 17,5 |
| 1861- | opady | 138 | 44 | 26 | 80 | 50 | 84 | 78 | 74 | 53 | 51 |
| 1870 | temp. | 20,7 | 17,8 | 16,9 | 16,4 | 21,4 | 16,9 | 17,2 | 20,4 | 18,5 | 19,1 |
| 1871- | opady | 76 | 30 | 48 | 6 | 54 | 63 | 20 | 74 | 81 | 45 |
| 1880 | temp. | 19,4 | 19,1 | 19,8 | 20,2 | 18,9 | 18,6 | 18,6 | 16,1 | 16,6 | 19,9 |
| 1881- | opady | 50 | 174 | 82 | 18 | 139 | 31 | 47 | 78 | 173 | 71 |
| 1890 | temp. | 18,8 | 20,1 | 18,9 | 19,4 | 19,3 | 17,6 | 19,5 | 16,4 | 18 | 18,6 |
| 1891- | opady | 72 | 20 | 85 | 19 | 35 | 39 | 120 | 92 | 172 | 98 |
| 1900 | temp. | 18,7 | 17,7 | 19,3 | 20,2 | 20 | 20,4 | 18,5 | 15,6 | 18,9 | 19,8 |
| 1901- | opady | 37 | 77 | 143 | 26 | 96 | 13 | 132 | 109 | 116 | 132 |
| 1910 | temp. | 20 | 15,9 | 17,8 | 18,5 | 18,4 | 19,4 | 16,6 | 18,9 | 17,1 | 17,1 |
| 1911- | opady | 37 | 32 | 91 | 101 | 77 | 87 | 99 | 181 | 166 | 68 |
| 1920 | temp. | 19,9 | 19,9 | 17,2 | 21,2 | 17,9 | 17,8 | 18,4 | 18,2 | 16,7 | 20,4 |
| 1921- | opady | 5 | 122 | 33 | 84 | 66 | 64 | 149 | 23 | 20 | 115 |
| 1930 | temp. | 19,5 | 18,2 | 18,5 | 17,6 | 18,9 | 19,6 | 20 | 19,3 | 18,3 | 17,9 |
| 1931- | opady | 71 | 108 | 55 | 166 | 124 | 115 | 53 | 71 | 49 | 73 |
| 1940 | temp. | 19 | 21,4 | 19,7 | 18,2 | 17,1 | 20,7 | 19 | 19,5 | 21 | 19 |
| 1941- | opady | 93 | 76 | 113 | 53 | 77 | 37 | 78 | 124 | 84 | 44 |
| 1950 | temp. | 20,1 | 18 | 18,2 | 20,2 | 18,8 | 20,5 | 20,2 | 17,9 | 18,7 | 18,5 |
| 1951- | opady | 68 | 26 | 65 | 136 | 154 | 82 | 88 | 38 | 80 | 159 |
| 1960 | temp. | 19 | 19,6 | 20,6 | 17,4 | 19,4 | 17,5 | 19,2 | 19,3 | 21,5 | 16,8 |

**1.5.** (Uwaga: Próby wygenerowane w tym zadaniu będą użyte w zad. 3.21). Wygenerować dwie niezależne próby o liczności 100 ze standardowego rozkładu normalnego.

**a)** Skonstruować wykres ramkowy i histogram (przy kilku długościach przedziału) dla otrzymanych prób. Obliczyć wartość średnią, medianę i wariancję prób.

Czynności z punktu a powtórzyć dla prób otrzymanych w następujący sposób:

**b)** przekształcić jedną z oryginalnych prób w próbę z rozkładu normalnego $N(1,1)$ i połączyć z drugą z oryginalnie otrzymanych prób w jedną próbę o liczności 200;

**c)** przekształcić jedną z oryginalnych prób w próbę z rozkładu normalnego

$N(5,1)$ i połączyć z drugą z oryginalnie otrzymanych prób w jedną próbę o liczności 200;

**d)** przekształcić jedną z oryginalnych prób w próbę z rozkładu normalnego $N(0,0{,}1)$, drugą w próbę z rozkładu normalnego $N(1,0{,}1)$ i połączyć obie próby;

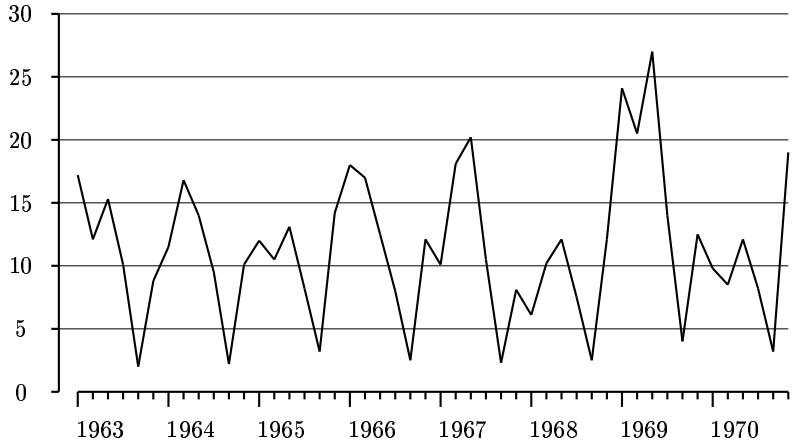
**e)** przeanalizować wyniki otrzymane w punktach a do d. Powtórzyć punkty b do d kilka razy dla nowo wygenerowanych prób i znaleźć zarówno cechy wspólne wykresów, jak i te, które zależą od konkretnie wylosowanych prób;

**f)** powtórzyć cały eksperyment (punkty a do e) rozpoczynając go z dwiema próbami ze standardowego rozkładu normalnego o liczności 50;

**g)** powtórzyć cały eksperyment (punkty a do e) rozpoczynając go z dwiema próbami ze standardowego rozkładu normalnego o liczności 25.

**1.6.** W przykładzie 1.6 analizowaliśmy histogramy dla 100 liczb losowo wybranych z odcinka $(0,1)$. Używając innej próby stuelementowej skonstruować analogiczne histogramy i porównać je z rys. 1.10. Powtórzyć całą operację dla próby zawierającej 200, 500 i 1000 liczb. Przeanalizować zmiany histogramu wraz ze zmianą liczności próby. Obliczyć wartość średnią, medianę, rozstęp międzykwartylowy i wariancję. Skomentować zachowanie się tych wartości jako funkcji liczności próby. Czy jest ono zgodne z intuicją? Uzasadnić swoje zdanie dla wartości średniej i rozstępu międzykwartylowego.

**1.7.** Wykres przebiegu przedstawiony na rys. 1.27 opisuje liczbę zachorowań (w tysiącach) na odrę w Polsce w latach 1963–1970 (okresy dwumiesięczne pomiaru; źródło: T. Dzierżykray-Rogalski (1986): *Rytmy i biorytmy biologiczne*. Wiedza Powszechna, Warszawa). Opisać dwie główne cechy tego wykresu i zinterpretować je.




Rys. 1.27. Liczba zachorowań (w tys.) na odrę w Polsce

**1.8.** Korzystając z tablic znaleźć wartość kwantyla $q_{0,05}$ i $q_{0,005}$ dla standardowego rozkładu normalnego. Ile wynoszą wartości odpowiednich kwantyli dla rozkładu normalnego $N(5,3)$?

**1.9.** W dużej fabryce samochodów średnia miesięczna płaca pracowników zatrudnionych bezpośrednio w produkcji wynosi 3300 zł, a jej odchylenie standardowe 400 zł. Pewien pracownik lakierni wie, że 65% pracowników produkcyjnych zarabia więcej od niego. Zakładając, że rozkład zarobków pracowników produkcyjnych jest w przybliżeniu normalny, obliczyć, ile wynosi miesięczna płaca tego pracownika.

**1.10.** Korzystając z tablic obliczyć wartość kwantyli $q_{0,25}$ i $q_{0,75}$ rzędu 0,25 i 0,75 dla standardowego rozkładu normalnego i sprawdzić, że poza przedziałem $[q_{0,25} - 1{,}5 \times (q_{0,75} - q_{0,25}), q_{0,75} + 1{,}5 \times (q_{0,75} - q_{0,25})]$ leży średnio 7 obserwacji na 1000.

ROZDZIAŁ 2

# Od modelu probabilistycznego do wnioskowania statystycznego

## 2.1. Model probabilistyczny – podstawy

Analiza danych nie może zostać zamknięta na etapie wstępnym. Po opisaniu rozkładu cechy w próbie, trudno nie spytać w jakim stopniu wyniki otrzymane na podstawie posiadanej próby opisują badaną cechę „w ogóle". Jeżeli interesuje nas rozkład dochodów świeżo upieczonych inżynierów informatyków w Polsce, ze względu na praktyczną niewykonalność takiego zadania nie badamy zarobków wszystkich absolwentów informatyki. Badamy natomiast zarobki pewnego podzbioru, czyli próby, interesujących nas absolwentów. W rezultacie niezbędne staje się rozwiązanie np. takiej kwestii: na ile otrzymana na podstawie próby mediana rozkładu zarobków opisuje medianę zarobków wszystkich absolwentów? Podobnie, gdy w przykł. 1.4 badamy latencję L3-N33 w próbie 62 zdrowych osobników, w rzeczywistości nie interesuje nas ta konkretna grupa osób, lecz rozkład tej latencji u wszystkich możliwych zdrowych osobników. Innymi słowy, trzeba umieć odpowiedzieć na pytanie, w jakim stopniu wnioski prawdziwe dla próby wolno uogólnić na własności cechy „w ogóle".

Powyższe pytania wymagają dogłębnego rozważenia kilku zagadnień, w tym jednego zupełnie fundamentalnego: jak poradzić sobie z tym, że z próby na próbę otrzymywane wyniki są różne? Najbardziej powszechnie stosowanym środkiem zaradczym jest odwołanie się do modeli probabilistycznych zjawiska, a właściwie do rachunku prawdopodobieństwa i następnie do opartych na nim modeli probabilistycznych.

Statystyczna analiza oraz statystyczne wnioskowanie oparte na modelach probabilistycznych odgrywa ogromną rolę w technice, ekonomii, naukach rolniczych, społecznych, medycznych i innych. W przypadku zjawiska, którego nie potrafimy opisać w ściśle deterministyczny sposób, co do którego przebiegu nie mamy zupełnej pewności, dobrą metodą jego poznania jest, jak wspomnieliśmy, odwołanie się do pomocy modelowania probabilistycznego.

Na tym jednak zadanie się nie kończy, bowiem najczęściej pewne parametry przyjętego modelu probabilistycznego mają nieznane wartości. Parametry te muszą być wywnioskowane na podstawie posiadanych danych, dzięki dokonaniu stosownej analizy statystycznej. Z tego powodu takie nie w pełni wyspecyfikowane modele probabilistyczne nazywamy modelami statystycznymi. Zajęcie się modelami statystycznymi jest zasadniczym celem tego podręcznika. Ażeby jednak móc się zająć realizacją owego celu, najpierw musimy zapoznać się z modelami probabilistycznymi.

### 2.1.1. Doświadczenia losowe i rachunek zdarzeń losowych

Niepewność towarzyszącą wynikowi doświadczenia lub obserwacji, np. zaobserwowanej wartości zarobków absolwenta lub latencji L3–N33 u zdrowego osobnika, przypisujemy **losowemu** charakterowi zjawiska. Ilościowe i ścisłe ujęcie owej losowości prowadzi do rachunku prawdopodobieństwa i w konsekwencji do budowy modeli probabilistycznych. (W dalszym ciągu, badając losowość zjawiska, będziemy zawsze mówić o doświadczeniach losowych, także wtedy, gdy faktycznie będziemy mieć na myśli zjawiska, których jesteśmy wyłącznie biernymi obserwatorami).

**Definicja 2.1.** *Doświadczenie nazywamy losowym, jeżeli może być powtarzane w (zasadniczo) tych samych warunkach, jego wynik nie może być przewidziany w sposób pewny oraz zbiór wszystkich możliwych wyników jest znany i może być opisany przed przeprowadzeniem doświadczenia.*

**Przykład 2.1.** Załóżmy, że interesuje nas czas jaki upłynął od momentu zainstalowania do momentu pierwszego uszkodzenia monitora pewnej firmy, pracującego w sieci komputerowej. Pojedyncze doświadczenie polega na obserwacji takiego monitora i zanotowaniu chwili jego pierwszego uszkodzenia. Powtarzanie doświadczenia to obserwacja większej liczby monitorów interesującej nas firmy. Uznajemy, że doświadczenia prowadzimy w *zasadniczo* takich samych warunkach, jeśli możemy nie zwracać uwagi na to, że nasze doświadczenia prowadzimy w nieco odmiennych okolicznościach. Mianowicie, obserwowane monitory mogą pracować w różnych pokojach, a zatem w trochę różnych temperaturach i nieco innej wilgotności powietrza. Niemniej jednak, warunki te uznajemy za wystarczająco podobne do siebie, aby pominąć wpływ istniejących różnic na sprawowanie się monitorów. Oczywiście, niemożliwe jest przewidzenie z góry, ile czasu dany monitor będzie pracować bez żadnej usterki. Niemożliwe jest zatem przewidzenie wyniku konkretnego doświadczenia. Wiemy natomiast, że czas pracy do

pierwszego uszkodzenia nie może być mniejszy od zera i na pewno jest skończony, choć nie potrafimy tego czasu ograniczyć od góry (nie możemy orzec zawczasu, że będzie to czas nie większy niż, powiedzmy, 1000 dni). W rezultacie, zakładając dla wygody, że czas liczymy z absolutną dokładnością, za przedział możliwych czasów bezawaryjnej pracy monitorów przyjmujemy przedział $[0, \infty)$. Przyjąwszy wszystkie te postulaty możemy nasze doświadczenie uznać za losowe.

Pozostając przy przykładach komputerowych, za doświadczenie losowe możemy uznać czas eksploatacji dysku twardego w komputerze danego typu, albo liczbę błędów w aplikacji (programie) zadanego typu. W pierwszym przypadku zbiorem możliwych wyników doświadczenia jest znowu przedział $[0, \infty)$, w drugim zbiór nieujemnych liczb całkowitych, $\{0, 1, 2, \ldots\}$. Obydwa przypadki są przedmiotem wielkiego zainteresowania nie tylko użytkowników komputerów, ale i firm oferujących sprzęt oraz oprogramowanie. Doświadczeniem losowym jest także liczba błędnie odczytanych bitów w paczce informacji o pojemności 1 megabitu, przesłanej przez sieć komputerową. W tym przypadku zbiorem możliwych wyników doświadczenia jest zbiór $\{0, 1, 2, \ldots, 10^6\}$. Przykłady ważnych z praktycznego punktu widzenia doświadczeń losowych można by mnożyć niemal w nieskończoność.

**Definicja 2.2.** *Zbiór wszystkich możliwych wyników doświadczenia losowego nazywamy* ***przestrzenią zdarzeń elementarnych*** *lub* ***przestrzenią próbkową*** *i oznaczamy symbolem $\mathcal{S}$. Pojedynczy element przestrzeni zdarzeń elementarnych, czyli pojedynczy wynik doświadczenia losowego, nazywamy* ***zdarzeniem elementarnym****. Dowolny podzbiór przestrzeni zdarzeń elementarnych nazywamy* ***zdarzeniem*** *(zdarzenia oznaczamy zwykle dużymi, początkowymi literami alfabetu, A, B, C itd).*

Przestrzeń zdarzeń elementarnych może być skończona lub nieskończona. W tym drugim przypadku może być przeliczalna, czyli jej elementy mogą być ponumerowane, lub może być nieprzeliczalna. Przestrzeń zdarzeń elementarnych z przykł. 2.1 jest nieprzeliczlna, $\mathcal{S} = [0, \infty)$. Gdy wynikiem pojedynczego doświadczenia jest liczba błędów w aplikacji zadanego typu, wygodnie jest przyjąć, że przestrzeń zdarzeń elementarnych jest nieskończona, ale przeliczalna, $\mathcal{S} = \{0, 1, \ldots\}$. Rzecz w tym, że liczba potencjalnych błędów może być dowolnie duża. Gdy z kolei doświadczeniem losowym jest liczba błędnie przesłanych bitów w paczce informacji o pojemności 1 megabitu, przestrzeń zdarzeń elementarnych jest skończona, $\mathcal{S} = \{0, 1, 2, \ldots, 10^6\}$. Zdarzenia elementarne w tych przykładach, to odpowiednio: konkretny, dokładny czas od instalacji do pierwszego uszkodzenia monitora, konkretna liczba błędów w aplikacji oraz konkretna liczba błędnie przesłanych bitów.

Jakkolwiek w powyższych przykładach zdarzenia elementarne są liczbami, nie ma to sugerować, że tak jest zawsze. W rzeczywistości natura zdarzeń elementarnych nie jest niczym ograniczona. Jeżeli doświadczenie losowe polega na stwierdzeniu czy badane lekarstwo leczy, czy też nie, przestrzeń zdarzeń elementarnych składa się z dwóch elementów: jednym zdarzeniem elementarnym jest wynik „leczy", drugim „nie leczy". Jeżeli interesuje nas jedna z trzech możliwości: czy lekarstwo „leczy i nie ma niepożądanych skutków ubocznych", czy „nie leczy i nie ma niepożądanych skutków ubocznych", czy też „daje niepożądane skutki uboczne", to na przestrzeń zdarzeń elementarnych składają się te trzy możliwe wyniki zaaplikowania leku. W doświadczeniu, w którym interesuje nas czy osobnik ma ciśnienie krwi „w normie", „poniżej normy" czy „powyżej normy",

$$\mathcal{S} = \{\text{ciśnienie w normie, ciśnienie poniżej normy, ciśnienie powyżej normy}\}.$$

W doświadczeniu, w którym podobnie jak w przykł. 1.1 pytamy o wyznanie mieszkańca Warszawy i rozróżniamy katolików, ewangelików, prawosławnych, żydów oraz inne wyznania,

$$\mathcal{S} = \{\text{katolik, ewangelik, prawosławny, żyd, inne wyznania}\}.$$

Niekiedy nawet proste doświadczenie ma zaskakująco bogatą przestrzeń zdarzeń elementarnych.

**Przykład 2.2.** Doświadczenie polega na rzucaniu monetą do wyrzucenia po raz pierwszy orła. Wówczas, oznaczając wynik kolejnego rzutu literą R, jeśli wyrzucono reszkę, oraz literą O, jeśli wyrzucono orła,

$$\mathcal{S} = \{\text{O, RO, RRO, RRRO}, \ldots\},$$

czyli jest to przestrzeń o nieskończenie wielu, chociaż przeliczalnie wielu, elementach.

Podkreślmy, że postać zdarzeń elementarnych i co za tym idzie postać przestrzeni tych zdarzeń wynika z konretnej postaci doświadczenia losowego, będącego przedmiotem naszego zainteresowania. Na przykład jednokrotny rzut monetą prowadzi do zupełnie innej przestrzeni zdarzeń elementarnych, choć wyrzucenie orła i reszki możemy oznaczyć tak samo jak poprzednio, czyli odpowiednio przez O oraz R. W tym bowiem przypadku mamy tylko dwa możliwe zdarzenia elementarne i

$$\mathcal{S} = \{\text{O}, \text{R}\}.$$

W podanym dalej przykł. 2.3 mamy znowu do czynienia z rzutami monetą, ale doświadczenie jest inne, a więc także przestrzeń zdarzeń elementarnych jest także inna.

Często nie interesuje nas żadne konkretne zdarzenie elementarne, a ich pewien zbiór, czyli pewne zdarzenie. Na przykład, możemy chcieć poznać szansę na zajście zdarzenia polegającego na tym, że czas pracy monitora do pierwszego uszkodzenia jest dłuższy niż sześć lat. Mamy zatem na myśli zdarzenie $A = (6, \infty)$, gdzie za jednostkę czasu przyjęliśmy 1 rok (oczywiście, $A \subset \mathcal{S}$). Jeżeli monitor ma gwarancję opiewającą na 3 lata, może nas interesować zdarzenie $B = [0, 3)$ lub zdarzenie $C = A \cup B$, czyli zdarzenie polegające na zajściu pierwszego uszkodzenia monitora albo w czasie objętym gwarancją, albo dopiero po sześciu latach jego niezawodnej pracy.

W tym miejscu należy jasno sobie uzmysłowić konwencję, jaką zawsze stosujemy mówiąc o zachodzeniu zdarzeń. **Mówimy, iż zaszło zdarzenie $D$, gdy wynik doświadczenia losowego należy do zbioru zdarzeń elementarnych** $D$ (tzn. gdy wynik doświadczenia losowego, będący zdarzeniem elementarnym, jest dowolnym elementem zbioru $D$, $D \subset \mathcal{S}$). Jeżeli np. monitor pracował bezawaryjnie przez 2,4 roku od chwili zakupu, to znaczy, iż zaszło zdarzenie $B = [0, 3)$.

Z definicji 2.2 oraz powyższych przykładów jasno wynika, że zdarzenia są utożsamiane ze zbiorami. Możemy zatem w naturalny sposób wprowadzić rachunek zdarzeń jako tożsamy z rachunkiem zbiorów. Przypomnijmy, że przez sumę zbiorów $D$ i $E$, oznaczaną $D \cup E$, rozumiemy zbiór, którego elementami są wszystkie elementy zbioru $D$ i wszystkie elementy zbioru $E$, i który nie zawiera innych elementów. Przez iloczyn zbiorów $D$ i $E$, oznaczany $D \cap E$, rozumiemy zbiór zawierający te i tylko te elementy, które należą zarówno do zbioru $D$, jak i do zbioru $E$. Różnicą zbiorów $D$ i $E$, oznaczaną $D - E$, jest zbiór złożony z tych i tylko tych elementów, które należą do $D$ i nie należą do $E$.

**Definicja 2.3.** ***Dopełnieniem zdarzenia*** *$A$, oznaczanym $A'$, nazywamy zdarzenie równe zbiorowi $\mathcal{S} - A$,*

$$A' = \mathcal{S} - A.$$

***Iloczynem zdarzeń*** *$A$ i $B$ nazywamy zdarzenie równe iloczynowi zbiorów $A$ i $B$, czyli równe zbiorowi*

$$A \cap B.$$

***Sumą zdarzeń*** *$A$ i $B$, nazywamy zdarzenie równe sumie zbiorów $A$ i $B$, czyli zdarzenie równe zbiorowi*

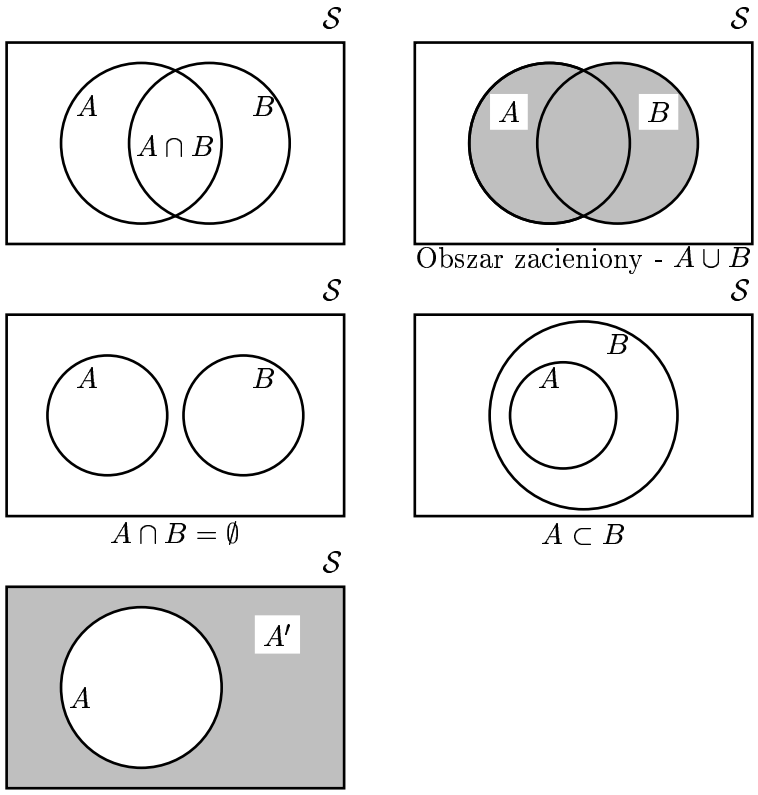
$$A \cup B.$$

*Zdarzenia $A$ i $B$* ***wzajemnie się wykluczają****, gdy ich iloczyn jest zbiorem pustym, czyli gdy ich iloczyn jest zdarzeniem niemożliwym,*

$$A \cap B = \emptyset,$$

*gdzie $\emptyset$ jest zbiorem pustym. Zdarzenie $A$* ***zawiera się w zdarzeniu*** *$B$, jeżeli z zachodzenia zdarzenia $A$ wynika zachodzenie zdarzenia $B$, czyli gdy*

$$A \subset B.$$




Rys. 2.1. Diagramy Venna ilustrujące def. 2.3

Zamiast $A \subset B$ pisze się niekiedy $A \subseteq B$, zaznaczając w ten sposób, że $A$ jest podzbiorem zbioru $B$ także wtedy, gdy $A = B$. Dobrą ilustrację powyższych definicji stanowią diagramy Venna (patrz rys. 2.1).

**Przykład 2.3.** Niech doświadczenie losowe polega na dwukrotnym rzucie monetą i odnotowaniu dwóch kolejnych wyników. Wówczas

$$\mathcal{S} = \{\text{OO},\ \text{OR},\ \text{RO},\ \text{RR}\},$$

gdzie OO oznacza wyrzucenie dwóch orłów, OR oznacza wyrzucenie najpierw orła i następnie reszki, RO najpierw reszki i następnie orła oraz RR wyrzucenie dwóch reszek. Niech

$$A = \{\text{orzeł w pierwszym rzucie}\} = \{\text{OO},\ \text{OR}\},$$

$$B = \{\text{reszka w drugim rzucie}\} = \{\text{RR},\ \text{OR}\},$$

$$C = \{\text{orzeł w drugim rzucie}\} = \{\text{OO},\ \text{RO}\}$$

oraz niech

$$D = \{\text{orzeł w obydwu rzutach}\} = \{\text{OO}\}.$$

Jeżeli rzucamy monetą dwa razy i otrzymujemy wynik OO, to zdarzenie $A$ zaszło; podobnie powiedzielibyśmy, że zdarzenie $A$ zaszło, gdybyśmy otrzymali wynik OR. Łatwo zauważyć, że

$$A \cap B = \{\text{OR}\},$$

$$A \cup B = \{\text{OO},\ \text{OR},\ \text{RR}\},$$

$$A \cap C = \{\text{OO}\},$$

$$A \cup C = \{\text{OO},\ \text{OR},\ \text{RO}\},$$

$$A \cup B \cup C = \mathcal{S},$$

$$B \cap C = \emptyset,$$

$$D \subset A,\ \ D \subset C.$$

**Przykład 2.4.** W dziekanacie wydziału informatyki technicznej są zainstalowane dwa komputery PC. Interesuje nas czas korzystania z każdego z komputerów w ciągu jednego, ośmiogodzinnego dnia roboczego. Przyjmijmy, że czas mierzymy z dokładnością do jednej minuty. Każde zdarzenie elementarne ma postać:

pierwszy komputer pracował $m$ minut, drugi komputer pracował $n$ minut,

gdzie $n \in \{0, 1, 2, \ldots, 480\}$ oraz $m \in \{0, 1, 2, \ldots, 480\}$. W języku algebry zbiorów możemy napisać

$$\mathcal{S} = \{(m, n):\ m \in \{0, 1, 2, \ldots, 480\}, n \in \{0, 1, 2, \ldots, 480\}\},$$

co oznacza, że $\mathcal{S}$ jest zbiorem wszystkich takich uporządkowanych par liczb $m, n$, że $m, n \in \{0, 1, 2, \ldots, 480\}$; np. para liczb $(430, 285)$ oznacza, że pierwszy komputer pracował (lub będzie pracować) danego dnia

430 minut, a drugi 285 minut. Niech $A$ oznacza zdarzenie polegające na tym, że pierwszy komputer będzie danego dnia pracować przynajmniej 300 minut. Zauważmy, iż takie zdarzenie oznacza, że drugi komputer może pracować dowolną, nam obojętną liczbę minut. Opis zdarzenia $A$ wyraźnie przedstawiający przedział czasu pracy komputera pierwszego oraz przedział pracy drugiego komputera, powinien zatem mieć postać

$$A = \{\text{czas pracy komputera I} \geq 300, \text{czas pracy komputera II dowolny}\};$$

w języku algebry zbiorów

$$A = \{(m, n) : m \in \{300, 301, \ldots, 480\}, n \in \{0, 1, 2, \ldots, 480\}\}.$$

Niech $B$ oznacza zdarzenie polegające na tym, że pierwszy komputer będzie pracować co najwyżej 400 minut, czyli że

$$B = \{\text{czas pracy komputera I} \leq 400, \text{czas pracy komputera II dowolny}\};$$

inaczej

$$B = \{(m, n) : m \in \{0, 1, \ldots, 400\}, n \in \{0, 1, 2, \ldots, 480\}\}.$$

Przecięcie zdarzeń $A$ i $B$ polega na pracy pierwszego komputera przez przynajmniej 300 minut, ale nie więcej niż 400 minut,

$$A \cap B = \{(m, n) : m \in \{300, 301, \ldots, 400\}, n \in \{0, 1, \ldots, 480\}\}.$$

Zdarzenie łączne $A \cup B$ polega na dowolnie długim czasie pracy zarówno pierwszego, jak i drugiego komputera. W przypadku drugiego komputera jest to oczywiste, w przypadku pierwszego wynika stąd, że

$$\{300, 301, \ldots, 480\} \cup \{0, 1, \ldots, 400\} = \{0, 1, \ldots, 480\}.$$

A zatem

$$A \cup B = \mathcal{S}.$$

Punkt ten zakończymy uogólnieniem definicji pary zdarzeń wzajemnie się wykluczających.

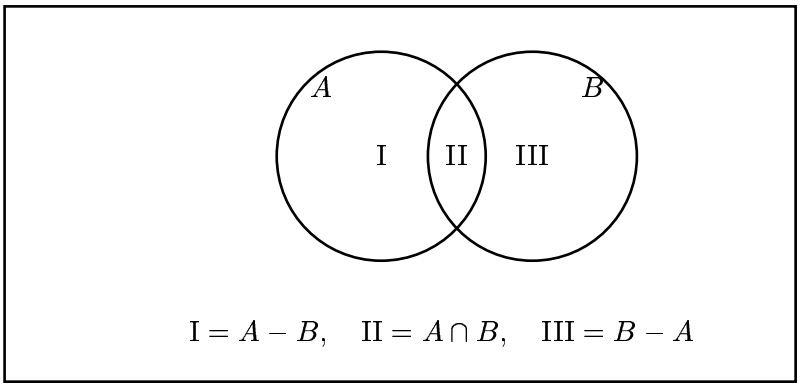
**Definicja 2.4.** *Mówimy, że* ***zdarzenia*** $A_1, A_2, A_3, \ldots$ ***wzajemnie się wykluczają****, jeżeli żadne dwa z nich nie mają wspólnych elementów, czyli jeżeli dla każdej pary różnych wskaźników* $i, j$, $i \neq j$,

$$A_i \cap A_j = \emptyset.$$

Powyższa definicja odnosi się w ogólności do nieskończonej, przeliczalnej rodziny zdarzeń, $A_1, A_2, \ldots$ Definicja ta obejmuje przypadek skończonej rodziny $n$ zdarzeń $A_1, A_2, \ldots, A_n$, tyle że w tej sytuacji żaden ze wskaźników $i, j$ nie może być większy niż $n$.

Zauważmy, że sumę zdarzeń $A$ i $B$, $A \cup B$, możemy przedstawić jako sumę zdarzeń wzajemnie się wykluczających (por. rys. 2.2)

$$A \cup B = (A \cap B) \cup (A - B) \cup (B - A). \tag{2.1}$$




Rys. 2.2. Diagram Venna ilustrujący równość $A \cup B = (A \cap B) \cup$ $\cup (A - B) \cup (B - A)$

### 2.1.2. Prawdopodobieństwo

W praktyce interesują nas nie same zdarzenia, ale przede wszystkim szanse ich zajścia. Model probabilistyczny musi zawierać w sobie możliwość mierzenia takich szans, inaczej mierzenia prawdopodobieństwa zajścia każdego zdarzenia. Jednym z wygodnych sposobów przypisywania zdarzeniom prawdopodobieństw jest oparcie się na częstościowej (inaczej częstotliwościowej) interpretacji pojęcia prawdopodobieństwa.

Chcąc omówić to podejście, skupimy się na najprostszym możliwym doświadczeniu losowym, a mianowicie na doświadczeniu jednokrotnego rzutu monetą. W tym przypadku $\mathcal{S} = \{\mathrm{O}, \mathrm{R}\}$, pytamy zaś np. o prawdopodobieństwo wyrzucenia orła. Jeżeli moneta jest „uczciwa", każdy odpowie, że prawdopodobieństwo to wynosi 1/2. Dlaczego?

Dobrego, choć nie jedynego, wyjaśnienia naszego przekonania dostarcza odwołanie się do wspomnianego już podejścia częstościowego. Wyobraźmy sobie, że monetę podrzucamy wiele, powiedzmy $N$, razy. Następnie obliczamy częstość z jaką otrzymywaliśmy orła, $n/N$, gdzie $n$ jest liczbą orłów

uzyskanych w $N$ rzutach. Gdybyśmy liczbie eksperymentów $N$ pozwolili dążyć do nieskończoności, gdybyśmy zatem doświadczenie powtarzali coraz większą liczbę razy, zauważylibyśmy, że częstość $n/N$ stabilizuje się wraz ze wzrostem $N$ i wydaje się dążyć do wartości $1/2$. Niektórzy badacze cierpliwie obserwowali to zjawisko. Na przełomie wieku XIX i XX Karl Pearson (1857-1936) przeprowadził opisany eksperyment 24000 razy ($N = 24000$) i otrzymał $n/N = 0,5005$. W drugiej połowie XX wieku użyto komputerów do symulacji powtórzeń naszego doświadczenia, przyjmując astronomicznie wielkie wartości $N$ i zawsze otrzymując $n/N$ praktycznie równe $1/2$.

Podejście częstościowe sprawia tylko jeden kłopot. W dalszym ciągu książki dowiemy się, że częstość $n/N$ jest rzeczywiście zbieżna do jakiejś wartości granicznej, ale nigdy nie będziemy w stanie wyznaczyć tej wartości doświadczalnie. Na to potrzeba by nieskończenie długiego czasu, ponieważ doświadczenie należałoby powtórzyć nieskończenie wiele razy. Niemniej jednak, wszyscy postępujemy rozsądnie, zawierzając posiadanym przybliżeniom wartości granicznej i, co może ważniejsze, zawierzając własnej intuicji. Na tej samej zasadzie, w rzucie „uczciwą" sześciościenną kostką, otrzymaniu dowolnego ustalonego wyniku przypisujemy prawdopodobieństwo $1/6$. I na tej samej zasadzie, czyli zasadzie *mądrego domniemania częstości zachodzenia dowolnego ustalonego zdarzenia*, przypisujemy temu zdarzeniu odpowiednie prawdopodobieństwo jego zajścia.

Podejście takie narzuca następujące naturalne aksjomaty, jakie spełniać musi prawdopodobieństwo (jak nieraz mówimy, miara prawdopodobieństwa).

**Definicja 2.5.** *Prawdopodobieństwo zdarzenia, oznaczane dla zdarzenia $A$ symbolem $P(A)$, jest liczbą rzeczywistą o następujących własnościach:*

1. *Dla każdego zdarzenia $A$, $A \subset \mathcal{S}$,*

$$0 \leq P(A) \leq 1.$$

2. *Dla przestrzeni zdarzeń elementarnych $\mathcal{S}$, stanowiącej zdarzenie pewne, oraz dla zbioru pustego $\emptyset$, stanowiącego zdarzenie niemożliwe, mamy*

$$P(\mathcal{S}) = 1, \quad P(\emptyset) = 0.$$

3. *Jeżeli zdarzenia $A_1, A_2, A_3, \ldots$ wzajemnie się wykluczają, to*

$$P(A_1 \cup A_2 \cup A_3 \cup \ldots) = P(A_1) + P(A_2) + P(A_3) + \ldots.$$

Sens wszystkich trzech aksjomatów, jakie na mocy powyższej definicji musi spełniać miara prawdopodobieństwa, jest jasny i zgodny z intuicją. Dwa pierwsze aksjomaty orzekają, że prawdopodobieństwo zdarzenia pewnego $\mathcal{S}$ jest równe 1, zdarzenia niemożliwego $\emptyset$ jest równe 0, natomiast prawdopodobieństwa wszystkich innych zdarzeń należą do przedziału $[0,1]$. Prawdopodobieństwo jest zatem miarą znormalizowaną (jest nieujemne i nie może być większe niż 1). Przestrzeń zdarzeń elementarnych jest rzeczywiście zdarzeniem pewnym, zajście bowiem tego zdarzenia oznacza zajście któregokolwiek zdarzenia elementarnego w sytuacji, gdy jakieś zdarzenie elementarne musi zajść. To, że $P(\emptyset) = 0$ nie wymaga komentarza. Aksjomat 3 orzeka, że prawdopodobieństwo sumy przeliczalnie, w ogólności nieskończenie, wielu zdarzeń wzajemnie się wykluczających jest równe sumie prawdopodobieństw tych zdarzeń. W przypadku, gdy mamy do czynienia ze skończoną rodziną $n$ zdarzeń wzajemnie się wykluczających $A_1, A_2, \ldots, A_n$, aksjomat 3 przyjmuje postać

$$P(A_1 \cup A_2 \cup \ldots \cup A_n) = P(A_1) + P(A_2) + \ldots + P(A_n). \qquad (2.2)$$

Sens tego aksjomatu wystarczy wyjaśnić w najprostszym przypadku dwóch zdarzeń wykluczających się, $A_1$ i $A_2$. Zdarzenie $A_1$ zachodzi wtedy tylko, gdy nie zachodzi zdarzenie $A_2$ i odwrotnie, $A_2$ zachodzi wtedy tylko, gdy nie zachodzi $A_1$. A zatem częstość zachodzenia któregokolwiek z tych zdarzeń ($A_1$ lub $A_2$) jest sumą częstości zachodzenia zdarzenia $A_1$ oraz częstości zachodzenia zdarzenia $A_2$, skąd prawdopodobieństwo zajścia któregokolwiek z nich powinno być równe sumie $P(A_1) + P(A_2)$.

Aksjomaty miary prawdopodobieństwa są podstawą do obliczania szans zajścia konkretnych zdarzeń. Fakty z rachunku prawdopodobieństwa i wnioskowania statystycznego, które nie zasługują na miano twierdzeń, ale które będziemy chcieli specjalnie wyodrębnić, będziemy niekiedy nazywać stwierdzeniami.

**Stwierdzenie 2.1.** *Niech wyniki doświadczenia losowego będą jednakowo prawdopodobne i niech możliwych wyników tego doświadczenia będzie $M$. Jeżeli zdarzenie $A$ składa się z $m$ elementów (czyli $m$ zdarzeń elementarnych), to*

$$P(A) = \frac{m}{M}.$$

Prawdziwość tego stwierdzenie wynika wprost z własności (2.1). Rzeczywiście, zdarzenia elementarne oczywiście wzajemnie się wykluczają i jeżeli są jednakowo prawdopodobne, to prawdopodobieństwo każdego z nich musi być równe $1/M$ (ponieważ prawdopodobieństwo sumy wszystkich $M$ zdarzeń elementarnych wynosi 1 i jest równe sumie $M$ takich samych prawdopodobieństw). Z tych samych powodów prawdopodobieństwo sumy $m$ takich

zdarzeń jest równe $m/M$. Na przykład, jeżeli doświadczenie polega na jednokrotnym rzucie „uczciwą" kostką o sześciu bokach oznaczonych liczbami $1, 2, \ldots, 6$, to prawdopodobieństwo wypadnięcia każdej z tych liczb wynosi $1/6$ i prawdopodobieństwo uzyskania w jednym rzucie wyniku parzystego jest równe $1/2$ (zdarzenie uzyskania wyniku parzystego składa się z $m = 3$ elementów, natomiast $M = 6$). Prostym uogólnieniem stwierdzenia 2.1 jest stwierdzenie następujące:

**Stwierdzenie 2.2.** *Niech przestrzeń zdarzeń elementarnych składa się z $M$ elementów, $\mathcal{S} = \{s_1, s_2, \ldots, s_M\}$, i niech prawdopodobieństwo zajścia zdarzenia elementarnego $s_i$ wynosi $P(s_i)$, $i = 1, 2, \ldots, M$. Wówczas, dla każdego zdarzenia $A \subset \mathcal{S}$,*

$$P(A) = \sum_{s_i \in A} P(s_i).$$

Rozumowanie dowodzące prawdziwości tego stwierdzenia przebiega podobnie jak poprzednio. Jedyna różnica polega na tym, że tym razem mogą być dodawane prawdopodobieństwa o różnych wartościach. Stwierdzenie 2.2 można na mocy własności 3 def. 2.5 uogólnić na przypadek przestrzeni zdarzeń elementarnych o nieskończonej, ale przeliczalnej liczbie elementów, $\mathcal{S} = \{s_1, s_2, s_3, \ldots\}$.

**Przykład 2.2 cd.** Jak wykażemy w następnym punkcie, naturalne jest przyjęcie, iż prawdopodobieństwo otrzymania po raz pierwszy orła w $i$-tym rzucie wynosi $1/2^i$, czyli

$$P(\{\mathrm{O}\}) = 1/2, \quad P(\{\mathrm{RO}\}) = 1/4, \quad P(\{\mathrm{RRO}\}) = 1/8$$

itd. Z podanego uogólnienia stwierdzenia 2.2 wynika na przykład, że prawdopodobieństwo otrzymania pierwszego orła nie później niż w czwartym rzucie jest równe

$$P(\{\mathrm{O}\} \cup \{\mathrm{RO}\} \cup \{\mathrm{RRO}\} \cup \{\mathrm{RRRO}\}) = 1/2 + 1/4 + 1/8 + 1/16 = 15/16.$$

**Twierdzenie 2.1.** *Niech $A$ będzie zdarzeniem, $A \subset \mathcal{S}$. Wówczas prawdopodobieństwo dopełnienia $A'$ zdarzenia $A$ jest równe*

$$P(A') = 1 - P(A),$$

*gdzie $P(A)$ jest prawdopodobieństwem zdarzenia $A$.*

Fakt ten jest zgodny z intuicją i wynika bezpośrednio z def. 2.5. Wystarczy w tym celu zauważyć, że (por. rys. 2.1)

$$A \cup A' = \mathcal{S},$$

obydwa zdarzenia się wykluczają oraz $P(\mathcal{S}) = 1$. Rzeczywiście, mamy stąd

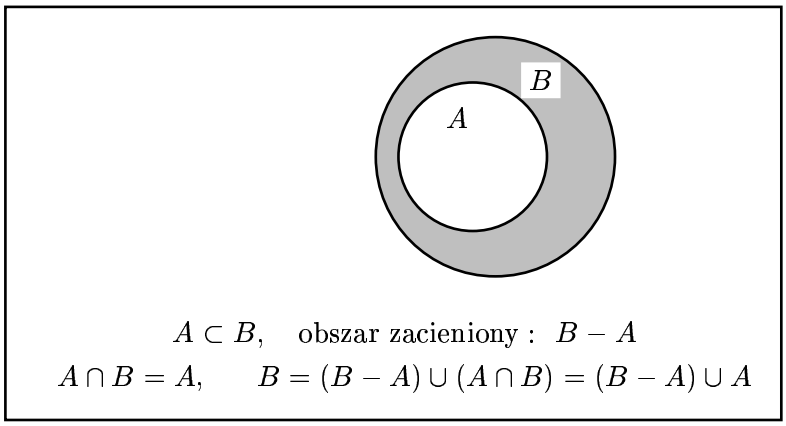
$$P(A \cup A') = P(A) + P(A') = 1,$$

co kończy dowód.

**Twierdzenie 2.2.** *Niech $A$ i $B$ będą zdarzeniami, $A \subset \mathcal{S}$ oraz $B \subset \mathcal{S}$, i niech $A \subset B$. Wówczas*

$$P(A) \leq P(B),$$

*gdzie $P(A)$ jest prawdopodobieństwem zdarzenia $A$ i $P(B)$ jest prawdopodobieństwem zdarzenia $B$.*




Rys. 2.3. Diagram Venna ilustrujący równość $B = A \cup (B - A)$

Fakt ten znowu jest zgodny z intuicją i wynika bezpośrednio z def. 2.5. Na mocy założeń (por. rys. 2.3)

$$B = A \cup (B - A).$$

Zdarzenia $A$ i $B - A$ są rozłączne,

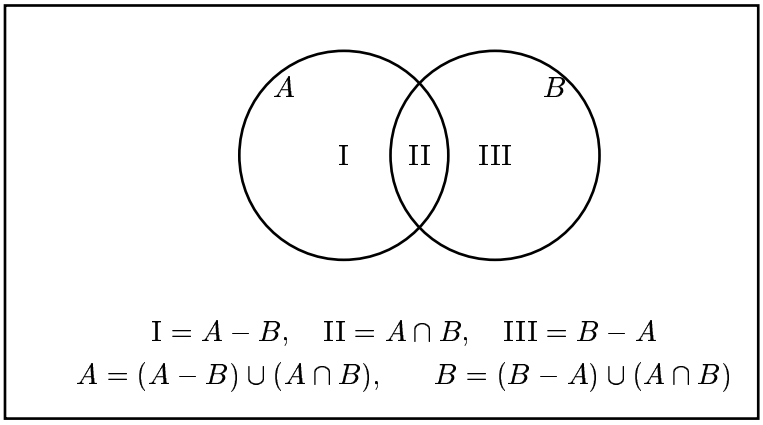
$$P(B) = P(A) + P(B - A)$$

i, ponieważ prawdopodobieństwo nie może być ujemne, twierdzenie jest udowodnione.

**Twierdzenie 2.3.** *Niech $A$ i $B$ będą zdarzeniami, $A \subset \mathcal{S}$ oraz $B \subset \mathcal{S}$. Wówczas*

$$P(A \cup B) = P(A) + P(B) - P(A \cap B),$$

*gdzie $P(A)$ jest prawdopodobieństwem zdarzenia $A$, $P(B)$ jest prawdopodobieństwem zdarzenia $B$ i $P(A \cap B)$ jest prawdopodobieństwem zdarzenia $A \cap B$.*




Rys. 2.4. Diagram Venna ilustrujący równość $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Tym razem spójrzmy przede wszystkim na diagram Venna, ilustrujący postać zdarzenia łącznego (patrz rys. 2.4). Widać natychmiast, że suma $P(A)+P(B)$ dwa razy uwzględnia prawdopodobieństwo przecięcia zdarzeń $A$ i $B$, $P(A \cap B)$. Z diagramu Venna wynika zatem przejrzyste uzasadnienie tezy tw. 2.3. Przy tym, $P(A \cup B) = P(A) + P(B)$ wtedy tylko, gdy $P(A \cap B) = 0$, w szczególności gdy zdarzenia $A$ i $B$ są rozłączne, czyli gdy $A \cap B = \emptyset$ (w zgodzie z własnością (2.2) dla $n = 2$). Twierdzenie 2.3 możemy ściśle udowodnić jeszcze raz, korzystając z trzeciej części def. 2.5. Mianowicie, pamiętając o równości (2.1) możemy napisać

$$P(A \cup B) = P(A \cap B) + P(A - B) + P(B - A). \tag{2.3}$$

Dalej, jeśli przedstawimy zdarzenie $A$ jako sumę zdarzeń rozłącznych $A \cap B$ i $A \cap B'$,

$$A = A \cap \mathcal{S} = A \cap (B \cup B') = (A \cap B) \cup (A \cap B'),$$

oraz zauważymy, że

$$A - B = A \cap B',$$

to otrzymamy

$$P(A) = P(A \cap B) + P(A - B),$$

skąd

$$P(A - B) = P(A) - P(A \cap B). \tag{2.4}$$

Podobnie otrzymujemy, że

$$P(B - A) = P(B) - P(A \cap B). \tag{2.5}$$

Równości (2.3)–(2.5) dowodzą tw. 2.3.

**Przykład 2.3 cd.** Rozsądnie jest założyć, że zdarzenia elementarne są jednakowo prawdopodobne. Stąd (por. komentarz do stwierdzenia 2.1)

$$P(\{\mathrm{OO}\}) = P(\{\mathrm{OR}\}) = P(\{\mathrm{RO}\}) = P(\{\mathrm{RR}\}) = 1/4.$$

Z twierdzenia 2.3 i stwierdzenia 2.1 wynika zatem, że

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 1/2 + 1/2 - 1/4 = 3/4.$$

Oczywiście ten sam wynik można uzyskać bezpośrednio ze stwierdzenia 2.1, korzystając z tego, że

$$A \cup B = \{\mathrm{OO},\ \mathrm{OR},\ \mathrm{RR}\}.$$

Kończąc ten punkt, wróćmy jeszcze do obliczania prawdopodobieństw na podstawie stwierdzenia 2.1. Zadanie sprowadza się do wyznaczenia liczby $M$ wszystkich zdarzeń elementarnych oraz liczby $m$ „sprzyjających" zdarzeń elementarnych, czyli tych które składają się na interesujące nas zdarzenie $A$. Podanie obydwu liczb nie zawsze jest trywialne.

Zanim zajmiemy się prawdopodobieństwami zdarzeń, rozważymy na ile możliwych sposobów mogą się zakończyć pewne doświadczenia losowe. Wyobraźmy sobie losowanie pierwszej szóstki spośród 30 sprinterów. Wylosowani sprinterzy wezmą udział w pierwszym biegu eliminacyjnym (pozostali biegacze wezmą udział w losowaniu drugiej szóstki, i potem następnych, ale kolejnymi losowaniami nie będziemy się zajmować). Kolejność, w jakiej zostanie wylosowanych sześciu biegaczy, będzie odpowiadać numerom torów, na których pobiegną (pierwszy wylosowany pobiegnie na torze pierwszym itd.).

Podany schemat losowania można opisać w sposób ogólny. Interesuje nas pytanie na ile sposobów można wylosować po kolei $k$ różnych obiektów, gdy obiekty są losowane bez zwracania spośród $n$ różnych obiektów ($k \leq n$) i gdy istotna jest kolejność, w jakiej obiekty zostają wylosowane. Losowanie bez zwracania oznacza, że obiekt wylosowany nie zostaje zwrócony do puli, z której obiekty są losowane (losowanie ze zwracaniem oznacza losowanie, w którym obiekty wylosowane są zwracane do puli – w takim przypadku

losuje się obiekty ciągle z tej samej puli o liczności $n$, i możliwe jest wielokrotne wylosowanie tego samego obiektu). Łatwo zauważyć, że pierwszy obiekt może być wybrany na $n$ sposobów – rzeczywiście, wylosowany może być każdy z $n$ elementów w puli. Po pierwszym losowaniu w puli pozostaje $n-1$ obiektów, z czego wynika, że drugi obiekt może być wylosowany na $n-1$ sposobów. Kontynuując to rozumowanie otrzymujemy, że $k$-ty obiekt może być wylosowany na $n-k+1$ sposobów. Ostatecznie zatem liczba szukanych sposobów wynosi

$$n(n-1)(n-2)\cdots(n-k+1) = \frac{n!}{(n-k)!},$$

gdzie $m! = 1 \times 2 \times \cdots \times (m-1) \times m$ (umawiamy się, że $0! = 1$). W szczególności, przyjmując $k = n$, otrzymujemy, że $n$ obiektów można uporządkować na $n! = n \times (n-1) \times \cdots \times 2 \times 1$ sposobów; mówimy, że można utworzyć $n!$ **permutacji** $n$ obiektów.

Rozpatrzmy następnie pytanie na ile sposobów można wylosować $k$ obiektów z $n$ obiektów, gdy losowanie odbywa się bez zwracania, ale nie interesuje nas kolejność, w jakiej obiekty zostały wylosowane. Liczba ta jest nazywana liczbą **kombinacji** $k$ obiektów spośród $n$ obiektów. Szukaną liczbę, oznaczmy ją $C(n,k)$, najłatwiej znaleźć zauważywszy, że

$$\frac{n!}{(n-k)!} = C(n,k) \times (k!).$$

Rzeczywiście, elementy każdej konkretnej kombinacji można uporządkować na $k!$ różnych sposobów – innymi słowy, jedna kombinacja daje $k!$ ciągów tych samych obiektów, różniących się tylko kolejnością ich ustawienia. Ostatecznie zatem

$$C(n,k) = \frac{n!}{k!(n-k)!}.$$

Otrzymaną liczbę kombinacji oznacza się symbolem $\binom{n}{k}$,

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}.$$

Zgodnie z wcześniejszą umową $0! = 1$, skąd $\binom{n}{0} = 1$.

Otrzymane wzory należą do dziedziny nazywanej kombinatoryką. Wyposażeni w nie możemy obliczać prawdopodobieństwa zdarzeń bardziej złożonych niż te, których prawdopodobieństwa obliczaliśmy do tej pory.

Fakt zajęcia się w dwóch poniższych przykładach grą hazardową nie wynika ani z naszego nadmiernego zainteresownia hazardem, ani stąd, że od takich zastosowań zaczęła się historia rachunku prawdopodobieństwa. Takie gry, a wśród nich

również rzuty monetą czy kostką, dostarczają szczególnie jasnych schematów zachowań losowych. Choć zatem nie są szczególnie praktyczne, wydają się dobrze służyć potrzebom dydaktyki.

**Przykład 2.5.** W pewnym typie gry w pokera, znanym z westernów, z pełnej talii 52 kart ciągnie się 5 kart. Liczba możliwych zestawów pięciu kart wynosi

$$\binom{52}{5} = \frac{52!}{5!47!} = 2598960.$$

Wyciągnięcie każdego z 2598960 możliwych zestawów pięciu kart jest jednakowo prawdopodobne. A zatem, korzystając ze stwierdzenia 2.1, prawdopodobieństwo wyciągnięcia dowolnego ustalonego zestawu obliczamy następująco:

$$P(\text{ustalony zestaw 5 kart}) =$$

$$= \frac{\text{liczba sposobów uzyskania ustalonego zestawu}}{\text{liczba wszystkich możliwych zestawów}}.$$

Załóżmy, że chcemy obliczyć prawdopodobieństwo wyciągnięcia zestawu pięciu pików. Mamy

$$\text{liczba sposobów uzyskania zestawu pięciu pików} = \binom{13}{5}$$

i ostatecznie

$$P(\text{zestaw 5 pików}) = \frac{\binom{13}{5}}{2598960} = \frac{(9)(10)(11)(12)(13)}{(5!)(2598960)} = 0{,}0000495.$$

Znalezienie prawdopodobieństwa wyciągnięcia czterech kart tej samej wysokości (np. czterech króli) i dowolnej piątej jest trochę trudniejsze. Ponieważ w talii są karty trzynastu różnych wysokości, mamy $\binom{13}{1}$, czyli 13 sposobów wybrania wysokości czterech kart. Dalej, mamy $\binom{4}{4}$, czyli jeden sposób wyciągnięcia czterech kart ustalonej wysokości. Wreszcie, mamy $\binom{48}{1}$, czyli 48 sposobów wyciągnięcia piątej karty. Jak z tego wynika, mamy (13)(1)(48) sposobów wyciągnięcia czterech kart tej samej wysokości i jednej karty dowolnej. Zatem

$$P(\text{4 karty tej samej wysokości}) = \frac{(13)(48)}{2598960} = 0,00024.$$

Podobnie możemy obliczyć prawdopodobieństwo wyciągnięcia dwóch par w pięciu kartach:

$$P(\text{dwie pary}) = \frac{\binom{13}{2}\binom{4}{2}\binom{4}{2}\binom{44}{1}}{\binom{52}{5}} = \\ = \frac{(78)(6)(6)(44)}{2598960} = 0,0475.$$

W tym przypadku najpierw obliczamy na ile sposobów możemy wybrać dwie różne wysokości kart, następnie na ile sposobów możemy wybrać każdą parę spośród czterech kart tej samej wysokości i w końcu na ile sposobów możemy wybrać piątą kartę, która musi być innej wysokości niż wysokości obydwu par.

**Przykład 2.6.** W urnie znajduje się 8 czerwonych kul oraz 4 białe kule. Nie widząc kul, z urny wyciągamy po kolei dwie kule. Jakie jest prawdopodobieństwo wyciągnięcia dwóch czerwonych kul (kolejność, w jakiej wyciągnięte zostają kule, jest obojętna)? Liczba możliwych kombinacji dwóch kul spośród dwunastu kul wynosi $\binom{12}{2}$. Ponieważ mamy 8 kul czerwonych, liczba możliwych kombinacji dwóch kul czerwonych spośród wszystkich kul tego koloru jest równa $\binom{8}{2}$. Ostatecznie zatem interesujące nas prawdopodobieństwo wynosi

$$\binom{8}{2}/\binom{12}{2} = \frac{(7)(8)}{(11)(12)} = \frac{14}{33}.$$

A jakie jest prawdopodobieństwo wyciągnięcia jednej kuli białej i jednej czerwonej (kolejność wyciągnięcia znowu jest obojętna)? Rozumowanie podobne do przeprowadzonych wcześniej daje

$$P(\text{jedna kula biała i jedna czerwona}) = \frac{\binom{4}{1}\binom{8}{1}}{\binom{12}{2}} = \frac{(4)(8)(2)}{(11)(12)} = \frac{16}{33}.$$

Czytelnik powinien zauważyć, że w przypadku ciągnienia dwóch kul czerwonych błędem byłoby napisać w liczniku wyrażenia na odpowiednie prawdopodobieństwo $\binom{8}{1}\binom{7}{1}$ czyli (8)(7), zamiast $\binom{8}{2}$; ponieważ czerwone kule nie różnią się między sobą i nie interesuje nas kolejność ich losowania (obliczanie prawdopodobieństwa nie zmieniłoby się, gdybyśmy losowo wybierali dwie kule jednocześnie), mamy do czynienia z kombinacją dwóch kul spośród ośmiu. Ale również, jeżeli chcemy wyraźnie zaznaczyć, że kule są wyciągane po kolei, tyle że nie interesuje nas kolejność,

w jakiej kule zostały wyciągnięte, zamiast $\binom{8}{2}$ możemy napisać (1/2)(8)(7). Taki zapis odpowiadałby właściwie następującej sytuacji: kule czerwone są odróżnialne, np. mają na sobie różne litery, $a, b, \ldots, h$, wyciągamy najpierw jedną z ośmiu kul, potem jedną z siedmiu (co możemy uczynić na (8)(7) różnych sposobów), a następnie utożsamiamy pary kul, które różnią się tylko kolejnością wyciągnięcia danych dwóch liter (np. utożsamiamy pary $(a, b)$ i $(b, a)$, w rezultacie czego zamiast 56 sposobów wyciągnięcia pary kul otrzymujemy 56/2 sposoby).

### 2.1.3. Prawdopodobieństwo warunkowe i zdarzenia niezależne

Często interesuje nas prawdopodobieństwo zdarzenia, powiedzmy zdarzenia $B$, gdy wiemy, że zaszło pewne inne zdarzenie $A$. Mówimy wówczas o prawdopodobieństwie zdarzenia $B$ pod warunkiem, że zaszło zdarzenie $A$. Prawdopodobieństwo takie oznaczamy symbolem $P(B|A)$.

Nim podamy definicję prawdopodobieństwa warunkowego $P(B|A)$ rozważymy postulaty, jakie prawdopodobieństwo to powinno spełniać. Po pierwsze, prawdopodobieństwo warunkowe zdarzenia $A$ pod warunkiem zajścia zdarzenia $A$ powinno być równe 1:

$$P(A|A) = 1;$$

można powiedzieć, że skoro zaszło zdarzenie $A$, to zbiór możliwych zdarzeń elementarnych został zredukowany z oryginalnego zbioru $\mathcal{S}$ do zbioru $A$. Po drugie, ponieważ $A$ możemy uznać za nową przestrzeń zdarzeń elementarnych, spośród zdarzeń elementarnych wchodzących w skład zdarzenia $B$ mogą zajść tylko te, które zarazem należą do zbioru $A$. Stąd powinien być spełniony postulat

$$P(B|A) = P(A \cap B|A).$$

Powyższe postulaty uzasadniają następującą definicję:

**Definicja 2.6.** *Niech $A$ i $B$ będą dowolnymi zdarzeniami, $A \subset \mathcal{S}$ i $B \subset \mathcal{S}$, przy czym prawdopodobieństwo zdarzenia $A$ jest dodatnie, $P(A) > 0$. Prawdopodobieństwo warunkowe zdarzenia $B$ pod warunkiem zajścia zdarzenia $A$ jest dane wzorem*

$$P(B|A) = \frac{P(A \cap B)}{P(A)}.$$

Łatwo zauważyć, że prawdopodobieństwo $P(B|A)$ spełnia obydwa podane postulaty.

**Przykład 2.3 cd.**

$$P(\text{orzeł w drugim rzucie} \mid \text{orzeł w pierwszym rzucie}) = P(C|A) = \\ = \frac{P(A \cap C)}{P(A)} = \frac{1/4}{1/2} = \frac{1}{2}.$$

**Przykład 1.1 cd.** Na podstawie danych z roku 1864 można orzec, że w owym roku prawdopodobieństwo bycia katolikiem w Warszawie wynosiło 0,591, bycia prawosławnym 0,014, bycia ewangelikiem 0,067, bycia żydem 0,326 oraz bycia warszawiakiem innego wyznania 0,002. Niech $A$ oznacza zdarzenie bycia chrześcijaninem, czyli bycia katolikiem lub prawosławnym lub ewangelikiem. Niech $B_1$ będzie zdarzeniem bycia katolikiem, niech $B_2$ oznacza bycie prawosławnym i $B_3$ niech oznacza bycie ewangelikiem. Wówczas

$$A = B_1 \cup B_2 \cup B_3$$

i, ponieważ zdarzenia $B_1, B_2, B_3$ wzajemnie się wykluczają,

$$P(B_1|A) = \frac{0,591}{0,591 + 0,014 + 0,067} = 0,879.$$

**Twierdzenie 2.4.** *Niech $A$ i $B$ będą takimi dowolnymi zdarzeniami, $A \subset \mathcal{S}$ i $B \subset \mathcal{S}$, że $P(A) > 0$ oraz $P(B) > 0$. Wówczas*

$$P(A \cap B) = P(B|A)P(A) = P(A|B)P(B).$$

Aby udowodnić to twierdzenie wystarczy zauważyć, że na mocy def. 2.6 $P(B|A) = P(A \cap B)/P(A)$ oraz $P(A|B) = P(A \cap B)/P(B)$. Jak zobaczymy w dalszym ciągu niniejszego punktu, choć jeszcze nie w poniższym przykładzie, podana tożsamość prowadzi często do radykalnego uproszczenia obliczeń (w przykł. 2.6 trudno o radykalne uproszczenie, bowiem sam przykład jest bardzo prosty).

**Przykład 2.6 cd.** Zadania postawione w przykładzie można rozwiązać nie odwołując się do pojęć kombinatoryki. Niech $C_1$ będzie zdarzeniem polegającym na otrzymaniu w pierwszym ciągnieniu kuli czerwonej.

Z kolei, niech $C_2$ będzie zdarzeniem polegającym na otrzymaniu czerwonej kuli w drugim ciągnieniu. Interesuje nas prawdopodobieństwo

$$P(\text{obie kule czerwone}) = P(C_1 \cap C_2).$$

Bez trudu możemy zauważyć, że

$$P(C_1) = \frac{8}{12} \text{ oraz } P(C_2|C_1) = \frac{7}{11},$$

skąd

$$P(C_1 \cap C_2) = P(C_2|C_1)P(C_1) = \frac{7}{11} \times \frac{2}{3} = \frac{14}{33}.$$

W przypadku, gdy interesuje nas prawdopodobieństwo wylosowania jednej kuli białej i jednej czerwonej, bez względu na kolejność, w jakiej kule zostały wylosowane, zadanie pozostaje koncepcyjnie równie proste jak poprzednio, chociaż trzeba wykonać dwa razy więcej rachunków. Niech dodatkowo $B_1$ będzie zdarzeniem polegającym na otrzymaniu w pierwszym ciągnięniu kuli białej, natomiast $C_2$ zdarzeniem polegającym na otrzymaniu czerwonej kuli w drugim ciągnieniu.

$$\begin{aligned} P(\text{jedna kula biała} \quad & \text{i} \quad \text{jedna czerwona}) = \\ = \quad & P[(B_1 \cap C_2) \cup (B_2 \cap C_1)] = \\ = \quad & P(B_1 \cap C_2) + P(B_2 \cap C_1) = \\ = \quad & P(C_2|B_1)P(B_1) + P(B_2|C_1)P(C_1) = \\ = \quad & \frac{8}{11} \times \frac{1}{3} + \frac{4}{11} \times \frac{2}{3} = \frac{16}{33}. \end{aligned}$$

Uzasadnienie powyższego rachunku pozostawiamy Czytelnikowi.

W przykładzie 2.6 zamiast np. obliczać bezpośrednio $P(C_1 \cap C_2)$, łatwiej jest najpierw obliczyć szansę zajścia zdarzenia $C_1$, równą $P(C_1)$ i następnie prawdopodobieństwo warunkowe $P(C_2|C_1)$. Uproszczenie obliczeń bierze się stąd, iż zajście zdarzenia $C_1$ ogranicza oryginalną przestrzeń zdarzeń elementarnych $\mathcal{S}$ do zbioru zdarzeń $C_1$. W tej nowej przestrzeni można znacznie łatwiej obliczyć prawdopodobieństwo zdarzenia $C_2$, czyli prawdopodobieństwo $P(C_2|C_1)$: nie potrzebujemy obliczać liczby kombinacji, by w nowej przestrzeni policzyć wszystkie zdarzenia elementarne oraz te z nich, które należą do zdarzenia $C_2$.

Z rozważań, które doprowadziły nas do definicji prawdopodobieństwa warunkowego, wynika bardzo ważna własność tego prawdopodobieństwa. Mianowicie, prawdopodobieństwo warunkowe obliczane pod warunkiem zajścia

ustalonego zdarzenia $A$ spełnia wszystkie aksjomaty nakładane na miarę prawdopodobieństwa. Aksjomaty te są dane def. 2.5, z tym że prawdopodobieństwa bezwarunkowe należy zastąpić prawdopodobieństwami warunkowymi oraz należy pamiętać, że funkcję przestrzeni zdarzeń elementarnych pełni tym razem zbiór $A$. Mamy zatem:

1. Dla każdego zdarzenia $B$

$$0 \leq P(B|A) \leq 1;$$

2.

$$P(A|A) = 1, \quad P(\emptyset|A) = 0.$$

3. Jeżeli zdarzenia $B_1, B_2, B_3, \ldots$ wzajemnie się wykluczają, to

$$P(B_1 \cup B_2 \cup B_3 \cup \ldots|A) = P(B_1|A) + P(B_2|A) + P(B_3|A) + \ldots.$$

**Przykład 1.1 cd.** Prawdopodobieństwo bycia katolikiem lub ewangelikiem pod warunkiem bycia chrześcijaninem jest równe

$$\begin{aligned} P(B_1 \cup B_3|A) &= P(B_1|A) + P(B_3|A) = \\ &= 0{,}591/0{,}672 + 0{,}067/0{,}672 = 0{,}979 \end{aligned}$$

Na mocy takiego samego rachunku otrzymujemy oczywiście

$$\begin{aligned} P(B_1 \cup B_2 \cup B_3|A) &= P(B_1|A) + P(B_2|A) + P(B_3|A) \\ &= 0{,}979 + 0{,}014/0{,}672 = 1. \end{aligned}$$

Należy zarazem przestrzec przed potraktowaniem prawdopodobieństw warunkowych względem różnych warunków jako opisujących tę samą miarę prawdopodobieństwa. Obliczając prawdopodobieństwa warunkowe pod różnymi warunkami, prawdopodobieństwa te faktycznie obliczamy względem różnych zbiorów możliwych zdarzeń elementarnych. Takie prawdopodobieństwa nie mają ze sobą nic wspólnego – raz obliczamy je przy założeniu, że zaszło jedno zdarzenie, innym razem przy założeniu, że zaszło inne zdarzenie.

Twierdzenie 2.4 mówi o tym jak prawdopodobieństwo iloczynu dwóch zdarzeń przedstawić za pomocą iloczynu prawdopodobieństwa warunkowego

i prawdopodobieństwa bezwarunkowego. Twierdzenie to można łatwo uogólnić na przypadek prawdopodobieństwa iloczynu dowolnej skończonej liczby zdarzeń. Rzeczywiście dla trzech zdarzeń mamy

$$\begin{aligned} P(A \cap B \cap C) &= P[C \cap (A \cap B)] = \\ &= P(C|A \cap B)P(A \cap B) = \\ &= P(C|A \cap B)P(B|A)P(A). \end{aligned}$$

Podobnie, dla zdarzeń $A_1, A_2, \ldots, A_k$ mamy

$$\begin{aligned} &P(A_1 \cap A_2 \cap \cdots \cap A_k) = \\ &= P(A_k|A_{k-1}, A_{k-2}, \ldots, A_1)P(A_{k-1}|A_{k-2}, \ldots, A_1) \cdots P(A_2|A_1)P(A_1). \end{aligned}$$

Podaną własność będziemy nazywać **regułą wielokrotnego warunkowania**.

**Przykład 2.7.** W pewnej uczelni technicznej prowadzi się badania mające na celu wczesne wykrywanie studentów zagrożonych nieukończeniem studiów. Liczy się, że wnioski z badań umożliwią między innymi zaproponowanie takich zmian toku nauczania, które zwiększą procent słuchaczy kończących studia bez obniżenia wymagań programowych. W przykładzie przedstawimy w uproszczeniu jedynie drobny fragment prowadzonych badań.

Analizie poddano populację słuchaczy, którzy albo już ukończyli studia, albo odpadli po którymś roku nauki. Tych, którzy ukończyli studia, podzielono dodatkowo na absolwentów dobrych i słabszych. Przedmioty nauczania podzielono na dwa bloki: blok przedmiotów ogólnych oraz blok przedmiotów specjalistycznych. Niech $A_1$ i $A_2$ oznaczają następujące zdarzenia:
$A_1$ = {ukończenie pierwszego semestru ze stopniem lepszym niż 3 z bloku przedmiotów ogólnych},
$A_2$ = {ukończenie drugiego semestru ze stopniem lepszym niż 4 z przedmiotów specjalistycznych}.
Niech dalej $B$, $C$ i $D$ oznaczają następujące zdarzenia:

$$\begin{aligned} B &= \{\text{dobre ukończenie studiów}\}, \\ C &= \{\text{słabsze ukończenie studiów}\}, \\ D &= \{\text{nieukończenie studiów}\}. \end{aligned}$$

Z badań wynika między innymi, że 76% studentów miało więcej niż trójkę z bloku przedmiotów ogólnych po pierwszym semestrze oraz, że

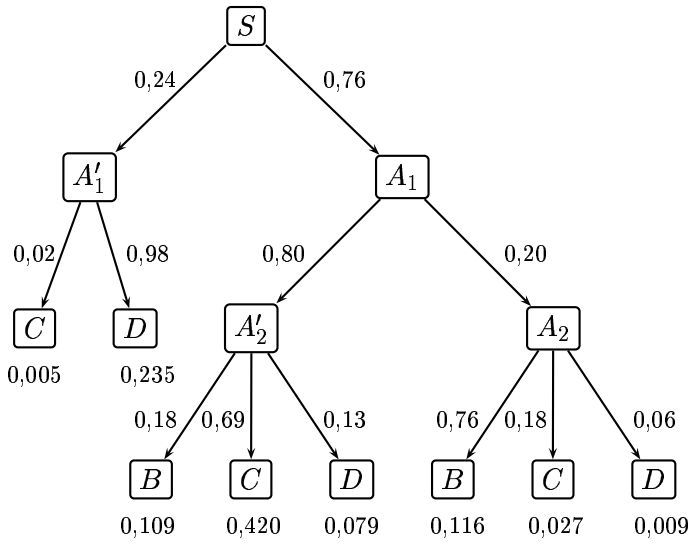
80% studentów w tej grupie miało najwyżej czwórkę z bloku przedmiotów specjalistycznych po drugim semestrze. Z kolei, 69% studentów, którzy zamknęli pierwszy semestr lepiej niż trójką z przedmiotów ogólnych i drugi semestr najwyżej czwórką z przedmiotów specjalistycznych, zakończyło studia z wynikiem słabszym. Mamy zatem

$$P(A_1) = 0{,}76,\ P(A_2'|A_1) = 0{,}80,\ P(C|A_1 \cap A_2') = 0{,}69.$$

Stąd, na mocy reguły wielokrotnego warunkowania, prawdopodobieństwo bycia studentem, który po pierwszym semestrze miał więcej niż trójkę z przedmiotów ogólnych, po drugim semestrze najwyżej czwórkę z przedmiotów specjalistycznych oraz zakończył studia z wynikiem słabszym wynosi

$$P(A_1 \cap A_2' \cap C) = 0{,}76 \times 0{,}80 \times 0{,}69 = 0{,}42.$$

Nie trzeba dodawać, że w ramach badania uzyskano także liczby oraz procenty studentów w innych grupach i tą drogą oszacowano inne interesujące prawdopodobieństwa, jak np. $P(B|A_1 \cap A_2')$, $P(B|A_1 \cap A_2)$ itd.




Rys. 2.5. Drzewo do przykł. 2.7

Przejrzysty sposób przedstawienia takich prawdopodobieństw, ich wzajemnych związków oraz opartych na nich obliczeń dają diagramy zwane **drzewami**. Drzewo odpowiadające danym z przykł. 2.7 jest pokazane na rys. 2.5.

Drzewo rozpoczyna się **korzeniem** umieszczonym na górze rysunku. Punkty $A_1$, $A_2$, $A_1'$, $A_2'$, $B$, $C$ i $D$, nazywane **węzłami** drzewa, odpowiadają zdarzeniom (lub podzbiorom zdarzeń) o tych samych symbolach. Łuki skierowane, łączące węzły, to **gałęzie**. Mówiąc obrazowo, ruch po drzewie rozpoczyna się od korzenia i biegnie jedną z możliwych **ścieżek**, tak jak to umożliwiają łuki skierowane (gałęzie), przez kolejne węzły aż do węzła końcowego. Wyobraźmy sobie, że w korzeniu drzewa jest umieszczona populacja wszystkich studentów uczelni. Każda ścieżka przedstawia możliwą drogę studenta. Z każdą gałęzią jest związane prawdopodobieństwo warunkowe osiągnięcia przez studenta węzła, do którego prowadzi gałąź, pod warunkiem, że student osiągnął wcześniej węzeł, z którego gałąź wychodzi. Zdarzenie $A_1$ jest zdarzeniem, którego zajście nie jest w drzewie uwarunkowane żadnym innym zdarzeniem, i dlatego $P(A_1) = 0{,}76$. Dalej mamy np. $P(A_2|A_1) = 0{,}20$ oraz $P(B|A_1 \cap A_2) = 0{,}76$; rzeczywiście, znalezienie się na omawianej ścieżce w węźle $A_2$ odbywa się pod warunkiem zajścia zdarzenia $A_1$, natomiast znalezienie się w węźle $B$ odbywa się pod warunkiem zajścia zdarzeń $A_1$ i $A_2$. Zauważmy, że reguła wielokrotnego warunkowania mówi, iż prawdopodobieństwo osiągnięcia liścia, czyli prawdopodobieństwo przebycia przez studenta danej ścieżki, jest równe iloczynowi prawdopodobieństw podanych przy ścieżce prowadzącej do tego liścia; np.

$$\begin{aligned} P(A_1 \cap A_2 \cap B) &= P(A_1)P(A_2|A_1)P(B|A_1 \cap A_2) \\ &= 0{,}76 \times 0{,}20 \times 0{,}76 = 0{,}116. \end{aligned}$$

(Prawdopodobieństwa odpowiadające ścieżkom są podane na rys. 2.5 pod liśćmi). Zauważmy też, że suma prawdopodobieństw przejścia wszystkich ścieżek kończących się liściem $B$ daje (bezwarunkowe) prawdopodobieństwo zajścia tego zdarzenia

$$P(B) = 0{,}109 + 0{,}116 = 0{,}225.$$

Podobnie,

$$P(C) = 0{,}452, \quad P(D) = 0{,}323.$$

W przykładzie 2.7 prawdopodobieństwa $P(B)$, $P(C)$ i $P(D)$ można oszacować bezpośrednio z posiadanych danych o studentach. W przykładzie tym interesujące są natomiast prawdopodobieństwa przejścia przez studenta danej ścieżki, np. prawdopodobieństwo $P(A_1 \cap A_2' \cap D)$, które mówi jaki odsetek studentów dobrych z przedmiotów ogólnych nie radzi sobie z przedmiotami specjalistycznymi i odpada ze studiów.

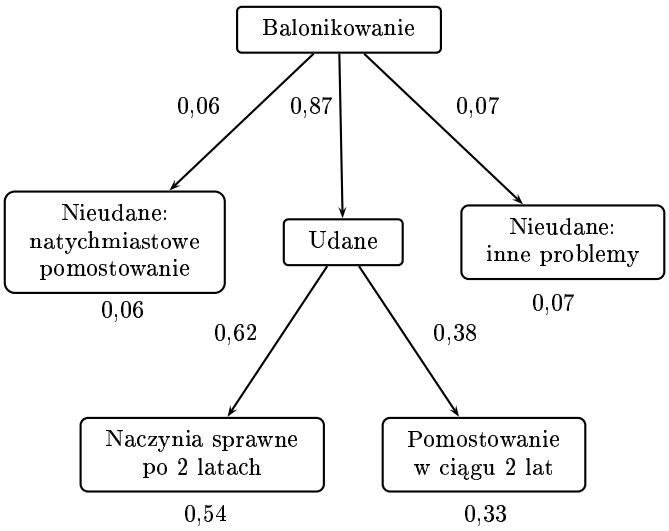
**Przykład 2.8.** W medycynie, w tym w kardiologii, prowadzi się rozliczne badania mające na celu porównanie różnych technik leczenia danej choroby. W przypadku choroby wieńcowej porównuje się na przykład skuteczność leczenia zwężenia naczyń wieńcowych za pomocą tzw.

balonikowania, czyli rozszerzenia naczyń, z operacją pomostowania naczyń (tzw. by-passem). Prowadzone porównanie wymaga dokładnej analizy skutków zastosowania każdej z dwóch technik. Na rysunku 2.6 jest przedstawione drzewo, które w pewnym uproszczeniu opisuje fragment analizy skutków balonikowania (R.A. Kloner, Y. Birnbaum [wyd.] (1999): *Cardiovascular Trials Review*, Le Jacq Communications). Jak wynika z rysunku, szansa zachowania dobrego wyniku balonikowania przez 2 lata wynosi

$$(0{,}87)(0{,}62) = 0{,}54.$$

Wykonanie balonikowania nie zawsze zapobiega operacji pomostowania naczyń wieńcowych. Szansa, że w ciągu 2 lat, albo bezpośrednio po zabiegu balonikowania, albo później będzie potrzebne pomostowanie wynosi

$$0{,}33 + 0{,}06 = 0{,}39.$$




Rys. 2.6. Drzewo do przykł. 2.8

Prawdopodobieństwo warunkowe umożliwia proste zdefiniowanie bardzo ważnej własności, jaką jest niezależność zdarzeń. Intuicyjnie, zdarzenia $A$ i $B$ o dodatnich prawdopodobieństwach ich zajścia są niezależne, jeśli informacja o tym, że jedno zaszło nie ma wpływu na prawdopodobieństwo zajścia drugiego:

$$P(B|A) = P(B) \text{ oraz } P(A|B) = P(A). \tag{2.6}$$

Wobec twierdzenia 2.4 podany postulat jest równoważny następującemu:

$$P(A \cap B) = P(A)P(B)$$

i, tym samym, uzasadnia taką definicję:

**Definicja 2.7.** *Zdarzenia $A$ i $B$ nazywamy* ***niezależnymi*** *wtedy i tylko wtedy, gdy*

$$P(A \cap B) = P(A)P(B).$$

*Zdarzenia $A_1, A_2, \ldots, A_k$ o dodatnich prawdopodobieństwach ich zajścia nazywamy niezależnymi wtedy i tylko wtedy, gdy dla każdego $m$, $2 \leq m \leq k$,*

$$P(A_{i_1} \cap A_{i_2} \cap \ldots \cap A_{i_m}) = P(A_{i_1})P(A_{i_2}) \cdots P(A_{i_m}), \tag{2.7}$$

*dla dowolnych różnych zdarzeń $A_{i_1}, A_{i_2}, \ldots, A_{i_m}$ z rodziny zdarzeń $\{A_{i_1}, A_{i_2}, \ldots, A_{i_k}\}$.*

**Przykład 2.2 cd.** Rozsądnie jest przyjąć, że kolejne rzuty monetą są od siebie niezależne. Prawdopodobieństwo uzyskania orła w wyniku jednego rzutu uczciwą monetą jest równe prawdopodobieństwu uzyskania reszki i wynosi 1/2. Zatem, $P(\mathrm{O}) = 1/2$, $P(\mathrm{RO}) = (1/2)^2$, $P(\mathrm{RRO}) = (1/2)^3$ i ogólnie prawdopodobieństwo uzyskania kolejno $(i-1)$ reszek, a następnie orła jest równe $(1/2)^i$.

Z definicji 2.7 wynika natychmiast, że zdarzenia $A$ i $\mathcal{S}$ są niezależne. To samo dotyczy zdarzeń $A$ i $\emptyset$. Z równości (2.6) wynika, że – zgodnie z intuicją – nie mogą być niezależne wzajemnie się wykluczające zdarzenia o dodatnich prawdopodobieństwach. Rzeczywiście, jeśli zdarzenia $A$ oraz $B$ mają dodatnie prawdopodobieństwa i wzajemnie się wykluczają, to zajście zdarzenia $A$ oznacza, iż nie mogło zajść zdarzenie $B$ i odwrotnie:

$$P(A|B) = 0 \neq P(A) \text{ oraz } P(B|A) = 0 \neq P(B).$$

Z nieco bardziej złożonego rachunku wynika, że jeżeli zdarzenia $A$ i $B$ są niezależne, to także są niezależne pary zdarzeń: $A'$ i $B$, $A$ i $B'$ oraz $A'$ i $B'$ (patrz zad. 2.4).

Na mocy drugiej części def. 2.7, jeżeli zdarzenia $A_1, A_2, \ldots, A_k$ są niezależne, to

$$P(A_1 \cap A_2 \cap \ldots \cap A_k) = P(A_1)P(A_2) \cdots P(A_k). \tag{2.8}$$

Warunek (2.8) jest słabszy od warunku (2.7). Innym szczególnym przypadkiem drugiej części def. 2.7 jest **niezależność parami**, o której mówimy, gdy zamiast warunku (2.7) użyć warunku

$$P(A_{i_1} \cap A_{i_2}) = P(A_{i_1})P(A_{i_2}),$$

spełnionego dla każdej pary różnych zdarzeń $A_{i_1}, A_{i_2}$. W ogólności, z niezależności parami nie wynika niezależność rodziny zdarzeń $A_1, A_2, \ldots, A_k$.

**Przykład 2.9.** Niech przestrzeń zdarzeń elementarnych składa się z czterech jednakowo prawdopodobnych elementów $s_1, s_2, s_3, s_4$. Określmy zdarzenia

$$A_1 = \{s_1, s_2\},\ A_2 = \{s_1, s_3\},\ A_3 = \{s_1, s_4\}$$

oraz zdarzenia

$$B_1 = \{s_1\},\ B_2 = A_1,\ B_3 = \emptyset.$$

Z jednej strony łatwo wykazać, że zdarzenia $A_1, A_2, A_3$ są parami niezależne, ale

$$P(A_1 \cap A_2 \cap A_3) \neq P(A_1)P(A_2)P(A_3).$$

Z drugiej strony

$$P(B_1 \cap B_2 \cap B_3) = P(B_1)P(B_2)P(B_3),$$

ale zdarzenia $B_1, B_2, B_3$ nie są parami niezależne.

**Przykład 2.10.** Utrzymanie łączności telefonicznej wymagało w przeszłości kładzenia wielożyłowych kabli podwodnych. Rozważmy kabel o łącznej długości 3008 km, składający się z odcinków 10-kilometrowych łączonych specjalnymi przekaźnikami wzmacniającymi sygnał. Zakłada się, że z prawdopodobieństwem 0,999 przekaźnik będzie pracować niezawodnie przez 10 lat. Zakłada się też, że uszkodzenia przekaźników są od siebie niezależne. Ponieważ na całej długości kabla zainstalowano 300 przekaźników, prawdopodobieństwo niezawodnej pracy tych wszystkich przekaźników przez 10 lat wynosi $(0,999)^{300}$, czyli tylko 0,74.

**Definicja 2.8.** *Mówimy, że zdarzenia $B_1, B_2, \ldots, B_k$ tworzą* ***podział przestrzeni zdarzeń elementarnych*** $\mathcal{S}$ *(inaczej układ zupełny zdarzeń elementarnych), jeżeli*

$$B_i \cap B_j = \emptyset, \quad gdy\ i \neq j$$

*oraz*

$$B_1 \cup B_2 \cup \ldots \cup B_k = \mathcal{S}.$$

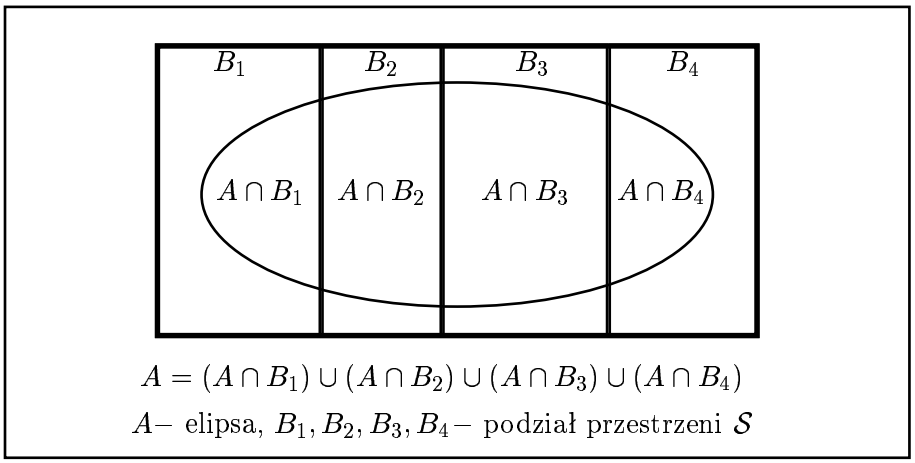
Podział tworzą zatem zdarzenia wzajemnie się wykluczające, których suma jest całą przestrzenią $\mathcal{S}$. Dwa następne twierdzenia, zwłaszcza drugie z nich, mają ogromne znaczenie w praktyce, są bowiem podstawą bardzo skutecznej metody obliczania prawdopodobieństw wielu ważnych zdarzeń. W obydwu korzystamy z pojęcia podziału.

**Twierdzenie 2.5.** (o prawdopodobieństwie całkowitym). *Jeżeli $B_1, B_2, \ldots, B_k$ tworzą podział przestrzeni zdarzeń elementarnych $\mathcal{S}$, $P(B_i) \neq 0$, $i = 1, 2, \ldots, k$, to dla każdego zdarzenia $A$ z $\mathcal{S}$*

$$P(A) = \sum_{i=1}^{k} P(A \cap B_i) = \sum_{i=1}^{k} P(A|B_i)P(B_i).$$

Dowodu wymaga tylko pierwsza z powyższych równości, druga jest bowiem natychmiastową konsekwencją tw. 2.4. Prawdziwość wspomnianej równości najłatwiej zauważyć konstruując odpowiedni diagram Venna (patrz rys. 2.7). Ścisły dowód jest prawie tak samo natychmiastowy, przy czym korzysta z tego, że zdarzenia $(A \cap B_1), (A \cap B_2), \ldots, (A \cap B_k)$ wzajemnie się wykluczają, ponieważ własność tę mają zdarzenia $B_1, B_2, \ldots, B_k$:

$$\begin{aligned} P(A) &= P(A \cap \mathcal{S}) = P(A \cap [B_1 \cup B_2 \cup \ldots \cup B_k]) = \\ &= P((A \cap B_1) \cup (A \cap B_2) \cup \ldots \cup (A \cap B_k)) = \\ &= \sum_{i=1}^{k} P(A \cap B_i). \end{aligned}$$




Rys. 2.7. Diagram Venna ilustrujący tw. 2.5

**Twierdzenie 2.6.** (reguła Bayesa). *Jeżeli $B_1, B_2, \ldots, B_k$ tworzą podział przestrzeni zdarzeń elementarnych $\mathcal{S}$, $P(B_i) \neq 0$, $i = 1, 2, \ldots, k$, to dla każdego takiego zdarzenia $A$ z $\mathcal{S}$, że $P(A) \neq 0$,*

$$P(B_m|A) = \frac{P(A|B_m)P(B_m)}{P(A)} = \frac{P(A|B_m)P(B_m)}{\sum_{j=1}^{k} P(A|B_j)P(B_j)},$$

*gdzie $B_m$ jest dowolnym ustalonym zdarzeniem spośród zdarzeń $B_1, B_2, \ldots$ $\ldots, B_k$.*

Dowód reguły Bayesa jest bardzo prosty. Na podstawie definicji prawdopodobieństwa warunkowego oraz twierdzenia o prawdopodobieństwie całkowitym otrzymujemy

$$P(B_m|A) = \frac{P(B_m \cap A)}{P(A)} = \frac{P(B_m \cap A)}{\sum_{j=1}^{k} P(A|B_j)P(B_j)}.$$

Aby udowodnić regułę Bayesa wystarczy teraz do licznika powyższego wyrażenia odpowiednio zastosować tw. 2.4.

Reguła Bayesa ma nieco tajemniczą postać, ponieważ sprowadza się do „żonglowania" zdarzeniami, względem których obliczamy prawdopodobieństwa warunkowe. Tymczasem takie „żonglowanie" jest chwytem o trudnej do przecenienia skuteczności przy obliczaniu interesujących prawdopodobieństw.

**Przykład 2.11.** Do pudełka włożono trzy normalne monety oraz jedną monetę fałszywą, której zarówno awers, jak i rewers są reszkami. Losowo wyciągamy z pudełka jedną monetę i ją rzucamy. Interesuje nas prawdopodobieństwo, że wybraliśmy fałszywą monetę, jeżeli wynikiem rzutu okazała się reszka. Interesuje nas zatem prawdopodobieństwo warunkowe

$$P(\text{moneta fałszywa}|\text{wypadła reszka}).$$

Tym razem obliczenie zdarzeń elementarnych składających się na zdarzenie {wypadła reszka} jest łatwe. Mianowicie, takich możliwości jest 5 i są to możliwości jednakowo prawdopodobne. Dwa z tych zdarzeń elementarnych pochodzą od monety fałszywej. Pod warunkiem wypadnięcia reszki zdarzenie {wypadła reszka} pełni funkcję nowej przestrzeni zdarzeń elementarnych i szukane prawdopodobieństwo warunkowe daje się obliczyć bezpośrednio,

$$P(\text{moneta fałszywa}|\text{wypadła reszka}) = 2/5.$$

Zauważmy jednak, że prostym zadaniem jest również obliczenie szukanego prawdopodobieństwa na podstawie reguły Bayesa. Rzeczywiście, natychmiast można podać prawdopodobieństwo wypadnięcia

reszki pod warunkiem, że moneta była normalna oraz prawdopodobieństwo wypadnięcia reszki pod warunkiem, że moneta była fałszywa. Niech zatem $A$ będzie zdarzeniem polegającym na wypadnięciu reszki. Niech dalej $B_1$ oznacza zdarzenie wybrania fałszywej monety, $B_2$ zaś oznacza zdarzenie wybrania normalnej monety. Zadaniem jest znalezienie prawdopodobieństwa $P(B_1|A)$. Zauważmy, że $B_2 = B_1'$, skąd

$$B_1 \cup B_2 = \mathcal{S} \text{ oraz } B_1 \cap B_2 = \emptyset.$$

Dalej

$$P(B_1) = 1/4, \quad P(B_2) = 3/4,$$
$$P(A|B_1) = 1 \text{ oraz } P(A|B_2) = 1/2.$$

Stosując regułę Bayesa, otrzymujemy ostatecznie

$$\begin{aligned} P(B_1|A) &= \frac{P(A|B_1)P(B_1)}{P(A|B_1)P(B_1) + P(A|B_2)P(B_2)} = \\ &= \frac{1/4}{1/4 + 1/2 \times 3/4} = \frac{2}{5}. \end{aligned}$$

Z kolejnego przykładu wynika, jak rzeczywiście mocnym narzędziem obliczeniowym jest reguła Bayesa. Jeśli nie znalibyśmy jej, to obliczenie szukanego tam prawdopodobieństwa byłoby zadaniem delikatnie mówiąc nietrywialnym. Dzięki przykładowi 2.12 można ponadto lepiej wniknąć w naturę prawdopodobieństwa warunkowego, w tym właściwie ocenić prawdopodobieństwa związane z rzadkimi zdarzeniami.

**Przykład 2.12.** Rozważmy rzadką chorobę, na którą szansa zapadnięcia wynosi 1/1000 (uważa się, że choroba dotyka 1 na 1000 osób). Test medyczny wykrywa chorobę u osoby chorej z prawdopodobieństwem 0,99 (orzeczeniu przez test istnienia choroby odpowiada uzyskanie dodatniego wyniku testu). W przypadku osoby zdrowej prawdopodobieństwo uzyskania wyniku dodatniego wynosi 0,02; takie prawdopodobieństwo „fałszywego alarmu" jest popularnie uznawane za dopuszczalne. Zapytajmy o prawdopodobieństwo, że osoba, w przypadku której test dał wynik dodatni, jest chora. Niech $A$ będzie zdarzeniem polegającym na uzyskaniu w badaniu testowym wyniku dodatniego. Dalej, niech

$$B_1 = \{\text{osoba chora}\} \text{ oraz } B_2 = \{\text{osoba zdrowa}\}.$$

Interesuje nas zatem prawdopodobieństwo $P(B_1|A)$. Skorzystanie tym razem z reguły Bayesa jest praktycznie koniecznością. Jak w poprzednim przykładzie $B_2 = B_1'$, skąd

$$B_1 \cup B_2 = \mathcal{S} \text{ oraz } B_1 \cap B_2 = \emptyset.$$

Oczywiście,

$$P(B_1) = 0{,}001, \quad P(B_2) = 0{,}999,$$

$$P(A|B_1) = 0{,}99 \text{ oraz } P(A|B_2) = 0{,}02.$$

Stosując regułę Bayesa, otrzymujemy zatem

$$\begin{aligned} P(B_1|A) &= \frac{P(A|B_1)P(B_1)}{P(A|B_1)P(B_1) + P(A|B_2)P(B_2)} = \\ &= \frac{0{,}99 \times 0{,}001}{0{,}99 \times 0{,}001 + 0{,}02 \times 0{,}999} = 0{,}047. \end{aligned}$$

Oczywiście

$$P(B_2|A) = 1 - P(B_1|A) = 0{,}953.$$

Okazuje się więc, że z prawdopodobieństwem 0,953 test daje wynik dodatni w przypadku osób zdrowych! Ten pozornie paradoksalny rezultat jest konsekwencją rzadkiego występowania choroby. Wyżej milcząco założyliśmy, że przed badaniem testowym nie prowadzi się żadnych badań wstępnych, eliminujących z tych badań ogromną większość spośród ludzi zdrowych. Innymi słowy, założyliśmy, że na 1000 osób poddawanych testowi aż 999 jest zdrowych. Gdyby badaniu testowemu poddawano 75 zdrowych i 25 chorych osób na 100, test dobrze spełniałby swoją funkcję (por. zad. 2.10).

Przykład 2.12 dobrze ilustruje znaczenie reguły Bayesa, ponieważ jest przykładem sytuacji, z jaką często spotykamy się w praktyce. Mianowicie, często interesuje nas hipoteza, której prawdziwość mogą sugerować pewne znane fakty, ale wnioskowanie o prawdziwości hipotezy na podstawie tych faktów jest co najmniej nieoczywiste (w przykładzie hipoteza dotyczy choroby osoby, natomiast odnotowywanym faktem jest wynik testu). Zarazem, zarówno pod warunkiem zachodzenia hipotezy, jak i pod warunkiem jej niezachodzenia, łatwo ocenić prawdopodobieństwo zdarzenia się wspomnianych faktów. Reguła Bayesa umożliwia przeprowadzenie wspomnianego wnioskowania.

## 2.2. Zmienne losowe

Przedstawione w poprzednich punktach podstawowe własności prawdopodobieństwa dają nam doskonałą możliwość bardziej ścisłej analizy cech losowych i ich rozkładów, o których mówiliśmy w rozdz. 1. Przyjrzyjmy się następującym przykładom.

**Przykład 2.13.** Rozpatrzmy liczbę trafień w dwóch wykonywanych przez koszykarza rzutach osobistych. Kodując trafienie jako 1, a chybienie jako 0, możemy wszelkie możliwe historie tego eksperymentu, czyli zbiór zdarzeń elementarnych $\mathcal{S}$, zapisać następująco:

$$\mathcal{S} = \{(0,1),\ (1,0),\ (1,1),\ (0,0)\}\,,$$

gdzie np. para $(0,1)$ oznacza 0 (chybienie) w pierwszym rzucie i 1 (trafienie) w drugim rzucie. Oznaczając liczbę trafień przez $X$, mamy

$$X((0,1)) = 1,\ X((1,0)) = 1,\ X((1,1)) = 2,\ X((0,0)) = 0\ .$$

Tak więc wynik eksperymentu polegającego na wykonaniu dwóch rzutów osobistych jest pewną funkcją zadaną na zbiorze zdarzeń elementarnych $\mathcal{S}$.

**Przykład 2.14.** Rozpatrzmy czas spędzony przez klienta w dużym banku w centrum Warszawy. Jeśli przyjmiemy, że bank jest otwarty w godzinach 9–17 (8 godzin, czyli 480 minut), możemy zbiór wszystkich możliwych czasów pobytu w banku zapisać jako $\mathcal{S} = \{x\colon\ 0 \leq x \leq \leq 480\ \text{min}\}$, gdzie $x = 0$ odpowiada sytuacji, gdy klient zrezygnował z załatwienia swojej sprawy natychmiast po przyjściu do banku, a $x = 480$ sytuacji, gdy klient spędził tam cały dzień. Oznaczając przez $X$ czas spędzony w banku przez klienta, możemy stwierdzić, że

$$X(x) = x \quad \text{dla dowolnego } x \in \mathcal{S}\,,$$

a więc interesujący nas czas może być traktowany jako funkcja identycznościowa określona na zbiorze zdarzeń elementarnych $\mathcal{S}$. Oczywiście, w tym przykładzie mogą nas interesować znacznie bardziej skomplikowne funkcje określone na $\mathcal{S}$: rozpatrzmy na przykład wielkość $Y$ określoną jako opłata uiszczona przez klienta za postój jego samochodu na parkingu przed bankiem. Załóżmy, że na parkingu płaci się 60 gr za pierwsze pół godziny postoju, 2 zł za czas pomiędzy 30 a 60 min, 4,60 zł za czas pomiędzy jedną a dwiema godzinami i 3,40 zł za każdą kolejną rozpoczętą godzinę. Wtedy na przykład dla $0 \leq x \leq 30, Y(x) = 0{,}6$ zł, dla $30 < x \leq 60$, $Y(x) = 2$ zł i tak dalej.

**Przykład 2.15.** Dwaj panowie, dawni koledzy ze szkoły, mieszkają na tym samym osiedlu we Wrocławiu i czasami przypadkowo się spotykają. Rozpatrzmy dzień ich pierwszego spotkania po 31 grudnia 1999.

Wówczas wszystkie możliwe momenty ich spotkania można przedstawić następująco:

$$\mathcal{S} = \{1, 2, \ldots\},$$

gdzie 1 oznacza, że panowie spotkali się pierwszego stycznia 2000 roku, 32: pierwszego lutego, 367: pierwszego stycznia 2001 roku itd. Załóżmy, że interesuje nas czas $Y$ (w minutach) ich rozmowy w trakcie pierwszgo spotkania i, że jest on zależny od czasu, który upłynął od początku roku w następujący sposób:

$$Y(s) = \min(10 \times s, 60),$$

gdzie $s$ jest liczbą dni, jaka upłynęła od początku roku do momentu ich spotkania oraz $\min(a, b)$ oznacza mniejszą z liczb $a$ i $b$. Znaczy to, że czas rozmowy wynosi 10 minut razy liczba dni od początku roku, ale nie dłużej niż 60 minut.

Powyższe przykłady prowadzą nas do następującej definicji.

**Definicja 2.9.** *Dowolną funkcję o wartościach rzeczywistych, określoną na zbiorze zdarzeń elementarnych $\mathcal{S}$, nazywamy* ***zmienną losową****.*

Zauważmy, że cecha losowa z przykł. 1.1, przypisująca każdemu mieszkańcowi Warszawy jego wyznanie, nie jest w sensie powyższej definicji zmienną losową, gdyż kategorie wyznaniowe nie są liczbami rzeczywistymi. Stanie się nią, gdy w jakikolwiek sposób przypiszemy liczby kategoriom. Rozpatrywanie cech losowych o wartościach rzeczywistych jest często wygodne, gdyż prowadzi to do łatwych rachunków. Zmienne losowe będziemy oznaczali dużymi literami z końca alfabetu, a ich konkretne watości małymi literami. Tak więc zapis $X = x$ oznacza, że zmienna losowa $X$ przyjmuje wartość $x$.

### 2.2.1. Zmienne dyskretne i ich rozkłady

Ze względu na zbiór wartości przyjmowanych przez zmienną losową wyróżniamy wśród nich klasę tak zwanych zmiennych losowych dyskretnych.

**Definicja 2.10.** *Zmienną losową nazywamy* ***dyskretną****, gdy przyjmuje wartości ze zbioru dyskretnego, to jest takiego, który jest albo skończony, albo przeliczalny, to jest taki, którego elementy można ponumerować kolejnymi liczbami naturalnymi.*

Oczywiście, zmienna losowa $X$ z przykł. 2.13 jest dyskretna, podobnie jak wartość opłaty parkingowej $Y$ rozpatrywana w przykł. 2.14. Zmienna losowa

$X$ z przykł. 2.14 nie jest dyskretna, choć ten intuicyjny fakt nie jest wcale łatwo ściśle uzasadnić. Skonstatujmy tu tylko, że zmienna losowa przyjmująca wszystkie możliwe wartości z odcinka nie może być dyskretna. Do dyskusji takich zmiennych wrócimy.

Zajmijmy się teraz pojęciem **rozkładu prawdopodobieństwa dyskretnej zmiennej losowej** mówiącego jakie wartości i z jakim prawdopodobieństwem są przez zmienną przyjmowane. Koncepcja ta jest bliska definicji rozkładu cechy w próbie, z tą różnicą, że zamiast częstości z jaką wartości dyskretnej zmiennej losowej pojawiły się w próbie interesuje nas teraz prawdopodobieństwo wystąpienia każdej z potencjalnych wartości tej zmiennej. Można powiedzieć, że w tym drugim przypadku interesuje nas rozkład prawdopodobieństwa rządzący możliwymi wynikami przyszłego zdarzenia losowego. W przypadku dyskretnej zmiennej losowej określenie rozkładu sprowadza się do podania następującej funkcji zwanej **funkcją prawdopodobieństwa rozkładu** lub krócej **funkcją prawdopodobieństwa**:

$$p(x) = P(\text{wszystkie } s \in \mathcal{S} \text{ takie, że } \ X(s) = x).$$

Zauważmy, że funkcja prawdopodobieństwa jest określona na zbiorze wszystkich możliwych wartości zmiennej losowej. Rozpatrzmy na przykład zmienną losową z przykł. 2.13 i przyjmijmy, że prawdopodobieństwo trafienia w jednym rzucie osobistym wynosi 0,8 i zdarzenie trafienia lub chybienia w drugim rzucie jest niezależne od analogicznych zdarzeń dla pierwszego rzutu. Wówczas

$$\begin{aligned} p(0) &= P(s \in \mathcal{S} : X(s) = 0) = P((0,0)) = (1 - 0{,}8) \times (1 - 0{,}8) = 0{,}04 \\ p(2) &= P(s \in \mathcal{S} : X(s) = 2) = P((1,1)) = 0{,}8 \times 0{,}8 = 0{,}64 \\ p(1) &= P(s \in \mathcal{S} : X(s) = 1) = P((0,1) \text{ lub } (1,0)) = \\ &= 0{,}2 \times 0{,}8 + 0{,}8 \times 0{,}2 = 0{,}32. \end{aligned}$$

Pamiętając, iż $X$ przyjmuje tylko jedną z trzech możliwych wartości 0, 1 lub 2, na przykład $p(1)$ można było obliczyć inaczej: $p(1) = 1 - p(0) - p(2) = = 1 - 0{,}64 - 0{,}04 = 0{,}32$. Obliczoną funkcję prawdopodobieństwa można przedstawić za pomocą następującej tabeli:

| $x$ | 0 | 1 | 2 |
|---|---|---|---|
| $p(x)$ | 0,04 | 0,32 | 0,64 |

Analogicznie obliczymy funkcję prawdopodobieństwa dla zmiennej $Y$ zdefiniowanej w przykł. 2.14. Załóżmy, że prawdopodobieństwo spędzenia przez klienta w banku

| | |
|---|---|
| od 0 do 30 minut włącznie wynosi | 0,5 |
| od 30 do 60 minut włącznie wynosi | 0,3 |
| od 60 do 120 minut włącznie wynosi | 0,15 |
| od 120 do 180 minut włącznie wynosi | 0,05 |
| powyżej 180 minut wynosi | 0 |

Wówczas wartości zmiennej losowej $Y$ odpowiadają opłatom za czas postoju od 0 do 180 minut. Funkcja prawdopodobieństwa zmiennej losowej $Y$ wynosi

$$p(0,6) = 0{,}5, \quad p(2) = 0{,}3, \quad p(4,6) = 0{,}15, \quad p(8) = 0{,}05\,.$$

Zauważmy, że podobnie jak poprzednio np. $p(0,6)$ można obliczyć alternatywnie jako $1-p(2)-p(4,6)-p(8)$. Pożytecznie jest pamiętać o następującej ogólnej własności funkcji prawdopodobieństwa, której w szczególnym przypadku użyliśmy wyżej.

**Stwierdzenie 2.3.** *Niech $x_1, x_2, \ldots$ oznaczają wszystkie różne wartości dyskretnej zmiennej losowej. Wówczas*

$$\sum_{i=1}^{\infty} p(x_i) = 1.$$

W przypadku, gdy zmienna losowa przyjmuje skończoną liczbę różnych wartości $x_1, \ldots, x_k$, powyższa suma jest interpretowana jako suma $k$ składników, $i = 1, \ldots, k$. Konwencję tę będziemy stosowali w przyszłości. Stwierdzenie wynika z własności 3 prawdopodobieństwa (def. 2.5) i faktu, że zbiory $\{s \in \mathcal{S}\colon p(s) = x_i\}$ stanowią podział zbioru zdarzeń elementarnych. Wprowadźmy jeszcze jedną definicję.

**Definicja 2.11.** *Niech $X$ będzie dowolną zmienną losową, niekoniecznie dyskretną.* ***Dystrybuantą zmiennej losowej*** *$X$ nazywamy funkcję $F$ określoną dla dowolnego rzeczywistego $x$ jako $F(x) = P(X \leq x)$.*

Tak więc dla każdego $x$ dystrybuanta $F(x)$ podaje prawdopodobieństwo zdarzenia, że $X \leq x$. Dla dyskretnej zmiennej losowej wartość $F(x)$ można obliczyć przez zsumowanie (skumulowanie[1]) funkcji prawdopodobieństwa dla wartości nie większych od $x$:

$$F(x) = \sum_{x_i : x_i \leq x} p(x_i), \tag{2.9}$$

gdzie $x_1, x_2, \ldots$ oznaczają wszystkie różne wartości $X$. W sytuacji, gdy dla żadnej wartości $x_i$ nie jest spełniona nierówność $x_i \leq x$, jako wartość $F(x)$

[1]Dlatego funkcja ta jest czasami nazywana **skumulowaną funkcją prawdopodobieństwa**.

przyjmuje się 0. Na podstawie wzoru (2.9) łatwo obliczymy dystrybuantę zmiennej $X$ z przykł. 2.13. Jest ona równa

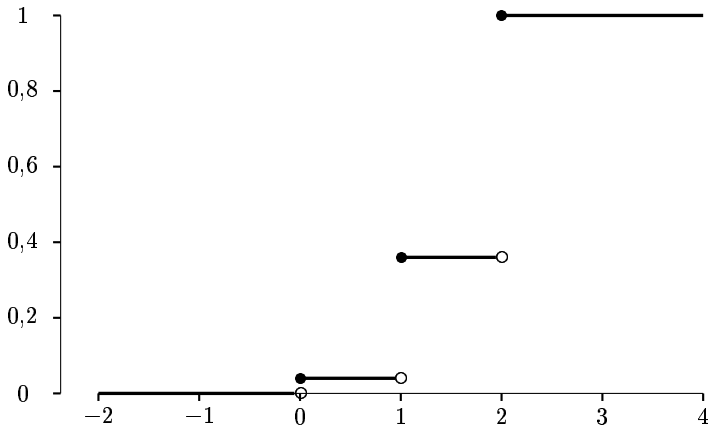
$$F(x) = \begin{cases} 0 & \text{dla } x < 0; \\ 0{,}04 & \text{dla } 0 \leq x < 1; \\ 0{,}36 & \text{dla } 1 \leq x < 2; \\ 1 & \text{dla } x \geq 2 \end{cases}$$

i można ją przedstawić graficznie jak na rys. 2.8. Z przykładu widać następujące ogólne własności dystrybuanty:

(1) dystrybuanta jest funkcją niemalejącą;

(2) wartości $F(x)$ dążą do 1 dla $x$ dążących do $+\infty$ i do 0 dla $x$ dążących do $-\infty$.

(3) dla każdego ustalonego $x_0$, wartości $F(y)$ dążą do wartości $F(x_0)$ dla $y$ dążących do $x$ i takich, że $y > x_0$.




Rys. 2.8. Dystrybuanta zmiennej losowej $X$ z przykł. 2.13

Faktycznie, własność, że $F(\cdot)$ jest funkcją niemalejącą, widać bezpośrednio z rysunku. Oczywiście, $F(\cdot)$ nie musi być funkcją ściśle rosnącą, na przykład dla przedziału $1 < x < 2$, w którym nie ma wartości zmiennej losowej $X$, dystrybuanta jest stała. Dla wartości mniejszych od zera dystrybuanta jest równa 0 i jest równa 1 dla $x \geq 2$, a więc własność (2) jest w sposób oczywisty spełniona. Podkreślmy, że sytuację taką jak w przykładzie, gdy wartości $F(x)$ są równe 1 dla odpowiednio dużych $x$ uważamy za szczególny przypadek sytuacji, gdy wartości $F(x)$ dążą do 1 dla $x$ dążących do $+\infty$. Również z wykresu widać, że na przykład $F(1) = 0{,}36$ jest granicą wartości dystrybuanty dla argumentów na prawo od $x$. Zauważmy jednocześnie, że wartości $F(x)$ dla $0 \leq x < 1$ są równe 0,04 i *nie* dążą do wartości $F(1)$, gdy $x$ dąży do 1.

**Przykład 2.15 cd.** Załóżmy, że spotkania w różne dni są zdarzeniami niezależnymi i, że prawdopodobieństwo spotkania się konkretnego dnia jest stałe i wynosi 0,3. Wówczas zmienna losowa $X(s) = s$ ma następującą funkcję prawdopodobieństwa $p(1) = 0{,}3$, $p(2) = 0{,}7 \times 0{,}3$ i ogólnie

$$p(s) = (0{,}7)^{s-1} \times 0{,}3 \quad \text{dla } s = 1, 2, \ldots,$$

gdyż zdarzenie spotkania się po raz pierwszy $s$-tego dnia oznacza, że w ciagu $s-1$ poprzednich dni do spotkania nie doszło i że odbyło się ono $s$-tego dnia. Tak więc odpowiadające prawdopodobieństwo wynosi $(1-0{,}3) \times \cdots \times (1-0{,}3)$ ($s-1$ razy) pomnożone przez 0,3. Zatem dla $x \geq 1$ wartość dystrybuanty $F(x)$ wynosi

$$F(x) = \sum_{s=1}^{[x]} (0{,}7)^{s-1} \times 0{,}3 = 0{,}3 \sum_{s=0}^{[x]-1} (0{,}7)^s = 1 - (0{,}7)^{[x]},$$

gdzie $[x]$ oznacza część całkowitą $x$, to znaczy największą liczbę całkowitą nie większą od $x$. Tak więc

$$F(x) = \begin{cases} 0 & \text{dla } x < 1; \\ 1 - (0{,}7)^{[x]} & \text{dla } x \geq 1. \end{cases}$$

W tym miejscu warto zastanowić się dlaczego mnożymy byty, wprowadzając nowe definicje, w sytuacji gdy dystrybuanta, przynajmniej dla zmiennej dyskretnej, jest wyznaczona całkowicie za pomocą równości (2.9), czyli za pomocą prawdopodobieństw zdarzeń elementarnych. Zaletą dystrybuanty $F(x)$ jest to, że kumuluje wszystkie wartości funkcji prawdopodobieństwa $p(s)$ dla $s \leq x$. Fakt ten można wykorzystać do obliczania prawdopodobieństwa, że zmienna losowa $X$ przyjmuje wartości z ustalonego przedziału

$$P(a < X \leq b) = P(X \leq b) - P(X \leq a) = F(b) - F(a), \tag{2.10}$$

oraz analogicznie

$$P(a \leq X \leq b) = F(b) - F(a) + p(a), \tag{2.11}$$

$$P(a < X < b) = F(b) - F(a) - p(b). \tag{2.12}$$

Dla przykładu 2.13, $P(0{,}5 < X < 1{,}5) = 0{,}36 - 0{,}04 - 0 = 0{,}32$, zgodnie ze wzorem (2.12), gdyż $p(1{,}5) = 0$. Podkreślmy, że jak widać ze wzorów (2.10), (2.11), (2.12) istotne jest, czy przedział wartości, jaki rozpatrujemy jest obustronnie otwarty czy domknięty. Nie ma to tylko znaczenia w sytuacji,

gdy prawdopodobieństwo przyjęcia wartości równej któremuś z końców jest równe 0, to jest $p(a) = p(b) = 0$. Zauważmy równocześnie, że $p(a)$ może być również obliczone przy znajomości dystrybuanty: jest to wartość skoku lewostronnego dystrybuanty w punkcie $a$. Tak więc dla przykł. 2.13 wartość skoku lewostronnego w punkcie 2 jest równa $1 - 0{,}36 = 0{,}64$ i jest równa $p(2)$.

Na zakończenie tego punktu zasygnalizujmy pewne często spotykane nieporozumienie. Mianowicie, różne zmienne losowe mogą mieć tę samą funkcję prawdopodobieństwa! Tak zdarza się bardzo często przy powtarzaniu obserwacji tego samego zjawiska losowego. Niech na przykład $X_1$ i $X_2$ oznaczają liczbę punktów zdobytych przez koszykarza w rzutach osobistych po pierwszym i drugim faulu na nim. Oczywiście, z reguły wartość $X_1$ nie jest równa wartości $X_2$, natomiast ich funkcje prawdopodobieństwa są takie same (przy założeniu, że skuteczność koszykarza nie zmieniła się między pierwszym a drugim faulem).

### 2.2.2. Wskaźniki położenia i rozproszenia dla dyskretnej zmiennej losowej

Zacznijmy od przykładu.

**Przykład 2.16.** W wyższej szkole prywatnej uczy się 1000 studentów. Badamy zmienną losową $X$ zdefiniowaną jako liczba podręczników przyniesionych na zajęcia przez losowo wybranego studenta w dniu 29 marca 2000 roku (przyjmujemy, że wszyscy studenci byli tego dnia w szkole). Załóżmy, że rozkład liczby podręczników wśród studentów wyglądał następująco:

| Liczba podręczników | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| Liczba studentów | 100 | 300 | 250 | 200 | 100 | 50 |

Oczywiście oznacza to, że prawdopodobieństwo, iż losowo wybrany student nie będzie miał ze sobą żadnej książki wynosi $100/1000 = 0{,}1$, że będzie miał jeden podręcznik wynosi 0,3, dwa podręczniki 0,25 itd. Tak więc funkcja prawdopodobieństwa zmiennej losowej $X$ jest następująca:

| $x$ | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| $p(x)$ | 0,10 | 0,30 | 0,25 | 0,20 | 0,10 | 0,05 |

W celu obliczenia średniej liczby podręczników przypadających na jednego studenta obliczmy całkowitą liczbę podręczników i podzielmy ją przez całkowitą liczbę studentów

$$\frac{0 \times 100 + 1 \times 300 + 2 \times 250 + 3 \times 200 + 4 \times 100 + 5 \times 50}{1000} = 2{,}05.$$

Ponieważ $100/1000 = 0{,}10 = p(0)$, $300/1000 = p(1)$ itd, powyższe wyrażenie można zapisać następująco:

$$0 \times 0{,}1 + 1 \times 0{,}3 + 2 \times 0{,}25 + 3 \times 0{,}2 + 4 \times 0{,}1 + 5 \times 0{,}05 = \sum_{i=0}^{5} i \times p(i)$$

i zinterpretować nieco inaczej. Powyższa suma odpowiada średniej liczbie książek przyniesionych przez losowo wybranego studenta, czyli **średniej wartości zmiennej losowej** $X$. Powyższe rozumowanie jest podstawą następującej definicji.

**Definicja 2.12.** *Dla dyskretnej zmiennej losowej $X$ o funkcji prawdopodobieństwa $p(\cdot)$* ***wartością średnią (oczekiwaną)*** *$X$ nazywamy liczbę*

$$\mu_X = \sum_{i=1}^{\infty} x_i p(x_i), \tag{2.13}$$

*gdzie $x_1, x_2, \ldots$ oznaczają wszystkie różne wartości zmiennej losowej $X$. W przypadku, gdy zmienna losowa przyjmuje skończoną liczbę różnych wartości $x_1, \ldots, x_k$, powyższa suma oznacza sumę $k$ składników, $i = 1, \ldots, k$.*

Często $\mu_X$ jest oznaczane jako $E(X)$; będziemy stosowali obie notacje. Zauważmy, że definicja ta odpowiada dokładnie definicji wartości średniej rozkładu cechy w próbie z częstościami wystąpienia kolejnych wartości zamienionymi na odpowiednie prawdopodobieństwa. Podkreślmy, że wartość średnia nie musi być równa żadnej faktycznej wartości przyjmowanej przez zmienną losową: w przykł. 2.16 $\mu_X = 2{,}05$ podręcznika! Podobnie wygląda definicja **mediany** zmiennej losowej $X$, którą definiujemy jako taki dowolny punkt $q_{0,5}$, że

$$F(x) \leq 0{,}5 \text{ dla } x < q_{0,5} \text{ i } F(x) \geq 0{,}5 \text{ dla } x \geq q_{0,5}.$$

Zauważmy, że tak zdefiniowana mediana zmiennej $X$ w przykł. 2.16 jest równa 2, jednakże może się zdarzyć, że mediana nie jest określona jednoznacznie. Oczywiście, mediana będzie jednoznaczna, jeśli istnieje dokładnie jeden punkt $q_{0,5}$ spełniający $F(q_{0,5}) = 0{,}5$. Gdy pamiętamy o definicji mody

rozkładu cechy w próbie, definicja **mody** zmiennej $X$ nie powinna nas zaskoczyć: jest to dowolne maksimum lokalne $p(\cdot)$, to znaczy taki dowolny punkt $x$, że funkcja prawdopodobieństwa dla wartości bezpośrednio poprzedzającej i następującej po $x$ jest mniejsza niż $p(x)$. Dyskusja dotycząca relacji między różnymi koncepcjami wskaźników położenia rozkładu w próbie pozostaje w mocy dla wskaźników położenia zmiennej losowej. Z jedną wszak różnicą: wartość średnia zmiennej losowej, w odróżnieniu od średniej z próby, może być nieskończona. Rozpatrzmy następujący przykład.

**Przykład 2.17. Paradoks petersburski.** Piotr i Paweł grają w następującą grę: Paweł rzuca monetą do momentu, gdy pojawi się pierwszy orzeł, a Piotr wypłaca mu $X = 1$ zł, gdy orzeł pojawi się w pierwszym rzucie, $X = 2$ zł , gdy pojawi się w drugim rzucie i ogólnie $X = 2^{i-1}$ zł, gdy orzeł pojawi się dopiero w $i$-tym rzucie. Zanim obliczymy wartość średnią $\mu_X$, zauważmy, że można ją interpretować jako sprawiedliwą cenę gry, to znaczy cenę, którą Paweł powinien wypłacić Piotrowi przed przystąpieniem do gry, aby była ona sprawiedliwa. Wszystkie wartości $X$ są równe $2^i$ dla pewnego $i = 0, 1, \ldots$. Zdarzenie $\{X = 2^i\}$ zachodzi, gdy orzeł pojawił się po początkowym ciągu $i - 1$ reszek. Oznacza to, że $p(2^i) = P(X = 2^i) = (1/2)^i$. Zatem wartość średnia $\mu_X$ jest równa

$$\mu_X = \sum_{i=1}^{\infty} 2^{i-1} \frac{1}{2^i} = \sum_{i=1}^{\infty} \frac{1}{2} = \infty,$$

przy czym ostatnia równość oznacza tyle, że suma wartości od pewnego momentu przekroczy dowolnie wielką liczbę. Paradoks ten można przeformułować w następujący sposób: bank gwarantuje nam wypłatę 1 miliona złotych, gdy w ciągu rzutów monetą pojawi się orzeł, pod dwoma warunkami. Po pierwsze, za pierwszy rzut zapłacimy 1 zł, gdy pojawi się w nim reszka, za drugą grę dodatkowo 2 zł, gdy znów pojawi się reszka itd. Po drugie, z gry nie możemy się wycofać: gdy nie mamy pieniędzy na następną rozgrywkę musimy ogłosić bankructwo i oddać wszysko co mamy bankowi. Bogaty sponsor liczący na reklamę swojej osoby zgadza się pożyczyć sumę $Y$ zł potrzebną do prowadzenia gry aż do jej wygrania. Oczywiście, podobnie jak powyżej wartość średnia $\mu_Y = \infty$. Jednakże z powodu prawostronnej skośności rozkładu $Y$ używanie wartości średniej jest tu mocno mylące. Zauważmy, że jeśli gra zakończy się przed dziewiętnastym rzutem, to po oddaniu pożyczonych pieniędzy, zyskamy (gdyż, jak łatwo sprawdzić opłata za 18 gier wynosi $1 + 2 + \cdots + 2^{18} = 524827$, jest zatem mniejsza od miliona i dopiero opłata za 19 gier przekracza milion). Ale zauważmy także, że prawdopodobieństwo, że orzeł nie pojawi się w pierwszych osiemnastu rzutach wynosi $2^{-18} = 0,000004$. Tak więc sponsor niekoniecznie nieroztropnie rozporządza swoimi pieniędzmi!

Na zakończenie dyskusji dotyczącej wskaźników położenia przytoczmy jeszcze jedną pożyteczną własność.

**Twierdzenie 2.7.** (**Wartość średnia** $f(X)$). *Niech $X$ będzie dyskretną zmienną losową o wartościach $x_1, x_2, \ldots$ i funkcji prawdopodobieństwa $p(\cdot)$, a $f$ dowolną funkcją rzeczywistą. Wówczas dyskretna zmienna losowa $f(X)$ ma wartość średnią równą*

$$\mu_{f(X)} = \sum_{i=1}^{\infty} f(x_i)p(x_i). \tag{2.14}$$

Własność ta jest oczywista, gdy funkcja $f(\cdot)$ jest różnowartościowa: wówczas zmienna losowa $f(X)$ przyjmuje wartości $f(x_i)$ z prawdopodobieństwem $p(x_i)$. W ogólnym przypadku jej udowodnienie pozostawiamy jako zadanie. Zauważmy, że dla zmiennej losowej z przykł. 2.16 i funkcji $f(x) = x^2$ mamy $\mu_{X^2} = \sum_{i=0}^{5} i^2 p(i) = 5{,}95$, gdy tymczasem $(\mu_X)^2 = (2,05)^2 = 4{,}2025$. Tak więc, ogólnie nie jest spełniona równość $\mu_{X^2} = (\mu_X)^2$. Co więcej jak dowiemy się dalej, jest ona spełniona bardzo rzadko. Zauważmy jeszcze, że równość (2.14) wyjaśnia następującą intuicyjną własność: wartość średnia zmiennej losowej poddanej zamianie jednostek jest wartością średnią zmiennej losowej obliczonej na oryginalnej skali i potem przeliczonej na nowe jednostki:

$$\mu_{aX+b} = a\mu_X + b, \tag{2.15}$$

gdzie $\mu_{aX+b}$ oznacza wartość średnią zmiennej losowej $Y = aX + b$. Aby uzasadnić równość (2.15) wystarczy rozpatrzeć równość (2.14) dla funkcji $f(x) = ax + b$ i skorzystać z własności $\sum_{i=1}^{\infty} p(x_i) = 1$.

Wprowadźmy teraz definicję wariancji i odchylenia standardowego dyskretnej zmiennej losowej $X$.

**Definicja 2.13.** ***Wariancją*** *dyskretnej zmiennej losowej o funkcji prawdopodobieństwa $p(\cdot)$ nazywamy liczbę*

$$\sigma_X^2 = \sum_{i=1}^{\infty} (x_i - \mu_X)^2 p(x_i). \tag{2.16}$$

***Odchylenie standardowe*** $\sigma_X$ *definiuje się jako* $\sqrt{\sigma_X^2}$.

Często $\sigma_X^2$ oznaczane jest jako $Var(X)$. Będziemy stosowali zamiennie obie notacje. Przeanalizujmy sens tej definicji. Rozpatrujemy w niej kwadrat odchylenia dowolnej wartości $x_i$ od wartości średniej $\mu_X$. Ponieważ $\mu_X$ ma interpretację przeciętnej wartości $X$, wartość $(x_i - \mu_X)^2$ odpowiada kwadratowi odchyłki od tej wartości przeciętnej. Ponieważ wartość $x_i$ pojawia się

z prawdopodobieństwem $p(x_i)$, zgodnie z def. 2.12 wyrażenie (2.16) jest wartością średnią kwadratu odchyłki wartości zmiennej losowej od swojej wartości przeciętnej i może być zapisane jako $\mu_{(X-\mu_X)^2}$. Jak widzimy, definicja wariancji zmiennej losowej $X$ jest podobna do definicji wariancji w próbie, zmieniamy tylko próbkowy wskaźnik położenia $\bar{x}$ na miarę położenia zmiennej losowej $X$, a czynnik $1/(n-1)$ zastępujemy przez prawdopodobieństwo wystąpienia wartości $x_i$ równe $p(x_i)$. Oczywiście, właściwy wskaźnik rozproszenia otrzymamy, rozpatrując zamiast $\sigma_X^2$ odchylenie standardowe $\sigma_X$, ponieważ to ostatnie jest już wyrażone w tych samych jednostkach co wartości zmiennej losowej $X$. Zauważmy, że ponieważ z definicji wynika, iż wariancja jest zawsze liczbą nieujemną, wzięcie pierwiastka ma sens: podkreślmy, że jako odchylenie standardowe rozpatrujemy zawsze pierwiastek *nieujemny* z wariancji. Zauważmy jeszcze, że również z definicji wynika, że wariancja może być równa 0 tylko wtedy, gdy wszystkie wartości zmiennej losowej są równe wartości średniej. Oznacza to, że taka zmienna losowa przyjmuje tylko jedną wartość (a więc jej losowość jest mocno problematyczna!).

W przykładzie 2.16

$$\begin{aligned}\sigma_X^2 = (0-2{,}05)^2 \times 0{,}1 \;&+\; (1-2{,}05)^2 \times 0{,}3 + (2-2{,}05)^2 \times 0{,}25 + \\ &+\; (3-2{,}05)^2 \times 0{,}2 + (4-2{,}05)^2 \times 0{,}1 + \\ &+\; (5-2{,}05)^2 \times 0{,}05 = 1{,}7475.\end{aligned}$$

Analogicznie jak w p. 1.3.2 możemy wprowadzić definicje **odchylenia przeciętnego** $\mu_{|X-\mu_X|}$ i **rozstępu międzykwartylowego** $q_{0,75} - q_{0,25}$ jako konkurencyjnych miar rozproszenia. Ich własności są analogiczne do własności ich próbkowych odpowiedników. Zauważmy tu tylko, że używanie odchylenia standardowego czy przeciętnego zakłada milcząco, że wartość średnia jest właściwą miarą położenia - stosowanie tych miar nie ma więc większego sensu na przykład dla bardzo skośnego rozkładu. Jedną z czysto technicznych zalet wariancji jest możliwość prostego jej obliczenia na podstawie tej własności.

**Stwierdzenie 2.4.** *Dla dyskretnej zmiennej losowej $X$ mamy*

$$\sigma_X^2 = \mu_{X^2} - (\mu_X)^2. \tag{2.17}$$

O prawdziwości równości (2.17) można się przekonać rozpisując definicję wariancji. Mianowicie, ponieważ $(x_i-\mu_X)^2 = x_i^2 - 2\mu_X x_i + \mu_X^2$ i $\sum_{i=1}^{\infty} p(x_i) = 1$, więc otrzymujemy

$$\begin{aligned}\sigma_X^2 &= \sum_{i=1}^{\infty} x_i^2 p(x_i) - 2\mu_X \sum_{i=1}^{\infty} x_i p(x_i) + \mu_X^2 \sum_{i=1}^{\infty} p(x_i) \\ &= \mu_{X^2} - 2(\mu_X)^2 + (\mu_X)^2 = \mu_{X^2} - (\mu_X)^2.\end{aligned}$$

Tak więc wariancję zmiennej losowej $X$ w przykł. 2.16 można było prościej obliczyć jako $5{,}95 - (2{,}05)^2 = 1{,}7475$.

Zauważmy jeszcze, że bezpośrednio z definicji wariancji wynika, że $\sigma^2_{aX} = a^2\sigma^2_X$ oraz $\sigma^2_{X+b} = \sigma^2_X$. Zatem łącząc te dwie własności, otrzymujemy

$$\sigma^2_{aX+b} = a^2\sigma^2_X \tag{2.18}$$

i pamiętając, że odchylenie standardowe jest dodatnim pierwiastkiem z wariancji, mamy

$$\sigma_{aX+b} = |a|\sigma_X.$$

W szczególności $\sigma_{-X} = \sigma_X$, czyli odchylenie standardowe zmiennych losowch $X$ i $-X$ jest takie samo.

Na zakończenie tego punktu zdefiniujmy moment rzędu $k$ i moment centralny rzędu $k$ dyskretnej zmiennej losowej $X$, dla $k = 1, 2, \ldots$. **Moment $m_k$ rzędu** $k$ zmiennej losowej $X$ jest zdefiniowany jako wartość średnia zmiennej losowej $Y = X^k$, a **moment centralny $\mu_k$ rzędu** $k$ jako wartość średnia zmiennej losowej $Z = (X - \mu_X)^k$

$$m_k = \mu_{X^k}, \qquad \mu_k = \mu_{(X-\mu_X)^k}.$$

Zauważmy, że $m_1 = \mu_X$, a $\mu_2 = \sigma^2_X$, zatem wartość średnia jest pierwszym momentem $X$, a wariancja jego drugim momentem centralnym. Podobnie jak w przypadku wartości średniej momenty wyższych rzędów mogą być nieskończone albo mogą nie istnieć.

### 2.2.3. Przykłady rozkładów dyskretnych

**1. Rozkład dwupunktowy.** Jest to najprostszy, poza rozkładem jednopunktowym, z możliwych rozkładów: zmienna losowa ma rozkład dwupunktowy, jeśli przyjmuje tylko dwie różne wartości $x$ i $y$. Wówczas, jeśli oznaczymy prawdopodobieństwo przyjęcia wartości $x$ przez $p$ i $y$ przez $q$, to mamy $q = 1 - p$. Odpowiednią funkcję prawdopodobieństwa możemy przedstawić w postaci diagramu

| wartość przyjmowana | $x$ | $y$ |
|---|---|---|
| prawdopodobieństwo | $p$ | $q$ |

Oczywistym przykładem zmiennej losowej o rozkładzie dwupunktowym jest wynik rzutu „uczciwą" monetą. Przyjmijmy, że zmienna losowa $X$ przyjmuje wartość 1, gdy wypadł orzeł i 0, gdy wypadła reszka. Wówczas $p_X(1)$=0,5 i $p_X(0)$=0,5. Gdyby moneta niekoniecznie była uczciwa i orzeł wypadał

z pewnym prawdopodobieństwem $p$, takim, dla którego $0 < p < 1$, wówczas $p_X(1) = p$ i $p_X(0) = 1 - p$. Zauważmy, że jest to w pewnym sensie jedyny przykład rozkładu dwupunktowego: zastępując orła przez dowolnie zdefiniowany „sukces" (dla którego zmienna losowa przyjmuje wartość $x$), a reszkę przez „porażkę" (zmienna losowa przyjmuje wartość $y$), otrzymujemy ogólne określenie zmiennej losowej o rozkładzie dwupunktowym. Oto przykłady „sukcesów": zdanie egzaminu na prawo jazdy (lub jego oblanie), wygrana (lub przegrana) w karty, przejechanie przez most w Warszawie w godzinach szczytu w czasie krótszym niż kwadrans. Niech

$$X = \begin{cases} 1 & \text{z prawdopodobieństwem } p, \\ 0 & \text{z prawdopodobieństwem } q = 1 - p. \end{cases}$$

Wówczas $X$ ma rozkład dwupunktowy o wartościach 0 i 1 oraz $\mu_X = 0 \times$ $\times q + 1 \times p = p$. W podobny sposób otrzymujemy

$$\sigma_X^2 = (0-p)^2 \times q + (1-p)^2 \times p = p^2 q + q^2 p = pq(p+q) = p \times q.$$

Podany rozkład dwupunktowy o wartościach 0 i 1 nazywamy **rozkładem Bernoulliego** z prawdopodobieństwem sukcesu $p$.

Gdy zamiast rzutu uczciwą monetą rozpatrzymy rzut uczciwą kostką do gry, otrzymamy typowy przykład rozkładu jednostajnego na sześciu punktach. Zmienna ma **rozkład jednostajny na $k$ punktach**, jeśli przyjmuje z jednakowym prawdopodobieństwem każdą z $k$ różnych wartości $x_1, \ldots, x_k$. Oczywiście, w tej sytuacji $p(x_1) = p(x_2) = \ldots = p(x_k) = 1/k$. Dla rzutu kostką możliwe wartości należą do zbioru $\{1, \ldots, 6\}$ i prawdopodobieństwo przyjęcia każdej z nich wynosi $1/6$.

**2. Rozkład dwumianowy.** W przykładzie 2.13 rozpatrzyliśmy zmienną $X$ będącą sumą punktów zdobytych przez koszykarza w dwóch rzutach osobistych. Zauważmy, że wartość $X$ można traktować jako liczbę sukcesów (trafień) w eksperymencie składającym się z dwóch prób, przy czym:

liczba prób (w rozpatrywanym przypadku równa 2) jest z góry ustalona;

każda próba kończy się jednym z dwóch możliwych wyników, które są takie same dla wszystkich prób;

wyniki prób nie zależą od siebie;

prawdopodobieństwo sukcesu $p$ (równe w tym przypadku 0,8) jest takie samo w każdej próbie.

Eksperyment spełniający powyższe warunki, przy dopuszczeniu dowolnej ustalonej liczby prób i dowolnego ustalonego prawdopodobieństwa sukcesu nazywa się eksperymentem lub schematem dwumianowym, a liczbę sukcesów w takim eksperymencie **zmienną losową o rozkładzie dwumianowym** lub **zmienną dwumianową**. Przymiotnik dwumianowy pochodzi

stąd, że w każdej próbie eksperymentu są możliwe tylko dwa wyniki, czyli dwa miana. Rozpatrzmy teraz sytuację, gdy koszykarz wykonuje nie dwa, ale dziesięć rzutów osobistych i rozpatrzmy liczbę zdobytych przez niego punktów. Oznaczamy sukces przez $S$, a porażkę przez $P$. Niech przykładowe zdarzenie elementarne, opisujące historię wykonywania rzutów osobistych wygląda następująco: $s_0 = (SPPSSSPSSP)$. Wówczas liczba trafień $X(s_0)$ jest równa 6. Zauważmy, że

$$\begin{aligned} P(\text{zaszło } s_0) &= 0{,}8 \times 0{,}2 \times 0{,}2 \times \cdots \times 0{,}8 \times 0{,}2 = (0{,}8)^6(0{,}2)^4 = \\ &= (0{,}8)^{X(s_0)}(0{,}2)^{10-X(s_0)} \end{aligned}$$

zależy tylko od liczby sukcesów $X$. Dla wszystkich zdarzeń elementarnych z sześcioma trafieniami prawdopodobieństwo to jest takie samo. Zatem

$$P(X = 6) = (\text{liczba zdarzeń z sześcioma trafieniami}) \times (0{,}8)^6(0{,}2)^{10-6}.$$

Jednocześnie liczba zdarzeń z sześcioma trafieniami jest równa liczbie wyboru sześciu rzutów, w których koszykarz trafił do kosza, z całkowitej liczby dziesięciu rzutów i wynosi $\binom{10}{6}$. Zatem

$$P(X = 6) = \binom{10}{6} \times (0{,}8)^6(0{,}2)^{10-6}$$

i ogólnie

$$P(X = k) = \binom{10}{k} \times (0{,}8)^k(0{,}2)^{10-k} \quad \text{dla } 0 \leq k \leq 10,$$

gdzie ograniczenie na $k$ wynika z faktu, że najmniejsza liczba trafień wynosi 0, a największa 10. Przyjmijmy oznaczenie $b(k; 10, 0{,}8)$ na $P(X = k)$ i ogólnie, jeśli $X$ oznacza liczbę sukcesów w $n$ próbach z prawdopodobieństwem sukcesu $p$, to oznaczmy $P(X = k)$ jako $b(k; n, p)$. Rozumując jak poprzednio otrzymujemy

$$P(X = k) = b(k; n, p) = \binom{n}{k} \times p^k(1-p)^{n-k} \quad \text{dla } 0 \leq k \leq n. \qquad (2.19)$$

Rozkład zadany wzorem (2.19) nazywamy rozkładem dwumianowym $Bin(n, p)$ z parametrami $n$ i $p$, a fakt, że zmienna losowa $X$ ma taki rozkład zapisujemy $X \sim Bin(n, p)$. Przypomnijmy, że pierwszy parametr rozkładu dwumianowego $n$ oznacza liczbę prób, a drugi parametr $p$ oznacza prawdopodobieństwo sukcesu w pojedynczej próbie.

Schemat dwumianowy może wydawać się na pierwszy rzut oka dosyć skomplikowany, lecz sytuacje, w których ma zastosowanie, pojawiają się wszędzie.

Nie bez powodu: w życiu ustawicznie powtarzamy te same czynności i często ich wyniki są losowe, przyjmują jedną z dwóch ustalonych wartości ze stałym prawdopodobieństwem i są niezależne od przeszłości. Oczywiście nie twierdzimy, że warunki te są zawsze dokładnie spełnione. Gdy rozważamy model dwumianowy interesuje nas, czy odstępstwa od nich nie są zbyt duże, ma tu zastosowanie nasza ogólna uwaga z początku rozdz. 2. Rozpatrzmy następujący przykład.

**Przykład 2.18.** Codziennie dojeżdżamy samochodem do pracy i często się do niej spóźniamy z powodu dużego rannego natężenia ruchu. Jeśli prawdopodobieństwo spóźnienia się dowolnego dnia wynosi 0,15, to jakie jest prawdopodobieństwo, że w ciągu trzech tygodni (15 dni roboczych) spóźnimy się nie więcej niż dwa razy?

Jeśli za sukces (swoiście pojmowany) uznamy spóźnienie się do pracy i przyjmiemy, że spóźnienia w różne dni są w przybliżeniu niezależne, to liczba spóźnień $Y$ może być interpretowana jako liczba sukcesów w schemacie dwumianowym z $n = 15$ i $p = 0,15$. Zatem

$$\begin{aligned} P(Y \leq 2) &= P(Y = 0) + P(Y = 1) + P(Y = 2) = \\ &= b(0; 15, 0{,}15) + b(1; 15, 0{,}15) + b(2; 15, 0{,}15) = \\ &= 0{,}0874 + 0{,}2312 + 0{,}2856 = 0{,}6042. \end{aligned}$$

Prawdopodobieństwa dwumianowe można obliczyć ze wzoru (2.19), można je również łatwo obliczyć za pomocą dowolnego pakietu statystycznego lub skorzystać z tablic statystycznych. Podamy teraz wzory na wartość średnią i wariancję zmiennej losowej $X$ o rozkładzie dwumianowym. Jeśli $X \sim$ $\sim Bin(n, p)$, to

$$\mu_X = n \times p, \qquad \sigma_X^2 = n \times p \times (1 - p).$$

Wzory te uzasadnimy na końcu podrozdz. 2.3, zauważmy tu jedynie, że wartość średnia $X$ i jej wariancja są równe $n$ razy odpowiednio wartość średnia lub wariancja sukcesu w jednej próbie.

**Przykład 2.19 (pobieranie próby ze skończonej populacji).** Rozpatrzmy partię towaru składającą się z $N$ jednostek, z których $p \times 100\%$ jest wadliwych. Chcemy przybliżyć nieznaną częstość $p$, pobierając kolejno $n$ elementów z partii i obliczając częstość elementów wadliwych w tak otrzymanej próbie. Załóżmy, że liczność próby $n$ nie przekracza liczby elementów wadliwych $Np$. Nasuwa się pytanie, jaki jest rozkład liczby $X$ pobranych elementów wadliwych. Nie jest to dokładnie

rozkład dwumianowy, bo choć liczba ciągnięć została ustalona na $n$ i każde losowanie kończy się sukcesem (pobranie elementu wadliwego) lub porażką (pobranie elementu sprawnego), wyniki ciągnień **nie** są niezależne, gdyż prawdopodobieństwo wylosowania elementu wadliwego w $k$-tym ciągnieniu zależy od tego, ile elementów wadliwych wylosowano w $k-1$ uprzednich ciągnieniach. Dokładniej, prawdopodobieństwo to wynosi $Np/(N-(k-1))$, gdy żaden z uprzednich elementów nie był wadliwy i $(Np-i)/(N-(k-1))$, gdy $i$ spośród nich było wadliwych (oczywiście $i \leq k-1$). Jednakże, jeśli $k$ (zmieniające się w zakresie od 1 do $n$) jest małe, powyższe prawopodobieństwo jest w przybliżeniu równe $p$ i w rezultacie wyniki poszczególnych ciągnień są w przybliżeniu niezależne. Tak więc w sytuacji, gdy $n$ jest małe w porównaniu z $N$, liczba elementów wadliwych w próbie ma w przybliżeniu rozkład dwumianowy $Bin(n,p)$. Przybliżenie przez rozkład dwumianowy uważa się za zadowalające, jeśli $n/N \leq 0{,}05$, to znaczy $n$ nie przekracza 5% liczności partii. Oczywiście, odrębną sprawą jest jak duża musi być liczność próbki, aby precyzyjnie przybliżyć nieznaną częstość $p$. Problemem tym zajmiemy się w rozdz. 3.

**3. Proces i rozkład Poissona**. Rozkład ten jest związany z sytuacją zliczania zdarzeń losowych określonego typu w pewnym odcinku czasu. Typowe przykłady to obserwacja, od pewnego momentu czasu przyjmowanego jako 0 do momentu $t$, samochodów przejeżdzających autostradą obok punktu obserwacyjnego. W czasie obserwacji rejestrujemy momenty pojawienia się kolejnych samochodów. Takie badania prowadzi się systematycznie w celu określenia intensywności użytkowania dróg. Możemy być również zainteresowani momentami pojawiania się kolejnych klientów w lombardzie lub kasie teatralnej, a czas liczymy od momentu ich otwarcia określonego dnia lub czasami wystąpienia kolejnych erupcji wulkanów lub trzęsień ziemi w określonym punkcie. W takich przypadkach często są spełnione trzy naturalne założenia:

– liczba zdarzeń, które zaszły między momentem $s$ a momentem $w$, czyli w przedziale czasowym $[s,w]$ jest niezależna od liczby zdarzeń, które zaszły przed momentem $s$;

– dla krótkiego odcinka czasowego długości $\Delta$, to jest takiego, że stosunek $\Delta/t$ jest mały, prawdopodobieństwo zajścia w tym czasie dokładnie jednego zdarzenia wynosi w przybliżeniu $\lambda\Delta$, gdzie $\lambda$ jest pewną stałą dodatnią zwaną intensywnością;

– prawdopodobieństwo zajścia co najmniej dwóch zdarzeń w krótkim odcinku czasu długości $\Delta$ jest zaniedbywalne w porównaniu z prawdopodobieństwem zajścia dokładnie jednego zdarzenia.

Okazuje się, że spełnienie tych postulatów określa jednoznacznie rozkład liczby zdarzeń w przedziale $[0,t]$, którą oznaczymy przez $X$. Rozkładem zmiennej losowej $X$ jest tak zwany rozkład Poissona z parametrem $\lambda t$, gdzie $\lambda$ jest intensywnością zdefiniowaną we własności drugiej. Zauważmy, że zmienna $X$ zależy od momentu $t$, dlatego możemy napisać $X = X_t$. Rodzina zmiennych $X_t$ dla $t$ przebiegających pewien przedział $(0,T)$ nosi nazwę **procesu Poissona**.

Mówimy, że zmienna losowa $X$ ma **rozkład Poissona $P(\lambda)$ z parametrem** $\lambda > 0$, jeśli zbiorem jej możliwych wartości jest zbiór wszystkich nieujemnych liczb całkowitych $\{0,1,2,\ldots\}$, a jej funkcja prawdopodobieństwa jest równa

$$p(k;\lambda) = \frac{e^{-\lambda}\lambda^k}{k!} \quad \text{dla } k = 0,1,2,\ldots$$

Zmienna losowa $X$ jest zatem zmienną przyjmującą przeliczalną, ale nie skończoną liczbę wartości. Okazuje się, że jeśli $X$ ma rozkład Poissona z parametrem $\lambda$, to wartość średnia $X$ jest równa wariancji $X$

$$\mu_X = \lambda \qquad \sigma_X^2 = \lambda. \tag{2.20}$$

**Przykład 2.20.** Z miesięcznej obserwacji małego skrzyżowania w Kaliszu wynika, że między godziną 11.00 a 12.00 pojawiają się tam średnio 4 ciężarówki o ładowności ponad 3,5 tony. Zakładając, że momenty ich pojawiania się w ustalonym dniu mogą być modelowane za pomocą procesu Poissona, obliczmy prawdopodobieństwo, że między 11.00 a 11.30 nie pojawi się żadna taka ciężarówka.

Oznaczając przez $X$ liczbę ciężarówek na skrzyżowaniu między 11.00 a 12.00 tego dnia, wiemy, że $X$ ma rozkład Poissona z parametrem $\lambda_1 = \lambda(12-11) = \lambda$, gdzie $\lambda$ jest intensywnością procesu Poissona przejazdu ciężarówek w ciągu dnia. Jednakże na podstawie równości (2.20) wiemy, że $\mu_X = \lambda_1$ i ponadto $\mu_X = 4$, a stąd $\lambda_1 = 4$. Rozważmy teraz liczbę $Y$ analogicznych zdarzeń między 11.00 a 11.30. Oczywiście, odcinek czasu między 11.00 a 11.30 jest zawarty w całkowitym czasie obserwacji tego dnia, a zatem zdarzenia w tym czasie tworzą również proces Poissona. Tak więc zmienna losowa $Y$ ma rozkład Poissona $P(4 \times 0,5) = P(2)$. Zatem

$$P(Y = 0) = e^{-2}2^0/0! = 0{,}1353.$$

Znaczenie rozkładu Poissona wynika również z jego związku z rozkładem dwumianowym. Mianowicie, rozważmy sytuację, w której parametry funkcji

prawdopodobieństwa $b(\cdot; n, p)$ zmieniają się w taki sposób, że $n \to \infty$, $p = p_n \to 0$, ale $np = \lambda > 0$. Wówczas dla ustalonego $k$

$$b(k; n, p) \to p(k; \lambda) \quad \text{gdy } n \to \infty.$$

Rzeczywiście,

$$\begin{aligned} b(k; n, p) &= \tfrac{n!}{k!(n-k)!} p^k (1-p)^{n-k} = \tfrac{n(n-1)\cdots(n-k+1)}{k!} (\tfrac{\lambda}{n})^k (1-\tfrac{\lambda}{n})^{n-k} = \\ &= 1(1-\tfrac{1}{n}) \cdots (1-\tfrac{(k-1)}{n}) \tfrac{\lambda^k}{k!} (1-\tfrac{\lambda}{n})^n (1-\tfrac{\lambda}{n})^{-k} \to \tfrac{\lambda^k}{k!} e^{-\lambda}, \end{aligned}$$

gdyż $k$ jest ustalone, $n$ dąży do $+\infty$ i $(1-\frac{\lambda}{n})^n$ dąży do $e^{-\lambda}$.

Ponadto, rozkład Poissona ma następującą ważną i ciekawą własność: jeśli $X$ i $Y$ są *niezależnymi* zmiennymi losowymi o rozkładzie Poissona $P(\lambda)$ i $P(\eta)$ odpowiednio, to suma $X + Y$ ma rozkład Poissona $P(\lambda + \eta)$. Z własności tej wynika przez indukcję, że jeśli $X_1, \ldots, X_n$ są zmiennymi niezależnymi i mają wszystkie rozkład Poissona z parametrem $\lambda$, to suma $X_1 + \ldots + X_n$ ma rozkład Poissona $P(n\lambda)$. Stąd już na podstawie Centralnego Twierdzenia Granicznego przedstawionego dalej w p. 2.4.2, blisko do ważnego wniosku:

*Jeśli $X$ ma rozkład $P(\lambda)$ dla dużego $\lambda$, to rozkład standaryzowanej zmiennej $(X - \lambda)/\sqrt{\lambda}$ jest w przybliżeniu normalny.*

**4. Rozkład geometryczny**. Rozkład geometryczny pojawił się w przykł. 2.3, gdy rzucaliśmy monetę aż do uzyskania pierwszego orła i w przykł. 2.15, gdzie opisywał czas od początku roku 2000 do momentu spotkania znajomych. Rozkład geometryczny jest czasami nazywany rozkładem czasu oczekiwania na pierwszy sukces i ta druga nazwa tłumaczy po części, dlaczego spotykamy się z nim w statystyce tak często jak z rozkładem dwumianowym: sytuacja ponawiania prób aż do skutku jest równie częsta jak niezależne powtarzanie tych samych czynności ustaloną liczbę razy. To drugie doświadczenie jest związane z rozkładem dwumianowym, natomiast pierwsze z rozkładem geometrycznym. Ogólnie, zmienna losowa $T$ ma rozkład geometryczny z parametrem $0 < p < 1$, jeśli zbiór jej możliwych wartości jest równy $\{1, 2, \ldots\}$ i funkcja prawdopodobieństwa ma postać

$$g(i, p) = (1-p)^{i-1} p \quad \text{dla } i = 1, 2, \ldots$$

Zmienna $T$ opisuje czas oczekiwania na sukces w ciągu niezależnych prób, z których każda kończy się sukcesem z prawdopodobieństwem $p$ i porażką z prawdopodobieństwem $1 - p$. Tak więc $T = 1$, jeśli sukces pojawił się w pierwszej próbie, $T = 2$, jeśli pojawił się w drugiej próbie itd.

Dla rozkładu geometrycznego

$$\mu_T = \frac{1}{p} \qquad \sigma_T^2 = \frac{1-p}{p^2}.$$

Zatem wartość średnia czasu oczekiwania na pierwszy sukces w eksperymencie, który może się powieść z prawdopodobieństwem 0,1 wynosi 10.

### 2.2.4. Ciągłe zmienne losowe

W punkcie 1.4.1 wprowadziliśmy intuicyjnie pojęcie gęstości rozkładu prawdopodobieństwa cechy ciągłej jako krzywej opisującej idealny histogram. Tak zdefiniowana gęstość ma następującą pożyteczną własność: pole pod nią nad dowolnym przedziałem jest równe częstości wpadania wartości cechy do tego przedziału. Intuicja ta jest bardzo bliska ścisłej definicji ciągłej zmiennej losowej i jej gęstości.

**Definicja 2.14.** *Zmienną losową $X$ nazywamy* ***ciągłą zmienną losową****, jeśli dla pewnej nieujemnej funkcji $f$ i takich dowolnych liczb $a$ i $b$, że $-\infty \leq a < b \leq \infty$ zachodzi równość*

$$P(a \leq X \leq b) = \int_a^b f(s)\, ds. \tag{2.21}$$

Zauważmy, że przyjmując w powyższej równości $a = -\infty$, otrzymujemy, że dystrybuanta $F_X(\cdot)$ spełnia równość

$$F_X(b) = \int_{-\infty}^b f(s)\, ds. \tag{2.22}$$

Funkcję $f$, dla której zachodzi równość (2.21) nazywamy **gęstością zmiennej losowej** $X$ lub **gęstością jej rozkładu**. Ta druga nazwa wynika stąd, że jak widać z def. 2.14, gęstość jest określona nie przez konkretne wartości zmiennej losowej $X$, ale przez wszystkie prawdopodobieństwa $P(a \leq X \leq b)$ dla zmieniających się $a$ i $b$, czyli przez rozkład zmiennej losowej $X$. Dwie różne ciągłe zmienne losowe mające ten sam rozkład będą zatem miały tę samą gęstość. Prostym przykładem jest losowa wartość $X$ wygenerowana przez generator liczb losowch z przedziału $(0, 1)$. Zmienna $1-X$ ma taki sam rozkład jak zmienna $X$, zostanie on omówiony w p. 2.2.6. Funkcja gęstości jest odpowiednikiem funkcji prawdopodobieństwa dla dyskretnej zmiennej losowej $X$. Aby się o tym przekonać, porównajmy (2.21) z analogiczną własnością dla zmiennej dyskretnej, w której zamiast całki występuje szereg: $P(a \leq X \leq b) = \sum_{i:a \leq x_i \leq b} p(x_i)$. Poczyńmy jeszcze jedną uwagę wynikającą z równości (2.21): jeśli gęstość $f$ jest równa 0 w pewnym obszarze, to prawdopodobieństwo przyjęcia przez zmienną losową wartości z tego obszaru jest też równe 0.

Zauważmy, że postulat, żeby gęstość była nieujemna jest intuicyjnie oczywisty: gdyby funkcja $f$ była ujemna na pewnym przedziale, to wartość całki z tej funkcji po tym przedziale byłaby też ujemna i w świetle pierwszego aksjomatu w def. 2.5 nie mogłaby być wartością prawdopodobieństwa. Z równości (2.21) wynika również, że

$$P(-\infty \leq X \leq \infty) = \int_{-\infty}^{\infty} f(s)\, ds = 1. \tag{2.23}$$

Tak więc krzywa gęstości powinna spełniać dwa naturalne postulaty: musi być nieujemna i pole pod nią musi wynosić 1. Są to w istocie jedyne warunki, które musi spełniać konkretna funkcja, aby była funkcją gęstości dla pewnego rozkładu ciągłego. Ta uwaga ma duże znaczenie w modelowaniu probabilistycznym: na podstawie danych można postulować pożądany kształt kandydata na funkcję gęstości, a następnie zmodyfikować ją tak, aby spełniała powyższe warunki.

Z jedną rodziną zmiennych losowych ciągłych już się zetknęliśmy: są to zmienne losowe normalne o gęstościach zadanych wzorem (2.21) w p. 1.4.2. Wrócimy do nich w p. 2.2.6. Teraz rozpatrzmy kilka podstawowych własności ciągłych zmiennych losowych.

Zauważmy, że jeśli przyjmiemy we wzorze (2.21) $a = b$ dla dowolnej ciągłej zmiennej losowej $X$, to otrzymamy $P(X = a) = 0$. Tak więc prawdopodobieństwo przyjęcia pojedynczej wartości przez ciągłą zmienną losową jest równe 0. Oczywiście oznacza to, że zachodzi następujące stwierdzenie.

**Stwierdzenie 2.5.** *Dla ciągłej zmiennej losowej $X$ o dystrybuancie $F$ zachodzi*

$$P(a < X < b) = P(a < X \leq b) = P(a \leq X < b) = P(a \leq X \leq b) =$$

$$= F(b) - F(a). \tag{2.24}$$

Otrzymujemy stąd ważny wniosek, że *żadna dyskretna zmienna losowa nie jest zmienną losową ciągłą*. Intuicyjnie, ciągła zmienna losowa powinna przyjmować wszystkie wartości z pewnego przedziału, ale to niestety nie wystarczy, aby była spełniona własność (2.21). Łatwo można wyobrazić sobie zmienną dyskretno-ciągłą, która nie jest ani dyskretna ani ciągła. Wystarczy z prawdopodobieństwem 1/2 wylosować liczbę z dowolnego rozkładu dyskretnego (np. dwupunktowego) a z prawdopodobieństwem 1/2 z dowolnego rozkładu ciągłego (np. jednostajnego). W dalszym ciągu będziemy z reguły ograniczali się do rozpatrywania zmiennych, które są albo dyskretne

albo ciągłe. Odnotujmy jeszcze, nie uzasadniając tego dokładniej, że z równania (2.22) wynika, iż dystrybuanta ciągłej zmiennej losowej jest funkcją różniczkowalną oraz zachodzi następujące twierdzenie.

**Twierdzenie 2.8.** *Jeśli gęstość $f$ zmiennej losowej $X$ jest funkcją ciągłą, to dla każdego $x$ zachodzi równość $F'(x) = f(x)$.*

Założenie o ciągłości gęstości $f$ jest potrzebne tylko po to, żeby równość w twierdzeniu zachodziła dla wszystkich $x$. Bez tego założenia może się zdarzyć, że pochodna $F'(x)$ nie będzie istniała, bądź nie będzie równa swojej gęstości w punkcie, ale brak równości nie może zachodzić we wszystkich punktach dowolnego odcinka.

**Przykład 2.21.** Niech gęstość $f$ ciągłej zmiennej losowej $X$ wynosi

$$f(x) = \begin{cases} \frac{2}{3} + x^2 & \text{dla } 0 < x < 1; \\ 0 & \text{w przeciwnym przypadku.} \end{cases}$$

Znajdźmy dystrybuantę $F$ zmiennej losowej $X$ oraz prawdopodobieństwo $P(X > 0{,}5)$.

Oczywiście, ponieważ $f$ jest równa zeru dla $x \leq 0$, $P(X \leq 0) = 0$ i $F(x) = 0$ dla $x \leq 0$. Podobnie, ponieważ $f$ jest równa zeru dla $x \geq 1$, $P(X \leq 1) = 1$ i $F(x) = 1$ dla $x \geq 1$. Dla $0 < x < 1$

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(s)\,ds = \int_{0}^{x} f(s)\,ds = \frac{1}{3}(2s + s^3)\Big|_0^x = \frac{1}{3}(2x + x^3).$$

Zatem

$$F(x) = \begin{cases} 0 & \text{dla } x \leq 0; \\ \frac{1}{3}(2x + x^3) & \text{dla } 0 < x \leq 1; \\ 1 & \text{dla } x > 1. \end{cases}$$

Możemy teraz łatwo obliczyć prawdopodobieństwo $P(X > 0{,}5)$. Jest ono równe

$$\int_{0{,}5}^{\infty} f(s)\,ds = \int_{0{,}5}^{1} f(s)\,ds = \frac{1}{3}(2s + s^3)\Big|_{0{,}5}^1 = 0{,}625.$$

Alternatywnie, $P(X > 0{,}5)$ możemy obliczyć, pamiętając, że $P(X > 0{,}5) = 1 - P(X \leq 0{,}5) = 1 - F(0{,}5) = 1 - \int_{-\infty}^{0{,}5} f(s)\,ds = 1 - 0{,}375 = 0{,}625$.

### 2.2.5. Wskaźniki położenia i rozproszenia dla ciągłych zmiennych losowych

Dla dyskretnej zmiennej losowej wartość średnia $\mu_X$ była zdefiniowana jako suma czynników $x_i p(x_i)$ dla wszystkich potencjalnych wartości $x_i$. W przypadku ciągłej zmiennej losowej definicję jej wartości średniej otrzymamy, zastępując funkcję prawdopodobieństwa gęstością zmiennej losowej oraz operację sumowania całkowaniem.

**Definicja 2.15.** ***Wartością średnią*** *ciągłej zmiennej losowej o gęstości $f$ nazywamy wielkość*

$$\mu_X = \int_{-\infty}^{\infty} s f(s)\, ds. \tag{2.25}$$

Wartość średnią łatwo zinterpretować intuicyjnie podobnie jak to uczyniliśmy w p. 1.4.1. Wartość średnia $\mu_X$ jest równa środkowi ciężkości pręta o tej własności, że gęstość masy w punkcie oddalonym o $s$ jednostek od początku układu współrzędnych wynosi $f(s)$.

W celu zdefiniowania mediany ciągłej zmiennej losowej rozważmy ogólniejszą definicję **kwantyla rzędu** $q$ dla $0 < q < 1$ jako dowolnego takiego punktu, że $F(x_q) = q$. W szczególności, **mediana** jest określona jako kwantyl rzędu 0,5, to jest taki punkt, że pole pod gęstością na lewo od niego (i oczywiście również na prawo) jest równe 0,5. Zauważmy, że kwantyle, a w szczególności mediana, nie muszą być wyznaczone jednoznacznie. Niejednoznaczność występuje, gdy gęstość jest równa 0 w otoczeniu pewnego kwantyla rzędu $q$ i odpowiada sytuacji, gdy dystrybuanta jest stała i równa $q$ na tym otoczeniu. **Modą** ciągłej zmiennej losowej jest dowolne maksimum lokalne jej gęstości.

**Przykład 2.21 cd.** Obliczmy wartość średnią i medianę zmiennej losowej rozpatrywanej w przykładzie. Z definicji 2.15 (wzór (2.25)) wynika, że

$$\mu_X = \int_{-\infty}^{\infty} s f(s)\, ds = \int_0^1 s\Big(\frac{2}{3} + s^2\Big)\, ds = \frac{s^2}{3} + \frac{s^4}{4}\Big|_0^1 = \frac{7}{12}.$$

Jednocześnie, mediana spełnia równanie $F(x_{0,5}) = (2x_{0,5} + x_{0,5}^3)/3 = 0{,}5$, po rozwiązaniu którego otrzymujemy $x_{0,5} = 0{,}63$.

Zauważmy, że podobnie jak dla dyskretnej zmiennej losowej wartość średnia ciągłej zmiennej losowej może być nieskończona lub może nie istnieć (por. przykł. 2.17). Przykładem może być zmienna losowa o gęstości Cauchy'ego

$$f(x) = \frac{1}{\pi}\frac{1}{1+x^2}, \qquad \text{dla} \quad -\infty < x < \infty.$$

Wynika to z faktu, że w tym przypadku funkcja podcałkowa $sf(s)$ w def. 2.15 (wzór (2.25)) zachowuje się dla dużych $s$ w przybliżeniu jak $g(s) = 1/(\pi s)$, a funkcja $g$ ma całkę nieskończoną.

Odnotujmy jeszcze jedną użyteczną własność, będącą odpowiednikiem tw. 2.7 dla zmiennych ciągłych.

**Twierdzenie 2.9.** *Niech $X$ będzie zmienną losową o rozkładzie ciągłym i $h$ dowolną funkcją określoną na zbiorze wartości $X$. Wówczas dla zmiennej losowej $Y = h(X)$ mamy*

$$\mu_Y = \int_{-\infty}^{\infty} h(s)f(s)\,ds. \tag{2.26}$$

*W szczególności $\mu_{aX+b} = a \times \mu_X + b$.*

Twierdzenie to jest użyteczne przede wszystkim wtedy, gdy mimo znajomości gęstości $f$ zmiennej losowej $X$ dokładne obliczenie postaci rozkładu zmiennej losowej $Y = h(X)$ jest trudne. Możemy wtedy starać się obliczyć na przykład parametr położenia zmiennej losowej $Y$. Dla innych miar położenia nie ma odpowiedników wzoru (2.26), zauważmy jednak, że gdy $h$ jest funkcją rosnącą, to medianą zmiennej losowej $h(X)$ będzie $h(q_{0,5})$, gdzie $q_{0,5}$ jest medianą zmiennej losowej $X$.

W komentarzu do def. 2.13 wariancji dyskretnej zmiennej losowej, zauważyliśmy, że może być ona zapisana jako $\mu_{(X-\mu_X)^2}$. Jeśli to ostatnie wyrażenie przyjmie się za ogólną definicję wariancji, to dla ciągłej zmiennej losowej przy uwzględnieniu równości (2.25), otrzymamy definicję następującą.

**Definicja 2.16.** ***Wariancją*** *ciągłej zmiennej losowej o gęstości $f$ nazywamy wielkość*

$$\sigma_X^2 = \int_{-\infty}^{\infty} (s-\mu_X)^2 f(s)\,ds. \tag{2.27}$$

***Odchylenie standardowe*** *$\sigma_X$ definiuje się jako nieujemny pierwiastek $\sqrt{\sigma_X^2}$.*

Łatwo można się przekonać, że stwierdzenie 2.4 oraz równość (2.18) z p. 2.2.2

pozostają w mocy dla rozkładu ciągłego. Zatem dla dowolnej ciągłej zmiennej losowej $X$ mamy

$$\sigma^2_{aX+b} = a^2\sigma^2_X$$

oraz

$$\sigma^2_X = \mu_{X^2} - (\mu_X)^2. \tag{2.28}$$

**Przykład 2.21 cd.** Ze wzoru (2.28) obliczymy wariancję zmiennej losowej zdefinowanej w przykładzie. Zgodnie z tw. 2.9 dla funkcji $h(x) = x^2$

$$\mu_{X^2} = \int_{-\infty}^{\infty} s^2 f(s)\,ds = \int_0^1 s^2\Big(\frac{2}{3} + s^2\Big)\,ds = \frac{2s^3}{9} + \frac{s^5}{5}\Big|_0^1 = \frac{19}{45}.$$

Z tego i wzoru (2.28) wynika, że $\sigma^2_X = 19/45 - (7/12)^2 = 0{,}0819$.

W nawiązaniu do przykładu zauważmy, że wariancję zmiennej $Y = h(X)$ można obliczyć na podstawie tw. 2.9 dla funkcji $h^2$ razem ze wzorem (2.28)

$$\sigma^2_Y = \int_{-\infty}^{\infty} h^2(s) f(s)\,ds - \Big\{\int_{-\infty}^{\infty} h(s) f(s)\,ds\Big\}^2. \tag{2.29}$$

Tak samo jak w przypadku dyskretnych zmiennych losowych określa się moment rzędu $k$ oraz moment centralny rzędu $k$ ciągłej zmiennej losowej.

### 2.2.6. Przykłady ciągłych zmiennych losowych

**1. Zmienna losowa o rozkładzie normalnym.** W punkcie 1.4.2 wprowadziliśmy zmienną losową o rozkładzie normalnym jako zmienną losową zadaną gęstością

$$f(x) = \phi_{\mu,\sigma}(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2} \qquad -\infty < x < \infty, \tag{2.30}$$

dla dowolnej rzeczywistej liczby $\mu$ i dodatniej liczby $\sigma$. Ściślej, zgodnie z komentarzem pod wzorem (2.23) w p. 2.2.4, aby stwierdzić, że $f$ jest gęstością, a zmienna losowa $X$ istnieje, powinniśmy wykazać, że $\int_{-\infty}^{\infty} f(s)\,ds = 1$ i $f$ jest

nieujemna. Drugi z tych warunków jest oczywisty, pierwszy przyjmiemy tutaj bez uzasadnienia, wymagającego dosyć zaawansowanych rozważań technicznych.

**Wartość średnia i odchylenie standardowe rozkładu normalnego** $N(\mu, \sigma)$

Udowodnimy teraz ważne stwierdzenie.

**Stwierdzenie 2.6.** *Jeśli $X$ jest zmienną losową o rozkładzie normalnym $N(\mu, \sigma)$, to wartość średnia $X$ jest równa $\mu$, a odchylenie standardowe równe $\sigma$: $\mu_X = \mu$ i $\sigma_X = \sigma$.*

Aby się przekonać o prawdziwości stwierdzenia zauważmy, że na mocy zasady standaryzacji $Z = (X - \mu)/\sigma$ ma rozkład standardowy normalny $N(0, 1)$. Jeśli zatem udowodnimy, że $\mu_Z = 0$ i $\sigma_Z = 1$, to na podstawie własności (2.26) i (2.29) otrzymamy

$$\mu_X = \mu_{\sigma Z+\mu} = \mu_{\sigma Z} + \mu = \sigma \times 0 + \mu = \mu,$$

gdzie $\mu_{\sigma Z+\mu}$ oznacza wartość średnią dla zmiennnej losowej $\sigma \times Z + \mu$ i

$$\sigma_{\sigma Z+\mu} = \sigma_{\sigma Z} = \sigma \times \sigma_Z = \sigma \times 1 = \sigma,$$

gdzie $\sigma_{\sigma Z}$ oznacza odchylenie standardowe zmiennej $\sigma \times Z$. Obliczmy wartość średnią $\mu_Z$. Z definicji 2.15 $\mu_Z = \int_{-\infty}^{\infty} z\phi(z)\,dz$, gdzie $\phi$ jest standardową gęstością normalną. Ponieważ funkcja $\phi$ jest parzysta, $\phi(z) = \phi(-z)$, więc funkcja podcałkowa $z\phi(z)$ jest nieparzysta. Zatem całka po półosi dodatniej z tej funkcji znosi się z całką po półosi ujemnej i $\mu_Z = 0$. W celu obliczenia wartości $\sigma_Z$, zauważmy, że $\sigma_Z^2 = \mu_{Z^2} - (\mu_Z)^2 = \mu_{Z^2}$ i

$$\mu_{Z^2} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^2 e^{-z^2/2}\,dz = -\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z(-ze^{-z^2/2})\,dz =$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-z^2/2}\,dz = 1,$$

gdzie przedostatnia równość wynika z całkowania przez części i zauważenia, że granica $-z\phi(z)$ jest równa zeru w plus i minus nieskończoności, a ostatnia równość wynika z równości (2.23). Stwierdzenie 2.6 tłumaczy po części popularność używania wartości średniej i odchylenia standardowego jako wskaźników położenia i rozproszenia rozkładu. Parametry często używanego w statystyce rozkładu normalnego odpowiadają bowiem tym właśnie wskaźnikom.

Rozpatrzmy teraz inny ważny przykład.

**2. Zmienna losowa o rozkładzie jednostajnym.** Niech $\mathcal{S} = [0,1]$ i zmienna losowa $X(s) = s$ dla $s \in \mathcal{S}$. Przyjmijmy, że prawdopodobieństwo $P$ na $\mathcal{S}$ jest określone w taki sposób, że wartość prawdopodobieństwa dla dowolnego odcinka jest równa jego długości, to znaczy $P([a,b]) = b - a$ dla dowolnego $[a,b] \subset [0,1]$. Wtedy $X$ możemy interpretować jako wynik eksperymentu polegającego na losowym i „jednostajnym" wyborze wartości z odcinka $[0,1]$. Zmienna $X$ nosi nazwę zmiennej o rozkładzie jednostajnym na odcinku $[0,1]$. Niech $f(x)$ będzie teraz zdefiniowana następująco:

$$f(x) = \begin{cases} 1 & \text{dla } 0 \le x \le 1; \\ 0 & \text{w przeciwnym przypadku.} \end{cases}$$

Wówczas dla $0 \le a \le b \le 1$

$$P(a \le X \le b) = b - a = \int_a^b f(s)\, ds.$$

Podobnie można przekonać się o spełnieniu równości (2.21) w def. 2.14 wtedy, gdy niekoniecznie obie liczby $a$ i $b$ są z odcinka $[0,1]$. Niech np. $0 \le a \le 1 < b$. Wówczas

$$P(a \le X \le b) = P(a \le X \le 1) = 1 - a = \int_a^b f(s)\, ds.$$

Zatem $f$ jest gęstością zmiennej losowej $X$. Powód, dla którego taki rozkład nazywa się jednostajnym, powinien być oczywisty z wykresu gęstości $f$. Ogólnie, nazwiemy zmienną losową o rozkładzie jednostajnym na odcinku $[a,b]$ zmienną losową $Y = a + (b-a)X$, gdzie $X$ jest zmienną losową o rozkładzie jednostajnym na $[0,1]$. Łatwo można przekonać się, że jej gęstość ma postać

$$f(x) = \begin{cases} \frac{1}{(b-a)} & \text{dla } a \le x \le b; \\ 0 & \text{w przeciwnym przypadku.} \end{cases}$$

Ponieważ $\int_0^1 s\, ds = 1/2$ i $\int_0^1 s^2\, ds = 1/3$, więc mamy, że $\mu_X = 1/2$ i $\sigma_X^2 = 1/3 - (1/2)^2 = 1/12$. Stąd stosując wzory (2.15) i (2.18) z p. 2.2.2, otrzymujemy dla zmiennej losowej $Y$ o rozkładzie jednostajnym na $[a,b]$

$$\mu_Y = a + (b-a) \times \frac{1}{2} = \frac{(b+a)}{2}, \qquad \sigma_Y^2 = \frac{(b-a)^2}{12}.$$

**3. Zmienna losowa o rozkładzie wykładniczym.** Zmienna losowa o rozkładzie wykładniczym pojawia się w sposób naturalny jako czas oczekiwania na wystąpienie pierwszego zdarzenia w ciągu zdarzeń opisanych procesem Poissona. Załóżmy, że od pewnego dnia, na przykład 1 maja 2001, rejestrujemy czas $T$, który upłynie od tego dnia do pierwszego po tym momencie trzęsienia ziemi w Kalifornii. Jeśli przyjmiemy, że momenty wystąpienia kolejnych trzęsień ziemi mogą być opisane procesem Poissona z intensywnością $\lambda$ na rok i oznaczymy liczbę trzęsień ziemi w przedziale czasowym $[0,t]$ przez $X$, to mamy, że $P(T > t) = P(X = 0)$. Pamiętając, że $X$ ma rozkład Poissona z parametrem $\lambda t$, otrzymujemy zatem

$$P(T > t) = \frac{e^{-\lambda t}(\lambda t)^0}{0!} = e^{-\lambda t}.$$

Oczywiście dla $t < 0$ mamy $P(T > t) = 1$. Dystrybuanta rozkładu zmiennej losowej $T$ ma zatem postać

$$F(t) = 1 - P(T \leq t) = \begin{cases} 0 & \text{dla } t < 0; \\ 1 - e^{-\lambda t} & \text{dla } t \geq 0. \end{cases} \tag{2.31}$$

Zmienną losową o dystrybuancie opisanej wzorem (2.31) nazywamy zmienną losową o rozkładzie wykładniczym. Jest to zmienna losowa o rozkładzie ciągłym. Łatwo się o tym przekonać rozpatrując funkcję (pochodną $F$)

$$f(t) = \begin{cases} 0 & \text{dla } t < 0; \\ \lambda e^{-\lambda t} & \text{w przeciwnym przypadku.} \end{cases}$$

Niech $a$ i $b$ będą dowolnymi liczbami rzeczywistymi i niech na przykład $0 \leq a \leq b$. Wówczas

$$P(a \leq T \leq b) = F(b) - F(a) = e^{-\lambda a} - e^{-\lambda b} = \int_a^b e^{-\lambda s}\, ds = \int_a^b f(s)\, ds.$$

Tak więc w tym przypadku jest spełniona definicja ciągłej zmiennej losowej. Podobnie sprawdzamy spełnienie definicji, gdy niekoniecznie obie spośród liczb $a$ i $b$ są dodatnie.

**Przykład 2.22.** Przyjmijmy, że czas rozmowy telefonicznej z aparatu publicznego jest zmienną losową o rozkładzie wykładniczym z parametrem $\lambda = 1/7$. Załóżmy, że ktoś ubiegł nas w dojściu do wolnego aparatu. Jakie jest prawdopodobieństwo, że będziemy czekali nie krócej niż 5 i nie dłużej niż 10 minut na zwolnienie telefonu?

Oznaczmy czas oczekiwania na zwolnienie aparatu przez $T$. Jest on równy czasowi rozmowy osoby telefonującej. Interesuje nas oczywiście prawdopodobieństwo

$$P(5 < T < 10) = e^{-\frac{1}{7}\times 5} - e^{-\frac{1}{7}\times 10} \approx 0{,}48954 - 0{,}23965 = 0{,}25.$$

Obliczmy teraz wartość średnią zmiennej losowej o rozkładzie wykładniczym. Zgodnie z definicją wartości oczekiwanej całkując przez części, otrzymujemy

$$\mu_T = \int_0^\infty x\lambda e^{-\lambda x}\,dx = -xe^{-\lambda x}\Big|_0^\infty + \int_0^\infty e^{-\lambda x}\,dx = 0 - \frac{e^{-\lambda x}}{\lambda}\Big|_0^\infty = \frac{1}{\lambda}.$$

Zauważmy, że wynik ten odpowiada w tym przypadku intuicji. Jeśli w przykładzie wstępnym intensywność wystąpienia trzęsień ziemi wynosi $\lambda = 2$ rocznie, to średni czas oczekiwania na pierwsze trzęsienie po 1 maja 2001 wyniesie 1/2 roku, czyli 6 miesięcy.

Analogicznie, stosując wzór na całkowanie przez części do obliczenia $\mu_X^2$, można obliczyć

$$\sigma_T^2 = \mu_{T^2} - (\mu_T)^2 = \frac{2}{\lambda^2} - \Big(\frac{1}{\lambda}\Big)^2 = \frac{1}{\lambda^2}.$$

Rozkład wykładniczy jest szczególnym przypadkiem **rozkładu gamma** rozpatrywanego w zad. 2.20.

Zmienne o rozkładzie wykładniczym pojawiają się, gdy mamy do czynienia z procesem Poissona nie tylko jako moment pierwszego zdarzenia tego procesu (liczony od momentu zerowego). Niech $T_1 = T$, $T_2$ niech oznacza czas między drugim a pierwszym zdarzeniem i ogólnie, $T_i$ czas między $i$–tym a $(i-1)$–szym zdarzeniem. Wówczas okazuje się, że wszystkie czasy $T_i$ między zdarzeniami mają rozkład wykładniczy, co więcej, są od siebie niezależne.

### 2.2.7. Nierówność Czebyszewa

Zakończymy przegląd własności zmiennych losowych omówieniem nierówności Czebyszewa, która umożliwia szacowanie prawdopodobieństwa, że bezwzględna odchyłka zmiennej losowej $X$ od swojej wartości średniej przekracza ustalony poziom. Nierówność ta zachodzi dla *dowolnej* zmiennej losowej o skończonej wariancji, bez względu na to, czy zmienna ta jest ciągła, czy nie.

**Twierdzenie 2.10. Nierówność Czebyszewa.** *Niech $X$ będzie zmienną losową o wartości średniej $\mu$ i skończonej wariancji $\sigma^2$. Wówczas*

$$P(|X-\mu| \geq \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2}$$

*dla dowolnego $\varepsilon > 0$.*

Zdefiniujmy zmienną losową

$$Y = \begin{cases} \varepsilon, & \text{gdy } |X-\mu| \geq \varepsilon; \\ 0, & \text{w przeciwnym przypadku.} \end{cases}$$

Zauważmy, że $|X-\mu| \geq Y$ i zatem $(X-\mu)^2 \geq Y^2$. Po obliczeniu wartości średniej obu stron ostatniej nierówności i skorzystaniu z definicji wariancji otrzymujemy $\sigma^2 \geq \mu_{Y^2}$. Zmienna losowa $Y^2$ przyjmuje tylko wartości $\varepsilon^2$ (z prawdopodobieństwem $P(|X-\mu| \geq \varepsilon)$) i 0, zatem $\mu_{Y^2} = \varepsilon^2 P(|X-\mu| \geq \varepsilon)$ i $\sigma^2 \geq \varepsilon^2 P(|X-\mu| \geq \varepsilon)$, co jest równoważne nierówności Czebyszewa.

Zauważmy, że zastępując $\varepsilon$ iloczynem $\varepsilon\sigma$ w nierówności Czebyszewa, otrzymamy jej następujące sformułowanie:

$$P(|X-\mu| \geq \varepsilon\sigma) \leq \frac{1}{\varepsilon^2}.$$

Tak więc prawdopodobieństwo, że wartość bezwzględna standaryzowanej zmiennej $Z = (X-\mu)/\sigma$ przekroczy wielkości $\varepsilon$ jest nie większe niż $1/\varepsilon^2$. Zastanówmy się, czy nierówność Czebyszewa umożliwia oszacowanie prawdopodobieństwa jednostronnej odchyłki $P(X-\mu > \varepsilon)$. Oczywiście tak, jeśli zastosujemy nierówność $P(X-\mu > \varepsilon) \leq P(|X-\mu| > \varepsilon)$. Czy możliwe jest uzyskanie lepszego oszacowania? Wiemy przecież, że

$$P(|X-\mu| > \varepsilon) = P(X-\mu > \varepsilon) + P(X-\mu < -\varepsilon).$$

Problem w tym, że jeśli nie wiemy nic o rozkładzie zmiennej losowej $X$, nie możemy powiedzieć, jak mają się do siebie składniki znajdujące się po prawej stronie ostatniej równości. Jeśli wiemy jednak, że rozkład $X$ jest symetryczny względem swojej wartości średniej, to oba te składniki są równe i mamy

$$P(X-\mu > \varepsilon) = \frac{1}{2} P(|X-\mu| > \varepsilon) \leq \frac{\sigma^2}{2\varepsilon^2}.$$

**Przykład 2.23.** Wartość średnia czasu bezawaryjnej pracy $X$ nowego monitora pewnej marki wynosi 3,5 roku, a odchylenie standardowe tego czasu wynosi 0,5 roku. Oszacujmy prawdopodobieństwo, że losowo

wybrany monitor tej marki popsuje się w czasie odległym o nie mniej niż 1,5 roku od wartości średniej bezawaryjnej pracy, to jest w ciągu 2 pierwszych lat pracy, albo po upływie 5 lat.

Interesuje nas prawdopodobieństwo $P(|X - 3{,}5| \geq 1{,}5)$. Ponieważ nie mamy żadnych informacji o rozkładzie $X$, więc stosujemy nierówność Czebyszewa

$$P(|X - 3{,}5| \geq 1{,}5) \leq \frac{(0{,}5)^2}{(1{,}5)^2} = 0{,}11.$$

Zauważmy, że z poprzedniej nierówności wynika, że $P(|X - 3{,}5| < 1{,}5) \leq 1 - 0{,}11 = 0{,}89$.

## 2.3. Para zmiennych losowych – rozkład łączny, rozkłady i parametry związane z rozkładem łącznym

Bardzo często interesuje nas nie jedna tylko, ale kilka zmiennych losowych określonych na tej samej przestrzeni zdarzeń elementarnych. W badaniach antropometrycznych wnioskowanie może opierać się na mierzeniu wzrostu, wagi ciała oraz grubości skóry przedramienia osobników pewnej populacji. Zdarzeniem elementarnym jest w tym przypadku wystąpienie osobnika o zadanych wartościach wzrostu, wagi i grubości skóry, natomiast za zmienne losowe wygodnie jest obrać te trzy wielkości, czyli wzrost, wagę i grubość skóry przedramienia. Oddzielne badanie rozkładu każdej z tych zmiennych może być interesujące, ale takie badanie nie mówi zupełnie nic o zależności między zmiennymi. Może się na przykład okazać, że zarówno wzrost, jak i waga mają w populacji rozkłady normalne o pewnych parametrach. Ale z oddzielnej analizy wzrostu i wagi w żaden sposób nie wyniknie, że wzrost i waga są ze sobą w pewien sposób powiązane, czyli że są od siebie zależne. Taka analiza nie może ani potwierdzić, ani zaprzeczyć temu, czego powinniśmy oczekiwać: mianowicie, że osobnicy wysocy ważą zwykle więcej niż osobnicy niscy. Ażeby móc wnioskować o zależności między zmiennymi, o tym, że wysokiemu wzrostowi osobnika towarzyszy zwykle jego duża waga, trzeba na rozkład wartości zmiennych losowych spojrzeć łącznie.

Potrzebny nam jest zatem łączny probabilistyczny opis kilku zmiennych losowych. W podrozdziale tym ograniczymy się do przypadku tylko dwóch zmiennych losowych. Czytelnik z łatwością zauważy, że wszystkie ogólne rozważania na temat pary zmiennych losowych mają swoje naturalne i proste uogólnienia na przypadek ich większej liczby. Zarazem, wszystkie rozwa-

żania na temat pary zmiennych losowych są naturalnym i zazwyczaj oczywistym uogólnieniem wcześniejszych rozważań dla pojedynczej zmiennej losowej. Z tego względu pojęcia wprowadzone w tym podrozdziale omówimy w sposób skrótowy, a przykłady ich zastosowań przedstawimy jedynie w zadaniach.

Tu podkreślimy jedynie, że z potrzebą wnioskowania o łącznym rozkładzie zmiennych losowych spotykamy się bardzo często. Interesować nas może: ustalony zestaw parametrów produkowanych obwodów scalonych; średnice wału głównego samochodowej skrzyni biegów w różnych przekrojach wału; wysokości dziennego oprocentowania kredytów różnego typu; pozycja na giełdzie ustalonej firmy, mierzona kilkoma wskaźnikami giełdowymi; wartości bieżącego deficytu płatniczego, stopy procentowej wyznaczanej przez Narodowy Bank Polski oraz inflacji; ciśnienie skurczowe i rozkurczowe oraz tętno zdrowych mężczyzn w wieku 40–50 lat w warunkach kontrolowanego stresu, itp.

Niech będą dane dwie dyskretne zmienne losowe $X$ oraz $Y$, określone na tej samej przestrzeni zdarzeń elementarnych. Ich łączny rozkład jest dany **funkcją prawdopodobieństwa łącznego** (gdy nie będzie to prowadziło do nieporozumień, będziemy pisać krótko **funkcją prawdopodobieństwa**)

$$f(x,y) = P(X = x, Y = y),$$

określającą prawdopodobieństwo jednoczesnego przyjęcia przez zmienną losową $X$ wartości $x$ i przez zmienną losową $Y$ wartości $y$. Dziedziną funkcji prawdopodobieństwa $f(x,y)$ jest zbiór par uporządkowanych $(x,y)$, gdzie $x$ przebiega wszystkie wartości, jakie przyjąć może zmienna $X$, natomiast $y$ przebiega wszystkie wartości, jakie przyjąć może zmienna $Y$. Funkcja prawdopodobieństwa ma następujące oczywiste własności:
(1)

$$f(x,y) \geq 0$$

dla wszystkich $x, y$;
(2)

$$\sum_x \sum_y f(x,y) = 1,$$

gdzie pierwsza suma przebiega wszystkie możliwe wartości zmiennej $X$ i druga suma przebiega wszystkie możliwe wartości zmiennej $Y$;
(3) dla każdego zbioru $A$ par uporządkowanych $(x,y)$

$$P((X,Y) \in A) = \sum_{(x,y)\in A} f(x,y),$$

gdzie sumowanie przebiega po wszystkich parach uporządkowanych $(x,y)$, należących do zbioru $A$.

Z własności (1) wynika, że prawdopodobieństwo żadnego zdarzenia nie może być ujemne, natomiast z własności (2), że prawdopodobieństwo zajścia jakiegokolwiek zdarzenia równe jest 1. Własność (3) jest podstawą obliczania prawdopodobieństwa zajścia zdarzenia $A$: ponieważ $A$ jest ustalonym zbiorem uporządkowanych par wartości zmiennych losowych $X$ i $Y$ (przy czym wartości zmiennej $X$ występują na pierwszym miejscu w parze), więc prawdopodobieństwo zaobserwowania którejkolwiek pary ze zbioru $A$ jest równe sumie wszystkich prawdopodobieństw $P(X = x, Y = y)$, gdzie $(x, y) \in A$.

**Przykład 2.24.** Niech funkcja prawdopodobieństwa łącznego będzie dana wzorem

$$P(X = x, Y = y) = f(x, y) = \begin{cases} \frac{1}{30}(x + y), & \text{gdy } x = 0, 1, 2 \\ & \text{oraz } y = 0, 1, 2, 3 \\ 0, & \text{w przypadku przeciwnym.} \end{cases}$$

Znajdziemy najpierw prawdopodobieństwo $P(X = 2, Y = 0)$:

$$P(X = 2, Y = 0) = f(2, 0) = \frac{1}{30}(2 + 0) = \frac{1}{15}.$$

Łączne prawdopodobieństwa przyjęcia określonej wartości przez pierwszą zmienną losową oraz określonej wartości przez drugą zmienną losową wygodnie jest przedstawić w postaci następującej tablicy prawdopodobieństw (zwanej tablicą kontyngencji):

| | $Y$ | | | |
|---|---|---|---|---|
| $X$ | 0 | 1 | 2 | 3 |
| 0 | 0 | 1/30 | 1/15 | 1/10 |
| 1 | 1/30 | 1/15 | 1/10 | 2/15 |
| 2 | 1/15 | 1/10 | 2/15 | 1/6 |

Wiersze tablicy odpowiadają konkretnym wartościom zmiennej losowej $X$, kolumny zaś zmiennej losowej $Y$. Przykładowo, w wierszu odpowiadającym wartości $X = 1$ i kolumnie odpowiadającej wartości $Y = 2$ odczytujemy wartość $P(X = 1, Y = 2) = f(1, 2) = 1/10$. (W tym miejscu warto przestrzec Czytelnika, że może się wprawdzie zdarzyć, iż $f(x, y) = f(y, x)$ dla pewnych wartości $x$ i $y$, ale w ogólności jest to nieprawda, w pierwszym przypadku bowiem chodzi o prawdopodobieństwo $P(X = x, Y = y)$, a w drugim o prawdopodobieństwo $P(X = y, Y = x)$; na przykład, $f(2, 3) = 1/6$, podczas gdy $f(3, 2) = 0$.) Tablica kontyngencji umożliwia łatwe sprawdzenie, czy podana funkcja $f(\cdot, \cdot)$ jest rzeczywiście funkcją prawdopodobieństwa –

wystarczy w tym celu wykazać, że suma wszystkich prawdopodobieństw jest równa 1.

Znajdziemy jeszcze prawdopodobieństwo $P(X \leq 2, Y \leq 1)$:

$$P(X \leq 2, Y \leq 1) = \sum_{x=0}^{2} \sum_{y=0}^{1} \frac{1}{30}(x+y) = \frac{1}{30}[(0+0)+(0+1)+ \\ + (1+0)+(1+1)+(2+0)+(2+1)] = 0{,}3.$$

Parę zmiennych losowych $X$ oraz $Y$, określonych na tej samej przestrzeni zdarzeń elementarnych, nazywamy **parą ciągłych zmiennych losowych**, jeżeli łączny rozkład takiej pary dany jest **funkcją łącznej gęstości prawdopodobieństwa** (krótko, **funkcją gęstości łącznej** lub **gęstością łączną**) $f(\cdot,\cdot)$ określoną dla wszystkich par uporządkowanych $(x,y)$, $-\infty < x < \infty$, $-\infty < y < \infty$, i mającą następujące własności:
(1)

$$f(x,y) \geq 0$$

dla wszystkich $x, y$;
(2)

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x,y)dxdy = 1;$$

(3) dla każdego zbioru $A$ par uporządkowanych $(x,y)$

$$P((X,Y) \in A) = \int\int_{(x,y)\in A} f(x,y)dxdy.$$

Dziedziną funkcji gęstości prawdopodobieństwa $f(\cdot,\cdot)$ jest zbiór par uporządkowanych $(x,y)$, gdzie $-\infty < x < \infty$ oraz $-\infty < y < \infty$. Podane własności są naturalnymi odpowiednikami własności funkcji prawdopodobieństwa pary dyskretnych zmiennych losowych. Zarazem, własności te są naturalnymi uogólnieniami przypadku jednej zmiennej losowej. Na przykład, w przypadku własności (3), jedyna różnica polega na tym, że całkowanie po zbiorze na prostej musi być zastąpione całkowaniem po zbiorze na płaszczyźnie.

Jeśli mamy, odpowiednio, funkcję prawdopodobieństwa oraz funkcję gęstości, to możemy określić łączną dystrybuantę dwóch zmiennych losowych. **Dystrybuantą łączną** dyskretnych zmiennych losowych $X$ i $Y$ nazywamy funkcję

$$F(x,y) \equiv P(X \leq x, Y \leq y) = \sum_{s \leq x} \sum_{t \leq y} f(s,t),$$

gdzie sumowanie odbywa się po wszystkich możliwych wartościach $s$ zmiennej losowej $X$, które są nie większe od $x$, oraz po wszystkich możliwych wartościach $t$ zmiennej losowej $Y$, które są nie większe od $y$. **Dystrybuantą łączną** ciągłych zmiennych losowych $X$ i $Y$ nazywamy funkcję

$$F(x,y) \equiv P(X \leq x, Y \leq y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(s,t)dsdt.$$

Zainteresowanie łącznym rozkładem zmiennych $X$ i $Y$ nie zmienia faktu, że interesujący może też być – i z reguły jest – rozkład każdej z tych zmiennych oddzielnie, gdy wartość drugiej zmiennej jest nam obojętna. W takiej sytuacji mówimy o rozkładzie brzegowym danej zmiennnej. **Rozkład brzegowy** zmiennej losowej $X$ jest dany funkcją prawdopodobieństwa

$$g(x) = \sum_{y} f(x,y),$$

gdy zmienne $X$ oraz $Y$ są dyskretne i $f(x,y)$ jest ich łączną funkcją prawdopodobieństwa (sumowanie odbywa się po wszystkich możliwych wartościach $y$ zmiennej $Y$) oraz funkcją gęstości

$$g(x) = \int_{-\infty}^{\infty} f(x,y)dy,$$

gdy zmienne $X$ oraz $Y$ są ciągłe i $f(x,y)$ jest gęstością łączną zmiennych $X$ i $Y$.

Podobnie, rozkład brzegowy zmiennej losowej $Y$ jest dany funkcją

$$h(y) = \begin{cases} \sum_{x} f(x,y), & \text{gdy } X \text{ i } Y \text{ są dyskretne} \\ \int_{-\infty}^{\infty} f(x,y)dx, & \text{gdy } X \text{ i } Y \text{ są ciągłe.} \end{cases}$$

**Przykład 2.24 cd.** Rozkład brzegowy zmiennej losowej $X$ jest dany funkcją prawdopodobieństwa

$$g(x) = P(X = x) = \sum_{y=0}^{3} f(x,y) = \frac{1}{30}\sum_{y=0}^{3}(x+y) = \frac{1}{15}(2x+3)$$

dla $x = 0, 1, 2$ oraz $g(x) = 0$ w przypadku przeciwnym. Podobnie,

$$h(y) = P(Y = y) = \sum_{x=0}^{2} f(x,y) = \frac{1}{10}(y+1)$$

dla $y = 0, 1, 2, 3$ oraz $h(y) = 0$ w przypadku przeciwnym. Rozkład brzegowy możemy łatwo przedstawić, korzystając z tablicy kontyngencji. Zauważmy mianowicie, że sumując prawdopodobieństwa w wierszach, otrzymujemy prawdopodobieństwa brzegowe zmiennej losowej $X$; sumując prawdopodobieństwa w wierszu odpowiadającym wartości $X = 0$, otrzymujemy $P(X = 0)$ itd. Podobnie, sumując prawdopodobieństwa w kolumnach, otrzymujemy odpowiednie prawdopodobieństwa rozkładu brzegowego zmiennej losowej $Y$. Wyniki są zebrane w poniższej tablicy:

| | $Y$ | | | | |
|---|---|---|---|---|---|
| $X$ | 0 | 1 | 2 | 3 | |
| 0 | 0 | 1/30 | 1/15 | 1/10 | 1/5 |
| 1 | 1/30 | 1/15 | 1/10 | 2/15 | 1/3 |
| 2 | 1/15 | 1/10 | 2/15 | 1/6 | 7/15 |
| | 1/10 | 1/5 | 3/10 | 2/5 | |

Łatwo wykazać, że wartości funkcji prawdopodobieństwa $g(x)$ dla $x = 0, 1, 2$ oraz wartości funkcji $h(y)$ dla $y = 0, 1, 2, 3$, wynikające ze wzorów analitycznych, są równe odpowiednim wartościom prawdopodobieństw brzegowych w tablicy.

Nierzadko interesuje nas rozkład jednej zmiennej losowej, gdy wiemy jaką wartość przyjęła druga zmienna. Mówimy wówczas o rozkładzie warunkowym zmiennej losowej. Załóżmy, że funkcja $f(\cdot, \cdot)$ jest łączną funkcją prawdopodobieństwa zmiennych $X$ i $Y$, gdy te zmienne losowe są dyskretne, lub że jest łączną gęstością prawdopodobieństwa zmiennych $X$ i $Y$, gdy te zmienne są ciągłe. Załóżmy też, że funkcja $h(\cdot)$ określa rozkład brzegowy zmiennej $Y$ (czyli, że jest, odpowiednio, brzegową funkcją prawdopodobieństwa lub brzegową gęstością zmiennej $Y$), przy czym $h(y) > 0$ dla pewnej ustalonej wartości $y$. **Rozkład warunkowy** zmiennej losowej $X$ pod warunkiem, że zmienna losowa $Y$ przyjęła wartość $y$, czyli że $Y = y$, jest dany funkcją

$$f(x|y) = \frac{f(x, y)}{h(y)}. \tag{2.32}$$

Jeżeli zmienna losowa $X$ jest dyskretna, funkcja $f(\cdot|y)$ jest (warunkową) funkcją prawdopodobieństwa zmiennej $X$ pod warunkiem, że zmienna $Y$ przyjęła wartość $y$. Jeżeli zmienna $X$ jest ciągła, funkcja $f(\cdot|y)$ jest (warunkową) gęstością prawdopodobieństwa zmiennej $X$ pod warunkiem, że $Y = y$. W obydwu przypadkach $f(\cdot|y)$ jest funkcją jednego argumentu (argumentu $x$), opisującą rozkład warunkowy zmiennej losowej $X$.

Zauważmy, że w przypadku dyskretnym funkcja prawdopodobieństwa warunkowego $f(\cdot|y)$ dokładnie odpowiada prawdopodobieństwu warunkowemu zdarzenia $A$ pod warunkiem, że zaszło zdarzenie $B$, $P(A|B)$. Mianowicie, funkcja ta opisuje prawdopodobieństwo przyjęcia przez zmienną losową $X$ wartości $x$ (jest to zdarzenie $A$) pod warunkiem, że zmienna losowa $Y$ przyjęła wartość $y$ (jest to zdarzenie $B$). W przypadku ciągłych zmiennych losowych otrzymanie takiej analogii dla gęstości warunkowej $f(\cdot|y)$ jest bardziej złożone (patrz zad. 2.21).

Czytelnikowi pozostawiamy podanie funkcji (prawdopodobieństwa oraz gęstości), opisujących rozkład warunkowy zmiennej losowej $Y$ pod warunkiem, że zmienna losowa $X$ przyjęła wartość $x$.

**Przykład 2.24 cd.** Łatwo obliczyć, że

$$f(x|y) = \begin{cases} \frac{\frac{1}{30}(x+y)}{\frac{1}{10}(y+1)} = \frac{x+y}{3(y+1)}, & \text{gdy } x = 0,1,2 \\ 0, & \text{w przypadku przeciwnym.} \end{cases}$$

Na przykład, gdy $Y = 2$,

$$f(x|2) = P(X = x|Y = 2) = \frac{x+2}{9},$$

gdy $x = 0,1,2$ oraz $f(x|2) = 0$, w przypadku przeciwnym. Stąd

$$P(X = 0|Y = 2) = \frac{2}{9}, \quad P(X = 1|Y = 2) = \frac{1}{3}, \quad P(X = 2|Y = 2) = \frac{4}{9}.$$

Wszystkie prawdopodobieństwa warunkowe łatwo jest także obliczyć korzystając z tablicy kontyngencji z dołączoną kolumną prawdopodobieństw brzegowych zmiennej losowej $X$ oraz dołączonym wierszem takich prawdopodobieństw zmiennej $Y$. Wystarczy w tym celu odczytaną w tablicy wartość prawdopodobieństwa łącznego podzielić przez odpowiednią wartość prawdopodobieństwa warunkowego.

W dalszym ciągu funkcje prawdopodobieństwa oraz funkcje gęstości niekiedy opatrywać będziemy wspólną nazwą **rozkładu zmiennej losowej** lub **(łącznego) rozkładu zmiennych losowych**. Umożliwi to jednoczesne omawianie przypadku dyskretnego i ciągłego.

Dwie zmienne losowe $X$ i $Y$ o łącznym rozkładzie $f(\cdot,\cdot)$ nazywamy **niezależnymi** wtedy i tylko wtedy, gdy dla wszystkich par uporządkowanych $(x,y)$ z zakresu wartości zmiennej losowej $X$ oraz zmiennej losowej $Y$

$$f(x,y) = g(x)h(y), \tag{2.33}$$

gdzie $g(\cdot)$ jest rozkładem brzegowym zmiennej losowej $X$ i $h(\cdot)$ jest rozkładem brzegowym zmiennej $Y$. Taka definicja niezależności dwóch zmiennych losowych jest w pełni zgodna z wcześniej wprowadzoną definicją niezależnych zdarzeń.

By ten fakt zilustrować, posłużymy się najpierw prostym przykładem. Gdy rozważaliśmy dwukrotny rzut uczciwą monetą, obliczaliśmy prawdopodobieństwo uzyskania dwóch orłów, mnożąc prawdopodobieństwo uzyskania orła w pierwszym rzucie przez prawdopodobieństwo takiego samego zdarzenia w drugim rzucie (zadanie było bardzo proste, ponieważ sprowadzało się do działania $(1/2)(1/2) = 1/4$). Dokonamy teraz tego samego obliczenia, posługując się zmiennymi losowymi oraz ich rozkładami. Niech $X$ będzie zmienną losową związaną z pierwszym rzutem i niech $X = 1$, gdy w pierwszym rzucie wypada orzeł oraz $X = 0$, gdy wypada reszka. Oczywiście,

$$g(1) = P(X = 1) = 1/2 \text{ oraz } g(0) = P(X = 0) = 1/2,$$

gdzie $g(\cdot)$ jest funkcją prawdopodobieństwa rozkładu brzegowego zmiennej losowej $X$.

(Dla ścisłości dodajmy, że w naszym eksperymencie przestrzeń zdarzeń elementarnych ma postać

$$\mathcal{S} = \{\text{(orzeł, orzeł), (orzeł, reszka), (reszka, orzeł), (reszka, reszka)}\}$$

oraz

$$X(\text{(orzeł, orzeł)}) = X(\text{(orzeł, reszka)}) = 1$$

i

$$X(\text{(reszka, orzeł)}) = X(\text{(reszka, reszka)}) = 0,$$

czyli $X$ jest rzeczywiście zmienną losową, określoną na przestrzeni $\mathcal{S}$.)

Niech, analogicznie, $Y$ będzie zmienną losową przyjmującą wartość 1, gdy w drugim rzucie wypada orzeł i $Y = 0$, gdy w drugim rzucie wypada reszka. Zauważmy, że funkcję prawdopodobieństwa łącznego obliczamy, **korzystając z założenia niezależności zmiennych losowych $X$ i $Y$, czyli korzystając ze wzoru (2.33)**:

$$\begin{aligned} f(1,1) = g(1)h(1) &= (1/2)(1/2) = 1/4, \\ f(1,0) = g(1)h(0) &= (1/2)(1/2) = 1/4 \text{ itd.}, \end{aligned}$$

gdzie $g(\cdot)$ i $h(\cdot)$ są funkcjami prawdopodobieństwa rozkładu brzegowego zmiennej losowej $X$ i zmiennej losowej $Y$.

W ogólności, odwołując się do ciągłego rozkładu łącznego i przyjąwszy, że zdarzenie $A$ oznacza przyjęcie przez zmienną losową $X$ dowolnej wartości z przedziału $[a, b]$ oraz zdarzenie $B$ oznacza przyjęcie przez zmienną losową $Y$ dowolnej wartości z przedziału $[c, d]$, z podanej definicji niezależności zmiennych losowych otrzymujemy

$$\begin{aligned} P(A \cap B) &= P(X \in [a,b], Y \in [c,d]) = \\ &= \int_c^d \int_a^b f(x,y)dxdy = \int_c^d \int_a^b g(x)h(y)dxdy = \\ &= \int_a^b g(x)dx \int_c^d h(y)dy = P(A)P(B), \end{aligned}$$

tak jak tego wymagamy w przypadku niezależności zdarzeń. Czytelnikowi pozostawiamy udowodnienie analogicznej własności dla pary dyskretnych zmiennych losowych.

Założywszy dodatkowo, że $g(x) > 0$ oraz $h(y) > 0$, niezależność (dyskretnych lub ciągłych) zmiennych losowych $X$ i $Y$ implikuje (por. (2.32) i (2.33))

$$f(x|y) = g(x) \text{ oraz } f(y|x) = h(y).$$

Jak tego należało sobie życzyć, przy założeniu niezależności zmiennych $X$ i $Y$, rozkład warunkowy zmiennej $X$ pod warunkiem, że zmienna $Y$ przyjęła ustaloną wartość $y$ jest równy rozkładowi brzegowemu zmiennej $X$; innymi słowy, fakt przyjęcia przez zmienną $Y$ jakiejś ustalonej wartości nie ma wpływu na rozkład zmiennej $X$ (analogiczna własność dotyczy rozkładu warunkowego zmiennej $Y$).

**Przykład 2.24 cd.** Pozostawiamy Czytelnikowi wykazanie, że

$$g(x)h(y) = \frac{1}{150}(2xy + 2x + 3y + 3),$$

$x = 0, 1, 2$, $y = 0, 1, 2, 3$, czyli że badane dyskretne zmienne losowe nie są niezależne.

Po dokładniejszym rozważeniu pojęcia niezależności pary zmiennych losowych natychmiast zauważymy, że na przykład gęstość

$$f(x,y) = \begin{cases} 8xy, & \text{gdy } 0 < x < y < 1 \\ 0, & \text{w przeciwnym przypadku} \end{cases}$$

nie może być gęstością pary niezależnych zmiennych losowych, mimo że gęstość ta ma postać pozornie sugerującą niezależność zmiennych $X$ i $Y$ (por.

zad. 2.22). Mianowicie, jeżeli zmienna losowa $Y$ przyjęła jakąś wartość $y$, to wiadomo, że zmienna $X$ musiała przyjąć wartość mniejszą od $y$ i przeto zmienna $X$ nie jest niezależna od zmiennej $Y$. Wniosek, płynący z podanego przykładu, możemy w języku matematyki i na podstawie formalnej definicji niezależności zmiennych losowych sformułować następująco. Jeżeli rozkład brzegowy zmiennej losowej $X$ przyjmuje wartości dodatnie na zbiorze $\mathcal{A}$, natomiast rozkład brzegowy zmiennej $Y$ przyjmuje wartości dodatnie na zbiorze $\mathcal{B}$, to własność (2.33) musi być spełniona dla każdej pary uporządkowanej $(x, y)$, dla której $x \in \mathcal{A}$ oraz $y \in \mathcal{B}$. W naszym przykładzie $\mathcal{A} = \mathcal{B} = (0, 1)$, ale, jak to już zauważyliśmy, gęstość $f(\cdot, \cdot)$ nie jest dodatnia dla każdej pary $(x, y)$, dla której są spełnione warunki $0 < x < 1$ i $0 < y < 1$.

W podrozdziale 2.2 wskazaliśmy na pożytki, jakie płyną z wprowadzenia momentów (zwykłych i centralnych) jednej zmiennej losowej. Gdy interesuje nas para zmiennych losowych, wartość oczekiwana zmiennej losowej, która jest odpowiednią funkcją zmiennnej losowej $X$ i zmiennej losowej $Y$, umożliwia skonstruowanie pewnego wskaźnika współzależności między zmiennymi $X$ i $Y$.

Niech $p(X, Y)$ będzie ustaloną (rzeczywistą) funkcją zmiennych losowych $X$ i $Y$ o łącznym rozkładzie $f(x, y)$. **Oczekiwaną wartością zmiennej losowej** $p(X, Y)$ nazywamy wielkość

$$\begin{aligned} \mu_{p(X,Y)} &\equiv E[p(X,Y)] = \\ &= \begin{cases} \sum_x \sum_y p(x,y) f(x,y), & \text{gdy } X \text{ i } Y \text{ są dyskretne} \\ \int\limits_{-\infty}^{\infty} \int\limits_{-\infty}^{\infty} p(x,y) f(x,y) dx dy, & \text{gdy } X \text{ i } Y \text{ są ciągłe.} \end{cases} \end{aligned} \tag{2.34}$$

Każdy z momentów pojedynczej zmiennej losowej, powiedzmy zmiennej $X$, może być przedstawiony jako wartość oczekiwana odpowiedniej funkcji $p(X, Y)$. Chcąc na przykład otrzymać wartość oczekiwaną zmiennej losowej $X$ wystarczy za funkcję $p(X, Y)$ przyjąć funkcję $p(X, Y) = X$:

$$\begin{aligned} \mu_X &= E(X) = \\ &= \begin{cases} \sum_x \sum_y x f(x,y) = \sum_x x g(x), & \text{gdy } X \text{ i } Y \text{ są dyskretne} \\ \int\limits_{-\infty}^{\infty} \int\limits_{-\infty}^{\infty} x f(x,y) dx dy = \int\limits_{-\infty}^{\infty} x g(x) dx, & \text{gdy } X \text{ i } Y \text{ są ciągłe,} \end{cases} \end{aligned}$$

gdzie $g(\cdot)$ jest rozkładem brzegowym zmiennej $X$.

Łatwo udowodnić następujące stwierdzenie:

**Stwierdzenie 2.7.** *Jeżeli $c$ jest stałą, to*

$$E[cp(X,Y)] = cE[p(X,Y)].$$

*Jeżeli $p_1(\cdot,\cdot)$ i $p_2(\cdot,\cdot)$ są dwiema funkcjami zmiennych losowych $X$ i $Y$, to*

$$E[p_1(X,Y) + p_2(X,Y)] = E[p_1(X,Y)] + E[p_2(X,Y)].$$

W szczególności, podstawiając $p_1(X,Y) = X$ i $p_2(X,Y) = Y$ otrzymujemy

$$E(X+Y) = E(X) + E(Y). \tag{2.35}$$

Wspomnieliśmy już, że wartość oczekiwaną (2.34) wprowadziliśmy (między innymi) po to, by skonstruować pewien wskaźnik współzależności między dwiema zmiennymi losowymi. Ponieważ wartość oczekiwana (2.34) umożliwia także uzyskanie momentów jednej zmiennej, więc możemy powiedzieć, iż wskaźnik taki potrafimy skonstruować w sposób spójny z zasadą budowy wartości oczekiwanych dla funkcji jednej zmiennej.

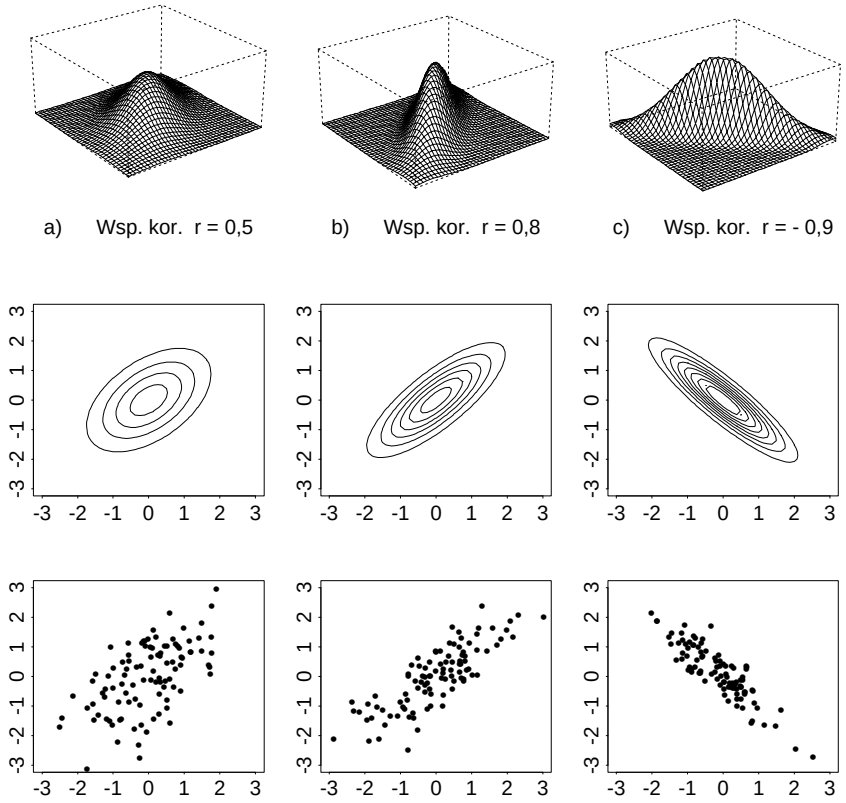
**Definicja 2.17.** *Niech $X$ i $Y$ będą zmiennymi losowymi o łącznym rozkładzie $f(\cdot,\cdot)$.* ***Kowariancją*** *zmiennych $X$ i $Y$ nazywamy wielkość*

$$\sigma_{XY} \equiv E[(X-\mu_X)(Y-\mu_Y)] =$$

$$= \begin{cases} \sum_x \sum_y (x-\mu_X)(y-\mu_Y) f(x,y), & \text{gdy } X \text{ i } Y \text{ są dyskretne} \\ \int\limits_{-\infty}^{\infty} \int\limits_{-\infty}^{\infty} (x-\mu_X)(y-\mu_Y) f(x,y)\,dxdy, & \text{gdy } X \text{ i } Y \text{ są ciągłe,} \end{cases}$$

*gdzie $\mu_X$ i $\mu_Y$ oznaczają odpowiednio wartość średnią (oczekiwaną) zmiennej $X$ i zmiennej $Y$. Zamiast $\sigma_{XY}$ piszemy też $Cov(X,Y)$. Zauważmy, że $\sigma_{XX} = \sigma_X^2$.*

Jeżeli „dużym" wartościom zmiennej $X$ (tzn. wartościom większym od wartości średniej $\mu_X$) towarzyszą zwykle „duże" wartości zmiennej $Y$ (większe od wartości $\mu_Y$) i jeżeli ponadto „małym" wartościom zmiennej $X$ towarzyszą zwykle „małe" wartości zmiennej $Y$ (w obydwu przypadkach chodzi o wartości mniejsze od odpowiedniej wartości średniej), to kowariancja zmiennych $X$ i $Y$ jest dodatnia. Z sytuacją taką mamy wyraźnie do czynienia w przypadku rozkładu z rys. 2.9b i, w mniejszej mierze, w przypadku rozkładu z rys. 2.9a. Jeżeli wartościom zmiennej $X$ większym od $\mu_X$ towarzyszą zwykle wartości zmiennej $Y$ mniejsze od $\mu_Y$ oraz jeżeli wartościom zmiennej $X$




Rys. 2.9. Gęstości dwuwymiarowego rozkładu normalnego oraz warstwice i przykładowe próby losowe dla różnych wartości współczynnika korelacji $r$

mniejszym od $\mu_X$ towarzyszą zwykle wartości zmiennej $Y$ większe od $\mu_Y$, to kowariancja tych zmiennych jest ujemna. Sytuacji tej odpowiada rozkład z rys. 2.9c. W pierwszym przypadku możemy mówić o „dodatniej" zależności między zmiennymi, w drugim o zależności „ujemnej".

Kowariancja umożliwia zatem skonstruowanie wskaźnika mówiącego o istnieniu (lub nieistnieniu) zależności „dodatniej" lub „ujemnej" między zmiennymi losowymi, przy czym owa dodatniość lub ujemność zależności rozumiana jest tak jak to podaliśmy wyżej. Zanim się jednak taką konstrukcją dokładniej zajmiemy, poznamy kilka własności kowariancji. Na mocy definicji

$$\begin{aligned} \mathrm{Cov}(X,Y) &= E(XY - X\mu_Y - Y\mu_X + \mu_X\mu_Y) = \\ &= E(XY) - E(X\mu_Y) - E(Y\mu_X) + \mu_X\mu_Y = \\ &= E(XY) - \mu_X\mu_Y. \end{aligned} \tag{2.36}$$

**Stwierdzenie 2.8.** *Jeżeli zmienne losowe $X$ i $Y$ są niezależne, to*

$$\mathrm{Cov}(X,Y) = 0.$$

Rzeczywiście, teza stwierdzenia 2.9 wynika wprost z równości (2.36) oraz z tego, że na mocy definicji niezależności pary zmiennych losowych (por. zad. 2.28)

$$E(XY) = E(X)E(Y).$$

**Uwaga:** *Stwierdzenie odwrotne do stwierdzenia 2.8 nie jest w ogólności prawdziwe.*

Mianowicie, łatwo jest podać przykłady zależnych zmiennych losowych, których kowariancja jest równa zeru. Niech np. zmienna losowa $X$ będzie ciągła i niech $Y = X^2$. Zmienna losowa $Y$ jest w oczywisty sposób zależna od zmiennej $X$ – obydwie zmienne są wszak związane zależnością funkcyjną, czyli najmocniejszym rodzajem wzajemnej zależności. Ale, jeżeli gęstość $g(\cdot)$ zmiennej losowej $X$ jest funkcją parzystą (czyli, dla każdej wartości $x$, $g(x) = g(-x)$), to nietrudno zauważyć, że $\mathrm{Cov}(X,Y) = 0$:

$$\begin{aligned}
\mathrm{Cov}(X,Y) &= E(XY) - \mu_X\mu_Y = E(X^3) - 0 = \\
&= \int_{-\infty}^{\infty} x^3 g(x)dx = \int_{-\infty}^{0} x^3 g(x)dx + \int_{0}^{\infty} x^3 g(x)dx = \\
&= \int_{-\infty}^{0} x^3 g(x)dx - \int_{-\infty}^{0} x^3 g(x)dx = 0.
\end{aligned}$$

Podany przykład uzmysławia fakt, iż kowariancja może być uznana za podstawę miary zależności określonego typu, takiego mianowicie, gdzie jest sens mówić o zależności dodatniej lub ujemnej we wcześniej podanym sensie. Zależność $Y = X^2$ jest zupełnie innego rodzaju i kowariancja może w takiej sytuacji być zerowa (w tym przypadku, zarówno małym (ujemnym), jak i dużym (dodatnim) wartościom zmiennej $X$ odpowiadają duże wartości zmiennej $Y$).

**Stwierdzenie 2.9.** *Jeżeli $a$ i $b$ są stałymi, to*

$$\mathrm{Var}(aX + bY) = a^2\mathrm{Var}(X) + b^2\mathrm{Var}(Y) + 2ab\mathrm{Cov}(X,Y).$$

Rzeczywiście,

$$\begin{aligned}
\mathrm{Var}(aX + bY) &= E\{[(aX + bY) - (a\mu_X + b\mu_Y)]^2\} = \\
&= E\{[a(X - \mu_X) + b(Y - \mu_Y)]^2\} = \\
&= a^2\mathrm{Var}(X) + 2abE[(X - \mu_X)(Y - \mu_Y)] + b^2\mathrm{Var}(Y).
\end{aligned}$$

Ze stwierdzeń 2.8 i 2.9 wynika następujący wniosek.

**Wniosek.** *Jeżeli zmienne losowe $X$ i $Y$ są niezależne, to*

$$\mathrm{Var}(aX + bY) = a^2\mathrm{Var}(X) + b^2\mathrm{Var}(Y).$$

Zauważyliśmy już, że kowariancja umożliwia skonstruowanie miary pewnego typu zależności między zmiennymi. Sama kowariancja za taką miarę nie może być uznana, ponieważ jest wielkością zależną od jednostek, w jakich są wyrażone zmienne $X$ oraz $Y$. Jeżeli kowariancja dwóch zmiennych losowych wyrażonych w centymetrach wynosi np. 300, to kowariancja zasadniczo tych samych zmiennych, ale wyrażonych w metrach wynosi 0,03. Wielkości kowariancji dla różnych par zmiennych losowych są nieporównywalne. Wspomnianej wady nie ma *współczynnik korelacji.*

**Definicja 2.18.** ***Współczynnikiem korelacji*** *między zmiennymi losowymi $X$ i $Y$ nazywamy wielkość*

$$\rho = \frac{\mathrm{Cov}(X, Y)}{[\mathrm{Var}(X)]^{1/2}[\mathrm{Var}(Y)]^{1/2}}.$$

Współczynnik korelacji $\rho$ między zmiennymi losowymi $X$ i $Y$ ma następujące ważne własności:

(1)
$$-1 \leq \rho \leq 1.$$

(2) Jeżeli $a$ i $b$ są stałymi, przy czym $b > 0$, oraz jeżeli

$$Y = a + bX,$$

to

$$\rho = 1.$$

(3) Jeżeli $a$ i $b$ są stałymi, przy czym $b < 0$, oraz jeżeli

$$Y = a + bX,$$

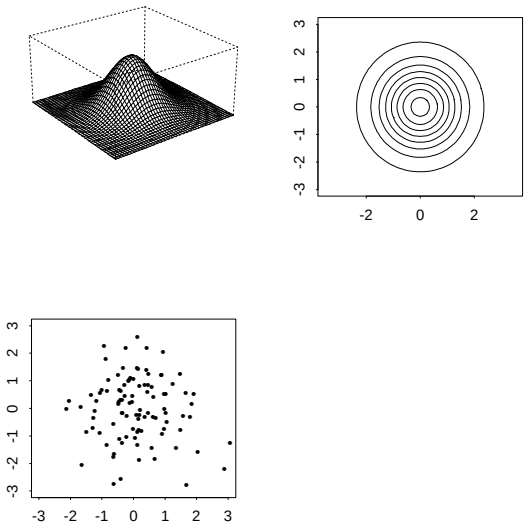
to

$$\rho = -1.$$

(4) Jeżeli zmienne losowe $X$ i $Y$ są niezależne, to

$$\rho = 0.$$

Współczynnik korelacji otrzymaliśmy przez stosowną normalizację kowariancji. Współczynnik ten osiąga wartość maksymalną, równą co do wartości




Rys. 2.10. Gęstość dwuwymiarowego rozkładu normalnego $N(0,0,1,1,0)$ oraz warstwice i przykładowa próba losowa

bezwzględnej 1, gdy zmienne losowe $X$ i $Y$ są związane liniową zależnością funkcyjną. Możemy zatem powiedzieć, że $\rho$ jest miarą zależności dodatniej lub ujemnej między zmiennymi losowymi (w sensie wcześniej już sprecyzowanym), wskazującej na „bliskość" tej zależności z zależnością liniową, gdy wartość $\rho$ jest bliska co do wartości bezwzględnej jedynce. Na rysunku 2.9 i 2.10 są pokazane łączne gęstości oraz warstwice łącznych gęstości pary zmiennych losowych $(X,Y)$, odpowiadające różnym wartościom $\rho$; na tym samym rysunku są również pokazane wylosowane z tych gęstości wartości zmiennych losowych $(X,Y)$. Dokładniejszą interpretację współczynnika korelacji znajdzie Czytelnik w rozdz. 4.

Kończąc omówienie łącznych rozkładów par zmiennych losowych przedstawimy tylko jeden, ale bardzo ważny, przykład ciągłego rozkładu takiej pary, a mianowicie dwuwymiarowy rozkład normalny, dany łączną gęstością

$$f(x,y) = (2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2})^{-1}e^{-q/2}, \tag{2.37}$$

gdzie $-\infty < x < \infty$, $-\infty < y < \infty$, stałe $\sigma_X$, $\sigma_Y$ i $\rho$ spełniają warunki $\sigma_X > 0$, $\sigma_Y > 0$, $-1 < \rho < 1$, a stała $q$ wynosi

$$q = \frac{1}{1-\rho^2}\Big[(\frac{x-m_X}{\sigma_X})^2 - 2\rho(\frac{x-m_X}{\sigma_X})(\frac{y-m_Y}{\sigma_Y}) + (\frac{y-m_Y}{\sigma_Y})^2\Big]$$

oraz $m_X$ i $m_Y$ są dowolnymi stałymi. Gęstość (2.37) nazywamy **dwuwy-**

**miarową gęstością normalną** i mówimy, że zmienne losowe $X$ i $Y$ o łącznym rozkładzie danym tą gęstością mają dwuwymiarowy rozkład normalny, oznaczany $N(\mu_X, \mu_Y, \sigma_X, \sigma_Y, \rho)$.

**Uwaga:** *Można udowodnić, że gęstość brzegowa zmiennej losowej $X$ ma rozkład normalny $N(\mu_X, \sigma_X)$, gęstość brzegowa zmiennej losowej $Y$ ma rozkład normalny $N(\mu_Y, \sigma_Y)$ oraz $\rho$ jest współczynnikiem korelacji między zmiennymi $X$ i $Y$.*

Poniższe twierdzenie orzeka, że w przypadku pary zmiennych losowych o dwuwymiarowym rozkładzie normalnym, zerowanie się współczynnika korelacji implikuje niezależność obydwu zmiennych losowych.

**Twierdzenie 2.11.** *Niech zmienne losowe $X$ i $Y$ mają dwuwymiarowy rozkład normalny. Wówczas zmienne $X$ i $Y$ są niezależne wtedy i tylko wtedy, gdy ich współczynnik korelacji $\rho$ jest równy 0.*

Ostatnie twierdzenie tego podrozdziału orzeka następującą ważną i ciekawą własność pary zmiennych losowych o łącznym rozkładzie normalnym.

**Twierdzenie 2.12.** *Para zmiennych losowych $X$ i $Y$ ma dwuwymiarowy rozkład normalny wtedy i tylko wtedy, gdy każda kombinacja liniowa tych zmiennych, $aX + bY$, gdzie $a$ i $b$ są dowolnymi stałymi, ma rozkład normalny.*

Ciekawą własnością pojedynczych niezależnych zmiennych losowych o rozkładzie normalnym jest fakt, że ich suma jest zmienną losową o rozkładzie normalnym:

**Stwierdzenie 2.10.** *Niech $X$ i $Y$ będą niezależne i mają rozkład normalny, $N(\mu_1, \sigma_1)$ i $N(\mu_2, \sigma_2)$ odpowiednio. Wówczas dla dowolnych liczb $a$ i $b$, z których obie nie są równocześnie 0, $aX + bY$ ma rozkład normalny $N(\mu, \sigma)$, gdzie $\mu = a\mu_1 + b\mu_2$ i $\sigma = \sqrt{a^2\sigma_1^2 + b^2\sigma_2^2}$.*

Oczywiście korzystając z własności kombinacji liniowych niezależnych zmiennych losowych, wiemy, że kombinacja $aX + bY$ ma wartość średnią $\mu = a\mu_1 + b\mu_2$ i wariancję $\sigma^2 = a^2\sigma_1^2 + b^2\sigma_2^2$. To co jest nowe w tym stwierdzeniu, to fakt, że kombinacja liniowa ma również rozkład normalny.

Na początku podrozdziału zaznaczyliśmy, iż ograniczamy się w nim do omówienia łącznego rozkładu tylko dwóch zmiennych losowych, ponieważ uogólnienie rozważań na przypadek większej liczby zmiennych jest proste i naturalne. Ze względu na dalsze zastosowania, kończąc niniejszy podrozdział, podamy tylko uogólnienie równości (2.35) oraz wniosku ze stwierdzenia 2.10

na przypadek $n \geq 2$ zmiennych losowych: jeżeli $a_1, a_2, \ldots, a_n$ są współczynnikami stałymi, to

$$E(a_1X_1 + a_2X_2 + \ldots a_nX_n) = a_1E(X_1) + a_2E(X_2) + \ldots + a_nE(X_n) \quad (2.38)$$

oraz, jeżeli zmienne losowe $X_1, X_2, \ldots, X_n$ są niezależne (niezależność $n$ zmiennych losowych jest naturalnym uogólnieniem pojęcia niezależności dwóch zmiennych losowych), to

$$\mathrm{Var}(a_1X_1 + a_2X_2 + \ldots + a_nX_n) = a_1^2\mathrm{Var}(X_1) + a_2^2\mathrm{Var}(X_2) + \ldots + a_n^2\mathrm{Var}(X_n). \quad (2.39)$$

Na mocy własności (2.38) i (2.39) można łatwo wyprowadzić wzory na wartość średnią i wariancję rozkładu dwumianowego $Bin(n, p)$. Wystarczy w tym celu zauważyć, że zmienna losowa o rozkładzie $Bin(n, p)$ jest sumą $n$ zmiennych losowych o rozkładzie Bernoulliego z prawdopodobieństwem sukcesu $p$.

## 2.4. Wnioskowanie statystyczne – podstawy

### 2.4.1. Podstawowe pojęcia

W podrozdziale 2.2 wprowadziliśmy pojęcie zmiennej losowej i jej rozkładu oraz staraliśmy się uzasadnić, że jest to użyteczna koncepcja do analizy własności cech losowych o wartościach liczbowych. Pokazaliśmy, jakie wskaźniki liczbowe opisują rozkład zmiennej losowej i jaka jest ich interpretacja. Wiemy, że wskaźniki te można z reguły łatwo obliczyć, a przynajmniej znamy odpowiedni ku temu algorytm, jeśli jest znany rozkład badanej zmiennej losowej. Zasadniczy problem polega na tym, że bardzo rzadko ten rozkład znamy dokładnie, gdyż naprawdę interesujące zmienne losowe są z reguły skomplikowane lub określone na przestrzeniach zdarzeń elementarnych o zbyt bogatej strukturze. Tym, czym dysponujemy jest pomiar wartości zmiennej losowej dla pewnej grupy jednostek. W rozdziale pierwszym staraliśmy się pokazać, jak przedstawić informację zawartą w tak uzyskanej próbie wartości. Zasadniczym i dotąd niedyskutowanym zagadnieniem jest jak związać informację z próby z informacją o nieznanym rozkładzie zmiennej losowej i jak oceniać wiarogodność tego powiązania. Tym problemem zajmuje się wnioskowanie statystyczne, któremu poświęcimy dużo miejsca w dalszej części książki.

Zacznijmy od wprowadzenia niezbędnych definicji. Przypuśćmy, że interesuje nas pewien wskaźnik zmiennej losowej $X$, która jest określona na

przestrzeni zdarzeń elementarnych $\mathcal{S}$, nazywanej często w statystyce **populacją**. Taką zmienną może być na przykład wzrost określony dla populacji wszystkich dorosłych Polaków, a interesującym nas wskaźnikiem może być wartość średnia wzrostu. Zamiast niemożliwego do wykonania pomiaru wzrostu wszystkich dorosłych Polaków i sporządzenia rozkładu wzrostu na tej podstawie, dokonujemy pomiaru na pewnej, specjalnie wybranej części populacji. Z reguły część ta jest tak wybrana, aby otrzymać reprezentatywną próbę potencjalnych wartości zmiennej. Nazwiemy ją tutaj prostą próbą losową.

**Definicja 2.19.** ***Prostą próbą losową*** *o liczności $n$ nazywamy ciąg niezależnych zmiennych losowych $X_1, X_2, \ldots, X_n$ określonych na przestrzeni $\mathcal{S}$ i takich, że każda z nich ma ten sam rozkład.*

Zmienna $X_1$ odpowiada potencjalnej wartości dla elementu populacji wybieranego jako pierwszy, $X_2$ dla drugiego elementu itd. Zauważmy, że możemy traktować $X_1$ jako zmienną losową tylko przed faktycznym wylosowaniem elementu próby, natomiast po jego wylosowaniu możemy mówić jedynie o konkretnej wartości zmiennej losowej dla wybranego elementu. Konkretny ciąg wartości $x_1, x_2, \ldots, x_n$ próby losowej $X_1, X_2, \ldots, X_n$ będziemy nazywać **realizacją próby losowej**. Jak wybierać reprezentatywną grupę elementów populacji, aby otrzymać prostą próbę losową? Problem ten omówimy dokładniej w podrozdz. 2.5 i rozdz. 7, tutaj poruszymy tylko kwestię najważniejszą.

*Wybór $n$ elementów populacji powinien być dokonany w taki sposób, żeby każdy podzbiór populacji, składający się z $n$ elementów miał taką samą szansę wybrania.*

Proces ten możemy przyrównać do testowania smaku świeżo ugotowanej zupy: aby mieć o nim dobre pojęcie, należy najpierw zupę bardzo dobrze wymieszać. Wtedy to, co skosztujemy, będzie miało smak adekwatny do smaku całości. W przypadku populacji o $m$ elementach, dla $m \geq n$, powyższy postulat jednakowego prawdopodobieństwa wyboru oznacza, że każdy $n$ elementowy podzbiór ma szansę wybrania równą $1/\binom{m}{n}$. Można zrealizować to tak, że wybieramy jakikolwiek element populacji w taki sposób, że każdy element populacji ma takie samo prawdopodobieństwo wyboru równe $1/m$. Następnie spośród pozostałych $m-1$ elementów wybieramy analogicznie drugi, przypisując mu prawdopodobieństwo wyboru równe $1/(m-1)$, itd. Zastanówmy się jednak dokładniej, czy przedstawiony sposób wyboru elementów populacji prowadzi zawsze do spełnienia def. 2.19. Oczywiście nie! Dobry przykład typowych trudności został omówiony w przykł. 2.19 dotyczącym próbkowania ze skończonej populacji. W tym przypadku prawdopodobieństwo otrzymania elementu wadliwego w $i$-tym kroku zależy od

historii poprzednich ciągnień. Cecha „bycia wadliwym" **nie** jest niezależna dla kolejno pobieranych elementów partii i **nie ma** dla nich tego samego rozkładu. Tak jest również w przypadku ogólnej sytuacji cechy o dowolnie wielu, a niekoniecznie dwu wartościach. Tylko gdy liczność populacji jest bardzo duża, wybór reprezentatywnej próby z populacji zapewnia **przybliżone** spełnienie def. 2.19. Tak więc z reguły definicja prostej próby losowej odpowiada idealizacji sytuacji rzeczywistej. W rozdziale 7 omówimy różne schematy wyboru prób z populacji skończonej: schematy te nie prowadzą do wyboru prostych prób losowych w sensie def. 2.19.

Zauważmy jednak, że bez wyboru reprezentatywnej części populacji nie mamy szans otrzymania prostej próby losowej z interesującego nas rozkładu. Aby to stwierdzić, powróćmy do przykładu oceny wzrostu dorosłych Polaków i załóżmy, że dokonujemy pomiaru wzrostu dla pewnej losowej grupy osób będacych studentami szkół wyższych. Bez względu na to, jak będziemy wybierać studentów, odpowiadająca próba nie będzie prostą próbą losową z rozkładu wzrostu *wszystkich dorosłych Polaków*. Dzieje się tak dlatego, że nie wszystkie jednostki z populacji dorosłych Polaków mają szansę wybrania: osoby, które nie są aktualnie studentami nigdy nie pojawią się w próbie. W dalszej części książki, jeśli nie jest wyraźnie powiedziane, że tak nie jest, będziemy zakładali, że mamy do czynienia z prostą próbą losową z rozkładu interesującej nas zmiennej. Będziemy również utożsamiali element populacji z odpowiadającą mu wartością cechy, gdyż tylko ona jest informacją rejestrowaną w doświadczeniu.

Przypuśćmy teraz, że dysponujemy prostą próbą losową $X_1, X_2, \ldots, X_n$ i na jej podstawie staramy się ocenić pewien wskaźnik rozkładu jednakowego dla wszystkich elementów próby. W omawianym powyżej przykładzie może interesować nas na przykład wartość średnia wzrostu. Oszacujemy ją, obliczając pewną funkcję próby $T(X_1, X_2, \ldots, X_n)$, która w naszym mniemaniu dobrze przybliża wartość nieznanego wskaźnika. Taką funkcję nazywamy **statystyką** i nazwa ta obowiązuje nie tylko w przypadku, gdy mamy do czynienia z prostą próbą losową. Zauważmy, że statystyka jest zmienną losową; odrębna nazwa ma podkreślać jej specjalną funkcję. Do oceny adekwatności użycia tej lub innej statystyki kapitalne znaczenie ma jej rozkład.

### 2.4.2. Rozkład średniej w prostej próbie losowej

Naszą analizę własności statystyk rozpoczynamy od omówienia rozkładu średniej w prostej próbie losowej. Powróćmy do przykł. 1.16 z p. 1.4.2, w którym stwierdziliśmy, że wzrost w populacji dorosłych Polaków jest cechą o rozkładzie normalnym o wartości średniej $\mu = 176$ cm i odchyleniu standardowym $\sigma = 6{,}5$ cm. Załóżmy na chwilę, że nie znamy dokład-

nie wartości średniej i postanowiliśmy ją oszacować na podstawie prostej próby losowej o liczności 100 z populacji wszystkich dorosłych Polaków. Oznaczmy naszą prostą próbę losową przez $X_1, \ldots, X_{100}$. Oczywiście, po pobraniu *prostej próby losowej* dysponujemy konkretnymi wartościami zmiennych $X_1 = x_1, \ldots, X_{100} = x_{100}$, które tworzą *próbę wartości* $x_1, \ldots, x_{100}$. Jak przekonaliśmy się w p. 1.3.1, jednym ze wskaźników położenia dla próby wartości jest wartość średnia w próbie, tutaj $\bar{x} = (x_1 + \ldots + x_{100})/100$. Ponieważ sposób konstrukcji wartości średniej nie zależy od zaobserwowanych wartości, możemy od tych konkretnych wartości abstrahować i przeprowadzić ogólną konstrukcję dla zmiennych losowych, tworząc statystykę $\bar{X} = (X_1 + X_2 + \ldots + X_{100})/100$. Otrzymana statystyka nosi nazwę średniej w prostej próbie losowej.

**Definicja 2.20.** ***Średnią w prostej próbie losowej*** $X_1, \ldots, X_n$ *o liczności* $n$ *nazywamy statystykę*

$$\bar{X} = \frac{(X_1 + X_2 + \ldots + X_n)}{n}. \tag{2.40}$$

Zauważmy, że def. 2.20 jest szczególnym przypadkiem ogólnej definicji statystyki dla funkcji $T$ zdefiniowanej w sposób następujący: $T(X_1, X_2, \ldots, X_n) = (X_1 + X_2 + \ldots + X_n)/n$. Jaki jest cel przeniesienia definicji średniej z poziomu próby wartości na poziom prostej próby losowej? Chodzi o to, że badając rozkład średniej $\bar{X}$, jesteśmy w stanie przeanalizować jej zachowanie dla *potencjalnej* próby w terminach rachunku prawdopodobieństwa. Wartość średnia $\bar{x}$ zależy od konkretnej próby wartości, na podstawie której została obliczona. Aby ją poprawnie zinterpretować, musimy wiedzieć, co to znaczy, że dla jednej próby $\bar{x} = 176{,}5$ a dla innej $\bar{x} = 177{,}8$. Podkreślmy tu jeszcze subtelność terminologiczną: mówimy o średniej $\bar{X}$ w odróżnieniu od *wartości* średniej $\bar{x}$ w próbie w celu zwrócenia uwagi na to, że pierwsza z nich jest zmienną losową, podczas gdy druga z nich jest konkretną liczbą.

Przedyskutujmy jeszcze kwestię, dlaczego rozpatrujemy średnią $\bar{X}$ jako naturalnego kandydata do oszacowania nieznanej wartości średniej $\mu_X$. Intuicyjnie jest oczywiste, że dla dużej liczności próby średnia $\bar{X}$ w prostej próbie losowej powinna dobrze przybliżać wartość średnią $\mu_X$. Jeśli rozpatrzymy zmienną $X_i$ równą 1, gdy w $i$-tym rzucie monetą pojawi się reszka i 0 w przeciwnym przypadku, to średnia $\bar{X}$ dla 1000 rzutów powinna dobrze przybliżać 0,5, czyli wartość średnią każdej zmiennej losowej $X_i$.

**Twierdzenie 2.13. Prawo wielkich liczb.** *Niech $X$ będzie zmienną losową o skończonej wariancji, $\sigma_X^2 < \infty$ i niech $X_1, \ldots, X_n$ będzie prostą próbą losową z rozkładu zmiennej losowej $X$. Wówczas dla dowolnie małej dodatniej liczby $\varepsilon$ prawdopodobieństwo*

$$P\Big(\bar{X} \text{ należy do } [\mu_X - \varepsilon, \mu_X + \varepsilon]\Big)$$

*jest bliskie 1 dla dużych liczności próby $n$.*

Prawo wielkich liczb uzasadnia słuszność użycia średniej $\bar{X}$ jako oszacowania wartości średniej $\mu_X$ w następującym sensie: dla ustalonej, ale dowolnie dużej precyzji $\varepsilon$, jesteśmy w stanie „prawie zawsze" oszacować nieznaną wartość średnią z dokładnością do $\varepsilon$, używając średniej $\bar{X}$, jeśli tylko próba będzie miała dostatecznie dużą liczność.

Przejdźmy teraz do badania podstawowych wskaźników rozkładu wartości średniej $\bar{X}$. Zajmiemy się obliczaniem wartości średniej rozkładu średniej $\bar{X}$ i odchylenia standardowego tego rozkładu. Przypuśćmy, że prosta próba losowa $X_1, \ldots, X_n$ pochodzi z rozkładu o wartości średniej $\mu$ i wariancji $\sigma^2$. Korzystając z własności kombinacji niezależnych zmiennych losowych, otrzymujemy

$$\mu_{\bar{X}} = \frac{1}{n}(\mu_{X_1} + \mu_{X_2} + \ldots + \mu_{X_n}) = \frac{1}{n}(\mu + \mu + \ldots + \mu) = \mu,$$

oraz

$$\sigma_{\bar{X}}^2 = \Big(\frac{1}{n}\Big)^2(\sigma_{X_1}^2 + \sigma_{X_2}^2 + \ldots + \sigma_{X_n}^2) = \frac{\sigma^2}{n}.$$

Otrzymaliśmy zatem następujące stwierdzenie.

**Stwierdzenie 2.11.** *Niech $\bar{X}$ będzie średnią w prostej próbie losowej o liczności $n$ z rozkładu o wartości średniej $\mu$ i wariancji $\sigma^2$. Wówczas*

$$\mu_{\bar{X}} = \mu \qquad i \qquad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}. \tag{2.41}$$

Z pierwszej z własności wynika, że położenie rozkładu średniej $\bar{X}$ pokrywa się z położeniem rozkładu pojedynczej obserwacji, a z drugiej, że rozproszenie rozkładu średniej *zmniejszyło się* $\sqrt{n}$ razy w stosunku do rozproszenia rozkładu pojedynczej obserwacji. Tak więc w naszym przykładzie wartość średnia rozkładu średniej $\bar{X}$ wynosi $\mu_{\bar{X}} = 176$ cm, a jej rozproszenie $\sigma_{\bar{X}} = 6{,}5/\sqrt{100} = 0{,}65$ cm. Zatem rozproszenie zmniejszyło się dziesięciokrotnie w porównaniu z rozproszeniem pojedynczej obserwacji! Na tym właśnie polega główna zaleta uśredniania: umożliwia znaczne zredukowanie zmienności oszacowania interesującego nas wskaźnika rozkładu cechy.

Zauważmy, że własność ta powinna być spełniona dla każdego sensownego oszacowania: dla większej liczności próby powinniśmy otrzymywać lepsze oszacowanie.

**Przykład 1.16 cd.** Obliczmy prawdopodobieństwo, że średnia $\bar{X}$ dla prostej próby losowej o liczności 100 różni się od prawdziwej wartości $\mu$ o więcej niż 1,5 cm.

Interesuje nas zatem prawdopodobieństwo, że średnia $\bar{X}$ znajdzie się poza przedziałem $[\mu - 1{,}5\,, \mu + 1{,}5]$. Wiemy, że wartość średnia $\mu_{\bar{X}} = \mu$ i rozproszenie $\sigma_{\bar{X}} = 0{,}65$ cm. Te informacje nie wystarczą jednak do obliczenia poszukiwanego prawdopodobieństwa. Wykorzystamy dodatkową informację o normalności rozkładu wzrostu. Ponieważ średnia $\bar{X}$ jest kombinacją liniową zmiennych losowych $X_1, \ldots, X_n$ o rozkładzie normalnym, więc sama ma również rozkład normalny. Wiemy, że wartość średnia tego rozkładu wynosi $\mu$, a rozproszenie 0,65 cm, zatem średnia $\bar{X}$ ma rozkład normalny $N(\mu, 0{,}65)$. Standaryzując średnią $\bar{X}$, otrzymamy

$$\begin{aligned} P(|\bar{X} - \mu| &> 1{,}5) = P(\bar{X} - \mu > 1{,}5 \text{ lub } \bar{X} - \mu < -1{,}5) = \\ &= P(\bar{X} - \mu > 1{,}5) + P(\bar{X} - \mu < -1{,}5) = \\ &= P\Big(\frac{\bar{X} - \mu}{0{,}65} > \frac{1{,}5}{0{,}65}\Big) + P\Big(\frac{\bar{X} - \mu}{0{,}65} < \frac{-1{,}5}{0{,}65}\Big) = \\ &= P(Z > 2{,}31) + P(Z < -2{,}31) = 2\Phi(-2{,}31) = 0{,}0208, \end{aligned}$$

gdzie $Z$ jest zmienną losową o standardowym rozkładzie normalnym. Zauważmy, że dla pojedynczej obserwacji $X_1$ prawdopodobieństwo analogicznej odchyłki wynosi

$$\begin{aligned} P(|X_1 - 176| > 1{,}5) &= P(X_1 - 176 > 1{,}5 \text{ lub } X_1 - 176 < -1{,}5) = \\ &= P\Big(\frac{X - 176}{6{,}5} > \frac{1{,}5}{6{,}5}\Big) + P\Big(\frac{X - 176}{6{,}5} < \frac{-1{,}5}{6{,}5}\Big) = \\ &= 2 \times P(Z < -0{,}231) = 0{,}8180. \end{aligned}$$

W powyższych obliczeniach korzystaliśmy z informacji, że średnia wzrostu w prostej próbie losowej ma rozkład normalny wynikający z normalności rozkładu poszczególnych obserwacji. Założyliśmy też znajomość wariancji pojedynczej obserwacji.

Okazuje się, że choć bez informacji o rozkładzie obserwacji nie jesteśmy z reguły w stanie obliczyć dokładnie poszukiwanego prawdopodobieństwa, to nie jesteśmy również całkowicie bezradni. Możemy wykorzystać interesujący fakt, że wraz ze wzrostem liczności próby $n$ rozkład średniej $\bar{X}$ coraz

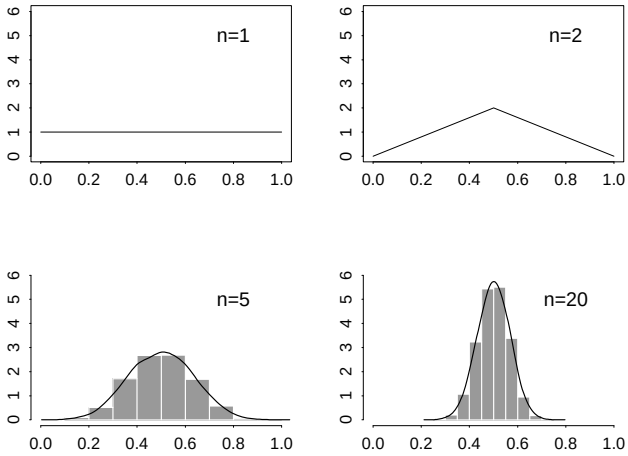
mocniej „zapomina" jaki był kształt rozkładu dla pojedynczej zmiennej i zaczyna coraz bardziej upodabniać się do rozkładu normalnego. Zachowanie to zilustrowane jest na rysunku. Na rysunku 2.11a przedstawiono gęstość rozkładu jednostajnego, czyli gęstość rozkładu średniej $\bar{X}$ dla jednej obserwacji z tego rozkładu, a na rysunku 2.11b-d przedstawiono odpowiednio rozkład średniej dla $n = 2, 5$ i 20 obserwacji. Widzimy, że dla $n = 5$ rozkład bardzo przypomina rozkład normalny, a dla $n = 20$ jest od niego praktycznie nieodróżnialny. Podobnie wygląda sytuacja dla większości rozkładów, nawet bardzo skośnych. Co najwyżej może zwiększyć się liczba obserwacji, przy której rozkład średniej $\bar{X}$ zaczyna przypominać rozkład normalny, *jednak praktycznie nigdy nie przekracza ona 25.* Na rysunku 2.11 przedstawiono działanie Centralnego Twierdzenia Granicznego zwanego również przybliżeniem normalnym.

**Twierdzenie 2.14. Centralne Twierdzenie Graniczne.** *Niech $X_1, \ldots, X_n$ będzie prostą próbą losową z rozkładu o średniej $\mu$ i skończonej wariancji $\sigma^2$. Wówczas dla dużych liczności próby $n$ rozkład standaryzowanej średniej $(\bar{X} - \mu)/(\sigma/\sqrt{n})$ jest bliski standardowemu rozkładowi normalnemu $N(0,1)$. Mianowicie, dla dowolnych liczb $a$ i $b$, $a \leq b$, i zmiennej losowej $Z$ o standardowym rozkładzie normalnym*

$$P\left(a \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq b\right) \quad \to \quad P(a \leq Z \leq b) = \Phi(b) - \Phi(a) \tag{2.42}$$

*dla $n$ dążących do nieskończoności. Równoważnie, rozkład średniej $\bar{X}$ jest w przybliżeniu równy rozkładowi normalnemu $N(\mu, \sigma/\sqrt{n})$.*

A zatem na podstawie powyższych rozważań i przybliżenia normalnego stwierdzamy, że dla prostej próby losowej z rozkładu normalnego jesteśmy zawsze w stanie, korzystając ze stwierdzenia 2.11, wyznaczyć dokładny rozkład średniej $\bar{X}$. Jest to rozkład normalny $N(\mu, \sigma/\sqrt{n})$. Bez informacji, że próba pochodzi z rozkładu normalnego, możemy stwierdzić, że rozkład średniej jest *w przybliżeniu* równy temu samemu rozkładowi normalnemu $N(\mu, \sigma/\sqrt{n})$, jeśli tylko liczność próby wynosi co najmniej 25. Sytuacja najbardziej kłopotliwa występuje dla mało licznej próby nie pochodzącej z rozkładu normalnego. Wtedy rozkład średniej może znacznie odbiegać od rozkładu normalnego. Możemy wtedy uciec się do oszacowania prawdopodobieństw odpowiadających ogonom rozkładu średniej $\bar{X}$, korzystając z nierówności Czebyszewa. Zaznaczmy również, że gdy jest znany rozkład, z którego pochodzi prosta próba losowa, rozkład średniej można wyznaczyć metodami symulacyjnymi: taka metoda zostanie omówiona w podrozdz. 8.3.




Rys. 2.11. Gęstości rozkładu (oraz wygenerowane histogramy) średniej $\bar{X}$ dla $n = 1, 2, 5, 20$ obserwacji z rozkładu jednostajnego

**Przykład 2.25.** Przyjmijmy, że rozkład naszego codziennego dojazdu do pracy jest w przybliżeniu rozkładem jednostajnym na odcinku [0,5 godz., 1 godz.] i że czasy dojazdów w różne dni są niezależne. Ile w przybliżeniu wynosi prawdopodobieństwo zdarzenia, że średni dzienny dojazd w ciągu 30 dni przekroczy 0,8 godz. = 48 min (lub równoważnie, że dojazdy w ciągu 30 dni zajmą mi więcej niż dobę)?

Oznaczmy przez $X_i$ czas dojazdu w $i$-tym dniu, $i = 1, \ldots, 30$. Przyjmijmy, że $X_i$ ma rozkład jednostajny na przedziale $[0{,}5, 1]$, a zatem

$$\mu_{X_i} = \frac{0{,}5+1}{2} = \frac{3}{4} \quad \text{i} \quad \sigma^2_{X_i} = \frac{(1-0{,}5)^2}{12} = \frac{1}{48}.$$

Interesuje nas prawdopodobieństwo $P(\bar{X} > 0{,}8)$. Wiemy z Centralnego Twierdzenia Granicznego, że średnia $\bar{X}$ ma w przybliżeniu rozkład normalny $N(\frac{3}{4}, \sqrt{\frac{1}{30\times 48}})$. Zatem standaryzując ją otrzymamy, że powyższe prawdopodobieństwo jest równe

$$P\left(\frac{\bar{X} - 3/4}{\sqrt{\frac{1}{30\times 48}}} > \frac{0{,}8 - 3/4}{\sqrt{\frac{1}{30\times 48}}}\right) \approx P(Z > 1{,}89) = 0{,}03,$$

gdzie $Z$ oznacza zmienną losową o rozkładzie standardowym normalnym.

**Poprawka w przybliżeniu normalnym.** Zauważmy, że w Centralnym Twierdzeniu Granicznym nie założyliśmy ciągłości rozkładu zmiennych $X_i$. W przypadku gdy zmienne $X_i$ w prostej próbie losowej przyjmują wartości całkowite, dokonuje się poprawki w przybliżeniu normalnym, opartej na następującym rozumowaniu. Jeśli każda wartość w próbie jest liczbą całkowitą, to suma $\sum_{i=1}^{n} X_i$ przyjmuje tylko wartości będące liczbami całkowitymi i dlatego dla dowolnych liczb *całkowitych* $a$ i $b$ możemy napisać

$$P(a \le \sum_{i=1}^{n} X_i \le b) = P(a - 0{,}5 \le \sum_{i=1}^{n} X_i \le b + 0{,}5). \tag{2.43}$$

Standaryzując sumę $\sum_{i=1}^{n} X_i$, otrzymamy

$$\begin{aligned} P(a - 0{,}5 \;&\le\; \sum_{i=1}^{n} X_i \le b + 0{,}5) = \\ &= P\Big(\frac{a - 0{,}5 - n\mu_X}{\sqrt{n}\sigma_X} \le \frac{\sum_{i=1}^{n} X_i - n\mu_X}{\sqrt{n}\sigma_X} \le \frac{b + 0{,}5 - n\mu_X}{\sqrt{n}\sigma_X}\Big) = \\ &\approx \Phi\Big(\frac{b + 0{,}5 - n\mu_X}{\sqrt{n}\sigma_X}\Big) - \Phi\Big(\frac{a - 0{,}5 - n\mu_X}{\sqrt{n}\sigma_X}\Big). \end{aligned}$$

Dlaczego wprowadzamy poprawkę do przybliżenia normalnego? Zauważmy, że przyjęcie $b = a$ oraz rezygnacja z poprawki $\pm 0{,}5$ w równości (2.43) doprowadziłyby nas do przybliżenia prawdopodobieństwa $P(\sum X_i = a)$ przez 0, ponieważ graniczny rozkład normalny jest ciągły.

Zauważmy, że na mocy powyższego rozumowania (por. p. 2.4.3)

$$\begin{aligned} P\Big(\frac{(a - 0{,}5)}{n} \le \bar{X} \le \frac{(b + 0{,}5)}{n}\Big) &\approx \\ \approx \Phi\Big(\frac{b + 0{,}5 - n\mu_X}{\sqrt{n}\sigma_X}\Big) - \Phi\Big(\frac{a - 0{,}5 - n\mu_X}{\sqrt{n}\sigma_X}\Big), \end{aligned} \tag{2.44}$$

tak więc wyrażenie (2.44) może być używane do szacowania prawdopodobieństw związanych ze średnią. Powyższa poprawka ma znaczenie przede wszystkim wtedy, gdy zależy nam na dokładnym oszacowaniu prawdopodobieństwa ogona rozkładu średniej. Taka sytuacja może na przykład wystąpić w ubezpieczeniach, gdy interesuje nas prawdopodobieństwo, że łączna suma wypłat z tytułu ubezpieczeń przekroczy pewien pułap.

**Przykład 2.26.** Rzucamy 30 razy kostką. Oszacujmy prawdopodobieństwo, że suma wyrzuconych oczek jest liczbą między 100 a 110.

Niech $X_i$ będzie zmienną losową równą liczbie oczek wyrzuconych w $i$-tym rzucie. Oczywiście $X_1, X_2, \ldots, X_{30}$ jest prostą próbą losową z rozkładu jednostajnego na zbiorze $\{1, 2, \ldots, 6\}$. Z zadania 2.13 wynika, że ich wspólna wartość oczekiwana $\mu$ jest równa 3,5, a wariancja $\sigma^2$ jest równa 35/12. Niech $X$ będzie sumą wyrzuconych oczek, $X = \sum_{i=1}^{30} X_i$. Oczywiście, $\mu_X = 30 \times 3{,}5 = 105$ i $\sigma_X^2 = 30 \times 35/12$. Interesuje nas prawdopodobieństwo

$$P(100 \leq X \leq 110) = P(99{,}5 \leq X \leq 110{,}5).$$

Standaryzując i stosując przybliżenie normalne z poprawką, otrzymujemy, że powyższe prawdopodobieństwo jest równe

$$P\left(\frac{99{,}5 - 105}{\sqrt{30\frac{35}{12}}} \leq \frac{X - 105}{\sqrt{30\frac{35}{12}}} \leq \frac{110{,}5 - 105}{\sqrt{30\frac{35}{12}}}\right) \approx \Phi(0{,}59) - \Phi(-0{,}59) = 2\Phi(0{,}59) - 1 = 0{,}4448.$$

Załóżmy teraz, że rzucamy kostką 10 razy i interesuje nas prawdopodobieństwo, że średnia $\bar{X}$ będzie odbiegała od swojej wartości średniej $\mu_{\bar{X}} = 3{,}5$ o więcej niż 1,5. Ze względu na zbyt małą liczność próby nie możemy zastosować Centralnego Twierdzenia Granicznego. Po skorzystaniu z nierówności Czebyszewa otrzymamy jednak oszacowanie nieznanego prawdopodobieństwa

$$P(|\bar{X} - 3{,}5| > 1{,}5) \leq \frac{\sigma_{\bar{X}}^2}{(1{,}5)^2} = \frac{\frac{35}{12 \times 10}}{(1{,}5)^2} = 0{,}13.$$

Powyższe oszacowanie ma tę zaletę, że można je stosować nie tylko w przypadku rozkładu liczby wyrzuconych oczek, ale ogólnie w przypadku dowolnego rozkładu pojedynczej obserwacji, jeśli ma on tylko skończoną wariancję.

### 2.4.3. Rozkład częstości

Przedyskutujmy teraz przypadek szczególny sytuacji opisanej w poprzednim punkcie, gdy zmienna losowa $X$, z rozkładu której pochodzi rozpatrywana próba losowa, przyjmuje tylko dwie wartości, 1 i 0. Oznaczmy $p = P(X = 1)$ i $q = 1 - p = P(X = 0)$. Przypadek ten odpowiada sytuacji, gdy część jednostek w populacji ma pewną własność i dla nich zmienna losowa $X$ przyjmuje wartość 1, a dla pozostałej części wartość $X$ jest równa 0. Na przykład rozpatrywaną cechą może być ukończenie studiów

rozpatrywane w populacji polskiej i $X$ jest równe 1, jeśli rozpatrywana osoba ukończyła studia. Liczba $p$, zwana **proporcją**, jest równa prawdopodobieństwu posiadania rozpatrywanej własności przez losowo wybraną jednostkę. W tym przypadku $\mu_X = 1 \times p + 0 \times (1 - p) = p$, a zatem rozpatrzony poprzednio problem szacowania wartości średniej jest w tym konkretnym przypadku problemem szacowania proporcji. Typowe przykłady, w których może interesować nas szacowanie proporcji, to:

– populacja elementów wyprodukowanych w pewnej fabryce w danym miesiącu, z których pewna proporcja $p$ jest wadliwa. Zmienna $X_i$ jest równa 1, jeśli $i$-ty element jest wadliwy i jest równa 0 w przeciwnym przypadku;

– populacja Polaków, z których pewna proporcja $p$ jest leworęczna;

– populacja studentów szkół wyższych w Polsce, z których pewna proporcja $p$ nie zdała przynajmniej jednego egzaminu w poprzednim semestrze.

Przeanalizujmy postać średniej $\bar{X}$ w prostej próbie losowej $X_1, X_2, \ldots, X_n$ w tym przypadku. Jest ona równa częstości występowania własności w prostej próbie losowej. Oznaczamy ją tradycyjnie przez $\hat{p}$ i nazywamy częstością występowania lub krótko częstością. Częstość jest naturalnym oszacowaniem nieznanej proporcji $p$.

**Definicja 2.21.** ***Częstością występowania*** *w prostej próbie losowej nazywamy statystykę*

$$\hat{p} = \frac{\sum_{i=1}^{n} X_i}{n},$$

*gdzie $X_1, X_2, \ldots, X_n$ jest prostą próbą losową z rozkładu dwupunktowego o wartościach 0 i 1. Statystykę $\hat{p}$ obliczoną dla konkretnych wartości w próbie nazywamy* ***wartością częstości****.*

Zauważmy, że w tym przypadku odstąpiliśmy od naszej konwencji oznaczania zmiennych losowych dużymi literami. Co więcej, wartość częstości jest oznaczana tym samym symbolem $\hat{p}$ co częstość. Będziemy starali się, by z kontekstu było jasne, czy mówimy o częstości $\hat{p}$ zmiennej losowej czy o przyjętej przez nią wartości. Powodem odrębnego rozpatrzenia problemu szacowania proporcji jest powszechność występowania takiej właśnie sytuacji jak również możliwość wyznaczenia dokładnego rozkładu częstości $\hat{p}$. Zauważmy mianowicie, że w rozpatrywanym przypadku nasza prosta próba losowa spełnia założenia eksperymentu dwumianowego: zmienną losową $X_i$ możemy traktować jako $i$-tą „próbę" w eksperymencie, kończącą się sukcesem ($X_i = 1$) z prawdopodobieństwem $p$. Liczba „prób" w eksperymencie jest z góry ustalona i równa $n$, a z założenia, że dysponujemy prostą próbą losową wynika, że wyniki tak określonych „prób" nie zależą od siebie. Zatem

suma sukcesów $\sum_{i=1}^{n} X_i$ ma rozkład dwumianowy $Bin(n,p)$. Otrzymaliśmy więc stwierdzenie.

**Stwierdzenie 2.12.** *Częstość występowania $\hat{p}$ pomnożona przez liczność próby $n$ ma rozkład dwumianowy $Bin(n,p)$. Ponadto*

$$\mu_{\hat{p}} = p \qquad oraz \qquad \sigma^2_{\hat{p}} = \frac{p(1-p)}{n}.$$

Druga część stwierdzenia wynika ze stwierdzenia 2.11 i postaci wartości średniej i wariancji dla rozkładu dwumianowego. Zauważmy, że w odróżnieniu od ogólnej sytuacji, gdy wartość średnia $\mu_{\bar{X}}$ i odchylenie standardowe $\sigma_{\bar{X}}$ nie są ze sobą związane, w tym przypadku obie z tych wielkości zależą od nieznanej proporcji $p$. Zauważmy ponadto, że z Centralnego Twierdzenia Granicznego, definicji częstości $\hat{p}$ i stwierdzenia 2.12 wynika twierdzenie.

**Twierdzenie 2.15.** *Dla dowolnych rzeczywistych $a$ i $b$*

$$P\left(a \leq \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \leq b\right) \quad \to \quad \Phi(b) - \Phi(a),$$

*gdy $n$ dąży do nieskończoności.*

W praktyce użyteczna jest wersja ostatniego twierdzenia, w której nieznaną wartość odchylenia standardowego $\sigma_{\hat{p}}$ zastąpiono jego oszacowaniem $(\hat{p}(1-\hat{p})/n)^{1/2}$.

**Twierdzenie 2.16.** *Dla dowolnych rzeczywistych $a$ i $b$*

$$P\left(a \leq \frac{\hat{p}-p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \leq b\right) \quad \to \quad \Phi(b) - \Phi(a),$$

*gdy $n$ dąży do nieskończoności.*

W przypadku dowolnego rozkładu ograniczyliśmy stosowalność przybliżenia normalnego do liczności próby $n \geq 25$. W przypadku rozkładu dwupunktowego możemy sformułować dokładniejszy warunek, przy spełnieniu którego możemy stosować przybliżenie normalne. Mianowicie, uznajemy je za zadowalające, gdy $np \geq 5$ i $n(1-p) \geq 5$. Ponieważ proporcja $p$ jest nieznana, w ostatnim warunku z reguły zastępujemy ją przez wartość częstości $\hat{p}$:

$$n\hat{p} \geq 5 \qquad oraz \qquad n(1-\hat{p}) \geq 5.$$

Zauważmy, że zgodnie z def. 2.21 $n\hat{p}$ oznacza liczebność elementów o rozpatrywanej własności i analogicznie $n(1-\hat{p})$ oznacza liczebność elementów, które tej własności nie mają. Odnotujmy jeszcze, że ponieważ rozkład dwupunktowy jest rozkładem dyskretnym, w przypadku stosowania przybliżenia normalnego dla proporcji, powinniśmy stosować poprawkę (2.44).

**Przykład 2.27.** W populacji dorosłych Polaków 39% ma kłopoty ze snem (źródło: Gazeta Wyborcza za Pentorem, 11 kwietnia 2000). Oszacujemy prawdopodobieństwo, że w prostej próbie losowej dorosłych Polaków o liczności 100, częstość osób mających kłopoty ze snem nie przekroczy 0,33.

Interesuje nas prawdopodobieństwo $P(\hat{p} \leq 0{,}33)$. Zgodnie ze wzorem (2.44) dla $a = -\infty$, $b = 33$ i $n = 100$, otrzymujemy

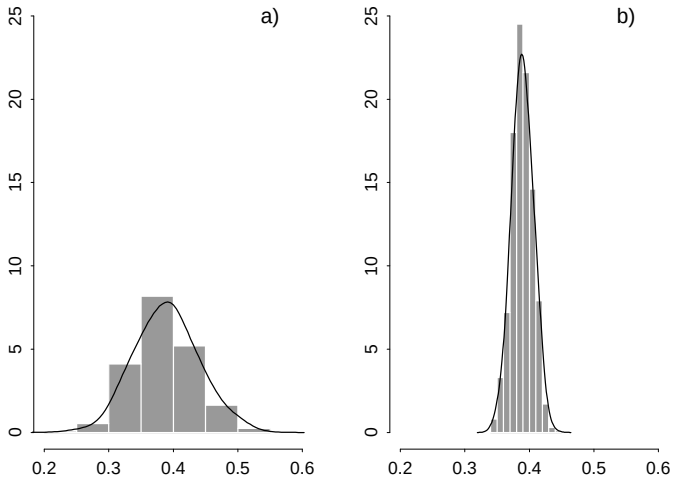
$$\begin{aligned} P(\hat{p} \leq 0{,}33) = P(\hat{p} \leq 0{,}335) &\approx \Phi\Big(\frac{33{,}5 - 39}{\sqrt{100 \times 0{,}39 \times 0{,}61}}\Big) = \\ &= \Phi(-1{,}13) = 0{,}1292. \end{aligned}$$

Przeanalizujmy sens otrzymanej wartości liczbowej prawdopodobieństwa. Oznacza ono, że rozpatrując dużą liczbę stuelementowych prostych prób losowych, powiedzmy 1000, z populacji dorosłych Polaków, możemy spodziewać się, że w około stu trzydziestu próbach liczba osób mających kłopoty ze snem nie przekroczy trzydziestu trzech.

### 2.4.4. Estymatory i ich podstawowe własności

W poprzednich dwóch punktach rozpatrzyliśmy własności dwóch często używanych statystyk: średniej i częstości. Służą one do oszacowania nieznanych parametrów populacji, odpowiednio wartości średniej i proporcji. Oczywiście, nie są to jedyne problemy szacowania nieznanych parametrów występujące w statystyce: możemy chcieć oszacować inny parametr położenia lub rozproszenia, np. odchylenie standardowe nieznanego rozkładu. W sytuacji, gdy statystyki są używane do przybliżenia nieznanych parametrów noszą one specjalną nazwę.

**Definicja 2.22.** *Statystykę $T(X_1, X_2, \ldots, X_n)$ służącą do oszacowania nieznanego parametru populacji nazywamy* ***estymatorem****. Dla konkretnych wartości próby $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$, liczbę $T(x_1, x_2, \ldots, x_n)$ nazywamy* ***wartością estymatora****.*




Rys. 2.12. Wykresy gęstości rozkładów częstości dla prostej próby losowej a) o liczności 100, b) o liczności 1000 (przykł. 2.27)

Na pierwszy rzut oka może się wydawać, że oszacowanie nieznanych parametrów jest jedynym celem wnioskowania, a zatem wyróżnianie estymatorów wśród statystyk jest niepotrzebne. Tak jednak nie jest: może na przykład interesować nas problem zaklasyfikowania rozpatrywanego elementu do jednej z kilku możliwych populacji. Statystyka skonstruowana w tym przypadku nie ma na celu możliwie precyzyjnego oszacowania nieznanego parametru populacji, a raczej przypisanie elementu do właściwej populacji z jak najmniejszym błędem. Inne przykłady związane z testowaniem hipotez poznamy w rozdz. 3.

**Nieobciążoność estymatorów.** Rozpatrzmy raz jeszcze zachowanie poprzednio wprowadzonych estymatorów.

**Przykład 2.27 cd.** Załóżmy, że rozpatrujemy proste próby losowe z populacji dorosłych Polaków, pierwszą o liczności 100, a drugą o liczności 1000 i w obu przypadkach badamy rozkład częstości ludzi mających kłopoty ze snem. Odpowiednie wykresy gęstości rozkładów częstości są przedstawione na rys. 2.12.

Oba rozkłady są symetryczne i jednomodalne oraz wartości średnie obu tych rozkładów, na mocy stwierdzenia 2.12, są równe dokładnie $p$: $\mu_{\hat{p}} = p$. To, co przede wszystkim różni oba rozkłady, to wielkość ich rozproszenia; drugi rozkład jest znacznie bardziej skupiony wokół swojej wartości średniej niż

pierwszy. Dla większej liczności próby estymator $\hat{p}$ jest znacznie bardziej precyzyjny. Rysunki te dobrze ukazują dwa warunki, na których opiera się jedno z podstawowych kryteriów doboru estymatorów. Kryterium to wymaga, aby po pierwsze, wskaźnik położenia rozkładu estymatora był równy parametrowi, który estymujemy. Po drugie, chcemy, aby dla dowolnej liczności próby rozproszenie takiego estymatora było możliwie małe. W przypadku, gdy jako wskaźnik położenia rozkładu estymatora rozpatrujemy wartość średnią, estymator mający pierwszą własność nazywamy **estymatorem nieobciążonym**. W języku matematyki wyrażamy ten postulat w następujący sposób. Nieznany parametr, który chcemy estymować będzie z reguły oznaczany grecką literą $\theta$. Parametr $\theta$ w zależności od sytuacji, która nas interesuje, może być np. wartością średnią, medianą lub rozproszeniem badanego rozkładu.

**Definicja 2.23.** *Niech $\theta$ będzie liczbą rzeczywistą oznaczającą nieznany parametr populacji i niech $\hat{\theta} = T(X_1, X_2, \ldots, X_n)$ będzie pewnym estymatorem parametru $\theta$. Różnicę między wartością średnią estymatora a nieznaną wartością parametru, $\mu_{\hat{\theta}} - \theta$, nazywamy* ***obciążeniem estymatora****. Estymator $\hat{\theta}$ nazywamy* ***nieobciążonym****, gdy jego obciążenie jest równe 0, tzn. $\mu_{\hat{\theta}} = \theta$ dla dowolnej wartości $\theta$, którą może przyjmować parametr i dowolnej liczności próby $n \geq 1$.*

Postarajmy się wyjaśnić, dlaczego żądamy, aby $\mu_{\hat{\theta}} = \theta$ dla dowolnej wartości $\theta$, którą może przyjmować parametr. Nawiązując do poprzedniego przykładu oznacza to, że wymagamy, aby położenie rozkładu częstości $\hat{p}$ było równe $p$ nie tylko dla populacji dorosłych Polaków, ale również między innymi dla populacji dorosłych Francuzów i Niemców, dla których wartości proporcji ludzi mających kłopoty ze snem są równe odpowiednio 0,46 i 0,36. Warunek ten ma być również spełniony dla każdej liczności próby: nieistotne jest, czy jest ona równa 2 czy 150. Podkreślmy, że warunek nieobciążoności nie niesie w sobie żadnych wymagań dotyczących wartości estymatora dla konkretnej realizacji prostej próby losowej. Może ona odbiegać, nawet bardzo znacznie, od wartości nieznanego parametru. Ważne jest dla nas jednak to, że estymator nieobciążony nie może *systematycznie* przeszacowywać lub niedoszacowywać estymowanej wartości: jego wartość średnia musi się z ową nieznaną wartością pokrywać.

*Stwierdzenie 2.12 gwarantuje, że warunek nieobciążoności jest spełniony dla częstości przy dowolnej wartości proporcji $p$. Podobnie, średnia $\bar{X}$ jest niebciążonym estymatorem wartości średniej $\mu$ dla dowolnej wartości $\mu$.*

Zauważmy, że w istocie średnia $\bar{X}$ jest estymatorem nieobciążonym *niezależnie* od rozkładu, z którego pochodzi prosta próba losowa. Jest to własność pożyteczna, zwłaszcza w sytuacji, gdy wiemy mało o rozkładzie, z którego

próbkujemy. Często zdarza się tak, że rozpatrywany estymator jest nieobciążony dla pewnej rodziny rozkładów, a dla innej już nie. Przedstawmy teraz estymator wariancji, który jest nieobciążony bez względu na rozkład, z którego pochodzi prosta próba losowa.

**Nieobciążoność estymatora wariancji.** W definicji 1.6 wprowadziliśmy pojęcie wariancji w próbie, oznaczając ją symbolem $s^2$. Przenosząc ją na poziom prostej próby losowej otrzymamy definicję **estymatora wariancji w prostej próbie losowej**

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2. \tag{2.45}$$

dla $n \geq 2$. Po pierwsze, zauważmy, że estymator wariancji jest faktycznie postaci $T(X_1, X_2, \ldots, X_n)$ dla pewnej funkcji $T$. Najłatwiej zauważyć to na podstawie równości, uzasadnienie której pozostawimy Czytelnikowi

$$S^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - \frac{(\sum_{i=1}^{n} X_i)^2}{n}\right). \tag{2.46}$$

Na podstawie równości (2.46) sprawdzimy, czy estymator $S^2$ jest estymatorem nieobciążonym. Ponieważ $\sigma_X^2 = \mu_{X^2} - (\mu_X)^2$, mamy $\mu_{X^2} = \sigma_X^2 + (\mu_X)^2$. Korzystając z ostatniej równości i wyrażenia na wartość średnią kombinacji liniowej niezależnych zmiennych losowych oraz oznaczając $\mu = \mu_{X_i}$, $\sigma^2 = \sigma_{X_i}^2$ i $Y = \sum_{i=1}^{n} X_i$, otrzymamy

$$\begin{aligned}
\mu_{S^2} &= \frac{1}{(n-1)}\Big(\sum_{i=1}^{n}\mu_{X_i^2} - \frac{1}{n}\mu_{Y^2}\Big) = \\
&= \frac{1}{(n-1)}\Big(\sum_{i=1}^{n}(\sigma^2 + \mu^2) - \frac{1}{n}(\sigma_Y^2 + (\mu_Y)^2)\Big) = \\
&= \frac{1}{(n-1)}\Big(n\sigma^2 + n\mu^2 - \frac{1}{n}(n\sigma^2) - \frac{1}{n}(n\mu)^2\Big) = \frac{1}{n-1}(n\sigma^2 - \sigma^2) = \sigma^2.
\end{aligned}$$

Wyprowadzenie to uzasadnia, dlaczego użyliśmy mnożnika $(n-1)^{-1}$ zamiast $n^{-1}$ w definicji $S^2$. Taki mnożnik gwarantuje nam nieobciążoność estymatora wariancji dla prostej próby losowej z *dowolnego* rozkładu. Okazuje się jednak, że estymator odchylenia standardowego $S$ zdefiniowany jako $S = \sqrt{S^2}$ jest już z reguły estymatorem obciążonym.

**Porównania estymatorów.** Z reguły istnieje wiele estymatorów nieobciążonych dla konkretnego nieznanego parametru. Na przykład w przypadku prób pochodzących z rozkładu normalnego średnia ucinana jest, podobnie

jak średnia, estymatorem nieobciążonym wartości średniej. Poza tym widać, że sama własność nieobciążoności nie określa koniecznie dobrego estmatora: estymator nieobciążony, ale mało precyzjny jest praktycznie bezużyteczny. Zauważmy, że dla dowolnej liczności próby $n \geq 2$ estymatory $\hat{\theta} = X_1$ i $\tilde{\theta} = (X_1 + X_2)/2$ są nieobciążonymi estymatorami wartości średniej rozkładu, ale ich rozproszenie nie maleje wraz z licznością próby. Dla dużej liczności $n$ średnia z próby $\bar{X}$ powinna być znacznie lepszym estymatorem. Dlatego żąda się, aby poza nieobciążonością estymator miał również możliwie małe rozproszenie. Ścisłe sformułowanie warunku jest następujące:

**Definicja 2.24.** *Niech $\theta$ będzie liczbą rzeczywistą oznaczającą nieznany parametr populacji. Nieobciążony estymator $\hat{\theta}(X_1, X_2, \ldots, X_n)$ parametru $\theta$ nazywamy* ***estymatorem nieobciążonym o minimalnej wariancji*** *(w skrócie* ***estymatorem NMW****), jeżeli wśród wszystkich estymatorów nieobciążonych parametru $\theta$ nie istnieje estymator, którego wariancja byłaby mniejsza dla jakiejś wartości $\theta$. Innymi słowy, dla wszystkich możliwych wartości $\theta$ i wszystkich nieobciążonych estymatorów $\tilde{\theta} = \tilde{T}(X_1, X_2, \ldots, X_n)$*

$$\mathrm{Var}(\hat{\theta}) \leq \mathrm{Var}(\tilde{\theta}).$$

Zwróćmy uwagę na pewien niuans tej definicji, który podkreślaliśmy już przy definicji nieobciążoności estymatora. Dla ustalonej populacji, czyli dla ustalonego rozkładu prawdopodobieństwa, z którego pochodzi próba losowa $X_1, X_2, \ldots, X_n$ wartość parametru $\theta$ jest ustaloną choć nieznaną liczbą. Ponieważ wartości liczby $\theta$ nie znamy, więc nasz estymator musi mieć pożądane własności dla każdej możliwej wartości tej liczby. Na przykład, gdy $\theta$ jest wartością średnią rozkładu, możemy chcieć by estymator $\hat{\theta}$ miał wymienione własności dla $\theta$ z przedziału obustronnie nieskończonego $(-\infty, \infty)$. Warto tu zauważyć, że model statystyczny tym właśnie się charakteryzuje, że niejako obejmuje całą rodzinę możliwych rozkładów prawdopodobieństwa (różnych dla różnych wartości nieznanego parametru lub nieznanych parametrów). Gdy np. wiemy, że populacja ma rozkład normalny o znanym odchyleniu standardowym $\sigma_0$ i nieznanej wartości średniej $\theta$, gdzie $-\infty < \theta < \infty$, to model statystyczny jest rodziną wszystkich rozkładów $N(\theta, \sigma_0)$ , $-\infty < \theta < \infty$. Można udowodnić, że średnia próbkowa jest estymatorem NMW w tej rodzinie rozkładów. Podobnie, gdy populacja ma rozkład normalny o nieznanej wartości średniej $\theta_1$ , $-\infty < \theta_1 < \infty$ oraz nieznanym odchyleniu standardowym $\theta_2 > 0$, to okazuje się, że w tej rodzinie estymatory $\bar{X}$ i $S^2$ są estymatorami NMW odpowiednio $\theta_1$ i $\theta_2$.

Zauważmy tu jeszcze, że jak widać z rys. 2.11 rozproszenie estymatora zależy od liczności próby; im większa liczność próby tym większa precyzja estymatora. Nadmieńmy, że warunek małego rozproszenia rozpatrywany samoist-

nie nie jest również wystarczający: estymator, który w precyzyjny sposób (czyli z małym rozproszeniem) estymuje nieprawidłową wartość parametru jest oczywiście mało wartościowy.

Odnotujmy jednak, że na problem jakości estymacji można spojrzeć w trochę inny sposób. Podejście to jest związane z faktem, że w wielu sytuacjach niewielkie obciążenie estymatora może być dopuszczalne, zwłaszcza gdy wraz z licznością próby staje się coraz mniejsze. Zamiast nakładania warunku nieobciążoności i możliwie małej wariancji, możemy zażądać, żeby odległość estymatora $\hat{\theta}$ od parametru $\theta$ była mała dla wszystkich lub dla większości prób. Okazuje się, że spełnienie tego postulatu dla wszystkich prób jest mało realistyczne i dlatego przyjmuje się często odpowiednio sformułowany warunek minimalizacji *wartości średniej* odległości $\hat{\theta}-\theta$ lub wartości średniej kwadratu tej odległości.

**Definicja 2.25.** *Wartość średnią kwadratu odległości* $(\hat{\theta}-\theta)^2$, $\mu_{(\hat{\theta}-\theta)^2}$, *nazywamy* ***błędem średniokwadratowym estymatora*** $\hat{\theta}$. *Estymator* $\hat{\theta}$, *dla którego nie istnieje estymator* $\tilde{\theta}$ *taki, że* $\mu_{(\tilde{\theta}-\theta)^2} \leq \mu_{(\hat{\theta}-\theta)^2}$ *dla każdego* $\theta$ *i nierówność ostra zachodzi dla pewnego* $\theta_0$, *nazywamy* ***estymatorem dopuszczalnym*** *(ze względu na błąd średniokwadratowy).*

Estymator dopuszczalny to zatem taki estymator, dla którego nie można znaleźć innego estymatora mającego mniejszy błąd średniokwadratowy dla pewnej wartości parametru i nie większy błąd średniokwadratowy dla jego pozostałych wartości. Ta nieco dziwna na pierwszy rzut oka definicja jest motywowana tym, że nie możemy zażądać, aby błąd średniokwadratowy „dobrego" estymatora $\hat{\theta}$ był najmniejszy dla każdego $\theta$ wśród wszystkich estymatorów. Dlaczego? Gdyż dla ustalonego $\theta_0$ żaden estymator $\hat{\theta}$ nie może „wygrać" z estymatorem $\tilde{\theta}$, faktycznie niezależnym od próby i zawsze przyjmującym tylko jedną wartość $\theta_0$. Błąd średniokwadratowy tego estymatora jest równy zeru przy wartości parametru $\theta_0$, zaś przy każdej innej wartości $\theta$ wynosi $(\theta-\theta_0)^2$. Estymator $\tilde{\theta}$ jest zatem „nie do pokonania", gdy $\theta=\theta_0$ i staje się coraz gorszy i w końcu coraz bardziej absurdalny, gdy prawdziwa wartość parametru oddala się od $\theta_0$. Stąd powstaje potrzeba określenia estymatora, który byłby „jednostajnie" (czyli dla wszystkich możliwych wartości $\theta$) do przyjęcia. Potrzebę taką spełnia np. właśnie zdefiniowany estymator dopuszczalny. Zauważmy na marginesie, że tego argumentu nie można użyć, gdybyśmy szukali estymatora o tej własności w klasie wszystkich estymatorów *nieobciążonych* (ponieważ estymator $\tilde{\theta}$ nie jest nieobciążony).

Związek między przedstawionymi podejściami można ustalić pisząc

$$(\hat{\theta}-\theta)^2 = (\hat{\theta}-\mu+\mu-\theta)^2 = (\hat{\theta}-\mu)^2 + (\mu-\theta)^2 + 2(\hat{\theta}-\mu)(\mu-\theta),$$

gdzie $\mu = \mu_{\hat{\theta}}$. Obliczmy wartość średnią obu stron powyższej równości.

Jeśli zauważymy, że wartość średnia ostatniego składnika jest 0, a wartością średnią pierwszego jest wariancja estymatora $\hat{\theta}$, to otrzymamy ważne stwierdzenie.

**Stwierdzenie 2.13.** *Dla dowolnego estymatora $\hat{\theta}$ jego* ***błąd średniokwadratowy*** *jest sumą jego wariancji i kwadratu obciążenia*

$$\mu_{(\hat{\theta}-\theta)^2} = \sigma^2_{\hat{\theta}} + (\mu_{\hat{\theta}} - \theta)^2.$$

Poszukiwanie estymatora dopuszczalnego nie wymusza zatem, aby był to estymator nieobciążony. Chcemy tylko, żeby wariancja i kwadrat obciążenia były możliwie małe. Może się zdarzyć, że estymator mający niezerowe ale niewielkie obciążenie ma błąd średniokwadratowy mniejszy niż dowolny estymator nieobciążony.

**Przykład 2.28.** Rozpatrzmy prostą próbę losową $X_1, X_2, \ldots, X_n$ pochodzącą z rozkładu jednostajnego na przedziale $(a, b)$, gdzie $a$ i $b$ są nieznanymi liczbami rzeczywistymi. W wielu zadaniach kontroli jakości interesuje nas rozstęp tego rozkładu zdefiniowany jako odległość „górnego" końca od „dolnego" końca rozkładu: $\theta = b - a$. Naturalnym estymatorem rozstępu rozkładu jednostajnego jest rozstęp w prostej próbie losowej

$$R = X_{(n)} - X_{(1)},$$

gdzie $X_{(n)}$ jest zmienną losową zdefiniowaną jako maksymalna, a $X_{(1)}$ jako minimalna spośród zmiennych $X_1, X_2, \ldots, X_n$. Okazuje się, że estymator $R$ jest estymatorem obciążonym rozstępu, $\mu_R \neq \theta$, ale wartość jego obciążenia zmierza do zera wraz z licznością próby. Estymator $R$ okazuje się być również przykładem estymatora największej wiarogodności (tego typu estymatory zdefiniujemy w rozdz. 3). Zauważmy, że konkurencyjny estymator parametru $\theta$ można skonstruować, pamiętając, że $\sigma^2_{X_i} = (b-a)^2/12$, zatem $\tilde{R} = \sqrt{12}S$ jest również naturalnym estymatorem rozstępu. Do tego przykładu powrócimy w rozdz. 3.

W rozpatrywanych ocenach dobroci estymatora, ważną rolę bezpośrednio (jak np. w def. 2.24) lub pośrednio (jak np. w def. 2.25) odgrywa jego wariancja lub równoważnie jego odchylenie standardowe. Odchylenie standardowe $\sigma_{\hat{\theta}}$ jest pewną funkcją parametrów rozkładu i liczności próby. Gdy parametry lub sama funkcja są nieznane, nie znamy dokładnie wartości odchylenia standardowego estymatora i musimy go oszacować w celu oceny zmienności badanego oszacowania. To prowadzi nas do następującej definicji.

**Definicja 2.26.** ***Błędem standardowym estymatora*** $\hat{\theta}$ *parametru* $\theta$ *nazywamy dowolny estymator jego odchylenia standardowego* $\sigma_{\hat{\theta}}$ *i oznaczamy go* $SE_{\hat{\theta}}$.

Zauważmy, że zgodnie z przyjęciem oznaczenia $\hat{\theta}$ na oznaczenie estymatora $\theta$, błąd standardowy moglibyśmy konsekwentnie oznaczać $\hat{\sigma}_{\hat{\theta}}$. Przyjmujemy raz jeszcze tradycyjnie używane oznaczenie. Rozpatrzmy przykładowo ocenę zmienności średniej w próbie. Na podstawie stwierdzenia 2.19 wiemy, że dla $\hat{\theta} = \bar{X}$ mamy $\sigma_{\hat{\theta}} = \sigma/\sqrt{n}$ i naturalnym kandydatem na błąd standardowy jest

$$SE_{\bar{X}} = \frac{S}{\sqrt{n}}, \tag{2.47}$$

gdzie $S$ jest estymatorem odchylenia standardowego w próbie. Wyrażenie (2.47) nazywa się często po prostu błędem standardowym $\bar{X}$. Podobnie, wyrażenie $\sqrt{\hat{p}(1-\hat{p})/n}$ nazywa się błędem standardowym częstości $\hat{p}$. Podkreślmy tutaj, że bardzo ważne jest określenie o błędzie standardowym *jakiego estymatora* mówimy. Samo określenie „błąd standardowy" jeszcze nic nie oznacza.

Powyższa przykładowa sytuacja jest typowa w praktyce statystycznej: jakkolwiek błąd standardowy $\hat{\theta}$ został zdefiniowany jako *dowolny* estymator $\sigma_{\hat{\theta}}$, z reguły istnieje naturalny i powszechnie używany estymator tej wielkości, który potocznie nazywa się błędem standardowym estymatora $\hat{\theta}$. Tak jest w przypadku, gdy $\sigma_{\hat{\theta}}$ jest znaną funkcją pewnych parametrów rozkładu i parametry te mają pewne powszechnie używane estymatory. Wtedy naturalną definicję błędu standardowego otrzymamy, podstawiając jako argumenty funkcji estymatory zamiast parametrów. W ten sposób postąpimy np. w rozdz. 4, definiując błędy standardowe współczynników równania regresji. Podkreślmy jednak, że istnieją inne metody definiowania błędu standardowego danego estymatora, nie oparte na estymacji nieznanych parametrów, które omówimy przy okazji przedstawienia metod Monte Carlo. Z kontekstu będzie jednak zawsze wynikało, o jak zdefiniowany błąd standardowy nam chodzi. Na zakończenie tego punktu podamy definicję wielkości, która często będzie się pojawiać w różnych miejscach w następnych rozdziałach.

**Definicja 2.27.** *Niech* $\hat{\theta}$ *będzie nieobciążonym estymatorem parametru* $\theta$. *Wówczas* ***studentyzowanym estymatorem*** $\theta$ *nazywamy wielkość* $(\hat{\theta} - \theta)/SE_{\hat{\theta}}$.

Studentyzowany estymator to estymator standaryzowany $(\hat{\theta} - \theta)/\sigma_{\hat{\theta}}$, w którym odchylenie standardowe estymatora zamieniono na jego błąd standardowy. Zgodnie z powyższymi komentarzami studentyzowaną średnią w próbie nazywamy wielkość $(\bar{X} - \mu)/(S/\sqrt{n})$. Nazwa odnosi się do Williama

Gosseta, używającego skromnego pseudonimu Student, który pierwszy rozpatrywał rozkład tak przekształconej średniej dla obserwacji pochodzących z rozkładu normalnego.

## 2.5. Metody zbierania danych

### 2.5.1. Podstawowy schemat eksperymentalny

Każdego dnia stykamy się z rozmaitymi zestawieniami danych, z których większość pochodzi z *obserwacji* sytuacji losowych: w gazetach i w Internecie można znaleźć wartości podstawowych indeksów giełdowych (obliczanych z reguły co kilka sekund), firmy podają do publicznej wiadomości swoje wskaźniki ekonomiczne, w bazach danych tych firm są zawarte informacje na temat zatrudnionych w nich pracowników, itd. Innym źródłem danych jest *eksperyment* losowy: w stosunku do grupy wyodrębnionych w pewien sposób jednostek eksperymentalnych (np. roślin, ludzi, maszyn itp.) wykonuje się pewne zaplanowane działanie (np. polegające na podaniu nowego leku pewnej grupie pacjentów) i obserwuje się interesującą nas reakcję lub jej brak. Są to dwa zasadnicze źródła danych. Poprzednio przedstawiliśmy podstawowe metody analizy statystycznej obserwacji losowych, nie wnikając dotąd głębiej w to, w jaki sposób zostały one zebrane. Źródło danych i sposób ich zebrania może mieć jednak podstawowe znaczenie dla poprawnej interpretacji wyników wnioskowania. Ilustruje to następujący przykład.

**Przykład 2.29.** Na rynku znajduje się wiele środków reklamowanych jako zwiększające energię i poprawiające koncentrację. Są one popularne m.in. wśród studentów w czasie trwania sesji egzaminacyjnych. W celu stwierdzenia skuteczności jednego z nich, nazwijmy go napojem XXL, zebrano średnie wyniki egzaminów pewnej szkoły wyższej w sesji letniej dla wszystkich studentów pierwszego roku regularnie używających tego środka podczas sesji. Średnie wyniki przedstawiono za pomocą histogramu i zestawiono go z analogicznym histogramem wyników studentów, którzy tego środka nie używali. Okazało się, że histogram dla pierwszej grupy jest wyraźnie przesunięty w prawo w stosunku do histogramu dla grupy drugiej, co wskazywałoby na to, że picie napoju istotnie poprawia efektywność działań studentów podczas egzaminów.

Zastanówmy się jednak, czy faktycznie tak musi być. Możliwa jest sytuacja, że studenci pijący napój XXL, są to najlepsi studenci na roku, którzy chcąc

stworzyć sobie optymalne warunki zdania egzaminów, oprócz intensywnej nauki sięgnęli po ten szeroko reklamowany środek. Inną możliwością jest to, że znaczna część spośród uczących się dużo studentów, odczuwała potrzebę poprawy swojej koncentracji i w tym celu używała napoju XXL. Tak więc jest możliwe, że grupy studentów pijących i niepijących napój XXL znacznie różnią się pewnymi cechami (jak np. ilość włożonej pracy, przywiązywanie wagi do dobrego zdania egzaminów), mogącymi mieć istotny wpływ na wyniki egzaminów. Inaczej mówiąc, nie jesteśmy w stanie wykluczyć w tej metodzie zbierania danych wpływu pewnych **zmiennych ukrytych**, których nie kontrolujemy, na interesującą nas zmienną (wynik egzaminu). Może się tak zdarzyć, że to zmienne ukryte mogą mieć zasadniczy wpływ na wynik egzaminu, większy od tych, które uważamy za istotne dla jego objaśnienia. Problem ten omówimy szerzej przy okazji rozważania modelu regresji wielokrotnej w podrozdz. 4.3. Wynika z tego również, że w tej sytuacji nie możemy stwierdzić, czy prawdopodobny jest związek przyczynowy między stosowaniem rozpatrywanego środka a wynikami egzaminów.

**Eksperymenty porównawcze dla jednego czynnika.** Wyjściem z sytuacji jest eksperyment, w którym określeni według pewnego klucza studenci zostaną poproszeni o picie napoju XXL dwa razy dziennie w sesji egzaminacyjnej, a wyniki egzaminu w tej grupie zostaną porównane z wynikami drugiej grupy wybranych studentów, którzy w tym czasie zgodzili się nie używać tego napoju. W ogólnym przypadku obiekty z pierwszej grupy nazywa się **jednostkami eksperymentalnymi**, a całą grupę **grupą eksperymentalną**, w odróżnieniu od drugiej grupy, zwanej **grupą kontrolną**, której obiekty nazywa się **jednostkami kontrolnymi**. Eksperyment, w którym staramy się ocenić wpływ pewnego działania porównując reakcję w grupie kontrolnej i eksperymentalnej nazywamy **eksperymentem porównawczym**. Cechę, której wartość kontrolujemy, lub terapię, której wpływ na wartość zmiennej objaśnianej chcemy badać, nazywamy **czynnikiem**. Oczywiście, można rozpatrywać eksperymenty porównawcze z więcej niż jednym czynnikiem, np. w przykł. 2.29 obok działania płynu XXL możemy rozważyć drugi czynnik będący ilością pracy włożonej w przygotowanie się do egzaminów. Omówimy najpierw krótko sposób wykonania eksperymentu porównawczego z jednym czynnikiem. Podstawową sprawą jest tutaj sposób przypisania jednostek do grupy eksperymentalnej lub kontrolnej. Metoda kwalifikacji powinna być taka, żeby obie grupy były możliwie podobne ze względu na wszystkie cechy, poza cechą lub cechami, wpływ których na wynik eksperymentu chcemy badać. W naszym przykładzie badaną cechą było picie lub nie napoju XXL. Najprostszą, a jednocześnie bardzo skuteczną metodą zapewnienia sobie tego podobieństwa jest randomizacja, to jest losowe przypisanie obiektów do grupy kontrolnej i eksperymentalnej.

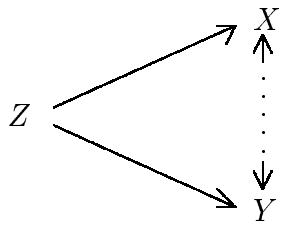
**Randomizacja.**

**Przykład 2.29 cd.** Załóżmy, że wylosowaliśmy prostą próbę losową składającą się z 50 studentów pierwszego roku szkoły wyższej. Próbę tę podzielimy w sposób losowy na dwie podpróby po 25 osób, z których pierwsza będzie spełniała rolę grupy eksperymentalnej, a druga grupy kontrolnej. Członkowie grupy eksperymentalnej spożywają w czasie sesji egzaminacyjnej dwa razy dziennie napój XXL, w odróżnieniu od członków grupy kontrolnej, którzy wstrzymują się od spożycia tego napoju. Po egzaminie porównujemy rozkłady średnich wyników w obydwu grupach i stwierdzamy, czy spożycie napoju XXL mogło mieć wpływ na wyniki egzaminów w sesji.

Eksperyment porównawczy, w którym przypisanie do grupy eksperymentalnej i kontrolnej jest losowe, nazywamy zrandomizowanym eksperymentem porównwawczym. Randomizacja jest naszym zabezpieczeniem przeciwko możliwym, istotnym z punktu widzenia eksperymentu różnicom charakterystyk obydwu grup. Wskutek zastosowania randomizacji, różnice w reakcjach dla grupy eksperymentalnej i kontrolnej mogą wynikać albo z działania badanego czynnika, albo być dziełem przypadku. Działanie przypadku staramy się zmniejszyć, zwiększając liczności obu prób. Jest to takie samo działanie jakie podejmujemy, aby zwiększyć precyzję badanego estymatora: w celu zmniejszenia jego odchylenia standardowego staramy się, o ile to możliwe, zwiększyć liczność próby, na podstawie której jest obliczany estymator.

Czasami stosuje się inne niż randomizacja metody utworzenia dwóch możliwie podobnych do siebie prób. Jedna z nich opiera się na szukaniu par jednostek mających bardzo zbliżone wartości cech mogących mieć wpływ na wynik eksperymentu. Po znalezieniu takiej pary, jedną z jednostek zalicza się do grupy eksperymentalnej, a drugą do kontrolnej. Operację tę powtarza się aż do uzyskania grup o odpowiedniej liczności. Metoda ta jest jednak znacznie trudniejsza do przeprowadzenia niż randomizacja, trudno również uniknąć arbitralności przy stwierdzaniu, które jednostki są rzeczywiście podobne do siebie. Zauważmy również, że nawet przy bardzo dużym doświadczeniu eksperymentatora, niełatwo jest ustalić jakie zmienne mogą potencjalnie wpływać na wynik eksperymentu. Może nas to w efekcie doprowadzić do błędnego wniosku o zależności przyczynowej. Rozpatrzmy przykładowo sytuację, gdy staramy się dociec, czy zmiany pewnej zmiennej $X$ są przyczyną zmian zmiennej $Y$. Jeśli istnieje **zmienna uwikłana** $Z$, której zmiany powodują zmiany zarówno zmiennej $X$ jak i zmiennej $Y$, to związek przyczynowy między zmiennymi $X$ i $Y$ jest pozorny (por. rys. 2.13). Jeśli rozkład zmiennej $Z$ w grupie kontrolnej i eksperymentalnej bardzo się

różni, to jest to faktycznym powodem różnic w rozkładach zmiennej $Y$ w obu grupach. *Randomizacja w eksperymencie porównawczym jest w efekcie najpewniejszą metodą służącą do wykrycia zależności przyczynowej.* Oczywiście, podobnie jak w przypadku testowania, nie możemy orzekać, że związek przyczynowy na pewno zachodzi, a co najwyżej możemy stwierdzić, że jest on bardzo prawdopodobny.




Rys. 2.13. Przyczynowa zależność zmiennych $X$ i $Y$ od zmiennej $Z$ jest powodem zależności $X$ i $Y$

Jak wykonać losowe przypisanie jednostek do grupy eksperymentalnej i grupy kontrolnej? Powróćmy do przykł. 2.29. Najprostsza metoda randomizacji rozpoczyna się od ponumerowania w dowolny sposób wszystkich pięćdziesięciu studentów liczbami od 1 do 50. Następnie losujemy 25 liczb ze zbioru $\{1, 2, \ldots, 50\}$ i studentów o wybranych numerach przypisujemy do grupy eksperymentalnej, a pozostałych 25 studentów kwalifikujemy do grupy kontrolnej. Sposób wyboru liczb losowych z pewnego ustalonego zbioru liczb przy użyciu generatora liczb losowych jest opisany w rozdz. 8 (porównaj w szczególności przykł. 8.5). W zastępstwie generatora liczb losowych możemy użyć tablic cyfr losowych, zamieszczonych w większości starszych podręczników statystyki. Są to ciągi cyfr będących realizacją prostej próby losowej z rozkładu jednostajnego na zbiorze $\{0, 1, \ldots, 9\}$. W celu wyboru 25 liczb losowych między 1 a 50 wybieramy na chybił trafił miejsce w tablicy cyfr losowych i poczynając od niego rozpatrujemy kolejne dwójki cyfr. Spośród nich wybieramy pierwszych 25 odpowiadających liczbom ze zbioru $\{1, 2, \ldots, 50\}$. Tak więc, jeśli ciąg cyfr losowych ma postać $1, 0, 9, 2, 8, 8, 4, 3, 3, 5, \ldots$ to numery pierwszych studentów zaliczonych do grupy eksperymentalnej są następujące $10, 43$ i $35$ (pary $(9, 2)$ i $(8, 8)$ zostały pominięte, bo nie odpowiadają liczbom ze zbioru $\{1, 2, \ldots, 50\}$). W przypadku, gdy łączna liczność grupy eksperymentalnej i kontrolnej przekracza 100, zamiast dwójek cyfr rozpatrujemy trójki itd. Podkreślmy raz jeszcze, że randomizacja oparta na eksperymencie losowym ma ogromną przewagę nad pozornie losowym przypisaniem do grup na chybił trafił przez eksperymentatora. Umożliwia ona wyeliminowanie czynnika nieuświadamianych preferencji występującego np. w sytuacji, gdy eksperymentator badający skuteczność nowego leku może

podświadomie kwalifikować osoby w lepszym stanie fizycznym do grupy eksperymentalnej, na której ten lek będzie testowany.

**Efekt placebo.** Powróćmy jeszcze raz do przykł. 2.29. W rzeczywistości wykonanie eksperymentu porównawczego w tej sytuacji jest nieco bardziej skomplikowane. Konieczne jest uwzględnienie możliwości wystąpienia tak zwanego efektu placebo, czyli odczuwania pozytywnego wpływu czynnika (zwanego placebo), który jest w rzeczywistości obojętny, jak np. płyn XXL pozbawiony składników mających zwiększać energię i koncentrację lub „pusty", pozbawiony substancji leczniczych lek. Jest to szczególnie ważne, jeśli wynik eksperymentu jest z konieczności oceniany przez osoby biorące udział w eksperymencie, jak np. w przypadku cierpiących na bezsenność pacjentów, którzy mają sami ocenić, czy pomógł im nowy środek nasenny. To, co musimy zrobić, to postarać się oddzielić rzeczywisty wpływ interesującego nas czynnika od możliwego i zaciemniającego obraz efektu placebo. W tym celu zapewniamy sobie, że zarówno jednostki eksperymentalne, jak i kontrolne nie wiedzą, do jakiej grupy należą. Taki eksperyment nazywa się często eksperymentem ślepym. W przykładzie 2.29 sprowadza się to do podawania płynu o nazwie XXL zarówno członkom grupy kontrolnej, jak i eksperymentalnej. W przypadku grupy eksperymentalnej jest to jednak prawdziwy napój XXL, natomiast w przypadku grupy kontrolnej podajemy jej członkom napój identyczny smakowo, ale pozbawiony wszelkich substancji aktywnych. W przypadku np. badania skuteczności leków wykonuje się często tzw. eksperymenty podwójnie ślepe, które planuje się w ten sposób, że osoba prowadząca również nie wie, które z jednostek są w rzeczywistości w grupie kontrolnej, a które w eksperymentalnej. Ma to na celu zapobieżenie podświadomej skłonności prowadzącego eksperyment do oceniania stanu pacjentów w grupie eksperymentalnej jako lepszego aniżeli stanu pacjentów w grupie kontrolnej.

### 2.5.2. Inne schematy eksperymentalne

W punkcie 2.5.1 omówiliśmy podstawowy schemat dokonania randomizowanego eksperymentu porównawczego dla jednego czynnika. Podobnie rzecz się ma dla eksperymentu porównawczego dla kilku czynników. W praktyce stosuje się kilka wariantów tych podstawowych schematów, które pokrótce omówimy.

Czasami grupa kontrolna nie jest potrzebna, gdyż standard związany z wartością czynnika stosowanego w przypadku grupy kontrolnej jest dobrze znany. Na przykład w celu stwierdzenia skuteczności przeprowadzonej reorganizacji w urzędzie gminnym możemy rozpatrzeć losowo wybraną grupę klientów tego urzędu, którzy załatwiają w nim sprawy co najmniej raz w miesiącu

i potraktować ich jako grupę eksperymentalną. Jednostkom z tej grupy zadajemy pytania dotyczące jakości obsługi: czy pozostała taka sama, czy się pogorszyła czy też polepszyła? Standard, do którego się odnosimy to, jakość obsługi w urzędzie gminnym przed reorganizacją oceniana przez te same osoby. W innym przypadku może nas przykładowo interesować czas reakcji ludzi zdrowych w godzinę po podaniu pewnego leku psychotropowego. W tym przypadku odniesieniem będzie rozkład czasu reakcji w populacji ludzi zdrowych określony na podstawie uprzednich badań.

**Schematy eksperymentalne dla wielu czynników.** W wielu przypadkach mamy do czynienia z wieloma czynnikami mogącymi mieć wpływ na interesującą nas cechę. Zagadnieniu analizy siły i struktury zależności są poświęcone trzy następne rozdziały. Rozpatrzmy tutaj tylko następujący przykład wprowadzający.

**Przykład 2.30.** Producent środka czyszczącego pewnej marki chce ocenić jak dobrze sprzedaje się on w supermarketach w pewnym mieście i jak zależy to od różnych czynników. W tym celu rozpatruje wszystkie 5 supermarketów w tym mieście, oznaczając je liczbami od 1 do 5 i traktując numer supermarketu jako pierwszy czynnik o wartościach ze zbioru $\{1, 2, \ldots, 5\}$. Następnymi rozpatrywanymi czynnikami są: cena (3 wielkości), opakowanie proszku (3 możliwe kolory) i wyeksponowanie proszku na półce (na niskim poziomie, w zasięgu ramion, ale nie na wysokości oczu i na wysokości oczu). W ten sposób otrzymuje się $5 \times 3 \times 3 \times 3 = 105$ możliwych wartości układu czynników

$$(\text{nr supermarketu},\ \text{cena},\ \text{kolor},\ \text{miejsce}).$$

Dla każdej ze 105 możliwych wartości układu czynników rozpatruje się liczbę opakowań proszku sprzedanych od poniedziałku do niedzieli przy zastosowaniu tego układu i następnie analizuje się zależności między rozpatrywanymi zmiennymi.

W innej sytuacji badania jakości stali jako interesujące czynniki technolodzy mogą uznać na przykład zawartość węgla w stopie, temperaturę surówki oraz metodę schładzania. W naukach rolniczych typowym przykładem takiego schematu eksperymentalnego jest badanie zależności wielkości plonu dla pewnej uprawy od intensywności nawadniania, nasłonecznienia i rodzaju nawozu. Zauważmy, że każdemu układowi wartości czynników warto przypisać więcej niż jedną jednostkę, która będzie dla tego układu testowana. Możemy wtedy ocenić wielkość naturalnej zmienności interesującej nas reakcji jednostek poddanych takim samym wpływom czynników. Mamy wtedy

do czynienia z **replikacjami** wyników dla ustalonego układu wartości czynników.

**Blokowanie.** W przykładzie 2.29 podkreśliliśmy konieczność kontrolowania w obu grupach, kontrolnej i eksperymentalnej, wszystkich czynników, których wpływ na zmienną objaśniającą chcemy badać. Czasami jednak trzeba zastosować inną metodę ujednolicania grup. Jej zastosowanie jest pożądane wtedy, gdy z góry wiemy, że istnieją zmienne, mające duży wpływ na badany wynik, które nie są rozpatrywane jako czynniki. Metoda ta polega na tzw. blokowaniu.

**Przykład 2.31.** Rozpatrzmy badanie prowadzone nad skutecznością reklamy określonego produktu, w którym jako możliwe czynniki rozpatrujemy długość reklamy oraz liczbę jej powtórzeń. Wiadomo jednak, że mężczyźni i kobiety odmiennie reagują na reklamy. Aby to uwzględnić rozpatruje się populację kobiet i mężczyzn oddzielnie, tworząc odpowiednie grupy kontrolne i eksperymentalne dla każdej z nich. W tym przypadku zmienną blokującą jest płeć. W eksperymencie dotyczącym niezawodności działania zastawki serca nowego typu zmienną blokującą może być wiek pacjenta lub jego ogólny stan zdrowia.

Celem blokowania jest zmniejszenie zmienności obserwowanej reakcji spowodowanej niejednorodnością grup. W przykładzie zmiennymi blokującymi były pewne zmienne związane z jednostkami. Przykładowe charakterystyki jednostek również często stosowane w tym celu to np. w przypadku ludzi wielkość dochodów, poziom wykształcenia lub doświadczenie. Zmiennymi blokującymi mogą być również charakterystyki samego eksperymentu takie jak: czas jego przeprowadzenia, rodzaj użytego instrumentu pomiarowego lub partia materiału, z której pobieramy jednostki.

Podkreślmy, że kwestie omówione w tym podpunkcie nie wyczerpują wszystkich istotnych problemów związanych z przeprowadzeniem eksperymentu. W praktyce możemy spotkać się z mniej typowymi trudnościami. Na przykład czas trwania całego eksperymentu może być znaczny i moment, w którym jest wykonywana jego kolejna faza może mieć wpływ na wielkość interesującej nas zmiennej. W przykładzie 2.30 ilość sprzedanego proszku może zależeć od okresu, w którym są prowadzone badania (np. okresu przedświątecznego lub poświątecznego). W takiej sytuacji efekt czasowy można uwzględnić, traktując czas jako zmienną blokującą. W przypadku badania wpływu więcej niż jednego poziomu czynnika wykorzystuje się czasami tę samą grupę eksperymentalną, poddając ją działaniu kolejnych wartości czynnika. Istotne jest wtedy, na ile grupa eksperymentalna, której czynnik był już aplikowany, zachowuje „pamięć" o tym fakcie. Może wystąpić

niepożądany efekt uczenia się jednostek z grupy eksperymentalnej, jak np. przy pomiarze czasu wykonywania określonej czynności po podaniu dawki leku poprawiającego koncentrację. Może także wystąpić efekt zmęczenia jednostek. Większości tych problemów można uniknąć określając wyraźnie cele i sposób dokonania eksperymentu. W szczególności istotne jest sprecyzowanie z góry, jakie czynniki badamy, jak zdefiniowana i jak mierzona jest zmienna wynikowa, jakiej populacji dotyczą badania i jak są wybierane jednostki do grupy eksperymentalnej i kontrolnej.

## 2.6. Zadania

**2.1.** Niech dany będzie zbiór $\mathcal{S} = \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$ oraz jego trzy podzbiory: $A = \{1, 2, 8, 9\}$, $B = \{2, 4, 6\}$ oraz $C = \{1, 3, 7\}$. Znaleźć zbiory

$$A \cap B, \quad B \cup C, \quad A \cap \{B \cup C\}'.$$

**2.2.** Krótkie komunikaty pogodowe mówią czy pogoda w danym miejscu jest słoneczna (oznaczana symbolem S), czy jest wietrzna (W) oraz czy pada deszcz (D). Używając symboli algebry zbiorów, zapisać zdarzenia odpowiadające następującym pogodom: pogodzie bezwietrznej, pogodzie słonecznej i wietrznej, pogodzie bezdeszczowej i jednocześnie bezsłonecznej, pogodzie deszczowej lub wietrznej, pogodzie deszczowej lub wietrznej, ale nie jednocześnie deszczowej i wietrznej.

**2.3.** Opisać przestrzeń zdarzeń elementarnych w następujących doświadczeniach:
**a)** Moneta jest podrzucana aż do wypadnięcia po raz pierwszy kolejno dwóch orłów.
**b)** Dane są dwie urny, jedna z 2 białymi kulami i 1 czerwoną kulą oraz druga z 4 białymi i 6 czerwonymi kulami. Doświadczenie polega na wybraniu urny i następnie wylosowaniu z niej jednej kuli.
**c)** Dane są dwie urny, jedna z 4 białymi kulami i 3 czerwonymi kulami oraz druga z 3 żółtymi i 2 czarnymi kulami. Doświadczenie polega na jednoczesnym wylosowaniu po jednej kuli z każdej z urn.

**2.4.** Wykazać, że jeżeli zdarzenie $A$ i $B$ są niezależne, to także niezależne są pary zdarzeń: $A'$ i $B$, $A$ i $B'$ oraz $A'$ i $B'$.
**Wskazówka:** W pierwszym przypadku skorzystać z tego, że $P(B) = P((A' \cup \cup A) \cap B)$ i odpowiednio zapisać prawą stronę tej równości. W następnych przypadkach zastosować podobne rozumowanie.

**2.5.** Satelita nadaje sygnały, które są niezależnie odbierane przez stację odbiorczą. Każdy sygnał jest odbierany z prawdopodobieństwem 0,999. Jakie jest prawdopodobieństwo, że stacja odbierze pierwszych 100 nadanych

sygnałów i nie odbierze sygnału sto pierwszego? Jakie jest prawdopodobieństwo, że stacja nie odbierze trzech pierwszych sygnałów nadanych przez satelitę i odbierze 97 następnych sygnałów?

**2.6.** Absolwent szkoły wyższej złożył swój życiorys oraz list intencyjny podjęcia pracy w 4 przedsiębiorstwach, A, B, C i D. Opierając się na danych z przeszłości, dotyczących absolwentów z podobnymi życiorysami, wie, że prawdopodobieństwo otrzymania oferty pracy z przedsiębiorstwa A wynosi 0,8, z przedsiębiorstwa B 0,4, z przedsiębiorstwa C 0,4 i z D 0,2. Podać prawdopodobieństwo otrzymania oferty z co najmniej jednego przedsiębiorstwa (zakładamy, że decyzje przedsiębiorstw są od siebie niezależne).
**Wskazówka:** Rozważyć najpierw zdarzenie, że absolwent nie otrzyma ani jednej oferty.

**2.7.** Doświadczenie polega na czterokrotnym rzucie monetą i odnotowaniu kolejnych wyników. Zakłada się, że rzuty są niezależne oraz, że prawdopodobieństwo otrzymania orła w danym rzucie wynosi 1/2. Jakie jest prawdopodobieństwo otrzymania w 4 rzutach 3 orłów?

**2.8.** Wiadomo, że w pudełku z 40 kulkami łożyskowymi o zadanej średnicy znajduje się 6 kulek z niewidocznymi wadami. Jakie jest prawdopodobieństwo, że wśród wybranych na chybił trafił 4 kulek nie ma ani jednej wadliwej? Jakie jest prawdopodobieństwo, że co najmniej jedna spośród nich jest wadliwa? Jakie jest prawdopodobieństwo, że dokładnie jedna jest wadliwa?

**2.9.** Do centralnego procesora są kierowane programy napływające z trzech różnych źródeł. Niech $X_1$, $X_2$ i $X_3$ oznaczają zdarzenie, odpowiednio, że program pochodzi ze źródła pierwszego, drugiego i trzeciego. Ilości programów napływających ze źródeł $X_1$, $X_2$ i $X_3$ mają się do siebie jak 1:2:2. Procesor kieruje programy na jeden z dwóch procesorów wykonujących obliczenia w układzie równoległym.

Niech $Y_1$ i $Y_2$ będą zdarzeniami polegającymi na tym, że napływający program jest kierowany odpowiednio na procesor pierwszy i drugi. Wiadomo, że

$$P(Y_1|X_1) = 0{,}6, \quad P(Y_1|X_2) = 0{,}3 \quad P(Y_1|X_3) = 0{,}4,$$

$$P(Y_2|X_1) = 0{,}4, \quad P(Y_2|X_2) = 0{,}7 \quad P(Y_2|X_3) = 0{,}6.$$

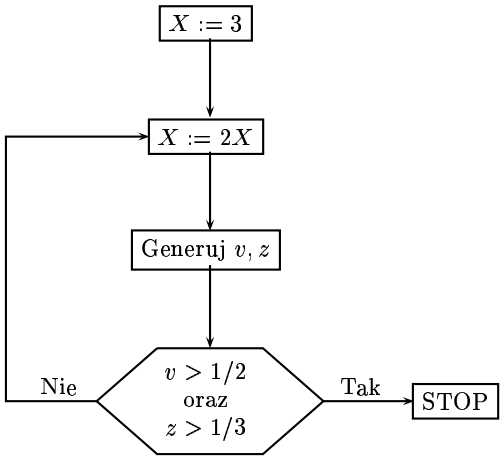
**a)** Jakie jest prawdopodobieństwo tego, że program skierowany do drugiego procesora wykonującego obliczenia pochodzi ze źródła pierwszego.
**b)** Obliczyć prawdopodobieństwo tego, że program skierowany do centralnego procesora zostanie następnie skierowany do drugiego procesora wykonującego obliczenia.

**2.10.** Rozważmy taki wariant przykł. 2.12, w którym osoby przed poddaniem testowi są wstępnie diagnozowane innymi metodami i tylko te, wobec

których istnieje wystarczająco mocne podejrzenie infekcji, przechodzą test. W rezultacie, na 100 osób poddanych testowi średnio 75 jest zdrowych oraz 25 chorych. Sprawdzić, jakie jest w tej sytuacji prawdopodobieństwo, że osoba, którą poddano testowi medycznemu i w przypadku której test dał wynik dodatni, cierpi na testowaną chorobę.

**2.11.** Prawdopodobieństwo zagrożenia pożarem w dużej fabryce chemicznej wynosi 0,02. W fabryce jest oczywiście zainstalowany system alarmowy. W sytuacji awaryjnej system zawodzi w 1% przypadków. Fałszywy alarm systemu zdarza się w 5% przypadków. Obliczyć warunkowe prawdopodobieństwo, że kolejny alarm włączony przez system okaże się fałszywy.

**2.12.** Rozpatrzmy schemat blokowy przedstawiony na rys. 2.14, w którym $v$ i $z$ oznaczają niezależne zmienne losowe o rozkładzie jednostajnym na przedziale $[0, 1]$:




Rys. 2.14. Schemat blokowy do zad. 2.12

Znaleźć rozkład prawdopodobieństwa $X$ na wyjściu układu.

**2.13.** Obliczyć wartość średnią i wariancję zmiennej losowej równej liczbie oczek w jednym rzucie kostką o sześciu bokach ponumerowanych od 1 do 6.

**2.14.** Dla zmiennej losowej $X$ o funkcji prawdopodobieństwa danej tabelką

| $x_i$ | 1 | 2 | 4 | 5 |
|---|---|---|---|---|
| $p_i$ | 0,4 | 0,2 | 0,1 | 0,3 |

wyznaczyć wartość średnią, odchylenie standardowe oraz skonstruować dystrybuantę zmiennej $X$.

**2.15.** Dystrybuanta $F$ zmiennej losowej $X$ jest określona następującą tabelką:

| $x$ | $(-\infty, 1)$ | $[1, 3)$ | $[3, 6)$ | $[6, \infty)$ |
|---|---|---|---|---|
| $F(x)$ | 0 | 0,3 | 0,6 | 1 |

Wyznaczyć funkcję prawdopodobieństwa tej zmiennej.

**2.16.** Zmienna losowa $X$ ma rozkład o dystrybuancie

$$F(x) = \begin{cases} 0 & \text{dla } x \leq 1 \\ \frac{x-1}{2} & \text{dla } 1 < x \leq 3 \\ 1 & \text{dla } x > 3. \end{cases}$$

Wyznaczyć wartość średnią oraz odchylenie standardowe zmiennej losowej $X$.

**2.17.** Zmienna losowa $X$ ma rozkład o gęstości

$$f(x) = \begin{cases} cx(1-x) & \text{dla } 0 < x < 1 \\ 0 & \text{poza tym,} \end{cases}$$

gdzie $c$ jest stałą dodatnią. Wyznaczyć wartość stałej $c$ i następnie wyznaczyć wartość średnią oraz odchylenie standardowe zmiennej losowej $X$.

**2.18.** Czas reakcji $X$ na pewien typ bodźca jest ciągłą zmienną losową o gęstości rozkładu

$$f(x) = \begin{cases} \frac{3}{2x^2}, & 1 \leq x \leq 3 \\ 0, & x < 1 \text{ lub } x > 3. \end{cases}$$

Obliczyć prawdopodobieństwo $P(1,5 \leq X \leq 2,5)$ i następnie wartość średnią oraz odchylenie standardowe zmiennej $X$.

**2.19.** Niech $T$ będzie zmienną losową o rozkładzie wykładniczym z parametrem $\lambda$, oznaczającą jak w przykł. 2.22 czas oczekiwania na zwolnienie publicznego aparatu telefonicznego. Wiadomo, że

$$P(T > t) = e^{-\lambda t}.$$

**a)** Udowodnić, że

$$P(T > t + s | T > t) = P(T > s) = e^{-\lambda s}.$$

**b)** Wyjaśnić dlaczego udowodniona własność uzasadnia nazywanie zmiennej losowej o rozkładzie wykładniczym zmienną *pozbawioną pamięci.*

**Wskazówka do p. a):** Skorzystać z tego, że na mocy definicji prawdopodobieństwa warunkowego dowodzona równość równoważna jest następującej:

$$P(T > t + s) = P(T > t)P(T > s).$$

**2.20.** Mówimy, że zmiena losowa $X$ ma rozkład gamma z parametrami $\alpha$ i $\beta$, gdzie $\alpha$ i $\beta$ są dowolnymi ustalonymi stałymi dodatnimi, gdy jej gęstość prawdopodobieństwa jest dana wzorem

$$f(x) = \begin{cases} \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta}, & \text{gdy } x > 0 \\ 0, & \text{gdy } x \leq 0, \end{cases}$$

gdzie $\Gamma(\alpha)$ jest tzw. funkcją gamma,

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx,$$

$\alpha > 0$.
**a)** Całkując przez części, wykazać, że $\Gamma(\alpha + 1) = \alpha\Gamma(\alpha)$.
**b)** Wykazać, że wartość średnia i wariancja zmiennej losowej $X$ wynoszą, odpowiednio, $\alpha\beta$ i $\alpha\beta^2$.
**c)** Sprawdzić, jaki jest kształt gęstości rozkładu gamma dla różnych wartości parametrów $\alpha$ i $\beta$. Zauważyć, że szczególną rolę odgrywa rozkład gamma z parametrem $\alpha = 1$ (pod jaką nazwą jest znany ten rozkład i jak się ma jego parametr do parametru $\beta$?). Wyjaśnić dlaczego parametr $\alpha$ jest zwany parametrem kształtu rozkładu, natomiast parametr $\beta$ nosi nazwę parametru skali.

**2.21.** Niech będą dane ciągłe zmienne losowe $X$ i $Y$ o łącznej gęstości prawdopodobieństwa $f(x, y)$ i gęstości brzegowej $f_y(y)$ zmiennej $Y$. Niech $x$ i $y$ będą ustalonymi liczbami, przy czym $f(y) > 0$. Wykazać, że dla dowolnej liczby dodatniej $h$

$$P(X \leq x | y - h < Y \leq y) = \frac{\int_{-\infty}^{x} \int_{y-h}^{y} f(s,t) dt ds}{\int_{y-h}^{y} f_y(t) dt}.$$

Powyższemu ułamkowi można nadać postać

$$\frac{\frac{1}{h} \int_{-\infty}^{x} \int_{y-h}^{y} f(s,t) dt ds}{\frac{1}{h} \int_{y-h}^{y} f_y(t) dt}.$$

Na podstawie całkowego twierdzenia o wartości średniej z analizy matematycznej zauważyć, że

$$\lim_{h \downarrow 0} \frac{\frac{1}{h} \int_{-\infty}^{x} \int_{y-h}^{y} f(s,t) dt ds}{\frac{1}{h} \int_{y-h}^{y} f_y(t) dt} = \frac{\int_{-\infty}^{x} f(s,y) ds}{f_y(y)},$$

czyli, że

$$\lim_{h\downarrow 0} P(X \leq x | y-h < Y \leq y) = \int_{-\infty}^{x} \frac{f(s,y)}{f_y(y)} ds.$$

Powyższa równość uzasadnia następujące rozumienie warunkowego prawdopodobieństwa zdarzenia $X \leq x$ pod warunkiem $Y = y$:

$$P(X \leq x | Y = y) = \int_{-\infty}^{x} \frac{f(s,y)}{f_y(y)} ds.$$

Zauważyć, że funkcja $\frac{f(x,y)}{f_y(y)}$, traktowana jako funkcja argumentu $x$, jest dla każdej ustalonej wartości $y$ nieujemna i spełnia własność

$$\int_{-\infty}^{\infty} \frac{f(x,y)}{f_y(y)} dx = 1.$$

Na podstawie powyższych obliczeń podać intuicyjne uzasadnienie wzoru (2.32) dla przypadku ciągłych zmiennych losowych.

**2.22.** Niech będą dane zmienne losowe $X$ i $Y$ o łącznej gęstości danej wzorem

$$f(x,y) = \begin{cases} 8xy, & \text{gdy } 0 \leq x \leq y \leq 1 \\ 0, & \text{w przypadku przeciwnym.} \end{cases}$$

Wykazać, że gęstość brzegowa zmiennej losowej $X$ ma postać

$$f_x(x) = \begin{cases} 4x(1-x^2), & \text{gdy } 0 \leq x \leq 1 \\ 0, & \text{w przypadku przeciwnym,} \end{cases}$$

natomiast gęstość brzegowa zmiennej losowej $Y$

$$f_y(y) = \begin{cases} 4y^3, & \text{gdy } 0 \leq y \leq 1 \\ 0, & \text{w przypadku przeciwnym.} \end{cases}$$

Wykazać, że gęstość warunkowa zmiennej losowej $X$ pod warunkiem, że zmienna losowa $Y$ przyjęła ustaloną wartość $y$ ma postać

$$f(x|y) = \begin{cases} \frac{2x}{y^2}, & \text{gdy } 0 \leq x \leq y \\ 0, & \text{w przypadku przeciwnym} \end{cases}$$

oraz że gęstość warunkowa zmiennej losowej $Y$ pod warunkiem, że zmienna losowa $X$ przyjęła ustaloną wartość $x$ wynosi

$$f(y|x) = \begin{cases} \frac{2y}{1-x^2}, & \text{gdy } x \leq y \leq 1 \\ 0, & \text{w przypadku przeciwnym.} \end{cases}$$

Podać gęstość warunkową zmiennej losowej $X$ pod warunkiem, że $Y = 1/2$.

**2.23.** Uzasadnić następującą definicję warunkowej wartości średniej dyskretnej zmiennej losowej $X$ pod warunkiem, że dyskretna zmienna losowa $Y$ przyjęła wartość $y$:

$$\mu_{X|Y=y} = \sum_{i=1}^{\infty} x_i f(x_i|y),$$

gdzie $x_1, x_2, \ldots$ oznaczają wszystkie różne wartości zmiennej losowej $X$, natomiast $f(\cdot|y)$ jest warunkową funkcją prawdopodobieństwa zmiennej $X$ pod warunkiem, że zmienna $Y$ przyjęła wartość $y$. Podać definicję warunkowej wartości średniej ciągłej zmiennej losowej $X$ pod warunkiem, że ciągła zmienna losowa $Y$ przyjęła wartość $y$.

**2.24.** Na podstawie zad. 2.22 obliczyć w przykł. 2.24 warunkową wartość średnią zmiennej losowej $X$ pod warunkiem, że zmienna losowa $Y$ przyjęła wartość 2.

**2.25.** Na podstawie zad. 2.22 obliczyć w zad. 2.21 warunkową wartość średnią zmiennej losowej $X$ pod warunkiem, że zmienna losowa $Y$ przyjęła wartość 1/3.

**2.26.** Pewna część chłodnicy samochodowej jest formowana na tłoczni wtryskowej. Z doświadczenia wynika, że najczęściej występującymi defektami na powierzchni tej części są zniekształcenia na jej końcach oraz tzw. wypływki. Prawdopodobieństwa wystąpienia zadanej liczby zniekształceń (zmienna losowa $X$) i zadanej liczby wypływek (zmienna losowa $Y$) na powierzchni części są podane w następującej tablicy:

| | | | $X$: Liczba zniekszt. | | | |
|---|---|---|---|---|---|---|
| | | | 0 | 1 | 2 | 3 |
| | | 0 | 0,53 | 0,05 | 0,02 | 0,01 |
| | | 1 | 0,17 | 0,03 | 0,02 | 0,01 |
| $Y$: | Liczba | 2 | 0,05 | 0,03 | 0,02 | 0,01 |
| | wypływek | 3 | 0,02 | 0,01 | 0,01 | 0,00 |
| | | 4 | 0,01 | 0,00 | 0,00 | 0,00 |

Obliczyć średnią liczbę zniekształceń oraz średnią liczbę wypływek. Obliczyć średnią liczbę defektów (łącznie zniekształceń i wypływek). Jaka jest średnia liczba zniekształceń wśród tych chłodnic, które mają jedną wypływkę (por. zad. 2.22)?

**2.27.** Udowodnić niezależność zmiennych losowych $X$ i $Y$ o łącznej gęstości

$$f(x,y) = \begin{cases} \frac{4}{3}x(1+y), & \text{gdy} \quad 0 \le x \le 1 \quad \text{oraz} \quad 0 \le y \le 1 \\ 0, & \text{w przypadku przeciwnym.} \end{cases}$$

**2.28.** Na podstawie wzorów (2.33) oraz (2.34) udowodnić, że jeżeli (ciągłe lub dyskretne) zmienne losowe $X$ i $Y$ są niezależne, to

$$E(XY) = E(X)E(Y).$$

**2.29.** Zmienna losowa o rozkładzie jednostajnym na przedziale $[0,1]$ ma wartość średnią $\mu = 1/2$ i odchylenie standardowe $\sigma = 1/\sqrt{12}$.
**a)** Wylosować 6 obserwacji z wyżej wspomnianego rozkładu. Niech

$$\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$$

oraz

$$z_n = \frac{\bar{x}_n - \mu}{\sigma/\sqrt{n}},$$

gdzie $x_i$, $i = 1, 2, \ldots, n$, oznacza wylosowane obserwacje i $n = 6$. Powtórzyć opisane doświadczenie losowania 6 obserwacji i obliczenia statystyk $\bar{x}_n$ i $z_n$ 100 razy. Skonstruować histogramy dla 100 uzyskanych wartości $\bar{x}_n$ i 100 uzyskanych wartości $z_n$, obrazujące rozkłady wylosowanych obserwacji.
**b)** Wylosować 12 obserwacji z wyżej wspomnianego rozkładu i obliczyć $\bar{x}_n$ oraz $z_n$, gdzie $n = 12$. Powtórzyć opisane doświadczenie losowania 12 obserwacji i obliczenia statystyk $\bar{x}_n$ i $z_n$ 100 razy. Skonstruować histogramy dla uzyskanych wartości $\bar{x}_n$ i wartości $z_n$.
**c)** Powtórzyć czynności z punktu b) dla 24, 48, 100 i 500 obserwacji z rozkładu jednostajnego na przedziale $[0,1]$.
Czy wyniki przeprowadzonych eksperymentów są zgodne z oczekiwaniami wynikającymi ze znajomości mocnego prawa wielkich liczb i Centralnego Twierdzenia Granicznego?

**2.30.** Zbiór składający się z 1000 znaków jest przesyłany między dwoma komputerami. Prawdopodobieństwo błędnej transmisji jednego znaku wynosi 0,02. Zdarzenia błędnej transmisji dla różnych znaków są niezależne. Oszacować prawdopodobieństwo, że podczas transmisji liczba błędów mieści się w granicach od 10 do 25.

**2.31.** Ocenia się, że liczba zgłoszeń usterek pewnego modelu samochodu w okresie gwarancyjnym, jakie w ciągu jednego tygodnia otrzymują warsztaty producenta na terenie całego kraju, jest zmienną losową o rozkładzie Poissona z parametrem $\lambda = 90$. Oszacować prawdopodobieństwo, że w danym tygodniu liczba zgłoszeń reklamacji mieści się w granicach od 80 do 110.

**2.32.** Książka składa się z 860 stron tekstu. Zdarzenie polegające na wystąpieniu przynajmniej jednego błędu literowego na jednej stronie jest niezależne od zajścia takich zdarzeń na innych stronach, jego zaś prawdopodobieństwo wynosi 0,007. Korzystając z pakietu statystycznego, podać prawdopodobieństwo znalezienia się w książce przynajmniej 6, ale nie więcej niż 10 stron z błędami literowymi. Oszacować to prawdopodobieństwo za pomocą przybliżenia rozkładu dwumianowego rozkładem Poissona. Ponieważ

heurystyczne warunki, umożliwiające użycie przybliżenia rozkładu dwumianowego rozkładem normalnym, są spełnione ($np = 6{,}02$ oraz $n(1-p) = 853{,}98$), zastosować to ostatnie przybliżenie. Czy otrzymany wynik daje oszacowanie równie dobre jak przybliżenie poissonowskie?

**2.33.** W celu stwierdzenia, czy stałe przyjmowanie aspiryny wpływa na zmniejszenie ryzyka kolejnych zawałów serca, planuje się eksperyment, w którym wezmą udział mężczyźni po przebytym zawale. Eksperyment ma trwać dwa lata. W tym czasie osobom z dwóch grup eksperymentalnych jest podawana raz dziennie odpowiednio dawka 150 mg i 300 mg aspiryny. Opisać dokładnie cały eksperyment, uwzględniając w jego planowaniu fakt, że prawdopodobieństwo nowego zawału zależy od liczby przebytych dotychczas zawałów.

ROZDZIAŁ 3

# Wnioskowanie statystyczne

## 3.1. Wprowadzenie

Niniejszy rozdział w całości poświęcamy rozpoczętemu w podrozdz. 2.4 systematycznemu wykładowi podstaw wnioskowania statystystycznego. Kilkakrotnie zwracaliśmy już uwagę na to, że do oceny wiarogodności wyników uzyskiwanych na podstawie zebranych danych jest potrzebny model statystyczny, opisujący rozkład prawdopodobieństwa mierzonej wielkości. Na przykład, bez modelu statystycznego nie wiemy, jaka rzeczywiście może być średnia wartość wzrostu dorosłego Polaka, mimo, że dysponujemy próbą $n = 100$ obserwacji, których średnia wartość $\bar{x}$ wynosi 176,5 cm. Jest dla nas doskonale oczywiste, że inna próba dałaby nam najpewniej inną średnią. Co gorsze jednak, bez poczynienia dodatkowych założeń, nie potrafimy powiedzieć jak bardzo otrzymana wartość średnia w próbie może się różnić od wartości średniej wzrostu w całej populacji dorosłych Polaków.

Już jednak założenie, że mamy do czynienia z prostą próbą losową z (nieznanego!) rozkładu zmiennej losowej $X$ o znanym odchyleniu standardowym $\sigma$, umożliwia skorzystanie z Centralnego Twierdzenia Granicznego, by na tej podstawie oszacować prawdopodobieństwo, że estymator $\bar{X}$ różni się od prawdziwej wartości średniej $\mu$ wzrostu o więcej niż zadaną liczbę centymetrów (takie postępowanie omówiliśmy w podrozdz. 2.4). Poczynione założenie skłania zarazem do zadania natychmiast pytania, na jakiej podstawie można założyć znajomość odchylenia standardowego rozkładu. Swego rodzaju modyfikacją tego pytania jest pytanie o możliwość zrezygnowania z wymagania znajomości wartości $\sigma$ i mimo to oszacowania prawdopodobieństwa uzyskania wyniku różniącego się o więcej niż zadaną liczbę centymetrów od prawdziwej wartości średniej rozkładu $\mu$.

Co więcej, jak na to zwróciliśmy już uwagę w p. 2.4.4, pytaniem podstawowym jest pytanie o to, jaką postać powinien mieć estymator wartości średniej rozkładu, by móc uznać go za w jakimś sensie najlepszy lub przynajmniej dobry. Innymi słowy powstaje pytanie, jak najpierw określić jakość estymatora

i następnie, jaki estymator jest w wybranym sensie najlepszy. W szczególności, chcemy umieć orzec, w świetle jakiego kryterium uzasadnione jest użycie średniej w próbie jako estymatora wartości średniej rozkładu.

W punkcie 2.4.4 podaliśmy już dwa kryteria oceny jakości estymatora, a mianowicie wprowadziliśmy pojęcia estymatorów nieobciążonych o minimalnej wariancji oraz estymatorów dopuszczalnych ze względu na błąd średniokwadratowy. Wspomnieliśmy także o istnieniu jeszcze innego sposobu wyboru estymatorów, zwanego estymacją największej wiarogodności (w skrócie mówimy o estymacji NW i o estymatorach NW). W podrozdziale 3.2 omówimy to ostatnie zagadnienie, jak też przedyskutujemy czwarty sposób konstrukcji estymatorów, tzw. estymatory oparte na metodzie momentów (estymatory MM). Wprowadzenie estymatorów NW wynika z ich dobrych własności oraz z tego, że łatwiej je uzyskać niż estymatory NMW. Często nie potrafimy podać estymatora NMW, wiemy zaś jak uzyskać estymator NW. Z kolei estymatory MM odgrywają, jak się dowiemy w p. 3.2.1, rolę w pewnym sensie pomocniczą.

Interesujące nas wartości parametrów są liczbami. Także przedyskutowane lub wspomniane dotąd estymatory przyjmują wartości liczbowe. Na przykład, liczbą jest zaobserwowana wartość średnia w próbie. Zauważmy, jednak że zamiast konstruować estymatory, których wartości powinny dawać możliwie dobre oszacowania nieznanych parametrów populacji, można by zastanowić się nad konstrukcją losowego przedziału pokrywającego z dużym prawdopodobieństwem prawdziwą wartość interesującego nas parametru. Ściślej, konstrukcja losowego przedziału sprowadza się do podania jego losowych brzegów, czyli do zaproponowania estymatora, którego wartościami są pary liczb. Taki losowy przedział nazywamy nieraz **estymatorem przedziałowym**. „Zwykłe", czyli wcześniej wspominane estymatory często nazywamy dla odróżnienia **estymatorami punktowymi**. Estymatorom przedziałowym jest poświęcony podrozdz. 3.3.

Problemem pokrewnym do zagadnienia estymacji parametru – zarówno estymacji punktowej, jak i zwłaszcza przedziałowej – jest problem testowania hipotez. Z tym ostatnim problemem mamy na przykład do czynienia, gdy nie chcemy estymować wartości średniej rozkładu, natomiast interesuje nas czy wartość ta jest mniejsza (lub większa, albo po prostu różna) od pewnej ustalonej wartości. Nie pytamy zatem ile wynosi wartość średnia $\mu$, ale czy $\mu < \mu_0$ (lub czy $\mu > \mu_0$, albo czy $\mu \neq \mu_0$), gdzie $\mu_0$ jest pewną z góry ustaloną liczbą. Z problemem testowania hipotezy mamy także do czynienia w sytuacji bardziej złożonej, gdy na przykład interesuje nas ocena prawdziwości przypuszczenia, że rozkład badanej populacji jest rozkładem normalnym (o dowolnych parametrach). Jest to zadanie trudniejsze od wcześniej poruszanych i jakościowo od nich różne – tym razem nie interesuje nas żaden

liczbowy parametr rozkładu, a ogólna postać tego rozkładu. Zagadnieniom testowania hipotez jest poświęcony podrozdz. 3.4.

Zanim zakończymy te uwagi wstępne, omówimy jeszcze krótko dwa zagadnienia. Po pierwsze, zwróćmy uwagę, że terminu parametr używamy właściwie w dwóch znaczeniach. Po drugie rozumiemy ten termin „szeroko", mając na myśli dowolną wielkość liczbową, charakteryzującą ustaloną własność rozkładu prawdopodobieństwa. Tak jest np. w przypadku dowolnego parametru położenia lub parametru rozproszenia. Tak jest np. także wtedy, gdy chcemy estymować górny kwartyl rozkładu albo kwantyl dowolnego ustalonego rzędu.

Pojęcie parametru można jednak rozumieć również „wąsko", gdy parametr ten występuje we wzorze określającym rozkład prawdopodobieństwa. Mówimy na przykład o rozkładzie Poissona z parametrem $\lambda$, mając na myśli, że wzór na funkcję prawdopodobieństwa rozkładu Poissona zawiera parametr $\lambda$ o ustalonej wartości. Przypomnijmy, że z rozkładem Poissona z parametrem $\lambda$ jest związany rozkład ciągły, opisujący czas między kolejnymi zdarzeniami procesu Poissona, a mianowicie rozkład wykładniczy także z parametrem $\lambda$ (parametr ów występuje w definicji gęstości rozkładu wykładniczego). Niekiedy definicja rozkładu zawiera więcej niż jeden parametr. Tak jest np. w przypadku rozkładu dwumianowego (z parametrami $n$ i $p$) oraz w przypadku rozkładu normalnego (z parametrami $\mu$ i $\sigma$). Oczywiście, w danym zadaniu statystycznym nie wszystkie parametry rozkładu muszą być nieznane – np. w przypadku zadania z rozkładem dwumianowym najczęściej jest znany parametr $n$ (w przypadku każdego zadania, z jakim będziemy mieć do czynienia, będziemy zawsze mówili, które parametry są nieznane, a które są znane).

Zauważmy, że każdy parametr w sensie „szerokim" musi być jakąś funkcją parametru lub parametrów występujących w definicji rozkładu prawdopodobieństwa populacji, czyli parametru lub parametrów w sensie „wąskim". Rzeczywiście, każda liczbowa charakterystyka dowolnej ustalonej własności rozkładu musi być jednoznacznie wyznaczona przez funkcję określającą ten rozkład, a zatem musi być funkcją parametru (lub parametrów) w sensie „wąskim". Na przykład, w p. 2.2.6 wykazaliśmy, że wartość średnia rozkładu wykładniczego z parametrem $\lambda$ jest równa $\lambda^{-1}$. Jest inną sprawą, że nie zawsze musimy wiedzieć, jaką konkretnie funkcją parametrów w sensie „wąskim" jest interesujący nas parametr w sensie „szerokim".

Ostatnim zagadnieniem, o jakim wspomnimy w tych uwagach, jest zauważenie, że na przykład estymacja parametru rozkładu wykładniczego albo estymacja wartości średniej i odchylenia standardowego rozkładu normalnego jest w istocie równoznaczna z estymacją rozkładu populacji, z której pochodzi próba losowa. Rzeczywiście, oszacowanie wspomnianych parame-

trów jest równoznaczne z oszacowaniem gęstości populacji. To, że do oszacowania funkcji gęstości wystarczy obliczenie skończenie wielu estymatorów o wartościach liczbowych wynika stąd, iż założyliśmy stosunkowo dokładną znajomość modelu probabilistycznego rządzącego badanym zjawiskiem – założyliśmy mianowicie znajomość tego modelu z dokładnością do skończenie wielu parametrów liczbowych. W omawianych przypadkach estymację parametrów w „wąskim" sensie, określających nieznany rozkład populacji, możemy zatem nazwać **parametryczną estymacją** rozkładu populacji.

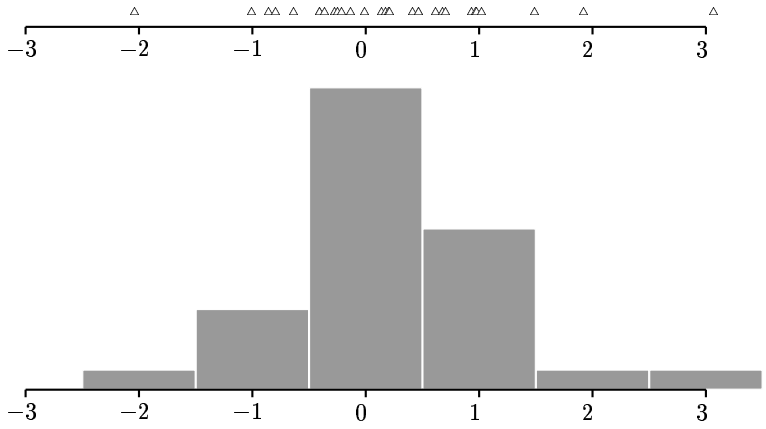
Zarazem, innym oszacowaniem nieznanej gęstości populacji jest opisany w rozdz. 1 histogram. Jak o tym wspomnieliśmy w p. 1.1.2, ze względu na nieciągłość histogramu, może on być zastąpiony estymatorem ciągłym (albo jeszcze gładszym, czyli mającym pochodną zadanego rzędu), np. odpowiednim estymatorem jądrowym lub estymatorem opartym na funkcjach sklejanych. Dla takich estymatorów nieznanej gęstości nie są potrzebne żadne założenia o postaci szacowanej funkcji i dlatego nazywa się je **estymatorami nieparametrycznymi**. Estymatorami tymi nie będziemy się jednak zajmować.

## 3.2. Estymacja punktowa

### 3.2.1. Estymatory największej wiarogodności

Wspomniane w p. 2.4.4 estymatory nieobciążone o minimalnej wariancji oraz estymatory dopuszczalne ze względu na błąd średniokwadratowy są oparte na dobrych podstawach metodologicznych, ale ich skonstruowanie może być praktycznie niemożliwe lub bardzo trudne. Warto zatem pokusić się o przedstawienie estymatora podobnie dobrze uzasadnionego, ale prostszego obliczeniowo.

Na rysunku 3.1 jest przedstawiona trzydziestoelementowa próba obserwacji ze standardowego rozkładu normalnego $N(0,1)$. Możemy powiedzieć, że mamy do czynienia z realizacją próby losowej z populacji $N(0,1)$. Wyobraźmy sobie jednak, że nie znamy wartości średniej rozkładu, $\mu = 0$, wiemy zaś tylko, iż populacja ma rozkład normalny o znanym, jednostkowym odchyleniu standardowym i wartości średniej równej jednej z zaledwie trzech możliwych wartości: $-1$, 0 lub 1. Wyobrażamy sobie zatem, że mamy do czynienia z bardzo prostym modelem statystycznym reprezentowanym przez rodzinę trzech rozkładów normalnych z $\sigma = 1$ i $\mu = -1$ lub $\mu = 0$ lub $\mu = 1$. Na rysunku 3.2 jest przedstawiona próba obserwacji z rys. 3.1 z naniesionymi trzema możliwymi gęstościami – wiemy, że obserwacje pochodzą z jednego z tych rozkładów.



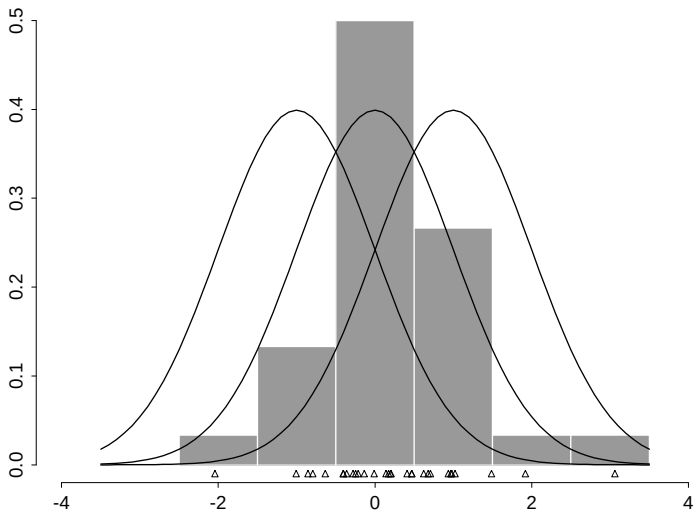

Rys. 3.1. Trzydziestoelementowa próba obserwacji ze standardowego rozkładu normalnego $N(0,1)$ i jej histogram

Zadanie polega na oszacowaniu prawdziwej wartości parametru $\mu$. Oczywiście, mając do wyboru na podstawie próby z rys. 3.1 wartości $-1$, 0 i 1, wybieramy oszacowanie $\hat{\mu} = 0$. Czynimy tak dlatego, że zdajemy sobie sprawę, iż realizacja próby powinna leżeć w tym przedziale, w którym masa prawdopodobieństwa danej gęstości jest „możliwie duża". Postulat ten daje intuicyjnie atrakcyjną podstawę do zaproponowania estymatora największej wiarogodności. By to uczynić, musimy jednak doprecyzować dwie kwestie.

Po pierwsze, skoro mamy do czynienia z próbą losową i jej realizacją, a więc zbiorem obserwacji, a nie z jedną zmienną losową i pojedynczą obserwacją, to powinniśmy poszukiwać „właściwej" gęstości całej próby, a nie „właściwej" gęstości pojedynczej zmiennej losowej. Sformułowany wcześniej postulat przyjmuje zatem taką postać: chcemy by łączna gęstość wybrana jako najlepiej odpowiadająca zaobserwowanej próbie przypisywała „możliwie dużą" masę prawdopodobieństwa obszarowi, w którym leżą obserwacje. Z kolei jednak pojęcie „możliwie dużej" masy prawdopodobieństwa nie jest precyzyjne i stąd rodzi się potrzeba drugiej modyfikacji naszego postulatu. Zamiast mówić o „możliwie dużej" masie, będziemy wymagać maksymalizacji pewnej wielkości.

Wróćmy do naszego przykładu, ale niech tym razem model statystyczny populacji, z której pochodzi próba losowa $X_1, X_2, \ldots, X_n$, będzie rodziną rozkładów normalnych o znanym odchyleniu standardowym równym 1 i nieznanej wartości średniej z przedziału nieskończonego $(-\infty, \infty)$, $\theta \in (-\infty, \infty)$.[1] Zadanie polega na estymacji parametru $\theta$.

[1]Zgodnie z naszą wcześniejszą konwencją estymowany parametr oznaczamy literą $\theta$. Tylko w niewielu przypadkach będziemy od tej konwencji odstępować.




Rys. 3.2. Trzy możliwe gęstości opisujące trzydziestoelementową próbę ze standardowego rozkładu normalnego $N(0,1)$

Zmienne losowe $X_1, X_2, \ldots, X_n$ są niezależne i dlatego zgodnie ze wzorem (2.33) ich łączna gęstość ma postać (opisując modele statystyczne zaznaczamy zwykle wyraźnie, że łączny rozkład próby – np. łączna gęstość – zależy od nieznanego parametru i piszemy $f(x_1, x_2, \ldots, x_n; \theta)$ zamiast $f(x_1, x_2, \ldots, x_n)$):

$$f(x_1, x_2, \ldots, x_n; \theta) = \frac{1}{(2\pi)^{n/2}} \exp(-\sum_{i=1}^{n}(x_i - \theta)^2/2). \tag{3.1}$$

Z chwilą zaobserwowania realizacji $x_1, x_2, \ldots, x_n$ próby losowej $X_1, X_2, \ldots$ $\ldots, X_n$ i wstawienia zaobserwowanych wartości do funkcji gęstości (3.1), funkcja ta przyjmuje konkretną wartość, zależną jednak od nieznanego parametru $\theta$. Przy zadanych wartościach $x_1, x_2, \ldots, x_n$, gęstość $f(x_1, x_2, \ldots$ $\ldots, x_n; \theta)$ możemy potraktować jako funkcję parametru $\theta$. Funkcję taką nazywamy **funkcją wiarogodności** i oznaczamy zwykle $L(\theta)$. W naszym przykładzie mamy zatem

$$L(\theta) = \frac{1}{(2\pi)^{n/2}} \exp(-\sum_{i=1}^{n}(x_i - \theta)^2/2). \tag{3.2}$$

Skoro przy tym zaobserwowaliśmy wartości $x_1, x_2, \ldots, x_n$, to prawdziwa wartość parametru $\theta$ musiała takiemu zdarzeniu „sprzyjać" i przeto – zgodnie z naszymi wcześniejszymi rozważaniami – wartość funkcji wiarogodności $L(\theta)$ powinna być „możliwie duża". Temu ostatniemu postulatowi nadajemy

taką ścisłą postać: **estymatorem największej wiarogodności** jest ta wartość parametru $\theta$, która (przy ustalonych, zaobserwowanych wartościach próby $x_1, x_2, \ldots, x_n$) maksymalizuje funkcję wiarogodności $L(\theta)$.

Ze względu na wygodę rachunkową maksymalizację funkcji wiarogodności zastępujemy zwykle maksymalizacją jej logarytmu naturalnego. Funkcja $L(\theta)$ jest oczywiście dodatnia, można ją więc zlogarytmować, przy czym logarytm jest funkcją rosnącą i przeto wartość $\hat{\theta}$ maksymalizująca funkcję $\ln L(\theta)$ zarazem maksymalizuje funkcję $L(\theta)$. Zlogarytmowanie funkcji wiarogodności jest zawsze wygodne, ponieważ rozkład łączny prowadzi do otrzymania dość złożonego iloczynu, logarytm iloczynu zaś daje łatwiejszą w obliczeniach sumę. Zwróćmy jeszcze uwagę, że interesuje nas zawsze maksimum globalne funkcji wiarogodności.

W naszym przykładzie maksymalizacja jest prosta. Wystarczy obliczyć pochodną funkcji $\ln L(\theta)$ i znaleźć wartość argumentu $\hat{\theta}$, dla której $\frac{d}{d\theta} \ln L(\theta) = 0$, by w ten sposób znaleźć punkt, w którym jest spełniony konieczny warunek maksimum funkcji. Następnie pozostaje już tylko sprawdzić, czy w punkcie tym jest spełniony także warunek wystarczający (ten drugi krok pozostawimy Czytelnikowi). Równania typu $\frac{d}{d\theta} \ln L(\theta) = 0$ nazywamy **równaniami wiarogodności** (w przypadku estymowania $k$ nieznanych parametrów mamy $k$ równań wiarogodności). Mamy

$$\frac{d}{d\theta} \ln L(\theta) = \frac{d}{d\theta} \left[ -\frac{n}{2} \ln(2\pi) - \sum_{i=1}^{n} (x_i - \theta)^2/2 \right] = 0,$$

czyli

$$\sum_{i=1}^{n} (x_i - \theta) = \sum_{i=1}^{n} x_i - n\theta = 0,$$

skąd wartość argumentu $\hat{\theta}$, maksymalizująca funkcję wiarogodności jest równa

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} x_i = \bar{x}. \tag{3.3}$$

(Jak już wspomnieliśmy, Czytelnikowi pozostawiamy sprawdzenie, czy funkcja $\ln L(\theta)$ rzeczywiście osiąga wartość największą w punkcie $\hat{\theta}$.) Ostatecznie zatem, estymatorem NW w problemie estymacji wartości średniej $\theta$ w rodzinie rozkładów normalnych o znanym odchyleniu standardowym jest estymator postaci (3.3). Jest to oczywiście dobrze nam znany i naturalny estymator wartości średniej rozkładu. Co więcej, jak wspomnieliśmy w p. 2.4.4, estymator (3.3) jest w naszym problemie zarazem estymatorem NMW. Fakty te potwierdzają racjonalność idei estymacji NW. Z potwierdzeniami tego typu spotkamy się jeszcze w tym punkcie parokrotnie.

Rozważona próba losowa pochodziła z rozkładu ciągłego. Idea estymacji NW nie zmienia się, gdy badana populacja jest opisywana rozkładem dyskretnym. Jedyna różnica polega na oczywistej konieczności zastąpienia łącznej gęstości próby losowej funkcją prawdopodobieństwa łącznego. Niech na przykład próba losowa pochodzi z rozkładu Poissona z nieznanym parametrem $\theta$, przy czym $\theta \in (0, \infty)$. Funkcja prawdopodobieństwa łącznego ma zatem postać (por. wzór (2.33))

$$f(x_1, x_2, \ldots, x_n; \theta) = \frac{e^{-n\theta}\theta^{(x_1+x_2+\cdots+x_n)}}{x_1!x_2!\cdots x_n!}. \tag{3.4}$$

Zaobserwowane wartości $x_i$, $i = 1, 2, \ldots, n$, pochodzą ze zbioru nieujemnych liczb całkowitych. Dla zadanych $x_i$, $i = 1, 2, \ldots, n$, wyrażenie (3.4) jest prawdopodobieństwem zdarzenia łącznego, polegającego na przyjęciu przez zmienną losową $X_1$ wartości $x_1$, przez zmienną losową $X_2$ wartości $x_2$ itd. Podane prawdopodobieństwo łączne, podobnie jak to uczyniliśmy w przypadku łącznej gęstości, możemy traktować jako funkcję parametru $\theta$, zwaną również funkcją wiarogodności:

$$L(\theta) = \frac{e^{-n\theta}\theta^{(x_1+x_2+\cdots+x_n)}}{x_1!x_2!\cdots x_n!}.$$

Analogicznie do przypadku ciągłego, zadanie znalezienia estymatora NW sprowadza się do znalezienia wartości argumentu $\theta$, maksymalizującej funkcję $L(\theta)$. Takie postępowanie ma następujące heurystyczne uzasadnienie: skoro zaszło zdarzenie $\{X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\}$, to prawdziwa wartość parametru $\theta$ musiała temu „sprzyjać", co w języku matematyki wyrażamy postulatem znalezienia estymatora $\hat{\theta}$, dla którego funkcja wiarogodności $L(\theta)$ osiąga wartość maksymalną.

Jak poprzednio, zadanie rozwiązujemy najpierw logarytmując funkcję wiarogodności i następnie przyrównując do zera pochodną funkcji $\ln L(\theta)$. Mamy

$$\ln L(\theta) = -n\theta + (\sum_{i=1}^{n} x_i) \ln \theta - \sum_{i=1}^{n} \ln(x_i!),$$

skąd otrzymujemy

$$\frac{d}{d\theta} L(\theta) = -n + \frac{\sum_{i=1}^{n} x_i}{\theta} = 0$$

i ostatecznie, po oznaczeniu rozwiązania powyższego równania symbolem $\hat{\theta}$,

$$\hat{\theta} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}.$$

A zatem (znowu pomijamy udowodnienie, że uzyskane rozwiązanie rzeczywiście maksymalizuje funkcję wiarogodności), estymatorem NW w problemie

estymacji parametru rozkładu Poissona jest średnia w próbie, $\bar{X}$. Ponieważ parametr rozkładu Poissona jest jego wartością średnią, otrzymany estymator NW jest estymatorem jak najbardziej naturalnym. Z bardziej złożonych rachunków wynika, że średnia w próbie jest także estymatorem NMW parametru rozkładu Poissona.

Omówione przykłady umożliwiają w sposób ogólny i lapidarny opisanie estymacji NW. Mając prostą próbę losową $X_1, X_2, \ldots, X_n$ z populacji o rozkładzie (danym gęstością lub funkcją prawdopodobieństwa) $f(x;\theta)$, gdzie $\theta$ jest nieznanym parametrem rozkładu, należącym do pewnego znanego zbioru $\Theta$, $\theta \in \Theta$, mamy zarazem rozkład łączny tej próby

$$f(x_1;\theta)f(x_2;\theta)\cdots f(x_n;\theta). \tag{3.5}$$

Przy ustalonych wartościach argumentów $x_1, x_2, \ldots, x_n$, funkcję tę traktujemy jako funkcję argumentu $\theta$ (i nazywamy funkcją wiarogodności),

$$L(\theta) = f(x_1;\theta)f(x_2;\theta)\cdots f(x_n;\theta).$$

Estymator NW jest statystyką odpowiadającą wartości argumentu $\theta$, dla której funkcja wiarogodności $L(\theta)$ osiąga wartość maksymalną (globalnie). Estymatory, w tym estymatory NW, oznaczać będziemy najczęściej symbolem $\hat{\theta}$.

Estymator NW jest rzeczywiście statystyką, jest bowiem funkcją próby, czyli funkcją zmiennych losowych $X_1, X_2, \ldots, X_n$, a zatem sam jest zmienną losową. Niekiedy piszemy zatem $\hat{\theta}(X_1, X_2, \ldots, X_n)$ zamiast krótko $\hat{\theta}$. Realizację estymatora NW oznaczamy odpowiednio symbolem $\hat{\theta}(x_1, x_2, \ldots, x_n)$. Zauważmy, że zastosowanie skrótowego symbolu $\hat{\theta}$ służy zarówno do oznaczenia estymatora jako zmiennej losowej, jak i jego realizacji, czyli zaobserwowanej wartości; nie prowadzi to jednak do nieporozumień, ponieważ z kontekstu zawsze wynika, czy chodzi nam o jedno czy drugie rozumienie estymatora.

W ogólnym przypadku estymator NW nie musi być nieobciążony. Prawdziwe jest jednak następujące stwierdzenie.

**Stwierdzenie 3.1.** *Wraz ze wzrostem liczności próby obciążenie estymatora NW dąży do zera.*

Mówimy, że estymatory NW są **asymptotycznie nieobciążone**. Stwierdzenie 3.1 obejmuje także te estymatory NW, których obciążenie jest równe zeru dla każdej liczności próby $n$ (obciążenie równe zeru dla każdego $n$ w oczywisty sposób dąży do zera, gdy $n$ rośnie). Prawdziwe jest ponadto stwierdzenie.

**Stwierdzenie 3.2.** *Jeżeli dla badanego modelu statystycznego istnieje estymator NMW, to przy spełnieniu pewnych ogólnych warunków dodatkowych estymator NW jest albo równy estymatorowi NMW, albo wraz ze wzrostem liczności próby jego wariancja coraz lepiej przybliża wariancję estymatora NMW.*

Wspomniane w stwierdzeniu warunki dodatkowe mają charakter techniczny i nie będą przez nas omawiane. Możemy w każdym razie powiedzieć, że w typowych problemach estymator NW jest dla dużych prób bliski estymatorowi NMW, o ile tylko ten ostatni istnieje.

Nieraz już wspominaliśmy, że zadanie estymacji może dotyczyć modelu statystycznego, w którym nie znamy więcej niż jednego parametru. Zadanie estymacji NW rozwiązuje się w takim przypadku w sposób całkowicie analogiczny do przypadku z jednym tylko nieznanym parametrem. By fakt ten wykazać, wystarczy odwołać się do przykładu.

Niech próba $X_1, X_2, \ldots, X_n$ pochodzi z populacji normalnej o nieznanej wartości średniej $\theta_1$ i nieznanej wariancji $\theta_2$. Gęstość każdej zmiennej losowej $X_i$ wyraża się przeto wzorem

$$f(x;\theta_1,\theta_2) = \frac{1}{(2\pi\theta_2)^{1/2}} \exp(-\frac{\sum_{i=1}^{n}(x_i-\theta_1)^2}{2\theta_2}).$$

Zakładamy, że $\theta_1 \in (-\infty,\infty)$ oraz $\theta_2 \in (0,\infty)$. Z pierwszego warunku wynika, że wartość średnia może być dowolną liczbą, z drugiego natomiast, że wariancja nie może z oczywistych względów być ujemna. Zadaniem jest znalezienie estymatorów NW parametrów $\theta_1$ i $\theta_2$. Bez trudu możemy skonstruować łączną gęstość próby losowej

$$f(x_1,x_2,\ldots,x_n;\theta_1,\theta_2) = \frac{1}{(2\pi\theta_2)^{n/2}} \exp(-\frac{\sum_{i=1}^{n}(x_i-\theta_1)^2}{2\theta_2}).$$

Dla ustalonych wartości $x_1, x_2, \ldots, x_n$ funkcję tę traktujemy jako funkcję nieznanych parametrów $\theta_1$ oraz $\theta_2$, czyli jako funkcję wiarogodności $L(\theta_1,\theta_2)$. Nasze zadanie polega na znalezieniu takich wartości $\hat{\theta}_1$ i $\hat{\theta}_2$, dla których funkcja $L(\theta_1,\theta_2)$ osiąga wartość maksymalną.

Zadanie to rozwiązujemy w sposób standardowy, czyli najpierw obliczamy logarytm naturalny funkcji wiarogodności:

$$\ln L(\theta_1,\theta_2) = -\frac{n}{2}\ln(2\pi\theta_2) - \frac{\sum_{i=1}^{n}(x_i-\theta_1)^2}{2\theta_2}.$$

Następnie tworzymy układ dwóch równań, których rozwiązanie spełnia warunki konieczne minimalizacji funkcji $\ln L(\theta_1, \theta_2)$, a zatem także funkcji $L(\theta_1, \theta_2)$:

$$\begin{aligned} \frac{\partial}{\partial \theta_1} \ln L(\theta_1, \theta_2) &= 0 \\ \frac{\partial}{\partial \theta_2} \ln L(\theta_1, \theta_2) &= 0, \end{aligned}$$

czyli w naszym problemie

$$\begin{aligned} -\frac{1}{\theta_2} \sum_{i=1}^{n} (x_i - \theta_1) &= 0 \\ -\frac{n}{2\theta_2} + \frac{\sum_{i=1}^{n} (x_i - \theta_1)^2}{2\theta_2^2} &= 0. \end{aligned}$$

Z pierwszego z równań wiarogodności otrzymujemy natychmiast

$$\hat{\theta}_1 = \bar{x}.$$

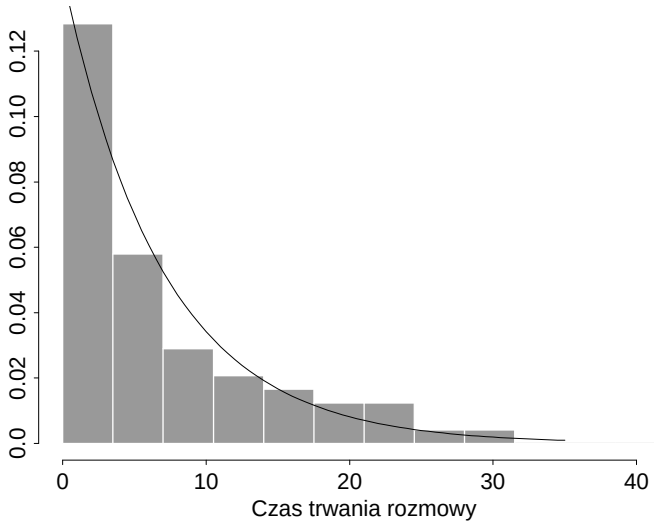
W rezultacie, po prostych przekształceniach z drugiego równania otrzymujemy

$$\hat{\theta}_2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

Otrzymane estymatory wartości średniej i wariancji są tyleż naturalne co ciekawe, zwłaszcza estymator NW wariancji rozkładu normalnego. Estymator $\hat{\theta}_1$ jest znowu średnią w próbie, czyli jest jednocześnie estymatorem NMW. Estymator $\hat{\theta}_2$ różni się od wariancji w próbie $S^2$ czynnikiem $n/(n-1)$ i jest przeto estymatorem obciążonym. Oczywiście, jest to estymator „bliski" wariancji w próbie $S^2$, tym bliższy im większa jest liczność próby $n$. Dla dużej liczności prób, obydwa estymatory są zgodnie ze stwierdzeniem 3.1 i 3.2 niemal identyczne.

Wszystkie rozważone dotąd przykłady estymatorów NW były estymatorami parametrów bezpośrednio występujących we wzorach na rozkłady populacji, z których pochodziły próby losowe. Tak oczywiście nie jest zawsze. W ostatnim przykładzie równie zasadnym jak pytanie o wariancję rozkładu jest pytanie o odchylenie standardowe, czyli pierwiastek wariancji. Prawdziwe jest jednak następujące stwierdzenie

**Stwierdzenie 3.3.** *Niech $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_m$ będą estymatorami NW parametrów $\theta_1, \theta_2, \ldots, \theta_m$ rozkładu populacji. Wówczas estymator NW dowolnej wzajemnie jednoznacznej funkcji $h(\theta_1, \theta_2, \ldots, \theta_m)$ tych parametrów ma postać $h(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_m)$.*




Rys. 3.3. Histogram czasu trwania rozmów telefonicznych (przykł. 3.1)

Często, mając estymatory NW $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_m$ parametrów $\theta_1, \theta_2, \ldots, \theta_m$ oraz *dowolną* funkcję $h(\theta_1, \theta_2, \ldots, \theta_m)$ tych parametrów, estymator NW tej funkcji definiuje się jako $h(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_m)$.

Na mocy stwierdzenia 3.3 oraz naszego ostatniego przykładu, estymatorem NW odchylenia standardowego w modelu normalnym z nieznaną wartością średnią oraz nieznaną wariancją jest statystyka

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

**Przykład 3.1.** W przykładzie 2.22 przyjęliśmy, że czas trwania rozmowy telefonicznej jest zmienną losową o rozkładzie wykładniczym z parametrem 1/7. W rzeczywistości założenie to było wynikiem eksperymentu, w którym pan Jan Ek zaobserwował następujące czasy trwania rozmów (w minutach):

0,48 0,50 0,53 0,57 0,58 0,63 0,67 0,68 0,72 0,78 0,82 0,87 0,92
1,00 1,10 1,22 1,28 1,35 1,42 1,53 1,65 1,83 1,93 2,02 2,18 2,40
2,60 2,75 2,93 3,03 3,32 3,53 3,75 3,95 4,25 4,33 4,45 4,73 5,00
5,33 5,55 5,75 6,03 6,60 6,87 7,37 7,77 8,12 8,72 9,26 9,88 10,50
11,25 11,82 12,80 13,38 13,95 14,38 15,60 16,08 16,72 17,90 19,08
19,83 21,13 22,33 23,35 27,20 28,63

Na podstawie histogramu uznaliśmy, że zaobserwowane wartości stanowią realizację próby losowej z rozkładu wykładniczego (o nieznanym parametrze). Nietrudno wykazać, że estymatorem NW parametru rozkładu wykładniczego jest statystyka (por. zad. 3.1)

$$\frac{1}{\bar{x}}.$$

(W naszym przypadku $\bar{x} \approx 7$ i dlatego za wartość parametru $\lambda$ w przykł. 2.22 przyjęliśmy 1/7.) Ze stwierdzenia 3.3 wynika jednocześnie, że estymatorem NW średniego czasu trwania rozmowy jest odwrotność estymatora NW parametru rozkładu, czyli 7 minut.

**Przykład 3.2.** Celem ankiety bywa uzyskanie informacji, której indagowani udzielają niechętnie i można podejrzewać, że uzyskane odpowiedzi nie zawsze są zgodne z prawdą. Przykładem jest pytanie: czy osoba ankietowana będzie głosować w najbliższych wyborach prezydenckich na kandydata powszechnie uważanego za osobę skrajnie kontrowersyjną. Innych przykładów dostarczają pytania dotyczące akceptacji lub nie postawy moralnie wątpliwej. W takich przypadkach dokonuje się tzw. randomizacji odpowiedzi, której celem jest ukrycie przed ankietującym prawdziwych poglądów ankietowanego.

Przyjmijmy, że interesuje nas częstość $p$, z jaką populacja dorosłych Polaków będzie głosować na szczególnie kontrowersyjnego kandydata o nazwisku Z. Ankietujący sporządza np. 100 kart, z których połowa zawiera pytanie *czy będziesz w najbliższych wyborach głosować na Z.* Druga połowa kart zawiera pytanie *czy ostatnia cyfra numeru twojego telefonu jest mniejsza od 5.* Milcząco tu zakładamy, że wszyscy ankietowani mają telefon. Zakładamy też, że częstość odpowiedzi *tak* na drugie pytanie wynosi 1/2, mamy bowiem do czynienia z rozkładem jednostajnym na przestrzeni zdarzeń elementarnych o 10 elementach (czyli 10 różnych cyfrach). Ankietowany otrzymuje talię 100 kart, tasuje je, wybiera jedną tak, aby nie widział jej ankietujący, i odpowiada. Ponieważ ankietujący nie zna pytania, na które słyszy odpowiedź, uznajemy, że ankietowany udziela odpowiedzi prawdziwej.

Ankiecie zostanie poddana próba losowo wybranych $n$ dorosłych Polaków, z których każdy udzieli odpowiedzi *tak* lub *nie*. Niech $X_i$, $i = 1, 2, \ldots, n$, oznacza zmienną losową przyjmującą wartość 1, gdy $i$-ta ankietowana osoba odpowiada *tak*, oraz wartość 0 w przypadku przeciwnym. Mamy zatem próbę losową $X_1, X_2, \ldots, X_n$ z rozkładu Bernoulliego o nieznanym prawdopodobieństwie sukcesu $\theta$ (sukcesem jest udzielenie odpowiedzi *tak*). Łatwo zauważyć, że po przeprowadzeniu ankiety, czyli dla ustalonych wartości

$x_1, x_2, \ldots, x_n$, otrzymujemy funkcję wiarogodności

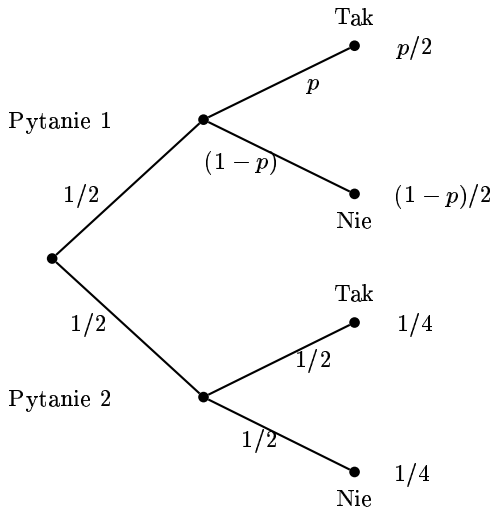
$$L(\theta) = \theta^{(x_1+x_2+\cdots+x_n)}(1-\theta)^{n-(x_1+x_2+\cdots+x_n)}.$$

Nietrudno też wykazać, że estymatorem NW parametru $\theta$ jest średnia w próbie $\bar{X}$ (por. zad. 3.2). Nas jednak interesuje częstość $p$, związana zależnością funkcyjną z $\theta$. Na rysunku 3.4 pokazano, że

$$\theta = \frac{p}{2} + \frac{1}{4},$$

ze stwierdzenia 3.3 otrzymujemy więc natychmiast, że estymatorem NW częstości $p$ jest estymator

$$\hat{p} = 2\hat{\theta} - 1/2.$$




Rys. 3.4. Prawdopodobieństwa odpowiedzi na pytania ankiety (przykł. 3.2)

Funkcja wiarogodności nie zawsze jest dostatecznie gładka by możliwe było analityczne znalezienie punktu ją minimalizującego. Z taką niegładką funkcją wiarogodności mamy do czynienia w przykł. 2.28.

**Przykład 2.19 cd.** Nieznanymi parametrami rozkładu są w tym przykładzie liczby $a$ i $b$. Stwierdzenie 3.3 umożliwia obliczenie estymatora NW rozstępu $b-a$ jako różnicy estymatorów NW parametru $a$ i parametru $b$.

Zajmijmy się zatem skonstruowaniem estymatorów NW parametrów $a$ i $b$. Funkcja gęstości populacji ma postać

$$f(x;a,b) = \begin{cases} \frac{1}{b-a}, & \text{gdy } a \le x \le b \\ 0, & \text{w przypadku przeciwnym,} \end{cases}$$

zaś łączna gęstość próby

$$f(x_1, x_2, \ldots, x_n; a, b) = \\ = \begin{cases} \frac{1}{(b-a)^n}, & \text{gdy } a \le x_1 \le b,\ a \le x_2 \le b, \ldots,\ a \le x_n \le b \\ 0, & \text{w przypadku przeciwnym.} \end{cases}$$

Funkcja wiarogodności $L(a,b)$ jest funkcją dwóch zmiennych, $a$ i $b$. Jeżeli

$$a \le \min\{x_1, x_2, \ldots, x_n\} \quad \text{oraz} \quad b \ge \max\{x_1, x_2, \ldots, x_n\},$$

to funkcja $L(a,b)$ jest dodatnia i równa $1/(b-a)^n$. Jeśli natomiast

$$a > \min\{x_1, x_2, \ldots, x_n\} \quad \text{lub} \quad b < \max\{x_1, x_2, \ldots, x_n\},$$

to funkcja $L(a,b)$ staje się równa 0. Przy tym na zbiorze, na którym funkcja wiarogodności jest dodatnia, jej wartości są tym większe im bliższe są sobie argumenty $a$ i $b$. Biorąc pod uwagę te wszystkie własności funkcji $L(a,b)$ otrzymujemy, że jej maksimum jest osiągane dla $b = \max\{x_1, x_2, \ldots, x_n\}$ oraz $a = \min\{x_1, x_2, \ldots, x_n\}$. W ten sposób otrzymaliśmy estymatory NW parametrów $a$ oraz $b$ i w konsekwencji wcześniej już podany estymator NW rozstępu, $R$.

Nawet jeżeli funkcja wiarogodności jest gładka, uzyskanie estymatorów NW może nastręczać poważne trudności np. z racji wielomodalności tej funkcji oraz niemożliwości uzyskania jawnych, analitycznych rozwiązań równań wiarogodności. W rezultacie, zwłaszcza gdy niefortunnie zostały wybrane warunki początkowe rozwiązania iteracyjnego, algorytm maksymalizacji może „ugrzęznąć" w maksimum lokalnym. Dotyczy to zwłaszcza modeli statystycznych z wieloma parametrami. Pomocne w wyborze „dobrych" warunków początkowych dla algorytmu iteracyjnej maksymalizacji funkcji wiarogodności może być obliczenie najpierw estymatorów opartych na metodzie momentów, którym jest poświęcony następny punkt.

Wspomnieć jeszcze wypada, choć problemem tym nie będziemy się zajmować, o istnieniu tzw. „trudnych" problemów, gdy równania wiarogodności

mają niestabilne rozwiązania (o niestabilności mówimy np. wtedy, gdy losowe usunięcie z próby niewielkiej części jej elementów może prowadzić do znacznych zmian otrzymywanych wartości estymatorów). Z sytuacją taką można się spotkać, gdy parametr kształtu rozkładu gamma (por. zad. 2.20) jest bliski jedności. Jest to typowy trudny problem, ponieważ w zależności od tego, czy parametr kształtu jest mniejszy, równy czy większy od 1, funkcje gęstości istotnie się jakościowo różnią.

### 3.2.2. Estymatory oparte na metodzie momentów

Estymatory te traktujemy jako estymatory pomocnicze, tak jak to zasygnalizowaliśmy w poprzednim punkcie. Same charakteryzują się nierzadko stosunkowo dużą wariancją i generalnie nie są polecane, zwłaszcza, gdy opierają się na momentach wyższych rzędów. Niemniej jednak są też znane przypadki, gdy podanie estymatora innego niż estymator oparty na metodzie momentów jest praktycznie niemożliwe. Posłużenie się wówczas takim estymatorem jest uzasadnione.

Rozważmy próbę losową $X_1, X_2, \ldots, X_n$ z rozkładu normalnego $N(\theta_1, \theta_2)$ o nieznanej wartości średniej $\theta_1$, przy czym $\theta_1 \in (-\infty, \infty)$, oraz nieznanym odchyleniu standardowym $\theta_2$, $\theta_2 \in (0, \infty)$. Zadanie polega na oszacowaniu na podstawie danych z próby, dwóch nieznanych parametrów populacji. Estymacja oparta na metodzie momentów (estymacja MM) polega na wykorzystaniu dwóch faktów: po pierwsze znamy naturalne estymatory momentów rozkładu prawdopodobieństwa i po drugie momenty te są pewnymi funkcjami nieznanych parametrów analizowanego modelu.

Dla wszystkich $p$, $p = 1, 2, \ldots$, dla których istnieją momenty rzędu $p$ zmiennych losowych $X_1, X_2, \ldots, X_n$, naturalnym estymatorem momentu rzędu $p$ jest moment próbkowy

$$\frac{1}{n}\sum_{i=1}^{n} X_i^p.$$

Oczywiście dla $p = 1$ mamy do czynienia ze zwykłą średnią w próbie $\bar{X}$. Z kolei, w przypadku populacji normalnej z wartością średnią $\theta_1$ i odchyleniem standardowym $\theta_2$, moment pierwszego rzędu każdej zmiennej losowej $X_i$ z tego rozkładu jest po prostu równy parametrowi $\theta_1$

$$EX_i = \theta_1, \tag{3.6}$$

natomiast moment drugiego rzędu jest następującą funkcją dwóch nieznanych parametrów (por. warunek (2.8))

$$EX_i^2 = \theta_1^2 + \theta_2^2. \tag{3.7}$$

Estymacja MM polega na zastąpieniu lewych stron równań (3.6) i (3.7) odpowiednio pierwszym i drugim momentem próbkowym oraz zastąpieniu nieznanych parametrów po prawej stronie tych równań ich estymatorami $\hat{\theta}_1$ i $\hat{\theta}_2$:

$$\bar{X} = \hat{\theta}_1$$

oraz

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 = \hat{\theta}_1^2 + \hat{\theta}_2^2.$$

Po rozwiązaniu otrzymanego układu równań z dwiema niewiadomymi ($\hat{\theta}_1$ i $\hat{\theta}_2$) otrzymujemy jawne wzory na szukane estymatory MM. W naszym przypadku pierwsze równanie daje natychmiast estymator MM parametru $\theta_1$,

$$\hat{\theta}_1 = \bar{X}$$

i w konsekwencji z drugiego równania otrzymujemy

$$\hat{\theta}_2 = \sqrt{(1/n)\sum_{i=1}^{n} X_i^2 - (\bar{X})^2} = \sqrt{(1/n)\sum_{i=1}^{n}(X_i - \bar{X})^2}\ .$$

(Czytelnikowi pozostawiamy sprawdzenie, czy ostatnia równość jest prawdziwa). Zatem, w przypadku estymacji MM, równania opisujące momenty populacji jako funkcje nieznanych parametrów zastępujemy ich próbkowymi odpowiednikami – zamiast symboli momentów wpisujemy momenty w próbie, zamiast zaś nieznanych parametrów wpisujemy estymatory, których szukamy. Postępowanie takie jest istotą metody MM niezależnie od tego o estymację jakich parametrów i jakiego rozkładu chodzi.

Omówiony przykład umożliwia zatem ogólne opisanie sposobu konstrukcji estymatorów MM. Niech będzie dana próba losowa $X_1, X_2, \ldots, X_n$ z populacji o rozkładzie (danym gęstością lub funkcją prawdopodobieństwa) $f(x; \theta_1, \theta_2, \ldots, \theta_k)$, gdzie wektor nieznanych parametrów $\theta_1, \theta_2, \ldots, \theta_k$ należy do pewnego znanego $k$-wymiarowego zbioru $\Theta$. Załóżmy, że zmienne losowe mają skończone momenty aż do $k$-tego rzędu. Załóżmy też, że potrafimy wyrazić te momenty jako funkcje $h_p(\theta_1, \theta_2, \ldots, \theta_k)$ nieznanych parametrów populacji

$$EX_i^p = h_p(\theta_1, \theta_2, \ldots, \theta_k), \tag{3.8}$$

dla $p = 1, 2, \ldots, k$. Estymatory MM nieznanych parametrów $\theta_1, \theta_2, \ldots, \theta_k$, które będziemy oznaczać symbolami $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k$, otrzymujemy, rozwiązując układ $k$ równań

$$\frac{1}{n}\sum_{i=1}^{n} X_i^p = h_p(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k), \tag{3.9}$$

$p = 1, 2, \ldots, k$. Zakładamy tu milcząco, że powyższy układ równań ma jednoznaczne rozwiązanie. Pod koniec tego punktu wspomnimy o możliwości uogólnienia zadania estymacji MM, umożliwiającego rezygnację z wymagania istnienia rozwiązania równań (3.9).

**Przykład 3.3.** Niech $X_1, X_2, \ldots, X_n$ będzie próbą losową, opisującą czasy pierwszych przeglądów gwarancyjnych samochodów Opel w pewnym warsztacie. Zakładamy, że zmienne losowe mają rozkład wykładniczy o nieznanym parametrze $\lambda$. Naszym zadaniem jest skonstruowanie estymatora MM tego parametru. Ponieważ wiemy, że wartość średnia rozkładu wykładniczego jest równa $1/\lambda$, otrzymujemy natychmiast, iż estymatorem MM parametru $\lambda$ jest wielkość $1/\bar{X}$. Jest to oczywiście dobrze nam znany estymator, będący zarazem estymatorem NW.

Niekiedy w metodzie MM opieramy się na centralnych momentach (począwszy od momentu rzędu drugiego), a nie na momentach zwykłych. Z układu równań (3.9) nie zmienione pozostaje pierwsze równanie (tzn. równanie dla $p = 1$), natomiast następne równania przyjmują postać

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^p = \tilde{h}_p(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k), \tag{3.10}$$

$p = 2, \ldots, k$, gdzie funkcje z prawych stron tych równań pochodzą z przedstawienia momentów centralnych badanej populacji jako funkcji jej nieznanych parametrów:

$$E(X_i - \mu_{X_i})^p = \tilde{h}_p(\theta_1, \theta_2, \ldots, \theta_k),$$

$p = 2, \ldots, k$. Poza tym metoda MM postaje taka sama (niekiedy, dla $p = 2$ dokonuje się kolejnej zmiany, estymator $(1/n)\sum_{i=1}^{n}(X_i - \bar{X})^2$ zastępując wariancją próbkową $S^2$).

**Przykład 2.19 cd.** Zauważmy, że drugi z estymatorów podanych w tym przykładzie można traktować jako estymator MM.

**Przykład 3.4.** W badaniach biomedycznych często przyjmuje się, że czas przeżycia jednostki doświadczalnej (np. po zarażeniu) jest zmienną losową o rozkładzie gamma danym gęstością (por. zad. 2.19)

$$f(x; \theta_1, \theta_2) = \begin{cases} \frac{1}{\theta_2^{\theta_1}\Gamma(\theta_1)} x^{\theta_1 - 1} e^{-x/\theta_2}, & \text{gdy } x > 0 \\ 0, & \text{gdy } x \leq 0, \end{cases} \tag{3.11}$$

gdzie $\Gamma(\cdot)$ oznacza funkcję gamma (wartość średnia zmiennej o tym rozkładzie wynosi $\theta_1\theta_2$, natomiast wariancja $\theta_1\theta_2^2$). Przyjmuje się przy tym, że parametr kształtu $\theta_1$ jest większy od 1. Jeżeli parametr ten może być stosunkowo bliski jedności, warto rozważyć posłużenie się estymatorami MM po to, by uzyskać warunki początkowe do obliczenia ostatecznie estymatorów NW (por. komentarz na końcu poprzedniego punktu). Mając próbę losową czasów reakcji $X_1, X_2, \ldots, X_n$ u $n$ osobników, możemy obliczyć średnią i wariancję próbkową. Dalej, jeśli wiemy, że

$$EX_i = \theta_1\theta_2$$

oraz

$$\sigma^2_{X_i} = \theta_1\theta_2^2,$$

to otrzymamy układ dwóch równań

$$\begin{aligned} \bar{X} &= \hat{\theta}_1\hat{\theta}_2 \\ S^2 &= \hat{\theta}_1\hat{\theta}_2^2. \end{aligned}$$

Stąd już łatwo otrzymujemy estymatory MM nieznanych parametrów gęstości (3.10):

$$\hat{\theta}_1 = \frac{\bar{X}^2}{S^2} \quad \text{oraz} \quad \hat{\theta}_2 = \frac{S^2}{\bar{X}}.$$

Dotąd rozważaliśmy w tym punkcie jedynie parametry w sensie „wąskim", czyli występujące we wzorze na rozkład populacji, z której pochodzi próba losowa. Nic jednak nie stoi na przeszkodzie zastosowaniu metody momentów do otrzymania dowolnych parametrów rozkładu, jeśli tylko umiemy zapisać dla nich równania typu (3.8). Zauważmy, że np. średnia oraz wariancja w próbie mogą zatem być traktowane jako estymatory MM w przypadku dowolnego rozkładu populacji – jeśli interesującym nas parametrem $\theta$ jest wartość średnia, mamy po prostu

$$EX_i = \theta \quad \text{i stąd} \quad \hat{\theta} = \bar{X}$$

(analogicznie postępujemy, definiując estymator MM wariancji dowolnego rozkładu).

Równania (3.8) możemy zapisać w postaci równoważnej

$$E[X_i^p - h_p(\theta_1, \theta_2, \ldots, \theta_k)] = 0,$$

$p = 1, 2, \ldots, k$. Równaniom tym możemy też nadać postać ogólniejszą

$$E[g_p(\theta_1, \theta_2, \ldots, \theta_k, X_i)] = 0, \tag{3.12}$$

$p = 1, 2, \ldots, k$, gdzie $g_p$ jest pewną funkcją nieznanych parametrów $\theta_1, \theta_2, \ldots$ $\ldots, \theta_k$ oraz zmiennej losowej $X_i$. Oczywiście, jeżeli

$$g_p(\theta_1, \theta_2, \ldots, \theta_k, X_i) = X_i^p - h_p(\theta_1, \theta_2, \ldots, \theta_k),$$

$p = 1, 2, \ldots, k$, to równania (3.12) są równoważne równaniom (3.8). Okazuje się jednak, choć takimi przypadkami nie będziemy się zajmować, że niekiedy jest uzasadnione inne określenie funkcji $g_p$. Odpowiednikami równań (3.9) są w rozważanym ujęciu równania (czynnik $1/n$ przed sumą jest z oczywistych względów zbędny)

$$\sum_{i=1}^{n} g_p(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k, X_i) = 0, \tag{3.13}$$

$p = 1, 2, \ldots, k$. Jak poprzednio, mówimy, że pierwiastki $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k$ układu równań (3.13) są estymatorami MM nieznanych parametrów $\theta_1, \theta_2, \ldots, \theta_k$.

Wspomnieliśmy już, że układ równań (3.9) może nie mieć rozwiązania. To samo dotyczy rzecz jasna równań (3.13). W takiej sytuacji poszukuje się estymatorów $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k$, które w pewnym ustalonym sensie matematycznym czynią wyrażenie

$$\sum_{i=1}^{n} g_p(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k, X_i)$$

możliwie bliskim zeru. Tak otrzymane estymatory nazywa się estymatorami opartymi na **uogólnionej metodzie momentów** i krótko nazywa estymatorami UMM.

Wspomnieliśmy już także, że w niektórych sytuacjach zaproponowanie estymatora innego niż estymator MM lub UMM jest trudne lub praktycznie niewykonalne. Tak jest np. w przypadku estymacji parametrów tzw. uogólnionego ujemnego rozkładu dwumianowego (por. zad. 3.4). Tak jest też np. w zadaniach regresji (por. rozdz. 4), jeśli zrezygnować z założenia normalności, oraz w zadaniach analizy szeregów czasowych.

### 3.2.3. M-estymatory

Estymatory tego typu umożliwiają uwzględnienie w ich konstrukcji problemu odporności na obserwacje odstające. Ich omówienie ograniczymy do przypadku estymacji parametru (inaczej wskaźnika) położenia $\theta$ rozkładu populacji, z której pochodzi prosta próba losowa $X_1, X_2, \ldots, X_n$. W zadaniach takich wygodnie jest niekiedy rozkład populacji zapisywać jako funkcję argumentu $x - \theta$, a nie funkcję argumentu $x$, czyli jako funkcję $f(x - \theta)$. Można powiedzieć, że różnice $X_i - \theta$ opisują losowe „odchyłki" oryginalnej

próby $X_1, X_2, \ldots, X_n$ od wielkości $\theta$ i same charakteryzują się zerową wartością parametru położenia. W przypadku ciągłego rozkładu symetrycznego względem $\theta$, takiego jak rozkład normalny, we wzorze na gęstość zmiennej losowej $X_i$ zawsze występuje wielkość $x_i - \theta$, przy czym $\theta$ jest naturalnym parametrem położenia tej gęstości (w tym przypadku zarówno wartością średnią, jak i medianą populacji).

Jak pamiętamy, idea estymacji NW sprowadza się do znalezienia takiej wartości $\hat{\theta}$, dla której logarytm funkcji wiarogodności $\sum \ln f(x_i - \theta)$, osiąga maksimum, co możemy zapisać równoważnie jako znalezienie wartości $\hat{\theta}$, minimalizującej funkcję

$$\sum_{i=1}^{n} \rho(x_i - \theta), \tag{3.14}$$

gdzie $\rho(x_i - \theta) = -\ln f(x_i - \theta)$.[2] Jeżeli funkcja $\rho(\cdot)$ jest różniczkowalna, zadanie zamieniamy na rozwiązanie (względem $\theta$) równania

$$\sum_{i=1}^{n} \psi(x_i - \theta) = 0, \tag{3.15}$$

gdzie $\psi(x_i - \theta) = \frac{d}{d\theta}\rho(x_i - \theta)$.

Idea M-estymatorów jest uogólnieniem powyższego postępowania, uzyskiwanym przez użycie funkcji $\rho(\cdot)$ różnej od funkcji $-\ln f(\cdot)$.[3] Estymacja metodą NW ma dobre podstawy metodologiczne, ale na konstrukcję interesującej funkcji $\rho(\cdot)$ można też spojrzeć zupełnie inaczej i w ten sposób uzasadnić wprowadzenie jej innych postaci. Mianowicie, jeżeli $\theta$ jest parametrem położenia, to można życzyć sobie, aby suma stosownie określonych odległości zmiennych $X_i$, $i = 1, 2, \ldots, n$, od wartości estymatora parametru $\theta$ była jak najmniejsza. Można zatem zaproponować, aby funkcja $\rho(\cdot)$ miała postać

$$\rho(x_i - \theta) = (x_i - \theta)^2 \quad \text{lub} \quad \rho(x_i - \theta) = |x_i - \theta|, \tag{3.16}$$

czyli by minimalizować (ze względu na $\theta$) funkcję

$$\sum_{i=1}^{n} (x_i - \theta)^2 \quad \text{lub} \quad \sum_{i=1}^{n} |x_i - \theta|. \tag{3.17}$$

Łatwo zauważyć, że w pierwszym przypadku równanie (3.15) przyjmuje postać

[2] Zadanie maksymalizacji funkcji $\sum \ln f(x_i - \theta)$ wygodnie jest zastąpić równoważnym mu zadaniem minimalizacji funkcji $\sum[-\ln f(x_i - \theta)]$ ze względu na dalsze zastosowania funkcji (3.14).

[3] Stąd nazwa *M-estymatory*; mianowicie, litera $M$ to pierwsza litera anglojęzycznej nazwy estymatorów NW, *maximum likelihood estimators*.

$$\sum_{i=1}^{n}(x_i - \theta) = 0,$$

którego rozwiązaniem jest po prostu średnia w próbie, $\hat{\theta} = \bar{x}$.

Estymatory powstałe przez minimalizację sumy kwadratów losowych odchyłek od wartości szukanego parametru nazywamy estymatorami opartymi na **metodzie najmniejszych kwadratów**. W bardziej złożonej sytuacji do metody tej wrócimy w następnym rozdziale. Czytelnikowi pozostawiamy zauważenie, że estymatory oparte na metodzie najmniejszych kwadratów są szczególnym przypadkiem estymatorów UMM.

Funkcja argumentu $\theta$ postaci $|x - \theta|$ jest różniczkowalna, gdy $x \neq \theta$. Jej pochodna jest równa $-1$, gdy $\theta < x$ oraz równa 1, gdy $\theta > x$. Ponieważ w punkcie nieciągłości tej funkcji (czyli dla $x = \theta$) jej wartość wynosi 0, możemy umówić się, że w tym punkcie funkcja $\psi(x - \theta)$, występująca w równaniu (3.15), jest także równa zeru oraz jest równa $-1$ lub 1 w zależności od tego czy $\theta < x$, czy $\theta > x$. Stąd, lewa strona równania (3.15) z $\psi(x_i - \theta) = |x_i - \theta|$ jest równa 0, gdy liczba wartości $x_i$ mniejszych od $\theta$ jest równa liczbie wartości $x_i$ większych od $\theta$. Ostatecznie zatem, M-estymatorem parametru położenia, gdy funkcja $\rho(\cdot)$ mierzy **bezwzględną wartość** odchyłek obserwacji $x_i$, jest mediana w próbie.

Jak pamiętamy z p. 2.1.1, medianę w próbie można uznać za estymator odporny na obserwacje odstające. Innymi estymatorami tego typu, zaproponowanymi tamże, jest średnia ucinana oraz średnia winsorowska. Nietrudno zauważyć, że jeśli rozwiążemy równanie (3.15) z funkcją

$$\psi(y) = \begin{cases} y, & \text{gdy } |y| < c \\ 0, & \text{w przypadku przeciwnym,} \end{cases}$$

gdzie $c$ jest dowolną ustaloną stałą dodatnią, to otrzymamy M-estymator, będący pewnym odpowiednikiem średniej ucinanej. Taki M-estymator niekiedy nazywa się estymatorem **metrycznie ucinanym**. Z kolei, jeśli rozwiążemy równanie (3.15) z funkcją

$$\psi(y) = \begin{cases} -c, & \text{gdy } x < -c \\ y, & \text{gdy } |y| < c \\ c, & \text{gdy } x > c, \end{cases}$$

gdzie $c$ jest dowolną ustaloną stałą dodatnią, to otrzymamy M-estymator, będący odpowiednikiem średniej winsorowskiej. Taki M-estymator niekiedy nazywa się estymatorem **metrycznie winsorowskim**. Wreszcie M-estymatorem, który jest kompromisem między średnią w próbie a medianą w próbie, jest estymator otrzymany przez minimalizację (względem argumentu $\theta$) wyrażenia (3.14) z

$$\rho(y) = \begin{cases} \frac{1}{2}y^2, & \text{gdy } |y| \leq c \\ c|y| - \frac{1}{2}c^2, & \text{w przypadku przeciwnym,} \end{cases}$$

gdzie $c$ jest dowolną ustaloną stałą dodatnią (ten estymator wygodniej jest zdefiniować za pomocą funkcji $\rho$). Podana funkcja jest określona tak, aby była ciągła i miała ciągłą pochodną. Gdy stała $c$ rośnie, otrzymany M-estymator coraz bardziej przypomina średnią w próbie. Gdy $c$ maleje, estymator staje się coraz bliższy medianie w próbie. Estymator ten nazywa się **estymatorem Hubera** od nazwiska statystyka, który wprowadził M-estymatory. Wszystkie trzy ostatnie M-estymatory (oraz mediana w próbie) mają prostą interpretację i każdy z nich może być uznany za estymator w jemu właściwy sposób odporny na obserwacje odstające.

## 3.3. Estymacja przedziałowa

Zadaniem estymacji przedziałowej jest skonstruowanie na podstawie próby losowej przedziału, o którym można z dużą dozą przekonania powiedzieć, iż zawiera prawdziwą wartość szacowanego parametru. Konstrukcja przedziału jest oczywiście równoznaczna z podaniem jego dwóch końców. Zatem, jeżeli próba losowa nie została jeszcze zaobserwowana, jest to przedział o losowych końcach, będących funkcjami tej próby. Jak z tego wynika, estymator przedziałowy jest wyznaczony przez dwie zmienne losowe, w przeciwieństwie do estymatora punktowego, który jest pojedynczą zmienną losową. Z kolei, zaobserwowana wartość estymatora przedziałowego, powstała na podstawie zaobserwowanej realizacji próby losowej, jest wyznaczona przez dwie liczby (dwa końce przedziału), a nie przez tylko jedną liczbę, jak w przypadku zaobserwowanej wartości estymatora punktowego.

Można powiedzieć, że idea estymacji przedziałowej wiąże w jednopunktową estymację nieznanego parametru ze znajomością rozproszenia tego estymatora punktowego. Jak zobaczymy w najprostszym przypadku estymacji wartości średniej rozkładu normalnego, mając estymator punktowy i jego rozproszenie można określić położenie środka estymatora przedziałowego oraz taką szerokość tego ostatniego estymatora, by z zadaną dozą przekonania móc orzec, iż utworzony na podstawie zaobserwowanej próby losowej przedział zawiera prawdziwą wartość parametru.

Ścisłe znaczenie sformułowania „zadana doza przekonania", które w statystyce zastępuje się pojęciem „zadanego poziomu ufności", zostanie wyjaśnione w dalszym ciągu tego podrozdziału. Otrzymane na podstawie zaobserwowanej próby wartości estymatorów przedziałowych będziemy nazywać **przedziałami ufności**.

### 3.3.1. Przedziały ufności dla wartości średniej rozkładu normalnego

**Przypadek rozkładu normalnego o znanym odchyleniu standardowym**

Rozważmy próbę losową $X_1, X_2, \ldots, X_n$ z rozkładu normalnego $N(\mu, \sigma)$ ze znanym odchyleniem standardowym $\sigma$. Zadanie polega na wyznaczeniu przedziału ufności dla nieznanej wartości średniej $\mu$. Wiadomo, że średnia w próbie

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

ma rozkład normalny $N(\mu, \sigma/\sqrt{n})$. Stąd, zmienna losowa

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \tag{3.18}$$

ma standardowy rozkład normalny $N(0,1)$. Z łatwością możemy wyznaczyć przedział, do którego wartości zmiennej losowej $Z$ należą z prawdopodobieństwem $1-\alpha$, gdzie $\alpha$ jest zadaną liczbą z przedziału $(0,1)$. Mianowicie, mamy na przykład

$$P(z_{\alpha/2} \leq Z \leq z_{1-\alpha/2}) = 1 - \alpha, \tag{3.19}$$

gdzie $z_{\alpha/2}$ jest kwantylem rzędu $\alpha/2$ i $z_{1-\alpha/2}$ jest kwantylem rzędu $1-\alpha/2$ standardowego rozkładu normalnego

$$P(Z \leq z_{\alpha/2}) = \frac{\alpha}{2} \quad \text{oraz} \quad P(Z \leq z_{1-\alpha/2}) = 1 - \frac{\alpha}{2}$$

(por. rys. 3.5). Ze względu na symetrię gęstości standardowego rozkładu normalnego mamy przy tym

$$z_{\alpha/2} = -z_{1-\alpha/2},$$

w konsekwencji czego równanie (3.19) przyjmuje postać

$$P(-z_{1-\alpha/2} \leq Z \leq z_{1-\alpha/2}) = 1 - \alpha. \tag{3.20}$$

Po podstawieniu prawej strony równości definicyjnej (3.18) zamiast $Z$ i dokonaniu prostych przekształceń otrzymujemy

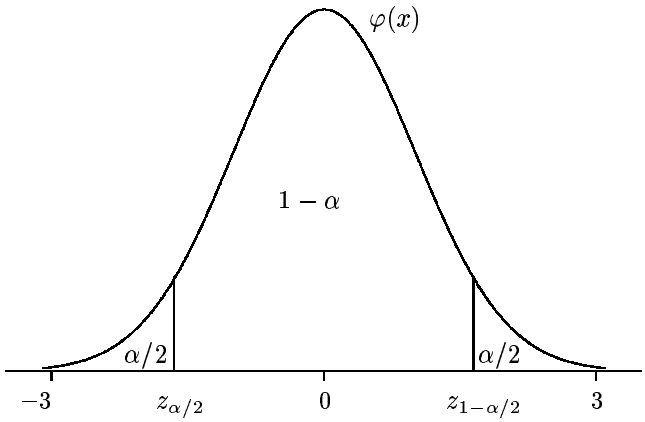
$$P(-z_{1-\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{1-\alpha/2}) =$$

$$= P(\bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}) = 1 - \alpha. \tag{3.21}$$

W ten sposób otrzymaliśmy przedział losowy, zawierający z zadanym prawdopodobieństwem $1-\alpha$ nieznaną wartość średnią $\mu$. Zaobserwowawszy próbę losową $X_1, X_2, \ldots, X_n$, czyli mając realizację tej próby $x_1, x_2, \ldots, x_n$, możemy obliczyć realizację średniej w próbie, $\bar{x}$ i podać **przedział ufności dla $\mu$ na poziomie ufności** $1-\alpha$

$$\left[\bar{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]. \tag{3.22}$$

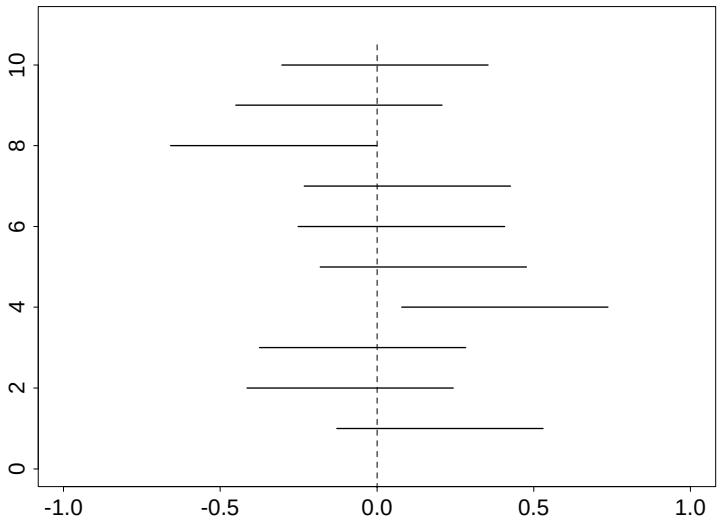
Wprowadzenie pojęcia poziomu ufności $1-\alpha$, niejako w miejsce prawdopodobieństwa $1-\alpha$, jest potrzebne i nie jest mnożeniem bytów ponad potrzebę. O prawdopodobieństwie można mówić tylko wtedy, gdy mamy do czynienia ze zmiennymi losowymi. Gdy mówimy o realizacjach zmiennych losowych, mówienie o prawdopodobieństwie traci sens. Przedział (3.22) nie jest już przedziałem losowym, jest zaś zwykłym przedziałem na osi liczbowej i albo zawiera nieznaną liczbową wartość średnią $\mu$, albo nie. Jak zatem rozumieć pojęcie poziomu ufności?




Rys. 3.5. Kwantyle rzędu $\alpha/2$ i $1-\alpha/2$ standardowego rozkładu normalnego

Aby odpowiedzieć na to pytanie, wróćmy do równości (3.21), która opisuje prawdopodobieństwo zajścia dobrze określonego zdarzenia losowego. Odwołajmy się do częstościowej interpretacji prawdopodobieństwa, która powiada, że gdybyśmy dysponowali nie jedną a 1 milionem prób losowych (o tej samej liczności $n$ i z tego samego rozkładu normalnego $N(\mu, \sigma)$) i, tym samym, gdybyśmy dysponowali nie jedną a 1 milionem średnich próbkowych $\bar{X}$, to oczekiwalibyśmy zajścia zdarzenia

$$\mu \in \left[\bar{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$




Rys. 3.6. Przedziały ufności dla $\mu$ dla 10 prostych prób losowych o liczności 25 z rozkładu $N(0,1)$

z częstością $(1-\alpha)10^6/10^6 = (1-\alpha)$. I tak właśnie należy rozumieć pojęcie poziomu ufności: dla około $100(1-\alpha)\%$ prób losowych obliczony przedział ufności zawiera szacowany parametr (z tego powodu przedział ufności na poziomie $1-\alpha$ nazywamy także $100(1-\alpha)\%$ przedziałem ufności). Na rysunku 3.6 przedstawiono sytuację dla zaledwie 10 prób losowych o liczności 25 z rozkładu $N(0,1)$ i 10 przedziałów ufności (3.22) na poziomie ufności 0,9.

Na mocy równości (3.21) mamy

$$P(|\bar{X} - \mu| \leq z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}) = 1-\alpha,$$

skąd wynika, że błąd średniej próbkowej $\bar{x}$ nie przekracza na poziomie ufności $1-\alpha$ wartości

$$z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

Długość przedziału ufności (por. (3.22)) jest równa podwojonej wartości podanego błędu. Zgodnie z intuicją i zdrowym rozsądkiem, długość przedziału ufności jest tym mniejsza im większa jest liczność próby $n$. Ściśle, długość przedziału ufności jest odwrotnie proporcjonalna do wartości $\sqrt{n}$. Wynika stąd, że dobierając odpowiednio dużą liczność próby, możemy uzyskać przedział ufności o dowolnie małej, ustalonej długości. Łatwo obliczyć, że jeżeli

chcemy, by przedział ufności nie był dłuższy od zadanej wartości, np. $2d$

$$2z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \leq 2d,$$

to liczność próby musi spełniać warunek

$$n \geq \left(\frac{z_{1-\alpha/2}\sigma}{d}\right)^2.$$

Przedział ufności (3.22), tak jak i przedziały podobnego typu, które dopiero omówimy, nazywa się niekiedy przedziałami **dwustronnymi** w przeciwieństwie do przedziałów **jednostronnych**, które są tylko jednostronnie ograniczone. W rozważanym problemie jednostronne przedziały ufności na poziomie ufności $1-\alpha$ mają postać

$$\mu \in \Big(-\infty, \bar{x} + z_{1-\alpha}\frac{\sigma}{\sqrt{n}}\Big] \tag{3.23}$$

oraz

$$\mu \in \Big[\bar{x} - z_{1-\alpha}\frac{\sigma}{\sqrt{n}}, \infty\Big). \tag{3.24}$$

Czytelnikowi pozostawiamy wykazanie, że przedziały te faktycznie mają poziom ufności $1-\alpha$ (por. zad. 3.5).

Przedział (3.22) nie jest oczywiście jedynym możliwym przedziałem dwustronnym, czyli ograniczonym, na poziomie ufności $1-\alpha$. Zamiast kwantyli $z_{\alpha/2}$ i $z_{1-\alpha/2}$ w równości (3.19) można by użyć dowolnych innych kwantyli rozkładu $N(0,1)$, $z_a$ i $z_b$, spełniających warunek

$$P(z_a \leq Z \leq z_b) = 1-\alpha.$$

Łatwo jednak zauważyć, że w przypadku rozkładu symetrycznego, a takim jest rozkład normalny, przedział ufności otrzymany na podstawie warunku (3.19) ma najmniejszą możliwą długość. Ponadto, jeżeli interesuje nas otrzymanie przedziału dwustronnego, naturalne jest wymaganie, by takie same były prawdopodobieństwa otrzymania wartości większych i mniejszych niż wartości zawarte w tym przedziale. Wymaganie to spełnia przedział (3.19). Dlatego w podobny sposób konstruuje się przedziały ufności, gdy rozkład prawdopodobieństwa, na którym się opieramy nie jest symetryczny (por. w następnym punkcie warunki (3.35) i (3.37)).

Kończąc omawianie przedziałów ufności dla wartości średniej rozkładu normalnego ze znanym odchyleniem standardowym, warto zwrócić uwagę na metodologię ich konstruowania. Punktem wyjścia do zbudowania przedziału (3.22) było znalezienie takiej funkcji punktowego estymatora wielkości nas

interesującej oraz parametrów rozkładu populacji, której rozkład prawdopodobieństwa nie zależy od żadnych nieznanych wielkości. Znaleziony w ten sposób rozkład prawdopodobieństwa nazywamy niekiedy **rozkładem odniesienia** dla danego problemu. W naszym przypadku funkcją wyjściowego estymatora jest zmienna losowa $Z$ dana równością (3.18), zaś rozkładem odniesienia standardowy rozkład normalny. Opisany zabieg posłużenia się zmienną losową o znanym rozkładzie odniesienia jest wspólny dla wszystkich zadań budowy przedziałów ufności i, jak się później przekonamy, dla problemu testowania hipotez.

**Przypadek rozkładu normalnego o nieznanym odchyleniu standardowym**

Najczęściej odchylenie standardowe rozkładu populacji nie jest znane. Nasuwa się zatem myśl zastąpienia zmiennej $Z$, danej równością (3.18), zmienną losową

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}. \tag{3.25}$$

Okazuje się, że jest to pomysł nie tylko naturalny, ale i trafny, ponieważ rozkład zmiennej losowej $T$ nie zależy od nieznanego parametru $\mu$ i jest znany. Można mianowicie udowodnić, że jest to tzw. *rozkład t* (zwany też rozkładem *Studenta*) *z* $n-1$ *stopniami swobody* (chcąc zaznaczyć o rozkład z iloma stopniami swobody nam chodzi, rozkład ten oznaczamy symbolem $t_{n-1}$).

Rozkład $t$ z $n$ stopniami swobody jest dany gęstością

$$f(x) = \frac{\Gamma((n+1)/2)}{\Gamma(n/2)\sqrt{\pi n}} \left(1 + \frac{x^2}{n}\right)^{-(n+1)/2},$$

$-\infty < x < \infty$, gdzie $n$ jest ustaloną stałą naturalną, zaś $\Gamma(\cdot)$ oznacza, jak zwykle, funkcję gamma (na końcu p. 3.3.2 wyjaśniamy ogólnie, kiedy iloraz dwóch zmiennych losowych ma rozkład $t$). Zauważmy, że gęstość rozkładu $t$ jest funkcją symetryczną. Jest to gęstość o ogonach grubszych od ogonów gęstości rozkładu $N(0,1)$, dążąca do tej ostatniej wraz ze wzrostem $n$ (por. niżej komentarz poprzedzający przedział (3.27)).

Mając zmienną losową $T$ i jej rozkład $t_{n-1}$ możemy przedział ufności dla $\mu$ zbudować w sposób zupełnie analogiczny do poprzedniego przypadku. Przedział ufności na poziomie $1-\alpha$ przyjmuje postać

$$\left[\bar{x} - t_{1-\alpha/2,n-1}\frac{s}{\sqrt{n}}, \bar{x} + t_{1-\alpha/2,n-1}\frac{s}{\sqrt{n}}\right], \tag{3.26}$$

gdzie $t_{1-\alpha/2,n-1}$ jest kwantylem rzędu $1-\alpha/2$ rozkładu $t_{n-1}$

$$P(T \leq t_{1-\alpha/2,n-1}) = 1 - \alpha/2.$$

Wartości kwantyli odczytujemy z tablic statystycznych dla rozkładu $t$ (tablica II jest zamieszczona na końcu książki).

O estymatorze $S^2$ wariancji $\sigma^2$ wiadomo, że jest nieobciążony oraz, że jego wariancja w przypadku, gdy próba losowa pochodzi z rozkładu normalnego wynosi

$$\frac{2\sigma^4}{n-1}.$$

Ponieważ rozproszenie estmatora $S^2$ maleje wraz ze wzrostem liczności próby $n$, estymator ten dąży w pewnym probabilistycznym sensie do prawdziwej wartości wariancji rozkładu $\sigma^2$ (nie będziemy tu wnikać w jakim matematycznie ścisłym sensie). Stąd, zmienne losowe $Z$ i $T$ stają się przy rosnącym $n$ nierozróżnialne, zaś gęstość rozkładu $t_{n-1}$ dąży do gęstości rozkładu $N(0,1)$, czyli kwantyle rozkładu $t$ dążą do kwantyli tego samego rzędu rozkładu $N(0,1)$. W konsekwencji, dla dostatecznie dużej liczności próby, zamiast przedziału ufności (3.26) można w przypadku nieznajomości odchylenia standardowego $\sigma$ traktować przedział

$$\left[\bar{x} - z_{1-\alpha/2}\frac{s}{\sqrt{n}}, \bar{x} + z_{1-\alpha/2}\frac{s}{\sqrt{n}}\right] \tag{3.27}$$

jako dobre przybliżenie przedziału ufności na poziomie ufności $1-\alpha$ dla wartości średniej $\mu$. Ten ostatni przedział tym lepiej przybliża przedział (3.26) im większa jest liczność próby $n$. W praktyce za wystarczająco dużą liczność próby uznaje się często $n \geq 30$.

**Przykład 3.5.** W przykładzie 1.14 zwróciliśmy uwagę na skośność rozkładu latencji L3-P60, związaną z jej zależnością od wzrostu osoby, na której dokonuje się pomiaru. Zarówno latencję L3-P60, jak i inne latencje, np. L5-P40, można traktować jako sumę pewnej deterministycznej funkcji wzrostu i składnika losowego, mającego ten sam rozkład prawdopodobieństwa dla wszystkich osób badanych (latencja L5-P40, której zaobserwowane wartości są podane w następnym rozdziale w tabeli do zad. 4.2, jest czasem od momentu wzbudzenia potencjału w korzeniu L5 do chwili osiągnięcia przez potencjał pierwszego minimum lokalnego). W następnym rozdziale zobaczymy (por. zad. 4.2), jak znaleźć funkcje uzależniające latencje od wzrostu i, tym samym, jak wyodrębnić składnik losowy każdego pomiaru. Otrzymane składniki losowe latencji L5-P40 są także podane w tabeli do zad. 4.2 (jest to kolumna tzw. rezyduów oznaczona symbolem R-P40). W punkcie 3.4.2 wykażemy, że otrzymaną próbę losowych rezyduów R-P40 można uznać za pochodzącą z rozkładu normalnego. Zbudujmy przedział ufności na poziomie ufności 0,95 dla wartości średniej rezyduów R-P40.

Średnia w próbie rezyduów wynosi 0,000, natomiast $s = 2{,}749$. Dokładny i przybliżony przedział ufności na poziomie ufności 0,95 ma na podstawie (3.26) i (3.27) postać, odpowiednio,

$$\mu \in [-0{,}698,\ 0{,}698] \quad \text{oraz} \quad \mu \in [-0{,}684,\ 0{,}684],$$

ponieważ $t_{61,0,975} = 2{,}000$ i $z_{0,975} = 1{,}960$.

**Porównanie dwóch wartości średnich: niezależne próby losowe oraz pary obserwacji**

Nierzadko interesuje nas nie pojedyncza wartość średnia i odpowiadający jej przedział ufności, ale chodzi nam o porównanie dwóch wartości średnich. Możemy na przykład chcieć sprawdzić czy jedna wartość średnia jest większa od drugiej. W takim przypadku można obliczyć przedział ufności dla różnicy obydwu wartości średnich (odejmujemy np. drugą wartość średnią od pierwszej i sprawdzamy, jakie jest położenie przedzialu ufności dla różnicy względem zera; inne podejście do problemu rozważymy w następnym podrozdziale). Porównania takie wymagają rozróżnienia dwóch zupełnie różnych sytuacji.

Po pierwsze możemy mieć do czynienia z **dwiema niezależnymi prostymi próbami losowymi** (o niekoniecznie tej samej liczności), $X_1, X_2, \ldots, X_{n_1}$ oraz $Y_1, Y_2, \ldots, Y_{n_2}$, z wartościami średnimi, odpowiednio, $\mu_1$ i $\mu_2$. Tak jest np. wtedy, gdy badamy ciśnienie skurczowe w grupie osób chorych na pewną ustaloną chorobę oraz, gdy to samo ciśnienie badamy w niezależnej grupie kontrolnej osób zdrowych. Pierwsza grupa stanowi w tym przypadku pierwszą próbę losową, a druga grupa stanowi próbę losową niezależną od pierwszej. Interesować nas może pytanie o przedział ufności dla różnicy $\mu_1 - \mu_2$.

Nadal zakładamy, że próby losowe pochodzą z rozkładów normalnych. Dodatkowo załóżmy na razie, że są znane odchylenia standardowe obydwu rozkładów $\sigma_1$ i $\sigma_2$. Mamy zatem do czynienia z dwiema próbami, z których pierwsza pochodzi z rozkładu $N(\mu_1, \sigma_1)$, natomiast druga z rozkładu $N(\mu_2, \sigma_2)$. Niech dalej $\bar{X}$ i $\bar{Y}$ oznaczają, odpowiednio, średnią w pierwszej i drugiej próbie losowej. Łatwo wykazać, że statystyka

$$Z = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \tag{3.28}$$

ma standardowy rozkład normalny. Statystyka ta w oczywisty sposób odpowiada statystyce (3.18) i przez analogię można natychmiast skonstruować przedziały ufności w interesującym nas obecnie przypadku. Na przykład

dwustronny przedział ufności dla różnicy $\mu_1 - \mu_2$ na poziomie ufności $1-\alpha$ ma postać

$$\left[ (\bar{x} - \bar{y}) - z_{1-\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}},\ (\bar{x} - \bar{y}) + z_{1-\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right].$$

Przypadek nieznanych odchyleń standardowych $\sigma_1$ i $\sigma_2$, znacznie częściej spotykany w praktyce, rozważymy jedynie przy założeniu równości obydwu odchyleń standardowych, $\sigma_1 = \sigma_2$ (przypadek nierównych odchyleń standardowych jest bardziej złożony, nieznany jest bowiem wówczas dokładny rozkład statystyki testowej). Jeżeli $\sigma_1 = \sigma_2 = \sigma$, wariancja różnicy $\bar{X} - \bar{Y}$ jest równa

$$\sigma^2(\frac{1}{n_1} + \frac{1}{n_2}).$$

Można wykazać, że oparty na obydwu próbach estymator postaci

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}, \tag{3.29}$$

gdzie $S_i^2$, $i = 1, 2$, jest wariancją w $i$-tej próbie, jest nieobciążonym estymatorem wariancji $\sigma^2$ (indeks $p$ w symbolu $S_p^2$ pochodzi od angielskiego słowa *pooled* i w ten sposób wskazuje na wykorzystanie w definicji obu estymatorów $S_1^2$ i $S_2^2$). Co więcej, okazuje się, że statystyka

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{3.30}$$

ma rozkład $t$ Studenta z $n_1 + n_2 - 2$ stopniami swobody. Jak poprzednio, możemy zatem natychmiast skonstruować potrzebny nam przedział ufności. Ograniczając się do dwustronnego przedziału ufności dla różnicy $\mu_1 - \mu_2$ na poziomie ufności $1-\alpha$, otrzymujemy oczywiście przedział

$$\Big[ (\bar{x} - \bar{y}) - t_{1-\alpha/2,n_1+n_2-2}s_p\sqrt{\tfrac{1}{n_1} + \tfrac{1}{n_2}},\ (\bar{x} - \bar{y}) + t_{1-\alpha/2,n_1+n_2-2}s_p\sqrt{\tfrac{1}{n_1} + \tfrac{1}{n_2}} \Big],$$

gdzie $t_{1-\alpha/2,n_1+n_2-2}$ jest kwantylem rzędu $1-\alpha/2$ rozkładu $t$ Studenta z $n_1 + n_2 - 2$ stopniami swobody.

Jakościowo inną sytuacją jest taka, w której mamy do czynienia z **parami obserwacji** $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$, gdzie pary mają taki sam dwuwymiarowy rozkład normalny i są wzajemnie niezależne, ale zmienne w parze mogą być zależne. Tak jest np. wtedy, gdy pacjentom z nadciśnieniem

tętniczym badamy ciśnienie skurczowe przed zastosowaniem terapii i po jej zastosowaniu. Każda para obserwacji odpowiada wówczas konkretnemu pacjentowi i zmienne w parze nie są oczywiście niezależne.

Zauważmy, że nawet jeżeli znamy wariancje zmiennych losowych $X_i$ oraz $Y_i$, $i = 1, 2 \ldots, n$, to ze względu na zależność między każdą taką parą zmiennych nie możemy na tej podstawie podać wariancji różnic $D_i = X_i - Y_i$. Możemy jednak podać oczywisty estymator tej wariancji, a mianowicie wariancję w próbie

$$S_D^2 = \frac{1}{n-1} \sum_{i=1}^{n} (D_i - \bar{D})^2,$$

gdzie $\bar{D} = \sum_{i=1}^{n} D_i$. Założywszy dalej, że różnice $D_i$ tworzą próbę niezależnych zmiennych losowych o rozkładzie normalnym z nieznaną wartością średnią $\mu_D$, zauważamy, że statystyka

$$T = \frac{\bar{D} - \mu_D}{S_D/\sqrt{n}}$$

jest oczywistym odpowiednikiem statystyki (3.25) i ma rozkład $t$ Studenta z $n-1$ stopniami swobody. W ten sposób zadanie konstrukcji przedziału ufności dla różnicy wartości średnich, gdy obserwacje występują w parach, sprowadza się do zadania wcześniej już omówionego. Na przykład dwustronny przedział ufności dla $\mu_D$ na poziomie $1 - \alpha$ ma postać (3.26) z $\bar{d}$ zamiast $\bar{x}$ i $s_D$ zamiast $s$.

Przykłady problemów obydwu typów, czyli konstrukcji przedziałów ufności w przypadku prób niezależnych oraz obserwacji występujących w parach, znajdzie Czytelnik w zadaniach.

### 3.3.2. Przedziały ufności dla wariancji rozkładu normalnego

#### Przedział ufności dla wariancji

Dobrze już przez nas zbadanym punktowym estymatorem wariancji, w szczególności wariancji rozkładu normalnego, jest oczywiście wariancja w próbie

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2.$$

W przypadku, gdy niezależne zmienne losowe $X_i$ pochodzą ze standardowego rozkładu normalnego, zmienna losowa

$$\sum_{i=1}^{n} X_i^2 \tag{3.31}$$

ma tzw. *rozkład $\chi^2$ o $n$ stopniach swobody.*

Rozkład $\chi^2$ o $n$ stopniach swobody jest dany gęstością

$$f(x) = \begin{cases} \frac{1}{2^{n/2}\Gamma(n/2)} x^{n/2-1} e^{-x/2}, & \text{gdy } x > 0 \\ 0, & \text{gdy } x \leq 0. \end{cases}$$

Łatwo zauważyć, że gęstość $\chi^2$ z $n$ stopniami swobody jest szczególnym przypadkiem gęstości gamma (3.11), a mianowicie jest to gęstość gamma z parametrami $n/2$ i 2 (zmienna losowa o rozkładzie $\chi^2$ z $n$ stopniami swobody ma wartość średnią $n$ i wariancję $2n$).

Rozkład sumy (3.31) implikuje, że zmienna losowa

$$\sum_{i=1}^{n} \left( \frac{X_i - \mu}{\sigma} \right)^2, \tag{3.32}$$

gdzie $X_i$, $i = 1, 2, \ldots, n$ są niezależnymi zmiennymi losowymi o rozkładzie $N(\mu, \sigma)$, ma także rozkład $\chi^2$ z $n$ stopniami swobody.

Naturalnym odpowiednikiem zmiennej losowej (3.32) jest zmienna losowa

$$\mathcal{X}^2 = \sum_{i=1}^{n} \left( \frac{X_i - \bar{X}}{\sigma} \right)^2 = \frac{(n-1)S^2}{\sigma^2}. \tag{3.33}$$

Okazuje się, że zmienna $\mathcal{X}^2$ ma rozkład $\chi^2$ z $n-1$ stopniami swobody. Rozkład zmiennej $\mathcal{X}^2$ nie zależy od nieznanych parametrów i jest znany, a zatem może być rozkładem odniesienia w rozważanym przez nas problemie (tablica rozkładu $\chi^2$ jest podana na końcu książki). Mamy

$$P(\chi^2_{\alpha/2,n-1} \leq \mathcal{X}^2 \leq \chi^2_{1-\alpha/2,n-1}) = 1 - \alpha, \tag{3.34}$$

gdzie $\alpha$ jest ustaloną liczbą z przedziału $(0, 1)$ oraz $\chi^2_{\alpha/2,n-1}$ i $\chi^2_{1-\alpha/2,n-1}$ są kwantylami odpowiednio rzędu $\alpha/2$ i $1 - \alpha/2$

$$P(\mathcal{X}^2 \leq \chi^2_{\alpha/2,n-1}) = \alpha/2 \quad \text{oraz} \quad P(\mathcal{X}^2 \leq \chi^2_{1-\alpha/2,n-1}) = 1 - \alpha/2,$$

rozkładu $\chi^2$ z $n-1$ stopniami swobody. Równość (3.34) możemy zapisać w postaci równoważnej

$$P(\chi^2_{\alpha/2,n-1} \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi^2_{1-\alpha/2,n-1}) = 1 - \alpha,$$

żeby po prostych przekształceniach otrzymać

$$P\left( \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}} \right) = 1 - \alpha. \tag{3.35}$$

Z równości (3.35) otrzymujemy szukany przedział ufności na poziomie ufności $1-\alpha$ dla wariancji rozkładu normalnego

$$\left[\frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}}, \frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}}\right].$$

Na tej samej podstawie otrzymujemy natychmiast przedział ufności na poziomie ufności $1-\alpha$ dla odchylenia standardowego rozkładu normalnego

$$\left[\sqrt{\frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}}}, \sqrt{\frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}}}\right].$$

Wprowadzony wcześniej rozkład $t$ z $\nu$ stopniami swobody został pierwotnie uzyskany jako rozkład ilorazu dwóch niezależnych zmiennych losowych: zmiennej losowej $Z$ o standardowym rozkładzie normalnym oraz pierwiastka $\sqrt{V/\nu}$, gdzie $V$ jest zmienną losową o rozkładzie $\chi^2$ z $\nu$ stopniami swobody

$$\frac{Z}{\sqrt{V/\nu}}.$$

Można udowodnić, że w przypadku próby losowej z rozkładu normalnego średnia w próbie $\bar{X}$ oraz wariancja w próbie $S^2$ są niezależne. Zatem statystyka (3.25) może być przedstawiona we właśnie podanej postaci:

$$T = \frac{\bar{X}-\mu}{S/\sqrt{n}} = \frac{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)S^2}{(n-1)\sigma^2}}} = \frac{Z}{\sqrt{V/(n-1)}},$$

gdzie $V = \mathcal{X}^2 = (n-1)S^2/\sigma^2$ jest zmienną o rozkładzie $\chi^2$ z $n-1$ stopniami swobody. Dlatego mogliśmy stwierdzić, że zmienna $T$, na której oparliśmy konstrukcję przedziału ufności dla wartości średniej, gdy jest nieznane odchylenie standardowe $\sigma$, ma rozkład $t$.

### Przedział ufności dla ilorazu wariancji

Wspomnieliśmy już, że nierzadko interesuje nas porównanie wartości średnich dwóch niezależnych prób z rozkładów normalnych. To samo dotyczy porównania dwóch wariancji. Różnica polega na tym, że podczas gdy w pierwszym przypadku narzuca się badanie różnicy między wartościami średnimi, w drugim naturalne jest badanie ilorazu wariancji. Z problemem porównania wariancji mamy do czynienia na przykład zawsze wtedy, gdy wprowadzamy nową technologię i zależy nam na zmniejszeniu wariancji mierzonego parametru naszego wyrobu (poziomu zanieczyszczeń produkowanego związku chemicznego, grubości walcowanej płyty, natężenia prądów błądzących itd.; ich wartości mają nie tylko być możliwie małe lub bliskie wartościom nominalnym, ale powinny z wyrobu na wyrób zachowywać stabilność).

Niech będą dane dwie niezależne próby losowe z populacji normalnych, odpowiednio $N(\mu_1, \sigma_1)$ i $N(\mu_2, \sigma_2)$, o licznościach $n_1$ i $n_2$. Ponieważ naturalne jest estymowanie wariancji obydwu populacji za pomocą odpowiednich wariancji próbkowych $S_1^2$ i $S_2^2$, narzuca się pytanie, czy można do konstrukcji przedziału ufności dla ilorazu $\sigma_2^2/\sigma_1^2$ użyć ilorazu $S_2^2/S_1^2$. Okazuje się, że tak, ponieważ z rachunku prawdopodobieństwa wiadomo, iż właściwie unormowany stosunek dwóch niezależnych zmiennych losowych o rozkładzie $\chi^2$ ma znany rozkład. Jeżeli mianowicie zmienna losowa $U$ ma rozkład $\chi^2$ z $\nu_1$ stopniami swobody oraz zmienna losowa $V$ ma rozkład $\chi^2$ z $\nu_2$ stopniami swobody, to zmienna

$$\frac{U/\nu_1}{U/\nu_2}$$

ma tzw. *rozkład $F$ (Snedecora) z $\nu_1$ i $\nu_2$ stopniami swobody.*

Rozkład ten jest dany gęstością

$$f(x) = \begin{cases} \frac{\Gamma((\nu_1+\nu_2)/2)(\nu_1/\nu_2)^{\nu_1/2}}{\Gamma(\nu_1/2)\Gamma(\nu_2/2)} \frac{x^{\nu_1/2-1}}{(1+\nu_1 x/\nu_2)^{(\nu_1+\nu_2)/2}}, & \text{gdy} \quad x > 0 \\ 0, & \text{gdy} \quad x \leq 0; \end{cases} \tag{3.36}$$

tablica rozkładu $F$ znajduje się na końcu książki.

Pamiętając o rozkładzie zmiennej losowej (3.33), otrzymujemy, że

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$$

jest zmienną losową o rozkładzie $F$ z $n_1 - 1$ i $n_2 - 1$ stopniami swobody. Stąd

$$P(f_{\alpha/2,\nu_1,\nu_2} \leq \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \leq f_{1-\alpha/2,\nu_1,\nu_2}) = 1 - \alpha, \tag{3.37}$$

gdzie $\alpha$ jest dowolną ustaloną liczbą z przedziału $(0, 1)$, $f_{\alpha/2,\nu_1,\nu_2}$ jest kwantylem rzędu $\alpha/2$ rozkładu $F$ z $\nu_1 = n_1 - 1$ i $\nu_2 = n_2 - 1$ stopniami swobody oraz $f_{1-\alpha/2,\nu_1,\nu_2}$ jest kwantylem rzędu $1-\alpha/2$ z tego samego rozkładu. Postępując jak we wcześniejszych problemach, po stosownych przekształceniach ostatecznie otrzymujemy postać przedziału ufności na poziomie ufności $1-\alpha$ dla ilorazu wariancji (dla uproszczenia oznaczeń w symbolach odpowiednich kwantyli nie uwzględniamy informacji o liczbie stopni swobody) :

$$\frac{\sigma_2^2}{\sigma_1^2} \in \left[\frac{s_2^2}{s_1^2} f_{\alpha/2}, \frac{s_2^2}{s_1^2} f_{1-\alpha/2}\right].$$

Ponieważ wiadomo, że kwantyle dowolnego ustalonego rozkładu $F$ są związane następującą zależnością

$$f_{1-\alpha/2} = (f_{\alpha/2})^{-1},$$

więc otrzymanemu przedziałowi ufności można nadać postać

$$\left[\frac{s_2^2}{s_1^2}(f_{1-\alpha/2})^{-1}, \frac{s_2^2}{s_1^2}f_{1-\alpha/2}\right].$$

Oczywista modyfikacja powyższego przedziału daje przedział ufności na poziomie $1-\alpha$ dla ilorazu odchyleń standardowych populacji normalnych.

### 3.3.3. Uwaga o przedziałach ufności w przypadku rozkładów ciągłych, innych niż normalny

W punktach 3.3.1 i 3.3.2 zakładaliśmy konsekwentnie, że próby losowe pochodzą z rozkładu normalnego. Przynajmniej jednak w sytuacjach rozważonych w p. 3.3.1 nasuwa się możliwość uznania uzyskanych tam przedziałów za przybliżone przedziały ufności dla wartości średniej, gdy rozkład, z którego pochodzą próby losowe jest inny niż normalny. O niektórych innych możliwościach poradzenia sobie z problemem rozkładu różnego od normalnego wspomnimy w rozdz. 8.

Zauważmy przede wszystkim, że zgodnie z Centralnym Twierdzeniem Granicznym statystyka (3.18) ma w przybliżeniu standardowy rozkład normalny bez względu na to, jaki jest rozkład zmiennych losowych $X_1, X_2, \ldots, X_n$, o ile tylko liczność próby $n$ jest dostatecznie duża. A zatem, jeżeli liczność $n$ jest wystarczająco duża, wszystkie przedziały ufności oparte na statystyce (3.18) i mające poziom ufności $1-\alpha$, gdy próba pochodzi z rozkładu normalnego, mają **w przybliżeniu** ten sam poziom ufności, gdy próba pochodzi z dowolnego innego, ustalonego rozkładu prawdopodobieństwa ze znaną wariancją $\sigma^2$. W praktyce, uznajemy, że liczność próby jest wystarczająco duża, gdy jest nie mniejsza od 30 i gdy rozkład nie odbiega zbyt daleko od jednomodalnego rozkładu symetrycznego. W przypadku, gdy rozkład jest wyraźnie skośny, przedziały (3.22)–(3.24) uznajemy za przybliżone przedziały ufności dla wartości średniej, jeżeli liczność $n$ jest przynajmniej nie mniejsza od około 40.

Zauważyliśmy już wcześniej, że statystyka (3.25) może być uznana za przybliżenie statystyki (3.18), tym lepsze im większa jest liczność próby. Wynika stąd, że skoro statystyka (3.25) dąży wraz ze wzrostem $n$ do statystyki (3.18), a ta ostatnia umożliwia konstrukcję przybliżonych przedziałów ufności, gdy rozkład próby odbiega od rozkładu normalnego, to taką samą własność aproksymowania dokładnego przedziału ufności mają też przedziały zbudowane na podstawie statystyki (3.25). Różnice między obydwiema konstrukcjami są dwie. Po pierwsze, w drugim przypadku niepotrzebna jest znajomość odchylenia standardowego rozkładu populacji. I po drugie, dla

małych liczności $n$, statystyki $Z$ i $T$ są istotnie różne. W mianowniku pierwszej występuje $\sigma$, drugiej natomiast $S$, pierwsza statystyka – jeżeli próba pochodzi z rozkładu normalnego – ma rozkład normalny, druga ma wówczas rozkład $t$ Studenta. Okazuje się, że w praktyce statystyka $T$ dobrze przybliża rozkład $t$ Studenta przy niezbyt dużych odchyleniach rozkładu próby od rozkładu normalnego już dla $n$ nawet nieco mniejszych od 20. Już zatem dla tak małych liczności prób można często skonstruować przybliżony przedział ufności oparty na statystyce (3.25).

Ostatecznie zatem, przedziały ufności oparte na statystyce $T$ mają zgoła uniwersalne zastosowanie, nie ograniczone do przypadku prób z rozkładu normalnego (oczywiście, jeśli liczność tych prób nie jest zbyt mała).

Nie jest to wszakże „uniwersalność bez granic". Czytelnik pamięta na pewno, że wariancja w próbie $S^2$ nie jest odpornym estymatorem wariancji populacji. Jest to mianowicie estymator nieodporny na obserwacje odstające. Nie będziemy się tym problemem zajmować dokładniej, ale musimy przynajmniej wspomnieć, że badania eksperymentalne potwierdzają, iż przedziały ufności wtedy tylko są precyzyjne, gdy w próbie nie ma elementów odstających (por. zad. 3.9).

Wykazana stosunkowo duża uniwersalność przedstawionej w p. 3.3.1 metody budowy przedziałów ufności dla wartości średniej rozkładu ciągłego nie ma swego odpowiednika w przypadku przedziałów ufności dla wariancji. Badania eksperymentalne potwierdzają, że odbieganie rozkładu próby od rozkładu normalnego ma z reguły poważne lub bardzo poważne konsekwencje dla precyzji przedziałów ufności dla wariancji. Dotyczy to zwłaszcza przedziałów ufności dla ilorazu wariancji – z reguły opisana konstrukcja przedziału ufności staje się bezużyteczna, gdy rozkłady prób nie są normalne lub bardzo bliskie normalnym (por. zad. 3.10). Zauważmy przy tym, że – paradoksalnie – nie jest to może wielki problem, poza tym, że metoda ma bardzo ograniczone pole zastosowań. Ów paradoks bierze się stąd, że badanie wariancji właściwie traci sens, gdy rozkład jest np. wyraźnie skośny – wariancja nie jest w takiej sytuacji dobrym wskaźnikiem rozproszenia rozkładu i tym samym nie warto się nią zajmować.

### 3.3.4. Przedziały ufności dla proporcji

W punkcie 2.4.3 rozważyliśmy własności częstości $\hat{p}$, czyli punktowego estymatora nieznanej proporcji $p$. Między innymi zwróciliśmy uwagę, że $\hat{p}$ jest nieobciążonym estymatorem proporcji. Znajomość teorii estymacji NMW umożliwia wykazanie, iż jest to także estymator NMW. Natomiast z bezpośrednich rachunków wynika (por. zad. 3.2), że $\hat{p}$ jest również estymatorem NW.

Opierając się na częstości $\hat{p}$, w tym punkcie skonstruujemy przedziały ufności dla proporcji $p$. Jest to zadanie proste w przypadku, gdy próba losowa niezależnych zmiennych o rozkładzie dwupunktowym, $P(X_i = 1) = 1 - P(X_i = 0) = p$, $i = 1, 2, \ldots, n$, jest dostatecznie liczna, by móc skorzystać z przybliżenia rozkładu statystyki

$$\frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \tag{3.38}$$

standardowym rozkładem normalnym (por. tw. 2.16 oraz akapit pod tym twierdzeniem, mówiący jakie są minimalne wymagania na liczność próby $n$ i proporcję $p$, by móc przybliżyć rozkład statystyki (3.38) rozkładem $N(0,1)$). Wówczas mamy bowiem ($z_{1-\alpha/2}$ oznacza, jak zwykle, kwantyl rzędu $1-\alpha/2$ rozkładu $N(0,1)$)

$$P\left(-z_{1-\alpha/2} \leq \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \leq z_{1-\alpha/2}\right) \approx 1 - \alpha$$

skąd

$$P\left(\hat{p} - z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) \approx 1 - \alpha.$$

Ostatecznie zatem, dla dostatecznie dużej liczności próby losowej, przybliżony dwustronny przedział ufności na poziomie ufności $1-\alpha$ dla proporcji $p$ ma postać (por. zad. 3.11 – obliczanie minimalnej liczności próby, przy której długość tego przedziału nie przekracza zadanej wielkości)

$$\left[\hat{p} - z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \ \hat{p} + z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]. \tag{3.39}$$

**Przykład 3.6.** Jedna z agencji badających opinię publiczną ogłosiła w czerwcu 2000 r., że przebadała reprezentatywną próbę 1000 dorosłych obywateli polskich, z których 57% poparło starania ich państwa o wejście do Unii Europejskiej. Uznając, że mamy do czynienia z rozkładem dwupunktowym (popieranie lub nie starań o wejście do UE), możemy skonstruować 95% przedział ufności dla proporcji obywateli popierających wejście Polski do UE. Próba jest wystarczająco liczna, by przy otrzymanej wartości częstości, $\hat{p} = 0{,}57$ móc posłużyć się przedziałem (3.39). Pamiętając, że $z_{0,975} = 1{,}96$ oraz $n = 1000$, na poziomie

ufności 0,95, otrzymujemy przedział [0,54, 0,60] (na podstawie stwierdzenia 2.8, wielkość $\sqrt{\hat{p}(1-\hat{p})/1000} = 0{,}0156$ możemy uznać za błąd standardowy otrzymanej częstości; w jednostkach procentowych, błąd standardowy estymatora wynosi zatem około 1,6%).

Konstrukcja przedziału ufności dla proporcji komplikuje się, gdy nie można skorzystać z przybliżenia normalnego. Jak to zobaczyliśmy we wszystkich wcześniejszych przypadkach, naszym celem jest znalezienie przy dowolnym ustalonym poziomie ufności $1-\alpha$ takich dwóch funkcji $h_1(\cdot)$ i $h_2(\cdot)$ próby losowej, aby była spełniona równość

$$P(h_1(X_1, X_2, \ldots, X_n) \leq \theta \leq h_2(X_1, X_2, \ldots, X_n)) = 1 - \alpha, \qquad (3.40)$$

gdzie $\theta$ jest parametrem, dla którego konstruujemy przedział ufności (w aktualnie rozpatrywanym przypadku $\theta = p$). Ponieważ zmienne losowe $X_1, X_2, \ldots, X_n$ przyjmują w interesującym nas tu przypadku tylko dwie wartości, więc mogą nie istnieć żadne takie funkcje $h_1(\cdot)$ i $h_2(\cdot)$, dla których równość (3.40) jest spełniona. Ten problem można by rozwiązać żądając spełnienia tej równości tylko w przybliżeniu (np. zamiast chcieć otrzymać przedział na poziomie ufności 0,95, możemy zadowolić się przedziałem, którego poziom ufności jest możliwie bliski wartości 0,95). Problem jednak w tym, że gdy próba losowa pochodzi z rozkładu dwupunktowego, nieznana jest analityczna postać szukanych funkcji $h_1(\cdot)$ i $h_2(\cdot)$. W rezultacie, jeżeli nie możemy skorzystać z przybliżenia normalnego, musimy się odwołać do specjalnych tablic statystycznych (patrz np. R. Zieliński, W. Zieliński (1990): *Tablice statystyczne*. Warszawa, PWN) lub do pomocy komputerowego pakietu statystycznego. Brzegi przedziału ufności są zwykle podane jako wartości zależne od zaobserwowanej sumy $\sum x_i = \hat{p}n$, liczności próby $n$ oraz żądanego poziomu ufności $1-\alpha$.

**Przykład 3.7.** Z bazy danych zawierającej 5000 wierszy (inaczej zapisów lub rekordów), przedstawiających wyniki pewnych badań medycznych wykonanych na 5000 pacjentów, wylosowano 500 wierszy i sprawdzono w ilu wierszach znajdują się błędy w zapisie danych. Okazało się, że wszystkie wiersze są zapisane bezbłędnie. Skonstruujemy przedział ufności na poziomie ufności 0,99 dla proporcji bezbłędnych wierszy w bazie danych.

Nasza próba losowa ma liczność $n = 500$, obserwowane zaś zmienne losowe $X_i$, $i = 1, 2 \ldots, 500$ przyjmują wartość 1, gdy $i$-ty wiersz jest poprawny oraz 0, gdy zawiera błąd. Ponieważ $n(1-\hat{p}) = 0$, nie możemy skorzystać z przybliżenia normalnego. Z tablic statystycznych

Zielińskiego i Zielińskiego (1990) dla parametru $p$ w rozkładzie dwumianowym odczytujemy, że przedział ufności na poziomie ufności 0,99 ma postać $p \in [0{,}989,\ 1]$. (Posłużenie się tablicą dla rozkładu dwumianowego wynika oczywiście stąd, iż $\sum X_i$ ma rozkład dwumianowy z parametrami $n$ i $p$.)

## 3.4. Testowanie hipotez

### 3.4.1. Testowanie hipotez w rodzinach rozkładów normalnych i rozkładów dwupunktowych

**Testy dla wartości średniej w rodzinie rozkładów normalnych – przypadek znanej wariancji**

Nierzadko nie tyle interesuje nas ocena wartości parametru populacji, co weryfikacja jakiejś o nim hipotezy. Wiadomo na przykład, że przy stosowanej technologii produkcji pewnego stopu metali średni poziom zanieczyszczeń tego stopu wyraża się liczbą $\mu_0$ promili. Wiadomo też, że zaproponowane zmiany technologii nie mogą zwiększyć tego poziomu. Istnieje zarazem nadzieja, że przy nowej technologii średni poziom zanieczyszczeń $\mu$ okaże się mniejszy od $\mu_0$. Zachodzi zatem naturalna potrzeba rozpoczęcia badań nad nową technologią od zweryfikowania możliwości odrzucenia hipotezy orzekającej, iż poziom zanieczyszczeń wynosi $\mu_0$ na rzecz przyjęcia hipotezy, że ów poziom jest mniejszy od $\mu_0$.

Mówimy tu o *średnim* poziomie zanieczyszczeń i o *hipotezach*, nie zaś po prostu o poziomie zanieczyszczeń i jego wartości, ponieważ badane zjawisko ma charakter losowy. Jeżeli, stosując nową technologię, wykonamy np. 15 próbek stopu, w każdym przypadku otrzymamy nieco inną wartość poziomu zanieczyszczeń. Dlatego nasz sąd musi dotyczyć właśnie *średniego* poziomu i dlatego ów sąd nie może być sądem kategorycznym, np. kategorycznie orzekającym, że średni poziom zanieczyszczeń *jest* mniejszy niż $\mu_0$. Na podstawie próby losowej możemy jedynie – z zadaną dozą pewności – *skłaniać się* do przyjęcia *hipotezy* o zmniejszeniu poziomu zanieczyszczeń stopu dzięki zastosowaniu nowej technologii.

Innym przykładem jest sytuacja, w której żądamy, aby w procesie produkcji obwodów elektrycznych napylana warstwa krzemu miała zadaną średnią grubość $\nu_0$. Dopuszczamy przy tym pewne wahania otrzymywanej grubości. Proces produkcji daje grubości napylanych warstw krzemu o ustalonym i dopuszczalnym odchyleniu standardowym, zachodzi jednak obawa, że ich wartość średnia $\nu$ jest różna od $\nu_0$. Interesuje nas zatem pytanie czy ist-

nieją podstawy do odrzucenia hipotezy stwierdzającej, że średnia grubość warstwy wynosi $\nu_0$ na rzecz hipotezy postaci $\nu \neq \nu_0$.

Jak już na to zwróciliśmy uwagę, w każdym przypadku wnioskowanie ma charakter statystyczny i opiera się na specjalnie w tym celu zebranej próbie pomiarów interesującej nas wielkości. Precyzyjniejsze sformułowanie naszego zadania przedstawimy na kolejnym przykładzie.

Dokładne toczenie tłoka pompy paliwa silnika samochodowego ma dawać średnicę pewnej części tłoka równą 7,5 mm. Celem eksperymentu jest sprawdzenie, czy zużycie noża tokarki nie spowodowało zwiększenia wartości średniej $\theta$ interesujących nas średnic. Pożądaną wartością średnią tych średnic jest oczywiście wartość $\theta_0 = 7{,}5$ mm.

Ogólnie biorąc, mamy zatem do czynienia z dwiema hipotezami, z których pierwsza podlega weryfikacji i może zostać odrzucona na korzyść drugiej hipotezy. Problem taki nazywamy problemem **testowania hipotez**. Pierwszą hipotezę nazywamy **hipotezą zerową**. W ostatnim przykładzie jest to hipoteza, że wartość średnia średnic wynosi $\theta_0 = 7{,}5$ mm. Hipotezę tę oznaczamy symbolem $H_0$ i piszemy

$$H_0\colon\ \theta = \theta_0, \tag{3.41}$$

przy czym w naszym przypadku $\theta_0 = 7,5$. Drugą hipotezę nazywamy z oczywistych względów **hipotezą alternatywną**. Oznaczamy ją symbolem $H_1$. W naszym przypadku hipotezę tę możemy zapisać krótko

$$H_1\colon\ \theta > \theta_0. \tag{3.42}$$

Podkreślenia wymaga tu następująca kwestia. **Hipotezie zerowej przypisujemy inną wagę niż hipotezie alternatywnej.** Można powiedzieć, że hipotez tych nie traktujemy symetrycznie. Mianowicie, **za hipotezę zerową przyjmujemy tę, której prawdziwość poddajemy w wątpliwość** i którą chcemy odrzucić, jeśli tylko znajdziemy do tego podstawę. W pewnym sensie ważniejsza jest dla nas hipoteza alternatywna, ponieważ nasze zadanie formułujemy jako zadanie szukania podstaw do odrzucenia hipotezy zerowej na korzyść przyjęcia jej alternatywy. Nasze postępowanie przypomina zachowanie prokuratora – jakkolwiek sąd musi opierać się na domniemaniu niewinności podsądnego (hipoteza zerowa), prokuratora interesuje uzasadnienie fałszywości tego domniemania i odrzucenie go na korzyść orzeczenia winy podsądnego (czyli na korzyść hipotezy alternatywnej).

Hipotezę, która jednoznacznie wyznacza rozkład prawdopodobieństwa, z którego jest losowana próba losowa, nazywamy **hipotezą prostą**. W naszym przypadku taką hipotezą jest hipoteza zerowa. W przypadku przeciwnym, gdy zbiór rozkładów opisywanych przez hipotezę zawiera więcej niż jeden

rozkład prawdopodobieństwa, hipotezę nazywamy **złożoną**. W naszym przykładzie hipoteza alternatywna jest właśnie hipotezą złożoną. Możliwe wartości parametru $\theta$, wyspecyfikowane przez hipotezę alternatywną, nazywamy wartościami alternatywnymi parametru będącego przedmiotem testowania.

Celem zweryfikowania hipotezy zerowej dokonano pomiarów odpowiedniej średnicy pięćdziesięciu tłoków. Pomiary można uznać za niezależne i pochodzące z tego samego rozkładu. Skądinąd wiadomo przy tym, iż średnice produkowanych tłoków mają rozkład normalny (zagadnienie weryfikacji hipotezy o normalności rozkładu prawdopodobieństwa rozważymy w dalszym ciągu tego podrozdziału). Jak wynika z wcześniejszego opisu problemu, wartością średnią tego rozkładu jest nieznana liczba $\theta$. Natomiast z wcześniej zdobytego doświadczenia wiadomo także, iż odchylenie standardowe rozkładu średnic jest znane i wynosi 0,05 mm. Reasumując, dysponujemy realizacją prostej próby losowej $X_1, X_2, \ldots, X_{50}$ z rozkładu $N(\theta, 0{,}05)$.

Posiadana informacja o rozkładzie oraz dysponowanie realizacją prostej próby losowej średnic z tego rozkładu umożliwiają rozwiązanie problemu weryfikacji hipotezy $H_0$ przy alternatywie $H_1$. Zważywszy, że nasze pytanie dotyczy wartości średniej rozkładu, naturalne jest oparcie się na własnościach średniej w próbie losowej. Niech zatem średnia w próbie $\bar{X}$ będzie naszą wstępną propozycją **statystyki testowej**, czyli tej funkcji próby, na której oprzemy nasz test hipotezy $H_0$ przy hipotezie alternatywnej $H_1$. Weźmy najpierw pod uwagę własności wstępnie wybranej statystyki testowej $\bar{X}$ przy założeniu, że ta właśnie hipoteza zachodzi. Umożliwi to opisanie „typowego" zachowania się statystyki testowej pod warunkiem zachodzenia $H_0$. Jeżeli zaobserwowana wartość $\bar{x}$ statystyki testowej $\bar{X}$ okaże się „typowa", powiemy, że nie mamy podstaw do odrzucenia hipotezy $H_0$. Jeżeli natomiast wartość $\bar{x}$ okaże się „nietypowa" przy założeniu zachodzenia hipotezy $H_0$ i zarazem owa „nietypowość" będzie wskazywać, iż prawdziwa wartość średnia rozkładu może spełniać hipotezę alternatywną $H_1$, powiemy, że mamy podstawy do odrzucenia hipotezy zerowej na korzyść przyjęcia hipotezy alternatywnej (hipotezę zerową uznajemy wówczas za fałszywą). Zauważmy, że **nieodrzucenie (czyli przyjęcie) hipotezy zerowej nie dowodzi jej prawdziwości, a jedynie wynika z braku podstaw do jej odrzucenia.** Podobnie jak w naszej analogii z procesem sądowym, gdzie zabezpieczamy się przed niesłusznym skazaniem podsądnego, tak tym razem będziemy się zabezpieczać przed błędnym odrzuceniem hipotezy zerowej. Podkreślmy przy tym jeszcze raz, że **za hipotezę zerową przyjmujemy tę, dla której chcemy szukać podstaw do jej odrzucenia.**

W rozważanym przez nas przypadku statystyka $\bar{X}$ ma pod warunkiem zachodzenia hipotezy $H_0$ rozkład normalny o wartości średniej $\theta_0$, przy czym

$\theta_0 = 7{,}5$, i odchyleniu standardowym $\sigma/\sqrt{n}$, gdzie $\sigma = 0{,}05$ oraz $n = 50$ (podane własności rozkładu statystyki $\bar{X}$ są natychmiastową konsekwencją tego, że próba losowa $X_1, X_2, \ldots, X_n$ pochodzi z rozkładu normalnego o odchyleniu standardowym $\sigma$ oraz z założenia prawdziwości hipotezy $H_0$). Jeżeli zatem prawdziwa jest hipoteza $H_0$, wartości statystyki $\bar{X}$ powinny być bliskie wartości $\theta_0$. Ich rozproszenie wokół tej wartości średniej jest „kontrolowane" przez odchylenie standardowe średniej w próbie, równe $\sigma/\sqrt{n}$.

Zamiast posługiwać się bezpośrednio statystyką $\bar{X}$, możemy ją oczywiście zastąpić jej standaryzowaną wersją

$$Z = \frac{\bar{X} - \theta_0}{\sigma/\sqrt{n}}. \tag{3.43}$$

W dalszym ciągu za właściwą statystykę testową wygodnie będzie uznać zmienną losową $Z$ zamiast średniej $\bar{X}$. Pod warunkiem zachodzenia hipotezy zerowej rozkład statystyki testowej jest zatem rozkładem $N(0,1)$, jej wartości powinny skupiać się wokół zera, zaś rozproszenie statystyki $Z$ jest kontrolowane przez jednostkowe odchylenie standardowe. Co więcej, jasne jest, że w przypadku gdy hipoteza zerowa jest fałszywa, prawdziwa natomiast jest hipoteza alternatywna (3.42), statystyka $Z$ powinna mieć tendencję do przyjmowania „dużych" wartości, tym przy tym większych im większa jest różnica między prawdziwą wartością średnią $\theta$ a wartością $\theta_0$. Rzeczywiście, w przypadku zachodzenia hipotezy $H_1$ standardowy rozkład normalny ma statystyka $\frac{\bar{X}-\theta}{\sigma/\sqrt{n}}$, skąd obliczana w toku testowania statystyka testowa $Z$ ma rozkład $N(\frac{\theta-\theta_0}{\sigma/\sqrt{n}}, 1)$, ponieważ

$$Z = \frac{\bar{X} - \theta_0}{\sigma/\sqrt{n}} = \frac{\bar{X} - \theta}{\sigma/\sqrt{n}} + \frac{\theta - \theta_0}{\sigma/\sqrt{n}}. \tag{3.44}$$

Pod warunkiem $H_1$ statystyka testowa $Z$ ma zatem rozkład normalny przesunięty względem rozkładu standardowego o $\frac{\theta-\theta_0}{\sigma/\sqrt{n}}$.

Możemy zatem sprecyzować co należy rozumieć przez „typowe" i „nietypowe" wartości statystyki testowej pod warunkiem zachodzenia $H_0$ i przy alternatywie $H_1$. Mianowicie, oznaczając jak zwykle kwantyl rzędu $1-\alpha$ rozkładu $N(0,1)$ symbolem $z_{1-\alpha}$, mamy

$$P_{H_0}(Z \geq z_{1-\alpha}) = \alpha, \tag{3.45}$$

gdzie napisaliśmy $P_{H_0}$ zamiast $P$, by wyraźnie zaznaczyć, że prawdopodobieństwo obliczamy przy założeniu prawdziwości hipotezy zerowej. Z równości (3.45) wynika, iż przy zachodzeniu hipotezy $H_0$ wartości statystyki

testowej $Z$ mogą znaleźć się w zbiorze $C = \{z\colon z \geq z_{1-\alpha}\}$, czyli w zbiorze wszystkich liczb nie mniejszych od $z_{1-\alpha}$, z prawdopodobieństwem $\alpha$. Na przykład, przyjmując $\alpha = 0{,}001$, z tablic otrzymujemy, że $z_{0{,}999} = 3{,}09$ i wartości statystyki $Z$ nie mniejsze od 3,09 będziemy na pewno skłonni uznać za nietypowe lub „nieprawdopodobne", gdy jest prawdziwa hipoteza $H_0$. Uznajemy, że takie „duże" wartości tej statystyki uzasadniają odrzucenie $H_0$ na korzyść przyjęcia hipotezy alternatywnej $H_1$. Zaobserwowanie mniejszej wartości statystyki testowej $Z$ niż 3,09 interpretujemy wówczas jako brak podstaw do odrzucenia hipotezy zerowej i hipotezę tę przyjmujemy jako prawdziwą. W postępowaniu tym uderza arbitralność wyboru konkretnej wartości parametru $\alpha$ i, w konsekwencji, podziału zbioru wszystkich możliwych wartości statystyki testowej na zbiór wartości „typowych" i „nietypowych" (przy założeniu prawdziwości hipotezy $H_0$). Do problemu tej arbitralności wrócimy w dalszym ciągu niniejszego punktu. Natomiast w tym miejscu krótko podsumujemy sposób przeprowadzenia testu.

Ustaliwszy statystykę testową, dzielimy zbiór wszystkich możliwych wartości tej statystyki na dwa dopełniające się podzbiory:

- zbiór wartości statystyki testowej, prowadzących do odrzucenia hipotezy $H_0$ na korzyść hipotezy $H_1$ (jest to zbiór „nietypowych" wartości statystyki testowej pod warunkiem prawdziwości $H_0$); zbiór ten nazywamy **zbiorem krytycznym** i zwykle oznaczamy literą $C$;
- zbiór wartości statystyki testowej, prowadzących do nieodrzucenia hipotezy $H_0$ (będziemy mówić krótko **zbiór przyjęć** hipotezy $H_0$), stanowiący dopełnienie zbioru krytycznego (zgodnie z naszą konwencją, jeśli zbiór krytyczny oznaczymy symbolem $C$, to zbiór przyjęć możemy oznaczyć $C'$).
- Wartości brzegowe zbioru $C$, graniczące ze zbiorem $C'$, nazywamy **wartościami krytycznymi** testu.

Jeżeli w naszym przykładzie zaobserwowana wartość $z$ statystyki $Z$ należy do zbioru $C = \{z\colon z \geq z_{0{,}999}\}$, czyli jeżeli $z$ jest liczbą nie mniejszą od wartości krytycznej $z_{0{,}999} = 3{,}09$, to hipotezę $H_0$ odrzucamy i przyjmujemy hipotezę $H_1$. Jeżeli natomiast $z \in C'$, przyjmujemy hipotezę $H_0$. Po dokonaniu pomiarów średnic 50 tłoków, $x_1, x_2, \ldots, x_{50}$, okazało się, że ich średnia $\bar{x} = 7{,}515$. Stąd

$$z = \frac{\bar{x} - \theta_0}{\sigma/\sqrt{n}} = 20\sqrt{50}(7{,}515 - 7{,}5) = 2{,}121.$$

Zatem, zaobserwowana wartość statystyki testowej $Z$ należy do zbioru przyjęć powstałego na mocy równości (3.45) z $\alpha = 0{,}001$. Zaobserwowana

wartość 2,121 jest „typowa" dla standardowego rozkładu normalnego, jeżeli uznać, że wartości „nietypowe" to wartości tak duże, że ich zbiór tworzy zdarzenie o prawdopodobieństwie równym zaledwie 0,001. Odwołując się do częstościowej interpretacji prawdopodobieństwa, możemy powiedzieć, że zgodziliśmy się uznać za nietypowe wartości, które mogą się średnio zdarzyć raz na 1000 50-elementowych prób badanych średnic.

Ściśle mówiąc, z równości (3.45) wynika, że prawdopodobieństwo znalezienia się wartości statystyki testowej w zbiorze krytycznym, gdy prawdziwa jest hipoteza zerowa $H_0$, wynosi $\alpha$. Innymi słowy, w analizowanym przez nas problemie testowania, $\alpha$ **jest prawdopodobieństwem odrzucenia hipotezy zerowej, gdy ta jest prawdziwa.**

**Definicja 3.1.** *Odrzucenie hipotezy zerowej, gdy ta jest prawdziwa, nazywamy* ***błędem pierwszego rodzaju****. Jeżeli hipoteza zerowa jest hipotezą prostą, prawdopodobieństwo $\alpha$ popełnienia błędu pierwszego rodzaju nazywamy* ***poziomem istotności testu****. Jeżeli w wyniku przeprowadzenia testu otrzymano wartość statystyki testowej należącą do zbioru krytycznego, to mówimy, że dane okazały się statystycznie istotne na poziomie $\alpha$.*

Jeśli, nadal przy zachodzeniu hipotezy zerowej, w omawianym przykładzie zgodzimy się powiększyć prawdopodobieństwo błędu pierwszego rodzaju do wartości 0,01, wartość krytyczna testu wyniesie 2,326, ponieważ $z_{0,99} = 2{,}326$. W takim przypadku zbiór krytyczny ma postać $C = \{z\colon z \geq 2{,}326\}$ i dla wartości $z = 2{,}121$ test przyjmuje hipotezę zerową na poziomie istotności 0,01. Zwiększenie poziomu istotności np. do wartości 0,05, powoduje odrzucenie hipotezy zerowej i przyjęcie hipotezy alternatywnej, ponieważ $z_{0,95} = 1,645$ (innymi słowy, dane są statystycznie istotne na poziomie 0,05). Konsekwencje podnoszenia poziomu istotności, czyli prawdopodobieństwa popełnienia błędu pierwszego rodzaju, są zupełnie naturalne. Tak musi się dziać, jeśli za nietypowe zgadzamy się uznać wartości z przedziału coraz mniej odległego od wartości średniej standardowego rozkładu normalnego. Sytuację tę zilustrowano na rys. 3.7a.

Czy to znaczy, że broniąc się przed popełnieniem błędu pierwszego rodzaju powinniśmy np. postulować jego minimalizację? Na pewno nie, ponieważ z naszego przykładu wynika, że nie tylko prawdopodobieństwo błędu pierwszego rodzaju (poziom istotności testu) **musi** być większe od zera, ale też że jego minimalizacja nie ma sensu. Po pierwsze, żeby otrzymać zerowe prawdopodobieństwo błędu pierwszego rodzaju, żeby więc nigdy nie odrzucać hipotezy zerowej, gdy ta jest prawdziwa, trzeba by przyjąć, że zbiór przyjęć hipotezy $H_0$ jest całą prostą $(-\infty, \infty)$. Wówczas jednak hipoteza $H_0$ nie byłaby nigdy odrzucana, bez względu na to, czy jest prawdziwa, czy też

fałszywa! I po drugie, za chwilę zobaczymy bez trudu, że zwiększanie prawdopodobieństwa błędu pierwszego rodzaju, czyli zwiększanie prawdopodobieństwa odrzucenia hipotezy zerowej, gdy jest prawdziwa, jest równoważne jednoczesnemu zwiększaniu prawdopodobieństwa odrzucenia tej hipotezy, gdy jest fałszywa. A zatem życząc sobie małego prawdopodobieństwa odrzucenia hipotezy zerowej, gdy jest prawdziwa, trzeba zarazem pamiętać, że chcielibyśmy mieć możliwie duże prawdopodobieństwo odrzucenia tej hipotezy, gdy jest fałszywa. Przyjęta ostatecznie strategia musi opierać się na kompromisie między obydwoma dążeniami.

Na rysunku 3.7a jest pokazana gęstość statystyki testowej $Z$ przy zachodzeniu hipotezy zerowej. Na rysunku 3.7b jest pokazana gęstość statystyki $Z$ w sytuacji, gdy prawdziwa wartość średnia średnic w próbie 50-elementowej wynosi $\theta = 7{,}51$, czyli gdy hipoteza zerowa jest fałszywa, natomiast prawdziwa jest hipoteza alternatywna $H_1$. Jak wynika z równości (3.44), statystyka $Z$ ma w tej sytuacji rozkład o gęstości normalnej z wartością oczekiwaną

$$\frac{\theta - \theta_0}{\sigma/\sqrt{n}} = 20\sqrt{50}(\theta - \theta_0) = 1{,}414$$

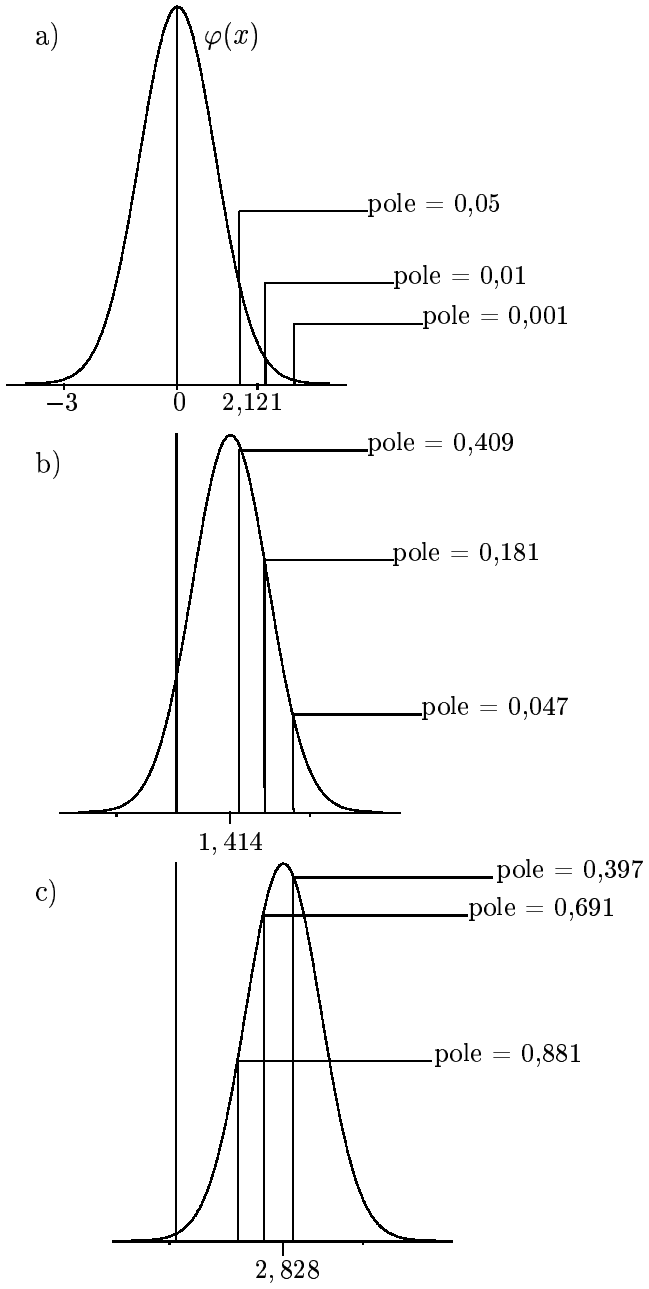
i odchyleniem standardowym 1. Prawdopodobieństwo odrzucenia fałszywej hipotezy $H_0$ i przyjęcia prawdziwej hipotezy $H_1$ jest równe całce z tej gęstości po zbiorze krytycznym $C$ (zauważmy, że zbiór, po którym całkujemy gęstość statystyki $Z$ dla prawdziwej wartości średniej, powstał na podstawie rozkładu odpowiadającego hipotezie zerowej). Dla zbioru krytycznego $C = (3{,}09,\ \infty)$ oraz $\theta = 7{,}51$ otrzymujemy (por. rys. 3.7b)

$$\begin{aligned} P_{\theta=7{,}51}(Z \geq 3{,}09) &= \frac{1}{\sqrt{2\pi}} \int_{3{,}09}^{\infty} \exp[-(u - 1{,}414)^2/2]du = \\ &= \frac{1}{\sqrt{2\pi}} \int_{1{,}676}^{\infty} \exp[-u^2/2]du = 0{,}047; \end{aligned}$$

symbol $P_{\theta=7{,}51}$ oznacza, że prawdopodobieństwo jest obliczane dla przypadku $\theta = 7{,}51$.

Wygodne jest wprowadzenie w tym miejscu pojęcia mocy testu.

**Definicja 3.2.** *Dla zadanej alternatywnej wartości parametru $\theta$, będącego przedmiotem testowania, prawdopodobieństwo odrzucenia (fałszywej) hipotezy zerowej i przyjęcia (prawdziwej) hipotezy alternatywnej nazywamy* ***mocą testu*** *dla tej wartości parametru $\theta$.*




Rys. 3.7. a) Zbiory krytyczne dla poziomów istotności: 0,05; 0,01; 0,001 i zaobserwowanej wartości statystyki testowej 2,121, b) wartości mocy dla $\theta = 7{,}51$, c) wartości mocy dla $\theta = 7{,}52$

Jak wynika z powyższych obliczeń, przyjęcie bardzo niskiego poziomu istotności testu (czyli bardzo małej wartości prawdopodobieństwa odrzucenia hipotezy zerowej, gdy jest prawdziwa), równego 0,001, prowadzi do bardzo nikłej mocy testu, gdy jest prawdziwa hipoteza alternatywna i prawdziwa wartość parametru $\theta$ jest większa od $\theta_0$ o 0,01. Jak to zilustrowano na rys. 3.7b, w sytuacji gdy $\theta - \theta_0 = 0{,}01$, moc testu można zwiększyć zwiększając poziom istotności $\alpha$. Przyjąwszy $\alpha = 0{,}01$ i stąd $C = [z_{0,99}, \infty)$, gdzie $z_{0,99} = 2{,}326$ jest kwantylem rzędu 0,99 standardowego rozkładu normalnego (czyli rozkładu statystyki testowej przy hipotezie zerowej), otrzymujemy moc testu równą 0,181. Podniesienie poziomu istotności do wartości $\alpha = 0{,}05$ daje zbiór krytyczny $C = [z_{0,95}, \infty)$, gdzie $z_{0,95} = 1{,}645$ oraz moc 0,409, gdy $\theta - \theta_0 = 0{,}01$. Podsumowując, **przy ustalonej wartości $\theta$ ze zbioru alternatyw** i, jak to dotąd cały czas zakładamy, przy ustalonej liczności $n$ próby losowej, **zwiększanie mocy testu może się odbywać jedynie kosztem zwiększania poziomu istotności.**

Jest intuicyjnie jasne, że – przy ustalonym poziomie istotności testu – moc zależy od wielkości przesunięcia rozkładu statystyki testowej $Z$ względem rozkładu standardowego. Przesunięcie to jest dane równością (3.44) i wynosi $\frac{\theta-\theta_0}{\sigma/\sqrt{n}}$. Zachowując $n = 50$, ale zakładając, że prawdziwa wartość parametru $\theta = 7{,}52$, czyli że $\theta - \theta_0 = 0{,}02$ otrzymujemy przesunięcie rozkładu statystyki $Z$, odpowiadającego alternatywie $\theta = 7,52$, równe 2,828 i stąd (por. rys. 3.7c) moc testu równą 0,397, gdy $\alpha = 0{,}001$; moc testu równą 0,691, gdy $\alpha = 0{,}01$; oraz 0,881, gdy $\alpha = 0{,}05$. Ogólnie biorąc, **przy zadanym poziomie istotności** (i ustalonej liczności próby losowej) **moc testu rośnie wraz ze wzrostem odległości między alternatywną wartością parametru $\theta$ oraz wartością tego parametru zadaną przez hipotezę zerową.**

Wszystkie obliczenia prowadzone dotąd w naszym przykładzie były związane z ustaloną licznością próby losowej, równą 50. Zwiększenie różnicy między parametrami $\theta$ i $\theta_0$ z 0,01 do 0,02 dało zwiększenie przesunięcia rozkładu statystyki testowej $Z$ względem rozkładu standardowego od wartości 0,141 do wartości 0,282. Łatwo wykazać, że **taki sam** efekt zwiększenia tego przesunięcia uzyskuje się zachowując różnicę $\theta - \theta_0 = 0,01$, ale zwiększając liczność próby losowej do $n = 200$. Mianowicie, dla $n = 200$ przesunięcie $\frac{\theta-\theta_0}{\sigma/\sqrt{n}}$ jest równe 2,828, czyli moc testu na poziomie istotności 0,01 jest znowu równa 0,691, mimo że $\theta - \theta_0 = 0{,}01$. Zatem, **nie zmieniając poziomu istotności, moc testu możemy zwiększyć, odpowiednio zwiększając liczność próby.** Jeżeli np. zależy nam na otrzymaniu w naszym przykładzie testu o poziomie istotności rzędu 0,01 i mocy większej niż 0,8, gdy prawdziwa wartość średnia średnic jest o 0,01 mm większa od wartości nominalnej 7,5 mm, to możemy oprzeć się na omówionym teście, ale pobrawszy próbę o liczności odpowiednio większej od 50 (por. zad. 3.13).

Nasze dotychczasowe rozważania pozwalają zatem zasugerować taki sposób testowania, który zapewnia rozsądny kompromis między naszym dążeniem do przyjmowania hipotezy zerowej, gdy jest prawdziwa i zarazem dążeniem do jej odrzucania, gdy jest fałszywa. Kompromis ten nie ma charakteru matematycznego czy statystycznego, lecz jest konsekwencją tego co w praktyce chcemy rzeczywiście osiągnąć. Jeżeli, jak wyżej zauważyliśmy, w naszym przykładzie chcemy wykrywać zwiększenie wartości średniej średnicy o 0,01 mm z mocą większą od 0,8, to na pewno powinniśmy zdecydować się na pobranie próby o liczności znacznie większej niż 50. Jeżeli natomiast wystarczy nam wykrywanie przesunięcia o 0,02 mm z mocą równą około 0,7 i odpowiada nam poziom istotności 0,01 (nie mówiąc o poziomie istotności 0,05), to wystarczy pobranie próby o liczności 50.

Kończąc dyskusję testu dla wartości średniej wybranej średnicy tłoka pompy paliwa należy podkreślić, że wszystkie uzyskane wnioski pozostają słuszne nie tylko dla innych problemów testowania hipotezy (3.41) przy alternatywie (3.42), ale także dla wszystkich innych problemów testowania rozważanych w dalszym ciągu tego punktu.

Wspomnijmy jeszcze, że moc testu przy zadanej alternatywnej wartości parametru $\theta$, będącego przedmiotem testowania, umożliwia natychmiastowe określenie prawdopodobieństwa odrzucenia hipotezy alternatywnej (czyli przyjęcia hipotezy zerowej), gdy ta jest prawdziwa. Mianowicie, z def. 3.2 wynika od razu, że prawdopodobieństwo to wynosi $1 - m$, gdzie $m$ oznacza moc przy zadanej alternatywnej wartości parametru $\theta$. Samo zdarzenie odrzucenia hipotezy alternatywnej, gdy jest prawdziwa zadana alternatywna wartość parametru $\theta$ nazywamy **błędem drugiego rodzaju**, odpowiadającym tej wartości parametru.

Wspomnijmy jeszcze też, że w niektórych zagadnieniach jest uzasadnione zastąpienie w problemie testowania (3.41)–(3.42) prostej hipotezy zerowej (3.41) hipotezą złożoną postaci

$$H_0\colon\ \theta \leq \theta_0. \tag{3.46}$$

Chwila zastanowienia wystarczy do zauważenia, że konstruując statystykę testową oraz zbiór krytyczny dla problemu testowania z hipotezą zerową (3.46) oraz hipotezą alternatywną (3.42) należy postępować **tak samo** jak postępowaliśmy rozwiązując problem (3.41)–(3.42). Rzecz w tym, że najtrudniej jest poprawnie wybrać między hipotezą zerową a hipotezą alternatywną wtedy, gdy prawdziwa wartość parametru $\theta$ z przedziału $(-\infty, \theta_0]$ jest wartością brzegową $\theta_0$. Opis własności testu o tyle się przy tym komplikuje, że prawdopodobieństwo popełnienia błędu pierwszego rodzaju jest przy hipotezie zerowej (3.46) i ustalonym zbiorze krytycznym funkcją prawdziwej wartości parametru $\theta$, gdzie $\theta \in (-\infty, \theta_0]$. Przyjąwszy w zaproponowanej konstrukcji $C = (z_{1-\alpha}, \infty)$, otrzymujemy prawdopodobieństwo błędu pierwszego rodzaju równe $\alpha$, gdy $\theta = \theta_0$. Dalej, prawdopodobieństwo to dla każdego parametru $\theta < \theta_0$ jest mniejsze od $\alpha$, i to tym mniejsze im większa jest

różnica $\theta_0 - \theta$. (W przypadku złożonej hipotezy zerowej poziomem istotności nazywamy największą wartość prawdopodobieństwa błędu pierwszego rodzaju, czyli w omawianej sytuacji poziom istotności testu wynosi $\alpha$.) Przy podanej konstrukcji statystyki testowej i zbioru krytycznego wszystko co zostało powiedziane o mocy testu zachowuje swoją ważność, co ostatecznie potwierdza słuszność użycia tego samego testu do testowania zarówno hipotezy (3.41), jak i (3.46) przy alternatywie (3.42).

Uważny Czytelnik zauważył zapewne ścisły związek zasady budowania zbioru krytycznego z zasadą budowy odpowiedniego przedziału ufności. Konstrukcja zbioru przyjęć hipotezy zerowej jest w oczywistym sensie dualna wobec konstrukcji zbioru krytycznego – ta ostatnia jest oparta na równości (3.45), natomiast konstrukcja zbioru przyjęć opiera się na równości

$$P_{H_0}(Z < z_{1-\alpha}) = 1 - \alpha, \tag{3.47}$$

gdzie statystyka $Z$ jest dana związkiem (3.43) i ma standardowy rozkład normalny. Ale równość (3.47) jest tą równością, na podstawie której został zbudowany przedział ufności (3.24), a zatem można uznać za dualne zadania konstrukcji przedziałów ufności oraz testowania.

Zarówno wspomniana dualność, jak i szczegółowo w tym punkcie omówiona naturalna zasada konstrukcji testu hipotezy (3.41) przy alternatywie (3.42) umożliwia natychmiastowe podanie właściwej postaci testu hipotezy zerowej (3.41) przy hipotezie alternatywnej

$$H_1\colon\ \theta \neq \theta_0, \tag{3.48}$$

w sytuacji gdy dysponujemy prostą próbą losową $X_1, X_2, \ldots, X_n$ z rozkładu normalnego o znanym odchyleniu standardowym $\sigma$ i nieznanej wartości średniej $\theta$. (Z oczywistych względów hipotezę (3.48) nazywamy **dwustronną**, podczas gdy hipotezę (3.42) nazywamy **jednostronną**.)

W takim przypadku statystyka testowa wyraża się znowu wzorem (3.43), natomiast zbiór krytyczny testu na poziomie istotności $\alpha$ przyjmuje postać (por. związek (3.20))

$$C = \{z\colon\ z \leq -z_{1-\alpha/2} \text{ lub } z \geq z_{1-\alpha/2}\}. \tag{3.49}$$

Zatem, jeżeli wartość statystyki testowej przyjmuje dla zaobserwowanej próby losowej wartość $z$ i $z \in C$, to hipotezę (3.41) odrzucamy i przyjmujemy hipotezę (3.48), jeżeli natomiast $z \in C'$, to hipotezę zerową przyjmujemy.

Jeżeli rozważone dotąd hipotezy alternatywne zastąpimy hipotezą

$$H_1\colon\ \theta < \theta_0 \tag{3.50}$$

i nie zmienimy pozostałych założeń dotyczących problemu testowania, to stosujemy tę samą statystykę testową (3.43), zaś zbiór krytyczny testu na poziomie istotności $\alpha$ przyjmuje postać

$$C = \{z\colon\ z \leq -z_{1-\alpha}\}. \tag{3.51}$$

**Przykład 3.8.** W procesie produkcji standardowych cyfrowych układów scalonych, znanych jako układy TTL, jeden z pierwszych etapów produkcyjnych polega na wytworzeniu tzw. warstwy epitaksjalnej typu $n$. Pożądane jest otrzymanie warstwy grubości 3,5 $\mu$m. Zakłada się, że proces jest stabilny, otrzymywane grubości warstwy epitaksjalnej mają rozkład normalny o wartości średniej $\theta$ i znanym odchyleniu standardowym 0,7 $\mu$m. Zachodzi podejrzenie, że średnia wartość grubości jest różna od wartości pożądanej $\theta = 3{,}5$ $\mu$m. Postanowiono poddać testowi hipotezę

$$H_0\colon\ \theta = 3{,}5$$

przy hipotezie alternatywnej

$$H_1\colon\ \theta \neq 3{,}5.$$

Pobrano zatem próbę losową 100 płytek podłożowych z wytworzoną wartswą epitaksjalną i otrzymano średnią grubość tej warstwy w próbie równą 3,58 $\mu$m. Wartość statystyki testowej (3.43) wynosi

$$z = \frac{3{,}58 - 3{,}5}{0{,}7/10} = 1{,}143,$$

co nie daje podstaw do odrzucenia hipotezy zerowej na żadnym racjonalnym poziomie istotności. W istocie, na poziomie istotności 0,1 wartość krytyczna testu wynosi 1,645, zaś na bardzo wysokim i w praktyce nie stosowanym poziomie istotności 0,2 wartość ta wynosi 1,282. (Zauważmy, że otrzymana konkluzja jest zawsze konsekwencją tego, jaką wielokrotnością odchylenia standardowego średniej w próbie jest różnica między zaobserwowaną średnią w próbie a nominalną wartością średniej – w rozpatrywanej sytuacji ta wielokrotność wynosi tylko 1,143.)

**Przykład 3.9.** Specjaliści sieci supermarketów sprzedających między innymi produkty spożywcze podejrzewają, że mleko pochodzące od jednego z producentów kooperujących z siecią ma niższą zawartość tłuszczu niż nominalna wartość 3,2%. Specjaliści zakładają przy tym,

że deklarowane przez producenta odchylenie standardowe zawartości tłuszczu w mleku nie zmieniło się i wynosi 0,05%. Ponadto zakładają, że faktyczna procentowa zawartość tłuszczu jest wielkością losową o rozkładzie normalnym.

Postanowiono zatem poddać testowi hipotezę

$$H_0\colon\ \theta = 3{,}2,$$

gdzie $\theta$ oznacza procentową zawartość tluszczu w mleku, przy alternatywie

$$H_1\colon\ \theta < 3{,}2.$$

Ponieważ zawartość tłuszczu powinna być taka sama w różnych kartonach tej samej partii produktu, do testu pobrano po jednym kartonie z dziesięciu losowo wybranych partii. Wybór liczności próby nie był przy tym przypadkowy: do takiej liczności próby doprowadziło postawione na wstępie wymaganie, by test na poziomie istotności 0,05 wykrywał alternatywę postaci $\theta = 3{,}17$ z mocą około 0,6 (por. zad. 3.14). Uzyskano następujące zawartości tłuszczu:

$$3{,}26,\ 3{,}12,\ 3{,}24,\ 3{,}16,\ 3{,}08,\ 3{,}14,\ 3{,}23,\ 3{,}11,\ 3{,}09,\ 3{,}24.$$

Średnia w otrzymanej próbie wynosi 3,167, skąd statystyka testowa (3.43) przyjmuje wartość

$$z = \frac{3{,}167 - 3{,}2}{0{,}05/\sqrt{10}} = -2{,}087. \tag{3.52}$$

Na podstawie wzoru (3.51) oraz tego, że $z_{0,95} = 1{,}645$, zdecydowanie się na test na poziomie istotności 0,05, prowadzi do przyjęcia hipotezy alternatywnej, orzekającej zbyt małą zawartość tłuszczu w mleku. Na poziomie istotności 0,01 przyjęta zostałaby hipoteza zerowa, w tym bowiem przypadku wartość krytyczna testu wynosi $-2{,}326$. Gdybyśmy zdecydowali się przyjąć poziom istotności 0,02, otrzymana wartość statystyki testowej także należałaby do zbioru krytycznego, chociaż leżałaby bardzo blisko wartości krytycznej równej w takim przypadku $-2{,}054$. Natomiast w przypadku poziomu istotności 0,0185, otrzymana wartość statystyki testowej, $z = -2{,}087$ jest równa wartości krytycznej takiego testu, czyli leży na brzegu zbioru krytycznego. Hipotezę zerową odrzucamy na poziomie istotności 0,0185, ale nie odrzucamy jej na żadnym mniejszym poziomie.

**$p$–wartości**

Ostatni przykład dobrze ilustruje problem właściwego doboru poziomu istotności i, tym samym, wartości krytycznej testu. Jak zobaczymy, przykład ten sugeruje zarazem, jak do pewnego stopnia uniknąć dylematów wynikających z arbitralności owego doboru.

Jeżeli w przykładzie 3.9 zażądać bardzo małego prawdopodobieństwa błędu pierwszego rodzaju, np. równego 0,01, hipoteza zerowa zostaje przyjęta. Oczywiście, gdybyśmy zażądali jeszcze mniejszego prawdopodobieństwa pierwszego rodzaju, np. 0,001, hipoteza zerowa „tym bardziej" musiałaby być przyjęta. Aby odrzucić hipotezę, statystyka testowa musiałaby w tym ostatnim przypadku przyjąć wartość z przedziału bardzo odległego od wartości zerowej (czyli swojej wartości średniej, gdy prawdziwa jest hipoteza zerowa), bo mającego prawdopodobieństwo równe zaledwie 0,001.

Spróbujmy w naszym przykładzie zwiększać prawdopodobieństwo błędu pierwszego rodzaju, by dojść do takiego jego poziomu (czyli dojść do takiego poziomu istotności), począwszy od którego hipoteza zerowa zostaje odrzucona na korzyść hipotezy alternatywnej. W przykładzie 3.9 najmniejszym prawdopodobieństwem błędu pierwszego rodzaju, przy którym zaobserwowana wartość statystyki testowej prowadzi do odrzucenia hipotezy zerowej jest 0,0185. Gdybyśmy zatem zgodzili się przyjąć np. poziom istotności testu równy 0,02 (i, tym bardziej, 0,05), hipotezę zerową także odrzucilibyśmy.

**Definicja 3.3.** *Najmniejszy poziom istotności, przy którym zaobserwowana wartość statystyki testowej prowadzi do odrzucenia hipotezy zerowej, nazywamy $p$–**wartością** przeprowadzonego testu.*

Zauważmy, że jeżeli testujemy hipotezę (3.41) przy hipotezie alternatywnej (3.42) i zaobserwowaliśmy wartość $z$ statystyki testowej, to $p$–wartość jest równa

$$P_{H_0}(Z \geq z),$$

gdzie $Z$ jest zmienną losową o standardowym rozkładzie normalnym. Jeżeli hipotezą alternatywną jest hipoteza (3.48), to $p$–wartość jest równa

$$2P_{H_0}(Z \geq |z|),$$

ponieważ rozkład normalny jest symetryczny i interesuje nas prawdopodobieństwo przyjęcia wartości nie mniejszej od $|z|$ lub nie większej do $-|z|$. Jeżeli wreszcie hipoteza alternatywna ma postać (3.50), to $p$–wartość wynosi oczywiście

$$P_{H_0}(Z \leq z).$$

Pytanie o $p$–wartość testu umożliwia w pewnej mierze uniknięcie problemu doboru poziomu istotności przed wykonaniem testu. Najpierw wykonuje się

test, w ten sposób otrzymuje wartość statystyki testowej i następnie sprawdza się dla jakiego poziomu istotności zaobserwowana wartość tej statystyki jest wartością krytyczną testu, czyli wartością leżącą na brzegu zbioru krytycznego (dla każdego wyższego poziomu istotności wartość statystyki testowej leży już nie na brzegu, lecz we wnętrzu zbioru krytycznego).

**Im mniejsza jest $p$–wartość, tym mocniejsze staje się przekonanie testującego o fałszywości hipotezy zerowej i prawdziwości hipotezy alternatywnej.** Mówiąc najogólniej, nikt nie odrzuci hipotezy zerowej otrzymawszy $p$–wartość rzędu 0,4 – zaobserwowaną wartość statystyki testowej należy w takiej sytuacji uznać za zdecydowanie typową przy zachodzeniu hipotezy zerowej. Jeżeli zależy nam na bardzo „pewnym" spełnianiu hipotezy zerowej, możemy ją odrzucić otrzymawszy $p$–wartość równą np. 0,12 – jeżeli np. hipoteza zerowa oznacza, że nowy konserwant nie zagraża zdrowiu, warto poddać w wątpliwość jej prawdziwość, gdy przeprowadzenie testu dało właśnie wspomnianą $p$–wartość 0,12. W innych problemach możemy stawać się skłonni do odrzucenia hipotezy zerowej dopiero, gdy $p$–wartość osiągnie wartość rzędu kilku setnych. Praktycznie zawsze odrzucimy hipotezę zerową otrzymawszy $p$–wartość rzędu 0,001. W każdym razie, $p$–wartość wskazuje jak bardzo nietypowa jest zaobserwowana wartość statystyki testowej.

Otrzymana w przykł. 3.9 $p$–wartość wynosi zaledwie 0,0185 i przeto powinna skłonić sieć handlową do odrzucenia hipotezy zerowej. Inna sprawa, że praktyczne konsekwencje odrzucenia hipotezy zerowej mogą być różne. Na przykład z odrzucenia w naszym przykładzie hipotezy zerowej nie wynika automatycznie, że sieć handlowa powinna natychmiast zaalarmować producenta mleka, domagając się dodatkowej kontroli procesu produkcji i najprawdopodobniej podjęcia stosownych kroków naprawczych. Rzecz w tym, że w praktyce niepożądana może być zdolność testu do odrzucenia hipotezy zerowej z dużą mocą, gdy prawdziwa wartość parametru jest bardzo bliska wartości odpowiadającej hipotezie zerowej (mówimy wówczas o zbytniej czułości testu). W naszym przypadku nie miałoby sensu wszczynanie alarmu, gdyby prawdziwa zawartość tłuszczu w mleku wynosiła powiedzmy 3,198, mimo że hipoteza zerowa byłaby wówczas fałszywa (por. zad. 3.14).

Kończąc omawianie $p$–wartości trzeba ostrzec Czytelnika przed częstym błędem interpretacyjnym, uznającym $p$–wartość za poziom istotności, czyli za prawdopodobieństwo popełnienia w przeprowadzonym teście błędu pierwszego rodzaju. Tymczasem $p$–wartość **nie jest poziomem istotności przeprowadzonego testu**, ponieważ wartość ta jest pewną funkcją zaobserwowanej wartości badanej statystyki testowej. A zatem $p$–wartość jest w rzeczywistości zmienną losową, a nie z góry i przed przeprowadzeniem testu ustaloną liczbą, jak to się dzieje przy ustalaniu poziomu istotności. Rzeczywiście, jak już wcześniej napisaliśmy, $p$–wartość powstaje przez sprawdzenie dla jakiego poziomu istotności zaobserwowana wartość

tej statystyki jest wartością krytyczną testu. Fakt, że $p$–wartość nie jest poziomem istotności, nie przeszkadza jednak w jej stosowaniu i takiemu jej intuicyjnemu rozumieniu, jakie wcześniej przedstawiliśmy.

**Testy dla wartości średniej w rodzinie rozkładów normalnych – przypadek nieznanego odchylenia standardowego**

To co dotąd powiedzieliśmy o testowaniu hipotez oraz w p. 3.3.1 o przedziałach ufności dla wartości średniej rozkładu normalnego, umożliwia natychmiastowe i bez zbędnych dyskusji zaproponowanie właściwego testu, gdy nie jest znane odchylenie standardowe $\sigma$.

Właściwą statystyką testową dla testowania hipotezy (3.41) o średniej $\theta$ przy hipotezie alternatywnej (3.41) lub (3.48) lub (3.50) jest oczywiście statystyka $T$ dana wzorem (3.25) z parametrem $\mu$ zastąpionym przez $\theta_0$,

$$T = \frac{\bar{X} - \theta_0}{S/\sqrt{n}}, \tag{3.53}$$

gdzie jak zwykle, $S$ jest odchyleniem standardowym w próbie oraz $n$ jest licznością próby, na podstawie której testujemy hipotezę.

W przypadku testowania na poziomie istotności $\alpha$ hipotezy (3.41) przy alternatywie (3.42) zbiór krytyczny testu przyjmuje postać

$$C = \{t\colon\ t \geq t_{1-\alpha,n-1}\},$$

gdzie $t$ jest zaobserwowaną wartością statystyki $T$ danej wzorem (3.53) i $t_{1-\alpha,n-1}$ jest kwantylem rzędu $1-\alpha$ rozkładu $t$ Studenta z $n-1$ stopniami swobody.

Jeżeli testujemy na poziomie istotności $\alpha$ hipotezę (3.41) przy hipotezie alternatywnej (3.48), zbiór krytyczny ma postać

$$C = \{t\colon\ t \leq -t_{1-\alpha/2,n-1} \ \text{ lub } \ t \geq t_{1-\alpha/2,n-1}\},$$

przy oczywistych oznaczeniach. Jeżeli natomiast testujemy na poziomie istotności $\alpha$ hipotezę (3.41) przy hipotezie alternatywnej (3.50), to

$$C = \{t\colon\ t \leq -t_{1-\alpha,n-1}\}.$$

Omówiony test nazywamy testem $t$.

Podobnie jak w przypadku konstrukcji przedziałów ufności, jeżeli liczność próby jest dostatecznie duża (w praktyce, gdy $n \geq 30$), w definicjach zbiorów krytycznych kwantyle rozkładu $t$ Studenta możemy zastąpić kwantylami tego samego rzędu standardowego rozkładu normalnego, by w ten sposób otrzymać testy na poziomie istotności równym $\alpha$ w przybliżeniu.

Jak już o tym wspominaliśmy, wszystkie ogólne rozważania przeprowadzone dla testów przy znanym odchyleniu standardowym i dotyczące poziomu istotności, mocy oraz $p$–wartości pozostają słuszne zarówno dla problemu omawianego obecnie, jak i dla problemów następnych.

**Przykład 3.10.** Matematyczne modele zachowania się neuronu bywają bardzo złożone i wymagają uproszczenia, jeśli chce się otrzymane równania efektywnie rozwiązać i na tej podstawie np. przewidywać dynamiczne zachowanie konkretnego układu neuronowego. Postanowiono zbadać adekwatność przybliżonego modelu pewnego układu, opisującego reakcję szczura na zadany bodziec. Czas reakcji na bodziec obliczono na podstawie tego modelu i następnie sprawdzono jakie są prawdziwe wartości czasu reakcji w próbie 24 szczurów. Zakładamy, że czasy reakcji możemy uznać za losowe i mające rozkład normalny. Przyjąwszy za $\theta_0$ wartość czasu reakcji otrzymaną z modelu uproszczonego, równą 3,54 milisekund, oraz zastosowawszy statystykę (3.53), testowi poddano hipotezę (3.41) przy alternatywie (3.48). Średnia w próbie wyniosła 3,59, natomiast odchylenie standardowe 0,18 milisekund. Stąd, statystyka $T$ przyjmuje wartość 1,361, zaś odpowiadająca jej $p$–wartość (dla rozkładu $t$ Studenta z 23 stopniami swobody i przy dwustronnej hipotezie alternatywnej) wynosi 0,192 (tak jak to się obecnie zwykle czyni, zamiast specyfikować poziom istotności, od razu znaleźliśmy $p$–wartość odpowiadającą przeprowadzonemu testowi). Nie ma zatem dostatecznych podstaw do odrzucenia modelu uproszczonego jako nieadekwatnego.

### Testy dla dwóch prób w rodzinie rozkładów normalnych

Ostatnią część punktu 3.3.1 poświęciliśmy dwóm ważnym zagadnieniom porównania wartości średnich. Skonstruowanym tam przedziałom ufności odpowiadają oczywiście stosowne procedury testowe, umożliwiające na zadanym poziomie istotności odrzucenie lub przyjęcie hipotezy o równości obydwu wartości średnich.

Rozważmy najpierw problem porównania wartości średnich **dwóch różnych populacji**, w przypadku gdy dysponujemy niezależnymi próbami losowymi z tych populacji, a mianowicie próbą o liczności $n_1$ z pierwszej populacji, $X_1, X_2, \ldots, X_{n_1}$, oraz próbą o liczności $n_2$ z drugiej populacji, $Y_1, Y_2, \ldots, Y_{n_2}$. Hipoteza zerowa ma postać

$$H_0\colon\ \mu_1 = \mu_2, \tag{3.54}$$

gdzie $\mu_1$ jest wartością średnią pierwszej populacji i $\mu_2$ jest wartością średnią drugiej populacji. Hipoteza alternatywna może mieć jedną z następujących

postaci

$$H_1\colon\ \mu_1 > \mu_2 \tag{3.55}$$

lub

$$H_1\colon\ \mu_1 < \mu_2 \tag{3.56}$$

lub

$$H_1\colon\ \mu_1 \neq \mu_2. \tag{3.57}$$

Załóżmy, że są znane odchylenia standardowe obydwu populacji, $\sigma_1$ i $\sigma_2$, czyli pierwsza ma rozkład $N(\mu_1, \sigma_1)$ natomiast druga $N(\mu_2, \sigma_2)$. Niech dalej $\bar{X}_1$ i $\bar{X}_2$ oznaczają, odpowiednio, średnią w pierwszej i drugiej próbie losowej. Wiemy już, że statystyka

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \tag{3.58}$$

ma standardowy rozkład normalny. Po oznaczeniu $\mu_1 - \mu_2 = \theta$ hipotezie (3.54) możemy nadać postać (3.41) z $\theta_0 = 0$, hipoteza (3.55) przyjmuje postać (3.42), zaś hipotezom (3.56) i (3.57) odpowiadają hipotezy (3.50) i (3.48), za każdym razem z $\theta_0 = 0$. Jeżeli jest spełniona hipoteza zerowa, statystyka (3.58) przyjmuje postać

$$Z = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}. \tag{3.59}$$

Statystyka ta w oczywisty sposób odpowiada statystyce (3.43), a zatem ostatecznie można testowanie równości średnich sprowadzić do wcześniej zbadanego problemu testowania pojedynczej średniej (ze statystyką (3.59) zamiast statystyki (3.43)).

Podobnie jak w punkcie 3.3.1, przypadek nieznanych odchyleń standardowych $\sigma_1$ i $\sigma_2$ rozważymy jedynie przy założeniu równości obydwu odchyleń standardowych, $\sigma_1 = \sigma_2$. Procedurę testową możemy oprzeć na statystyce (3.30), która dla hipotezy zerowej (3.54) przyjmuje postać

$$T = \frac{\bar{X}_1 - \bar{X}_2}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{3.60}$$

i która ma wówczas rozkład $t$ Studenta z $n_1 + n_2 - 2$ stopniami swobody. Jeśli mamy podaną statystykę testową i znamy jej rozkład, łatwo już możemy skonstruować zbiory krytyczne (lub sposób obliczania $p$–wartości) dla testu hipotezy (3.54) przy hipotezie alternatywnej (3.55) lub (3.56) lub (3.57).

Na przykład przy hipotezie alternatywnej (3.57) zbiór krytyczny na poziomie istotności $1-\alpha$, oparty na zaobserwowanej wartości statystyki (3.60), przyjmuje postać

$$C = \{t\colon t \le -t_{1-\alpha/2,n_1+n_2-2} \text{ lub } t \ge t_{1-\alpha/2,n_1+n_2-2}\}.$$

Natomiast $p$–wartość obliczamy w tym przypadku ze wzoru (por. def. 3.3 i komentarz do niej)

$$2P_{H_0}(T \ge |t|),$$

gdzie $t$ jest zaobserwowaną wartością statystyki (3.60).

**Przykład 3.11.** Występujące w układach scalonych klasyczne tranzystory domieszkowane złotem mają tzw. czas magazynowania ładunku rzędu 7 ns. Producent ma nadzieję, że pewna zmiana technologii doprowadziła do zmniejszenia czasu magazynownia w nowych tranzystorach i chciałby przetestować hipotezę (3.54) przy hipotezie alternatywnej (3.55), gdzie $\mu_1$ oznacza średni czas magazynowania przy starej technologii oraz $\mu_2$ ten sam czas po zmianie technologii.

Na podstawie niezależnych badań producent doszedł do wniosku, że przy stosowaniu obydwu technologii otrzymuje się w przybliżeniu normalne rozkłady czasu magazynowania, oraz że odchylenia standardowe obydwu rozkładów można uznać za takie same. Zdecydował się więc abstrahować od ewentualnych małych odstępstw rozkładów od rozkładu normalnego i zastosował test oparty na statystyce (3.60).

W tym celu pobrał dwie niezależne próby losowe po 50 tranzystorów każda, pierwszą składającą się z tranzystorów produkowanych zgodnie ze starą technologią i drugą składającą się z nowych tranzystorów. Z pomiarów czasów magazynowania otrzymał $\bar{x}_1 = 6{,}6$, $\bar{x}_2 = 6{,}3$ oraz $s_p = 0{,}5$ i stąd wartość statystyki $T$ równą $t = 3{,}0$. Ostatecznie zatem $p$–wartość przeprowadzonego testu wyniosła 0,002. Hipoteza zerowa powinna zostać odrzucona na korzyść hipotezy (3.55). Jest sprawą producenta rozważyć, czy na zmniejszeniu średniej wartości czasu magazynownia można poprzestać, czy też należy także dążyć do zmniejszenia odchylenia standardowego interesującego nas czasu.

Rozważmy teraz sytuację, w której mamy do czynienia z **parami obserwacji** $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$, gdzie pary są wzajemnie niezależne, ale zmienne w parze mogą być zależne. Jak pamiętamy, typową sytuacją tego rodzaju jest dwukrotne mierzenie wartości pewnej cechy tego samego obiektu.

Niech, jak w punkcie 3.3.1, różnice $D_i = X_i - Y_i$ tworzą próbę niezależnych zmiennych losowych o rozkładzie normalnym z nieznaną wartością średnią $\mu_D$. Hipoteza zerowa przyjmuje wówczas postać

$$H_0\colon\ \mu_D = 0, \tag{3.61}$$

natomiast możliwe hipotezy alternatywne (3.53)–(3.56), odpowiednio, $\mu_D > > 0$, $\mu_D < 0$ oraz $\mu_D \neq 0$. Statystyka

$$T = \frac{\bar{D}}{S_D/\sqrt{n}}, \tag{3.62}$$

gdzie zachowujemy oznaczenia z p. 3.3.1, jest oczywistym odpowiednikiem statystyki (3.53) z $\theta_0 = 0$ i ma rozkład $t$ Studenta z $n-1$ stopniami swobody. W ten sposób zadanie konstrukcji testów dla porównania wartości średnich par obserwacji sprowadza się do analogicznego zadania dla pojedynczej wartości średniej (mianowicie wartości średniej różnic $D_i$ przy nieznajomości ich odchylenia standardowego).

**Przykład 3.12.** Jednym z testów, którymi rozpoczęto analizę nowego leku na nadciśnienie tętnicze było zaaplikowanie go próbie 22 chorych pacjentów, u których ciśnienie skurczowe było bliskie wartości 144 mmHg (różnice ciśnienia u pacjentów z próby sięgały około 1 mmHg). Ponieważ górna granica normy tego ciśnienia wynosi 140, chciano sprawdzić, czy zastosowanie określonej terapii badanym lekiem daje obniżenie ciśnienia o około 5 mmHg. Takie postępowanie testowe wynika ze sposobu prowadzenia terapii w leczeniu nadciśnienia – przy zadanej wartości ciśnienia, ustalona dawka leku powinna spowodować jego obniżenie mniej więcej do poziomu górnej granicy normy.

Każdemu pacjentowi zmierzono ciśnienie skurczowe przed rozpoczęciem terapii oraz po jej zakończeniu. W ten sposób dla $i$-tego pacjenta, $i = 1, 2, \ldots, 22$, dysponowano parą wyników $(x_i, y_i)$ przed terapią i po niej. Celem było poddanie testowi hipotezy

$$H_0\colon\ \mu_D = 5$$

przy hipotezie alternatywnej

$$H_1\colon\ \mu_D \neq 5.$$

Interesująca nas tu hipoteza zerowa jest modyfikacją hipotezy (3.61). Jasne jest przy tym, jak powinna wyglądać odpowiednia modyfikacja

statystyki (3.62). Mianowicie, zamiast zastosować statystykę (3.62), należy oprzeć się na statystyce

$$T = \frac{\bar{D} - d_0}{S_D/\sqrt{n}},$$

gdzie, w naszym przypadku, $d_0 = 5$ i $n = 22$. Dla próby 22 pacjentów otrzymano $\bar{d} = 5{,}3$ oraz $s_D = 0{,}4$. W rezultacie statystyka $T$ przyjęła wartość $t = 3{,}518$, co dało $p$–wartość (dla rozkładu $t$ Studenta z 21 stopniami swobody i przy dwustronnej hipotezie alternatywnej) 0,002. Hipotezę zerową należało zatem zdecydowanie odrzucić – innymi słowy, terapia okazała się nie spełniać nałożonych wymagań.

Czytelnikowi pozostawiamy zauważenie, że również w problemie z dwiema niezależnymi próbami losowymi łatwo można równość średnich zastąpić hipotezą zerową $\mu_1 - \mu_2 = d_0$ (i odpowiednio dla hipotez alternatywnych: $\mu_1 - \mu_2 > d_0$ lub $\mu_1 - \mu_2 < d_0$ lub $\mu_1 - \mu_2 \neq d_0$), gdzie $d_0$ jest ustaloną liczbą.

**Testy dla wariancji w rodzinie rozkładów normalnych**

Niech będzie dana próba losowa o liczności $n$ z rozkładu normalnego o nieznanej wariancji $\sigma^2$. To co wiemy o przedziałowym estymatorze dla wariancji, sugeruje użycie statystyki

$$\mathcal{X}^2 = \frac{(n-1)S^2}{\sigma_0^2} \tag{3.63}$$

do testowania hipotezy zerowej

$$H_0\colon\ \sigma^2 = \sigma_0^2.$$

Rzeczywiście, wystarczy w tym celu przypomnieć, że przy zachodzeniu $H_0$ statystyka $\mathcal{X}^2$ ma rozkład $\chi^2$ $n-1$ stopniami swobody. Łatwo obliczyć, że użycie statystyki (3.63) do testowania na poziomie istotności $\alpha$ podanej hipotezy zerowej przy hipotezie alternatywnej

$$H_1\colon\ \sigma^2 \neq \sigma_0$$

prowadzi do zbioru krytycznego

$$C = \{x^2\colon\ x^2 \leq \chi^2_{\alpha/2,n-1}\ \text{ lub }\ x^2 \geq \chi^2_{1-\alpha/2,n-1}\},$$

gdzie

$$x^2 = \frac{(n-1)s^2}{\sigma_0^2},$$

$s^2$ jest zaobserwowaną wartością wariancji w próbie oraz $\chi^2_{\gamma,n-1}$ jest kwantylem rzędu $\gamma$ rozkładu $\chi^2$ z $n-1$ stopniami swobody.

Jeżeli dwustronną hipotezę alternatywną zastąpić hipotezą jednostronną

$$H_1\colon\ \sigma^2 < \sigma_0^2,$$

to wykażemy, że zbiór krytyczny przyjmuje postać

$$C = \{x^2\colon x^2 \leq \chi^2_{\alpha,n-1}\}.$$

Możemy mianowicie napisać

$$\mathcal{X}^2 = \frac{(n-1)S^2}{\sigma_0^2} = \left(\frac{(n-1)S^2}{\sigma^2}\right)\left(\frac{\sigma^2}{\sigma_0^2}\right).$$

Dalej, jeśli $\sigma^2$ jest prawdziwą wariancją rozkładu, to pierwszy czynnik po prawej stronie ostatniej równości ma rozkład $\chi^2$. Przy prawdziwości podanej hipotezy alternatywnej drugi czynnik prawej strony równości implikuje, że statystyka $\mathcal{X}^2$ musi mieć tendencję do przyjmowania wartości mniejszych niż wynikające z rozkładu $\chi^2$, co uzasadnia przyjętą postać zbioru krytycznego.

Jeżeli przyjąć jednostronną hipotezę alternatywną postaci

$$H_1\colon\ \sigma^2 > \sigma_0^2,$$

to zbiór krytyczny przyjmuje postać

$$C = \{x^2\colon\ x^2 \geq \chi^2_{1-\alpha,n-1}\}.$$

Uzasadnienie tego faktu pozostawiamy Czytelnikowi.

Niekiedy interesuje nas porównanie dwóch wariancji. Niech zatem dane będą dwie niezależne próby losowe z populacji normalnych, o licznościach $n_1$ i $n_2$. Niech $S_1^2$ i $S_2^2$ oznaczają, odpowiednio, wariancję w próbie pierwszej i drugiej. Niech wreszcie hipoteza zerowa orzeka równość wariancji obydwu populacji

$$H_0\colon\ \sigma_1^2 = \sigma_2^2.$$

Wiemy, że przy tej hipotezie statystyka (por. p. 3.3.2)

$$F = \frac{S_1^2}{S_2^2} \tag{3.64}$$

ma rozkład $F$ Snedecora z $n_1 - 1$ i $n_2 - 1$ stopniami swobody. Zatem, testując na poziomie istotności $\alpha$ hipotezę zerową przy dwustronnej hipotezie alternatywnej

$$H_1\colon\ \sigma_1^2 \neq \sigma_2^2,$$

za zbiór krytyczny przyjmujemy

$$C = \{f\colon\ f \le f_{\alpha/2,\nu_1,\nu_2}\ \text{ lub }\ f \ge f_{1-\alpha/2,\nu_1,\nu_2}\},$$

gdzie

$$f = \frac{s_1^2}{s_2^2},$$

$s_i^2$ jest zaobserwowaną wartością wariancji w próbie $i$-tej, $i = 1, 2$, oraz $f_{\gamma,\nu_1,\nu_2}$ jest kwantylem rzędu $\gamma$ rozkładu Snedecora z $\nu_1 = n_1 - 1$ i $\nu_2 = n_2 - 1$ stopniami swobody.

Chcąc określić zbiór krytyczny, gdy dwustronną hipotezę alternatywną zastąpić hipotezą jednostronną

$$H_1\colon\ \sigma_1^2 < \sigma_2^2,$$

zauważmy najpierw, że jeżeli $\sigma_1^2 < \sigma_2^2$, to

$$F < \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}. \tag{3.65}$$

Następnie, jak stwierdziliśmy w p. 3.3.2, jeżeli prawdziwe wariancje w populacjach wynoszą odpowiednio $\sigma_1^2$ oraz $\sigma_2^2$, to statystyka występująca z prawej strony nierówności (3.65) ma rozkład Snedecora z $\nu_1 = n_1 - 1$ i $\nu_2 = n_2 - 1$ stopniami swobody. A zatem, pod warunkiem zachodzenia hipotezy alternatywnej, statystyka $F$, dana wzorem (3.64) i występująca po lewej stronie nierówności (3.65), będzie miała tendencję do przyjmowania wartości mniejszych niż wynikające z rozkładu Snedecora z $\nu_1 = n_1 - 1$ i $\nu_2 = n_2 - 1$ stopniami swobody. Przy hipotezie alternatywnej $H_1\colon \sigma_1^2 < \sigma_2^2$, uzasadnia to przyjęcie następującego zbioru krytycznego

$$C = \{f\colon\ f < f_{\alpha,\nu_1,\nu_2}\},$$

gdzie $\nu_1 = n_1 - 1$ i $\nu_2 = n_2 - 1$. Czytelnikowi pozostawiamy skonstruowanie testu przy hipotezie alternatywnej $H_1\colon\ \sigma_1^2 > \sigma_2^2$.

Omówiony test dla ilorazu wariancji nazywamy testem $F$.

**Uwagi o testowaniu wartości średniej i wariancji w rodzinach rozkładów innych niż normalny oraz o nadużywaniu testów**

Wszystko co w p. 3.3.3 powiedzieliśmy o odporności przedziałów ufności dla wartości średnich i dla wariancji pozostaje oczywiście prawdziwe w odniesieniu do testów odpowiadających tym przedziałom. Jedyna różnica polega na zastąpieniu poziomu ufności poziomem istotności. Jeśli zatem dany przedział w przybliżeniu zachowuje założony poziom ufności mimo odstępstwa próby od normalności rozkładu, to takie samo stwierdzenie dotyczy

poziomu istotności analogicznego testu. Krótko reasumując wcześniejszy komentarz, testy dla wartości średnich są odporne na odstępstwa od normalności rozkładu, ale nie na wartości odstające. Omówione testy dla wariancji są bezużyteczne przy istnieniu wyraźnych odstępstw od normalności rozkładu.

Brak odporności testu $F$ na odstępstwa od normalności skłonił badaczy do poszukiwania odpornego testu równości wariancji. Najpopularniejszym takim testem, często sprawdzającym się w praktyce, gdy zawodzi test $F$, jest test Levene'a i jego odmiany. Ściśle biorąc nie jest to test porównujący wariancje dwóch populacji. Jest to pewna wersja testu $t$ danego statystyką (3.60), gdzie zamiast oryginalnych prób bierze się pod uwagę wartości bezwzględne różnic między oryginalnymi elementami próby a ich średnią (lub medianą) w próbie. Idea testu sprowadza się do zastąpienia porównania wariancji porównaniem odchyleń przeciętnych.

W kontekście przykł. 3.9 zwróciliśmy już uwagę (por. zad. 3.14), że odrzucenie hipotezy zerowej nie powinno prowadzić do podejmowania pochopnych decyzji. Jeżeli hipoteza zerowa jest fałszywa, ale prawdziwa wartość interesującego nas parametru jest jej tak bliska, że różnica między obydwiema wartościami jest **praktycznie** nieistotna, otrzymany wynik testu nie powinien mieć żadnych konsekwencji praktycznych. Innymi słowy, otrzymany wynik może być statystycznie istotny, ale nieistotny z praktycznego punktu widzenia. Z taką sytuacją spotykamy się wtedy, gdy zaprojektowaliśmy test zbyt czuły w stosunku do naszych potrzeb. Przez nadmierną czułość testu rozumiemy jego zbyt dużą moc przy bardzo małej (praktycznie nieistotnej) odległości prawdziwej wartości parametru od wartości odpowiadającej hipotezie zerowej. Wracając raz jeszcze do przykł. 3.9: otrzymana średnia w próbie jest wyraźnie mniejsza od wartości nominalnej i, co równie ważne, błąd standardowy tej średniej jest bardzo mały. W tym zatem przypadku, otrzymany wynik jest istotny zarówno statystycznie, jak i praktycznie – mamy rzeczywiście prawo wierzyć, że produkuje się mleko o niedopuszczalnie zaniżonej zawartości tłuszczu.

Bezkrytyczne utożsamianie istotności statystycznej z praktyczną jest błędem, tak jak błędem jest stosowanie testu wtedy, gdy niespełnienie założeń o próbie losowej odbiera wynikom ich wiarogodność. Test powinien być starannie zaprojektowany. Warto też zalecić graficzną analizę rozkładu testowanej próby oraz porównanie estymatora testowanej wielkości z jego błędem standardowym.

**Testy dla proporcji**

Ograniczymy się do przypadku dostatecznej liczności próby, by móc skorzystać z przybliżenia normalnego statystyki

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}, \tag{3.66}$$

gdzie $p$ jest prawdziwą wartością prawdopodobieństwa sukcesu, $n$ jest licznością próby, na podstawie której obliczamy częstość $\hat{p}$ (por. p. 2.4.3 i zwłaszcza tw. 2.15 oraz komentarz po tw. 2.16).

Przy założeniu prawdziwości hipotezy zerowej

$$H_0\colon\ p = p_0,$$

statystyka (3.66) z $p_0$ zamiast $p$,

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}, \tag{3.67}$$

ma w przybliżeniu standardowy rozkład normalny i zadanie testowania hipotezy zerowej przy hipotezie alternatywnej

$$H_1\colon\ p > p_0$$

lub

$$H_1\colon\ p < p_0$$

lub

$$H_1\colon\ p \neq p_0$$

sprowadza się do odpowiedniego zadania testowania hipotez o wartości średniej rozkładu normalnego przy znanym odchyleniu standardowym (statystyka (3.67) jest oczywistym odpowiednikiem statystyki (3.43)). Zauważmy, że przy testach dla proporcji nie ma problemu oceny wariancji częstości, gdyż zależy ona tylko od prawdziwej wartości proporcji, a ta jest znana przy założonej hipotezie zerowej.

**Przykład 3.6 cd.** Pewne ugrupowanie polityczne było przekonane, że poparcie Polaków dla wejścia ich kraju do Unii Europejskiej nigdy nie przekroczy 53%. Po przeprowadzeniu w czerwcu 2000 r. ankiety wśród 1000 dorosłych Polaków, z których 57% poparło starania Polski do UE, możemy przetestować hipotezę zerową

$$H_0\colon\ p = 0{,}53$$

przy hipotezie alternatywnej

$$H_1\colon\ p > 0{,}53;$$

właściwie interesuje nas poddanie w wątpliwość prawdziwości założonej hipotezy zerowej $H_0\colon p \leq 0{,}53$, ale z dyskusji w pierwszej części p. 3.4.1 (por. (3.46)) wiemy już, że tę ostatnią hipotezę możemy zastąpić podaną hipotezą prostą. Bez trudu otrzymujemy, że

$$Z = \frac{0{,}57 - 0{,}53}{\sqrt{\frac{0{,}53(1-0{,}53)}{1000}}} = 2{,}534,$$

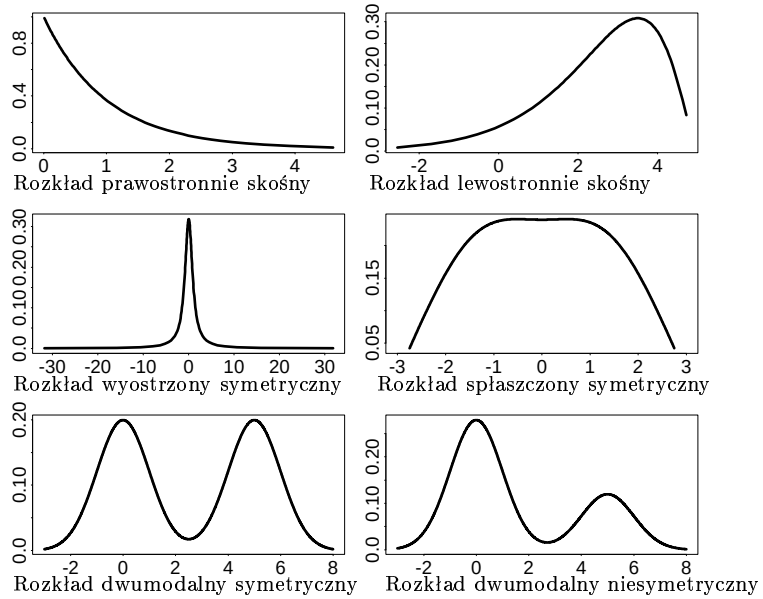
co daje $p$–wartość przeprowadzonego testu równą 0,006. W tej sytuacji z przekonaniem odrzucamy hipotezę wspomnianego ugrupowania politycznego.

### 3.4.2. Testowanie zgodności

**Testowanie normalności rozkładu: wykres kwantylowy**

Do wielu zagadnień omówionych dotychczas w tej książce konieczne jest założenie, aby rozkład, z którego pochodzi prosta próba losowa, był przynajmniej w przybliżeniu normalny. Także w rozdz. 4 pożądaną własnością analizowanych tam błędów losowych jest ich normalność. Bardzo istotną rolę w całości wnioskowania statystycznego odgrywa zatem testowanie hipotezy o normalności rozkładu. Jest to przy tym hipoteza złożona – nie pytamy o to, jakie są parametry rozkładu, chcemy natomiast wiedzieć czy populacja, którą badamy, ma rozkład normalny (o dowolnych parametrach). Ogólnie, testy, których zadaniem jest orzeczenie, czy dana próba pochodzi czy nie z danego rozkładu prawdopodobieństwa nazywamy testami zgodności. Taka hipoteza zerowa może być, jak to właśnie zasugerowaliśmy, złożona, ale może też być prosta – w tym drugim przypadku hipoteza zerowa dotyczy pojedynczego, konkretnego rozkładu, np. rozkładu normalnego o zadanej wartości średniej i zadanym odchyleniu standardowym.

W naszej książce nieraz już zwracaliśmy uwagę, jak dobrym środkiem analizy danych są metody graficzne. Do takich nader użytecznych narzędzi należy także wykres kwantylowy, będący wyjątkowo skuteczną, graficzną metodą testowania normalności rozkładu. Co więcej, gdy ta hipoteza zerowa powinna być odrzucona, wykres kwantylowy często wskazuje na typ odstępstwa badanego rozkładu od normalności. Jeżeli mianowicie próba losowa jest dostatecznie liczna, spoglądając na wykres kwantylowy, potrafimy nie tylko orzec, że hipoteza o normalności powinna być odrzucona, ale także,




Rys. 3.8. Wybrane typy rozkładów

czy prawdziwy rozkład, z którego pochodzi próba jest skośny (prawo- czy lewostronnie), symetryczny i wyostrzony, symetryczny i spłaszczony, czy też dwumodalny (por. rys. 3.8).

Wykres kwantylowy nie jest testem w ścisłym sensie, nie można bowiem obliczyć $p$–wartości czy też powziąć decyzji na zadanym poziomie istotności. Wykres dostarcza jedynie wizualnej informacji o badanej próbie losowej. Jednak taka dobrze zinterpretowana informacja jakościowa okazuje się mieć bardzo duże znaczenie praktyczne.

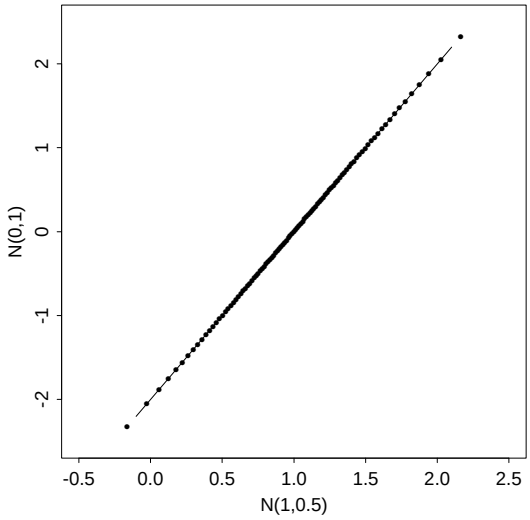
Niech $n$ będzie ustaloną liczbą naturalną. Niech $z_{i/n}$ oznacza kwantyl rzędu $i/n$, $i = 1, 2, \ldots, n-1$, standardowego rozkładu normalnego. Zauważmy, że jeżeli $x_{i/n}$, $i = 1, 2, \ldots, n-1$, oznaczają kwantyle tego samego rzędu rozkładu normalnego o wartości średniej $\mu$ i odchyleniu standardowym $\sigma$, to prawdziwa jest zależność liniowa

$$x_{i/n} = \mu + \sigma z_{i/n}$$

i równoważna jej zależność

$$z_{i/n} = \frac{x_{i/n} - \mu}{\sigma}.$$

Innymi słowy, punkty $(x_{i/n}, z_{i/n})$ leżą na prostej o równaniu $y = \sigma^{-1}(x - \mu)$; por. rys. 3.9, gdzie $n = 100$. Ta prosta obserwacja stanowi podstawę wykresu kwantylowego.




Rys. 3.9. Wykres punktów $(x_{i/n}, z_{i/n})$, $z_{i/n} = \frac{1}{0,5}(x_{i/n} - 1)$

Niech mianowicie $x_1, x_2, \ldots, x_n$ będą zaobserwowanymi elementami próby losowej o liczności $n$. Uporządkujmy elementy próby od najmniejszego do największego, przy czym niemalejąco uporządkowane elementy próby oznaczmy w następujący sposób:

$$x_{1:n}, x_{2:n}, \ldots, x_{n:n}$$

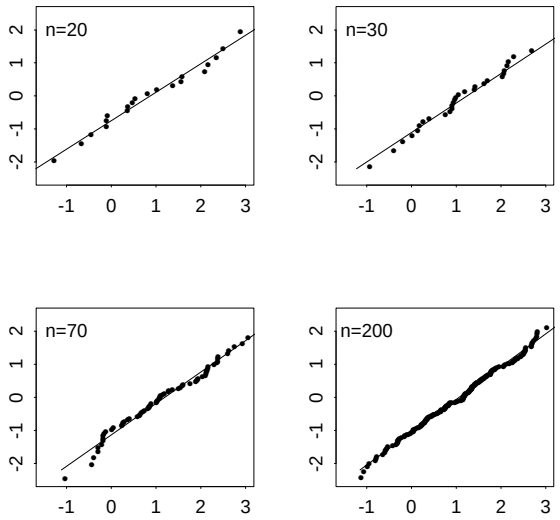
(w rozdz. 1 pisaliśmy $x_{(i)}$ zamiast $x_{i:n}$, $i = 1, 2, \ldots, n$; tym razem chcemy wyraźnie zaznaczyć, że kolejne co do wielkości elementy pochodzą z próby o liczności $n$). Element próby $x_{i:n}$ nazywamy $i$-tą **statystyką pozycyjną** próby (czyli pierwsza statystyka pozycyjna jest równa najmniejszemu elementowi próby, natomiast $n$-ta statystyka pozycyjna oznacza największy element próby).

Zauważmy, że $i$-ta stystyka pozycyjna, $x_{i:n}$, $i = 1, 2, \ldots, n-1$, jest pewnym przybliżeniem kwantyla rzędu $i/n$ rozkładu, z którego pochodzi badana próba losowa. Rzeczywiście, przy ustalonym indeksie $i$, w przedziale $(-\infty, x_{i:n}]$ leży $100(i/n)\%$ elementów próby losowej, czyli przedział ten jest przybliżeniem przedziału $(-\infty, x_{i/n}]$, dla którego zachodzi związek

$$\int_{-\infty}^{x_{i/n}} f(x)dx = \frac{i}{n},$$

gdzie $f(\cdot)$ oznacza gęstość rozkładu elementów próby losowej.

Problemem statystyka jest oczywiście to, że nie wie, z jakiego rozkładu pochodzi badana próba losowa i przeto nie potrafi wyznaczyć kwantyli tego

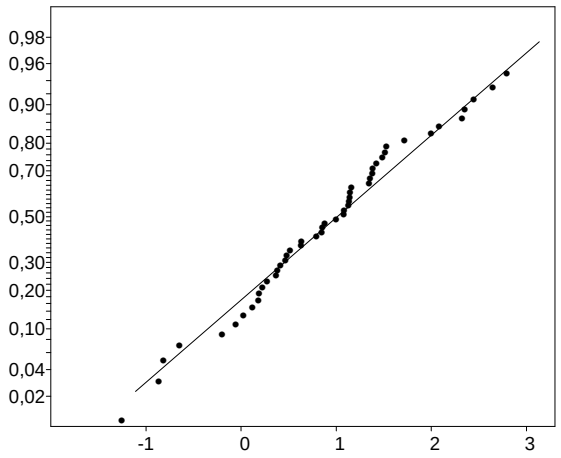



Rys. 3.10. Wykresy kwantylowe dla 4 prób z rozkładu normalnego $N(1,1)$

rozkładu. Ale, jak właśnie zauważyliśmy, pewnymi przybliżeniami tych kwantyli są statystyki pozycyjne. A zatem, jeżeli próba losowa pochodzi z rozkładu normalnego, to punkty $(x_{i:n}, z_{i/n})$, gdzie $z_{i/n}$ są znanymi kwantylami rozkładu $N(0,1)$, powinny na płaszczyźnie „układać się" wokół prostej określonej przez (nieznane) punkty $(x_{i/n}, z_{i/n})$. Na rysunku 3.10 są podane przykłady takiej sytuacji: próby o licznościach 20, 30, 70 i 200 pochodzą z rozkładu $N(1,1)$.

Dokładnie mówiąc, statystycy są przebiegli i dążąc do możliwie najlepszego przybliżenia przez nanoszone punkty pewnej linii prostej zauważyli, że zamiast kwantyli $z_{i/n}$ powinno się brać kwantyle rozkładu $N(0,1)$ nieco innego rzędu. Nie będziemy tu wnikać w racje, jakie stoją za takim postępowaniem, wspomnimy jedynie o dwóch sprawach. Po pierwsze, tak zmodyfikowane wykresy nadal opierają się na tych podstawach intuicyjnych, jakie tu przedstawiliśmy, i nimi się w rzeczywistości posługujemy. I po drugie, najczęściej dziś stosowanym kwantylem zamiast kwantyla $z_{i/n}$ jest kwantyl rozkładu $N(0,1)$ rzędu $(i-3/8)/(n+1/4)$; często stosuje się też kwantyl rzędu $(i-1/2)/n$ lub $i/(n+1)$.

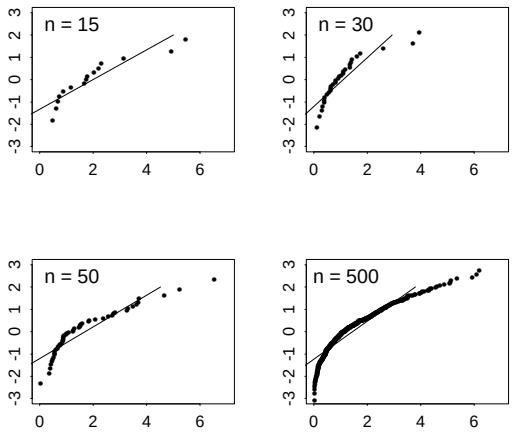
Tak powstałe wykresy nazywamy **wykresami kwantylowymi**. Często zamiast kwantyli rozkładu $N(0,1)$ na osi podaje się odpowiadające im rzędy (czyli zamiast wartości $z_{i/n}$ na osi podaje się wartość $i/n$ lub wartość $100(i/n)\%$). Tak jest na rys. 3.11 gdzie, jak poprzednio, wielkości związane z rozkładem $N(0,1)$ są umieszczone na osi pionowej. To, na której osi którą wielkość umieścić, jest oczywiście kwestią umowy i równie dobrze można by




Rys. 3.11. Wykres kwantylowy; na osi pionowej są zaznaczone rzędy kwantyli ($n = 50$)

na pionowej osi umieścić statystyki pozycyjne $x_{i:n}$ (w różnych pakietach statystycznych przypisanie zmiennych osiom jest różne lub uzależnione od woli użytkownika).

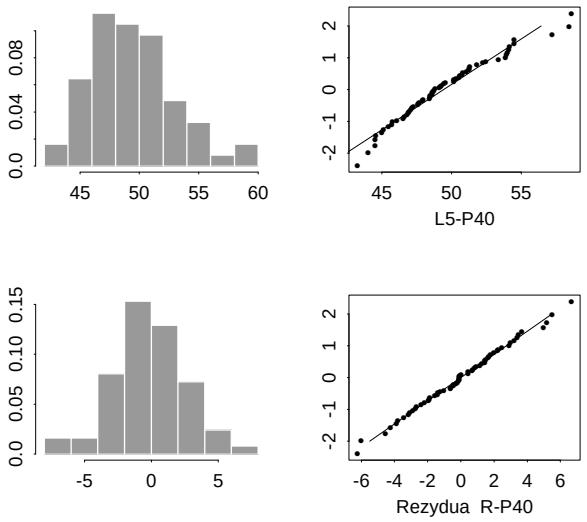
Prawidłowa interpretacja wykresów kwantylowych wymaga doświadczenia. Próba losowa nie może być przy tym zbyt mało liczna. Jeżeli rozkład jest normalny lub wyraźnie odbiega od normalnego, do prawidłowego odczytania wykresu może wystarczyć parenaście obserwacji. Aby jednak tak było, odchylenie standardowe rozkładu nie może być zbyt duże. Rzecz w tym, że przy dostatecznie małym odchyleniu standardowym rozkładu normalnego, już parenaście obserwacji może utworzyć na wykresie wyraźną konfigurację punktów, przypominającą linię prostą. Podobnie, przy dostatecznie małym odchyleniu standardowym rozkładu istotnie skośnego, punkty mogą utworzyć wyraźną konfigurację przypominającą łuk. Na rysunku 3.12 są pokazane przykładowe wykresy kwantylowe dla próby o liczności 15, 30, 50 oraz 500 z rozkładu prawostronnie skośnego. Interpretując wykresy kwantylowe trzeba pamiętać, że statystyki pozycyjne są tym gorszymi estymatorami kwantyli im bardziej skrajne (bliskie 0 lub 1) są rzędy szacowanych kwantyli – wraz ze zbliżaniem się rzędu kwantyla do 0 lub 1 rośnie odchylenie standardowe jego estymatora (proponujemy Czytelnikowi poszukać intuicyjnych racji wyjaśniających ten fenomen).




Rys. 3.12. Wykresy kwantylowe dla prób o licznościach 15, 30, 50 oraz 500 z rozkładu prawostronnie skośnego

**Przykład 3.5 cd.** Na rysunku 3.13 są pokazane wykresy kwantylowe oraz histogramy dla latencji L5-P40 i rezyduów R-P40. Na ich podstawie można stwierdzić, że rozkład latencji L5-P40 nie jest rozkładem normalnym, charakteryzuje się bowiem prawostronną skośnością. W przypadku rezyduów R-P40 można z dużą dozą pewności orzec normalność rozkładu tej zmiennej losowej.

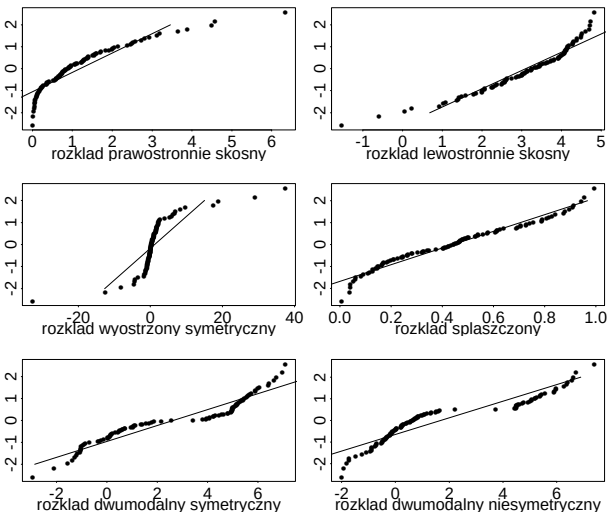
Jeśli dokładnie porówna się kształty gęstości spłaszczonych, wyostrzonych oraz skośnych, to można przewidzieć kształt odpowiadających im wykresów kwantylowych (trzeba jednak pamiętać, że wyraźne ujawnienie się tych kształtów wymaga, by próba losowa była dostatecznie liczna). Przykładowe wykresy kwantylowe, dotyczące szczególnie wyrazistych sytuacji, są pokazane na rys. 3.14. W omówieniu tym ograniczymy się do przykładu gęstości spłaszczonej i krótko wspomnimy o gęstości wyostrzonej. Porównajmy najpierw odległość między dwoma, bardzo bliskimi mediany kwantylami rozkładu $N(0,1)$ z odległością między kwantylami tego samego rzędu gęstości spłaszczonej. Taką odległość otrzymujemy, obliczając odpowiednie pole pod właściwą gęstością. W sytuacji z rys. 3.15 nietrudno zauważyć, że w otoczeniu wspólnej mediany obydwu gęstości odległość między kwantylami rozkładu $N(0,1)$ powinna okazać się mniejsza niż odpowiadająca jej odległość między kwantylami rozkładu spłaszczonego. Wynika stąd, że wykres kwantylowy (z dystrybuantą rozkładu $N(0,1)$ na osi pionowej) będzie spłaszczony w części środkowej. Wraz z oddalaniem się (w lewo lub w prawo) od wspól-



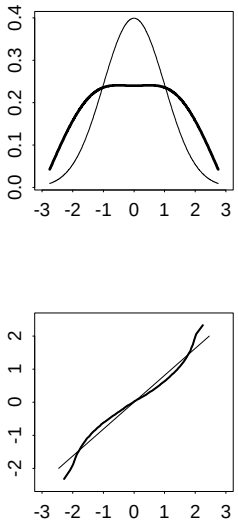

Rys. 3.13. Histogramy i wykresy kwantylowe dla latencji L5-P40 i rezyduów R-P40 (przykł. 3.5 cd.)

nego środka dwóch rozkładów przyrost pola pod gęstością normalną staje się coraz wolniejszy aż dochodzimy do obszaru, w którym odległość między kwantylami gęstości rozkładu spłaszczonego staje się mniejsza niż odległość między kwantylami tego samego rzędu rozkładu $N(0,1)$. Dalsze oddalanie się od środka rozkładów sprawia, że stosunek obydwu odległości rośnie, prowadząc do widocznego zwiększenia nachylenia wykresu kwantylowego. Taki właśnie obraz jest przedstawiony na rys. 3.15 (jeżeli liczność próby nie jest duża, wykres kwantylowy zwykle traci swój regularny charakter zwłaszcza w ogonach badanej gęstości – wynika to stąd, iż w rzeczywistości nie dysponujemy kwantylami badanego rozkładu, a jedynie ich estymatorami o jakości tym gorszej im dalej jesteśmy od środka rozkładu).

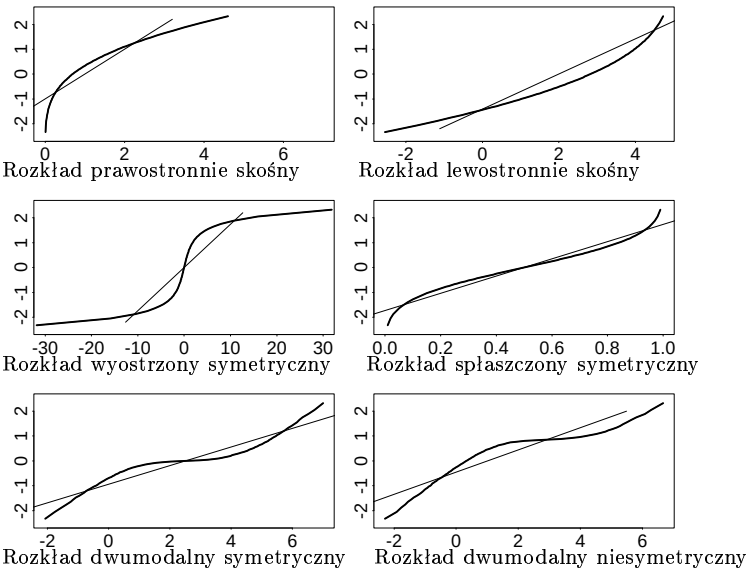
W przypadku rozkładu wyostrzonego sytuacja jest w pewnym sensie odwrotna do właśnie opisanej i stąd taki, a nie inny „idealny" wykres kwantylowy na rys. 3.16. Zachęcamy Czytelnika do uzasadnienia podanego na rys. 3.16 kształtu „idealnych" wykresów kwantylowych dla rozkładów skośnych i rozkładu dwumodalnego, a także wyjaśnienia dlaczego wykres kwantylowy jest zwykle dobrym narzędziem do wykrywania obserwacji odstających. Dodajmy na koniec, że wykresy kwantylowe nie przypadkiem mają zwykle różne skale na obydwu osiach układu współrzędnych; skala na osi, na której są odkładane statystyki pozycyjne próby, zależy od położenia i rozproszenia rozkładu i powinna być tak dobrana, by dać możliwie najlepszy efekt wizualny.




Rys. 3.14. Wykresy kwantylowe dotyczące szczególnie wyrazistych sytuacji



Rys. 3.15. Wykres gęstości spłaszczonej oraz odpowiadający jej wykres kwantylowy




Rys. 3.16. „Idealne" wykresy kwantylowe dla rozkładów z rys. 3.8

**Testowanie normalności rozkładu: testy kierunkowe**

W ostatnich częściach tego punktu krótko omówimy ścisłe testy normalności rozkładu. Pomijać będziemy wszelkie szczegóły, z jednej bowiem strony wiemy już dobrze, jak konstruuje się testy i z drugiej najczęściej przeprowadzamy je, korzystając z komputerowych pakietów statystycznych. Krótkie przynajmniej ich omówienie jest zarazem niezbędne, jeśli chcemy z owych pakietów korzystać w sposób odpowiedzialny.

Wykres kwantylowy nie jest oczywiście testem w sensie ścisłym, ponieważ nie daje liczbowo wyrażonego uzasadnienia podejmowanej decyzji. Inna sprawa, że to iż wykres kwantylowy nie jest takim testem, opiera się natomiast na inteligentnie pomyślanej wizualizacji danych, nie jest jego żadną wadą.

Test w sensie ścisłym wymaga podania hipotezy alternatywnej (oczywiście hipoteza zerowa brzmi: rozkład, z którego pochodzi badana próba losowa jest normalny). Rozróżnia się tu dwie ogólne możliwości. Jeżeli hipoteza alternatywna specyfikuje typ odstępstwa od normalności rozkładu, test nazywamy **kierunkowym**. Jeżeli natomiast w hipotezie alternatywnej nie jest określona postać odstępstwa od rozkładu normalnego, test nazywamy **uniwersalnym**.

W punkcie tym rozważymy dwa testy kierunkowe: test skośności oraz wyostrzenia. Żadnemu z tych pojęć nie nadawaliśmy dotąd ścisłego znaczenia. Rozumieliśmy je zgodnie z intuicją. Obecnie, chcąc formalnie zapisać alternatywę kierunkową, musimy tym pojęciom nadać precyzyjne znaczenie.

**Definicja 3.4.** *Niech*

$$\beta_1 = \frac{\mu_3}{\sigma^3}$$

*oraz*

$$\beta_2 = \frac{\mu_4}{\sigma^4},$$

*gdzie $\mu_i$, $i = 3, 4$, jest momentem centralnym $i$-tego rzędu oraz $\sigma$ jest odchyleniem standardowym zmiennej losowej o zadanym rozkładzie prawdopodobieństwa. Wielkość $\beta_1$ nazywamy* ***współczynnikiem skośności*** *(inaczej asymetrii) zadanego rozkładu prawdopodobieństwa. Wielkość $\beta_2$ nazywamy* ***kurtozą*** *tego rozkładu prawdopodobieństwa.*

Współczynnik skośności $\beta_1$ może być większy od 0, równy 0 lub mniejszy od 0. W pierwszym przypadku mówimy, że rozkład jest prawostronnie (lub dodatnio) skośny (asymetryczny), w drugim, że jest symetryczny i w trzecim, że jest lewostronnie (lub ujemnie) skośny (asymetryczny).

Można wykazać, że w przypadku rozkładu normalnego $\beta_2 = 3$. Wielkość

$$\beta_2 - 3$$

nazywamy **współczynnikiem wyostrzenia** rozkładu. Jeżeli współczynnik wyostrzenia jest dodatni, rozkład nazywamy **wyostrzonym**. Jeżeli jest ujemny, rozkład nazywamy **spłaszczonym**.

Obecność odchylenia standardowego w odpowiedniej potędze w mianownikach definicji współczynnika skośności oraz kurtozy czyni z tych współczynników wielkości niemianowane. W rezultacie o ich wartości decyduje tylko kształt rozkładu, a nie jednostki, w jakich wyraża się wartości zmiennej losowej (jeśli np. interesuje nas współczynnik skośności ciągłego rozkładu zmiennej losowej oznaczającej wzrost, to wartość tego współczynnika nie zależy od tego czy jednostką pomiaru są centymetry czy metry). Podane definicje pokrywają się przy tym z intuicją. Na przykład, wystarczy przyjrzeć się definicji momentu centralnego trzeciego rzędu, by zauważyć, że gęstość, którą nazywamy prawostronnie skośną, powinna mieć dodatnią wartość momentu $\mu_3$ (i stąd współczynnika $\beta_1$) ze względu na jej wydłużony prawy ogon. Z kolei wyostrzenie pociąga za sobą posiadanie przez gęstość długich i grubych obydwu ogonów, co powinno dawać dużą wartość momentu $\mu_4$ i przeto dodatnią wartość współczynnika wyostrzenia.

Jeśli znamy już definicje, możemy nadać ścisłą postać kierunkowym testom normalności. W teście skośności hipotezie zerowej nadajemy postać

$$H_0\colon\ \beta_1 = 0 \ \text{ lub równoważnie } \mu_3 = 0,$$

natomiast hipoteza alternatywna przyjmuje jedną z następujących postaci:

$$H_1\colon\ \beta_1 > 0\ \text{ lub równoważnie } \mu_3 > 0$$

albo

$$H_1\colon\ \beta_1 < 0\ \text{ lub równoważnie } \mu_3 < 0.$$

W pierwszym przypadku testujemy zatem symetrię rozkładu przy alternatywie jego prawostronnej skośności. W drugim hipotezą alternatywną jest lewostronna skośność. Zauważmy, że ściśle biorąc nie mamy w tym przypadku do czynienia z testowaniem normalności rozkładu, lecz jego symetrii. Innymi słowy, mówiąc w tej sytuacji, iż testujemy normalność, wychodzimy z założenia, że jeśli rozkład jest symetryczny, to jest to rozkład normalny. Przesłanki, które doprowadziły nas do tego założenia nie są przedmiotem testowania.

W teście wyostrzenia hipoteza zerowa ma postać

$$H_0\colon \beta_2 = 3.$$

Hipoteza alternatywna przyjmuje jedną z następujących postaci:

$$H_1\colon \beta_2 > 3$$

albo

$$H_1\colon \beta_2 < 3.$$

Tym razem testujemy hipotezę, że kurtoza jest równa 3, przy alternatywie wyostrzenia lub – w drugim przypadku – spłaszczenia tego rozkładu. Jak poprzednio, nie jest to test normalności w ścisłym sensie.

Naturalną statystyką testową w przypadku testu skośności oraz dysponowania próbą losową $X_1, X_2, \ldots, X_n$ jest statystyka

$$B_1 = \frac{M_3}{M_2^{3/2}}, \tag{3.68}$$

gdzie

$$M_j = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^j, \tag{3.69}$$

$j = 2, 3$, jest momentem centralnym odpowiedniego rzędu w próbie. Oczywiście, we wzorze (3.68) można by użyć wariancji $S^2$ zamiast momentu $M_2$.

W przypadku testu wyostrzenia naturalną statystyką testową jest

$$B_2 = \frac{M_4}{M_2^2}, \tag{3.70}$$

gdzie moment $M_4$ jest dany wzorem (3.67) z $j = 4$. Podobnie jak poprzednio, we wzorze (3.70) można by użyć wariancji $S^2$ zamiast momentu $M_2$.

Sposób przeprowadzenia każdego z omówionych testów kierunkowych nie powinien już budzić żadnych wątpliwości i dlatego kwestię tę pomijamy. Oczywiście, niezbędne jest dysponowanie tablicą rozkładu prawdopodobieństwa statystyki (3.68) – lub odpowiednio (3.70) – przy założeniu normalności rozkładu. Tablice takie można znaleźć np. we wspomnianych już *Tablicach statystycznych* R. i W. Zielińskich oraz w normie statystycznej PN-ISO 5479 (w normie są podane tablice rozkładu statystyki $B_1$ oraz statystyki $B_2$, natomiast w pierwszej z zacytowanych pozycji są podane tablice rozkładu statystyki $M_3/S^3$ oraz $M_4/S^4$). Także niektóre komputerowe pakiety statystyczne zawierają takie tablice.

Zakres zastosowań testów kierunkowych jest ograniczony. Ich użycie można postulować jedynie wtedy, gdy znany jest typ ewentualnego odstępstwa od normalności rozkładu.

**Testowanie normalności rozkładu: Uniwersalne testy Shapiro–Wilka oraz Eppsa–Pulleya**

Szczególnie ciekawe i ważne są testy uniwersalne, nie zakładające znajomości rodzaju ewentualnego odstępstwa od normalności.

Skuteczność posługiwania się wykresami kwantylowymi nasuwa myśl wyznaczenia linii prostej, która byłaby w jakimś sensie możliwie najlepiej dopasowana do punktów wykresu, i następnie orzeczenia ewentualnej statystycznej istotności odstępstw tych punktów od otrzymanej prostej. Taki właśnie test zaproponowali w roku 1965 Shapiro i Wilk. Do dziś ich test można uznawać za najlepszy test uniwersalny, co nie znaczy, że niektóre inne testy nie mogą go przewyższać dla pewnych typów alternatyw, lub też ogólnie mu dorównywać (jakość testu przy zadanej alternatywie i na zadanym poziomie istotności jest określona przez jego moc przy tej alternatywie). Ogólnie mówiąc, test Shapiro–Wilka charakteryzuje się szczególnie dużą mocą, gdy prawdziwy rozkład jest wyraźnie skośny ($|\beta_1| > 1/2$) lub spłaszczony i w przybliżeniu symetryczny ($\beta_2 < 3$ oraz $|\beta_1| < 1/2$).

Co ciekawe, zarysowana idea testu Shapiro–Wilka i jego kilku znanych odmian sprowadza się do rozwiązania zadania regresji liniowej, która jest przedmiotem następnego rozdzialu. Ponieważ test Shapiro–Wilka można znaleźć w każdym pakiecie statystycznym, rozwiązaniem wspomnianego zadania regresji nie będziemy się dalej zajmować. Tablice dla tego testu zawarte są np. w *Tablicach statystycznych* R. i W. Zielińskich (1990).

Ostatnio pewną popularnością cieszy się opublikowany w roku 1983 test Eppsa i Pulleya. Idea, na której jest oparty jest ciekawa i odwołuje się do pojęcia tzw. funkcji charakterystycznych (w teście analizuje się ważoną całkę

kwadratu różnicy funkcji charakterystycznej z próby i funkcji charakterystycznej rozkładu normalnego). Idea ta wykracza poza zakres tematyczny książki i dlatego pomijamy omówienie testu Eppsa–Pulleya. W normie statystycznej PN-ISO 5479 sugeruje się, by test Eppsa–Pulleya stosować w sytuacjach innych niż te, w których szczególnie dobrymi własnościami charakteryzuje się test Shapiro–Wilka. Cytowana norma nie uwzględnia jednak badań z ostatnich lat, wskazujących na atrakcyjność testu adaptacyjnego, o którym wspominamy w dalszym ciągu tego punktu. Jeżeli zatem zdecydowaliśmy się zamieścić wzmiankę o teście Eppsa–Pulleya, to dlatego, że przewidujemy, iż znajdzie się w pakietach statystycznych przed testami adaptacyjnymi i że jest uniwersalnym testem normalności ogólnie uznawanym za dobry. Tablice dla tego testu można np. znaleźć w normie PN-ISO 5479.

**Uwagi o uniwersalnych testach zgodności dla rozkładu normalnego i innych rozkładów**

Przedstawione dotąd uniwersalne testy zgodności rozkładu zostały skonstruowane z myślą o konkretnej rodzinie rozkładów, mianowicie o rodzinie rozkładów normalnych (mających gęstość postaci (1.1) i różniących się tylko parametrami $\mu$ i $\sigma$).

Nie będziemy zajmować się konstrukcją uniwersalnych testów zgodności, projektowanych specjalnie dla jakiejś ściśle wybranej rodziny gęstości, innej niż rodzina gęstości normalnych. Wspomnimy jedynie, że skoro opisany wcześniej wykres kwantylowy można traktować jako swego rodzaju graficzny i nieformalny, ale tym niemniej bardzo skuteczny test normalności rozkładu, to chciałoby się mieć analogiczne wykresy kwantylowe dla innych, ciekawych z praktycznego punktu widzenia rozkładów. Można by mieć nadzieję, że taki wykres dla innego rozkładu powinien mieć podobnie dobre własności, jak wykres kwantylowy dla rozkładu normalnego. Tak jest w istocie i dlatego pakiety statystyczne z reguły umożliwiają produkowanie wykresów kwantylowych, służących np. do oceny czy próba losowa pochodzi z rozkładu wykładniczego. Innymi najczęściej badanymi rozkładami, dla których można konstruować wykresy kwantylowe, są między innymi popularne rozkłady teorii niezawodności i analizy przeżycia: tzw. rozkład logarytmiczno-normalny, czyli rozkład takiej zmiennej losowej, której logarytm ma rozkład normalny, oraz tzw. rozkład Weibulla, którego gęstość dla pewnych wartości swoich parametrów przypomina gęstość wykładniczą, a dla innych gęstość logarytmiczno-normalną. Jest inną sprawą, że konstrukcja wykresu kwantylowego dla rozkładu innego niż normalny nie jest już tak prosta jak w tym ostatnim przypadku.

Krótko omówimy teraz trzy typy uniwersalnych testów zgodności, z których każdy można stosować do testowania zgodności z zadaną, niekoniecznie normalną rodziną rozkładów. Ogólnie biorąc, w przypadku testowania

normalności, tylko pierwszy z nich można uznać za konkurencyjny wobec już przedstawionych testów uniwersalnych. Testy drugiego i trzeciego typu, ciągle szeroko stosowane do testowania zgodności zarówno z rozkładem normalnym, jak i z rozkładami ciągłymi innymi niż normalny, mają ogólnie biorąc gorsze własności niż test typu pierwszego.

**Adaptacyjny test Neymana.** Zaczniemy od wspomnienia o tzw. gładkim teście adaptacyjnym, służącym do testowania złożonej hipotezy o zgodności z zadaną rodziną rozkładów. Jest to test oparty na znacznie od niego starszym teście Neymana.[4] Omawianie w niniejszym podręczniku testu adaptacyjnego można uznać za przedwczesne, bowiem interesująca nas tu jego wersja została opublikowana zaledwie 3 lata temu (W.C.M. Kallenberg, T. Ledwina (1997): Data-driven smooth tests when the hypothesis is composite. *J. American Statist. Assoc.* **92**, s. 1094–1104). Musi jeszcze upłynąć trochę czasu nim test ten zostanie spopularyzowany oraz znajdzie się w szeroko dostępnych pakietach statystycznych. Także omówienie jego idei wymagałoby wyjścia poza podstawy statystyki i przeto musimy z takiego omówienia zrezygnować. Zarazem test ten zasługuje przynajmniej na wzmiankę, bo choć nie został specjalnie skonstruowany do testowania normalności, to z dotychczasowych badań wynika, że przy różnych typach alternatyw dorównuje lub niewiele ustępuje testowi Shapiro–Wilka, niekiedy zaś charakteryzuje się większą mocą niż ten ostatni. Co więcej, moc gładkiego testu adaptacyjnego przewyższa niekiedy moc testów kierunkowych.

Najogólniej mówiąc, złożona hipoteza alternatywna testu jest przedstawiona jako bardzo bogata rodzina pewnych gęstości. Adaptacyjny charakter testu polega na wyborze z rodziny alternatyw podrodziny dobrze opisującej testowaną próbę. Później, po wybraniu podrodziny alternatyw, stosuje się klasyczny test Neymana (którego omówienie również pomijamy). Zwróćmy uwagę, że ponieważ podrodzina alternatyw jest wybrana na podstawie próby, test ma szansę dobrze rozpoznać typ odstępstwa od hipotezy o normalności rozkładu i, tym samym, mieć dużą moc. Wybór podrodziny jest oparty na stosownie zmodyfikowanej metodzie największej wiarogodności. Jak już zaznaczyliśmy, idea adaptacji testu do próby okazała się bardzo trafna.

**Test Kołmogorowa.** Podobnie jak pozostałe do omówienia uniwersalne testy zgodności, test Kołmogorowa jest testem klasycznym. Testy te nie mają własności adaptacyjnych i przeto nie należy oczekiwać, że ich jakość okaże się generalnie porównywalna z jakością testów uniwersalnych, skonstruowanych z myślą o testowaniu jedynie normalności.

[4] Jerzy Neyman (1894 – 1981) był wybitnym polskim statystykiem, jednym z twórców teorii testowania hipotez.

Ażeby móc przedstawić test Kołmogorowa, musimy najpierw zdefiniować **dystrybuantę empiryczną**.

**Definicja 3.5.** ***Dystrybuantą empiryczną*** *prostej próby losowej $X_1, X_2, \ldots, X_n$, opartą na zaobserwowanych wartościach $x_1, x_2, \ldots, x_n$ tej próby, nazywamy funkcję $F_n(x)$, określoną na prostej, $-\infty < x < \infty$, przyjmującą w punkcie $x$ wartość*

$$F_n(x) = \frac{\text{liczba elementów } x_1, x_2, \ldots, x_n, \text{ które są nie większe od } x}{n}$$

Posługując się statystykami pozycyjnymi $x_{1:n}, x_{2:n}, \ldots, x_{n:n}$, odpowiadającymi zaobserwowanym wartościom $x_1, x_2, \ldots, x_n$, dystrybuancie empirycznej możemy nadać równoważną postać

$$F_n(x) = \begin{cases} 0, & \text{gdy } x < x_{1:n} \\ \frac{k}{n}, & \text{gdy } x_{k:n} \leq x < x_{k+1:n} \\ 1, & \text{gdy } x_{n:n} \leq x, \end{cases}$$

gdzie $k = 1, 2, \ldots, n-1$. Innymi słowy, dystrybuanta $F_n(x)$ jest funkcją schodkową, równą zeru dla wszystkich $x$ mniejszych od najmniejszej zaobserwowanej wartości w próbie i mającą skoki o wartości $1/n$ dla $x$ równych kolejnym statystykom pozycyjnym. Dystrybuanta empiryczna jest próbkowym przybliżeniem prawdziwej dystrybuanty rozkładu, z którego pochodzi próba losowa.

Załóżmy, że dysponujemy próbą losową $X_1, X_2, \ldots, X_n$, pochodzącą z rozkładu o nieznanej dystrybuancie $F$ i że chcemy poddać testowi hipotezę

$$H_0\colon\ F(\cdot) = F_0(\cdot), \tag{3.71}$$

orzekającą, iż dystrybuanta $F$ jest dla wszystkich $x \in (-\infty, \infty)$ równa pewnej ustalonej dystrybuancie $F_0(\cdot)$. Za hipotezę alternatywną uznajemy

$$H_1\colon\ F(\cdot) \neq F_0(\cdot). \tag{3.72}$$

Powyższy problem testowania różni się od ostatnio rozważanych na dwa sposoby. Po pierwsze nie zakładamy, że $F_0(\cdot)$ jest dystrybuantą rozkładu normalnego. I po drugie zakładamy, że jest to dowolna *ustalona* dystrybuanta, a zatem mamy tym razem do czynienia z *prostą* hipotezą zerową – nie testujemy przynależności do rodziny rozkładów, lecz zgodność z jednym konkretnym rozkładem. Niedługo powrócimy do omawiania złożonej hipotezy zerowej.

Naturalną statystyką testową jest w podanej sytuacji *statystyka Kołmogorowa*

$$D_n = \sup_x |F_n(x) - F_0(x)|, \tag{3.73}$$

czyli kres górny (po wszystkich możliwych wartościach argumentu $x$) różnicy między dystrybuantą empiryczną a dystrybuantą przyjmowaną w hipotezie zerowej. Oczywiście, hipoteza zerowa powinna zostać odrzucona, gdy statystyka Kołmogorowa przyjmie zbyt dużą wartość. Należy podkreślić, że przy zachodzeniu hipotezy zerowej rozkład statystyki $D_n$ nie zależy od rozkładu $F_0(\cdot)$. Innymi słowy, rozkład statystyki $D_n$ jest rozkładem odniesienia dla rozważanego problemu. W tablicach statystycznych są podane wartości krytyczne dla zadanych poziomów istotności testu Kołmogorowa. W pakietach komputerowych są podane również $p$–wartości, odpowiadające otrzymanym wartościom statystyki $D_n$.

Powróćmy teraz do problemu testowania złożonej hipotezy o normalności rozkładu. Odpowiednia statystyka testowa ma znowu postać (3.73), $F_n(x)$ oznacza jak poprzednio dystrybuantę empiryczną, obliczoną na podstawie próby, ale tym razem $F_0(x)$ jest dystrybuantą rozkładu normalnego z wartością średnią równą $\bar{x}$ i wariancją równą $s^2$. A zatem, ponieważ prawdziwa wartość średnia oraz wariancja nie są znane i nie wymagamy w hipotezie zerowej, by przyjmowały ustalone wartości, parametry te estymujemy, a otrzymane ich oszacowania wstawiamy do wzoru na gęstość $F_0(x)$.

Takie postępowanie ma doniosłe znaczenie dla właściwego zrozumienia procedury testowej. Jak wiemy, obliczanie wartości krytycznej oraz $p$–wartości testu opiera się na znajomości rozkładu statystyki testowej przy założeniu prawdziwości hipotezy zerowej. Tak jest i tym razem, trzeba jednak pamiętać, że statystyka $D_n$ ma inny rozkład w przypadku, gdy hipoteza zerowa jest prosta (czyli, gdy zakłada się dokładną znajomość rozkładu przy zachodzeniu $H_0$) oraz w przypadku, gdy jest złożona i parametry rozkładu są estymowane. Przy tej samej zaobserwowanej wartości statystyki $D_n$ w obydwu przypadkach otrzymuje się inne wartości krytyczne oraz inne $p$–wartości testu.

Intuicyjnie powinno być jasne, że test Kołmogorowa może się charakteryzować stosunkowo dużą mocą jedynie wtedy, gdy dystrybuanta empiryczna oraz dystrybuanta $F_0(\cdot)$ z estymowanymi parametrami rozkładu istotnie się różnią w otoczeniu wartości średniej prawdziwego rozkładu. Problem w tym, że w ogonach rozkładu (czyli dla wartości $x$ odległych od wartości średniej) każda dystrybuanta jest bliska zera dla małych wartości $x$ oraz bliska 1 dla dużych wartości $x$. W rezultacie dystrybuanty zgoła różnych rozkładów mogą być sobie bliskie z dala od ich (prawie takiej samej) wartości średniej.

Dokładniejsze porównanie testu Kołmogorowa, jako testu normalności rozkładu, z innymi testami odkładamy do zad. 3.23.

**Test Craméra–von Misesa.** Zaczniemy od przypadku prostej hipotezy zerowej (3.71). Test, który teraz krótko omówimy jest ideowo podobny do testu Kołmogorowa, ale jest oparty na analizie innej statystyki testowej, a mianowicie *statystyki Craméra–von Misesa*

$$W_n^2 = \int_{-\infty}^{\infty} (F_n(x) - F_0(x))^2 f_0(x) dx, \tag{3.74}$$

gdzie $F_n(\cdot)$ jest dystrybuantą empiryczną próby losowej, natomiast $F_0(\cdot)$ jest dystrybuantą i $f_0(\cdot)$ gęstością odpowiadającą prostej hipotezie zerowej (3.71). Tym razem analizie poddaje się całkę z ważonego (funkcją $f_0(\cdot)$) kwadratu różnicy między dystrybuantą empiryczną a dystrybuantą $F_0(\cdot)$. W przypadku testowania złożonej hipotezy o normalności badanego rozkładu, we wzorach na $F_0(\cdot)$ i $f_0(\cdot)$ pojawiają się estymatory nieznanej wartości średniej i wariancji rozkładu, z którego pochodzi próba losowa. Wartości krytyczne oraz $p$–wartości dla takiego testu znajduje się w tablicach statystycznych albo wykorzystuje się pakiet statystyczny.

Najogólniej mówiąc, z badań empirycznych wynika, że własności testu Craméra–von Misesa są generalnie lepsze od własności testu Kołmogorowa. Niemniej jednak i ten test jest stosunkowo nieczuły na odstępstwa od normalności w ogonach testowanego rozkładu. Pewną modyfikację podejścia Craméra–von Misesa, przypisującą większą wagę różnicom między rozkładami w ich ogonach, zaproponowali Anderson i Darling. Modyfikacji tej nie będziemy jednak omawiać.

## 3.5. Zadania

**3.1.** Niech $X_1, X_2, \ldots, X_n$ będzie próbą losową z rozkładu wykładniczego z parametrem $\lambda$. Znaleźć estymator NW tego parametru.

**3.2.** Niech $X_1, X_2, \ldots, X_n$ będzie próbą losową z rozkładu Bernoulliego z prawdopodobieństwem sukcesu $\theta$. Znaleźć estymator NW tego prawdopodobieństwa.

**3.3.** Niech $X_1, X_2, \ldots, X_n$ będzie próbą losową z rozkładu dwumianowego ze znaną liczbą powtórzeń doświadczenia Bernoulliego $m$ i nieznanym prawdopodobieństwem pojedynczego sukcesu $p$. Wykazać, że estymatorem NW parametru $p$ jest średnia w próbie podzielona przez $m$. Czy otrzymany estymator ma postać zgodną z intuicją?

**3.4.** Ujemny rozkład dwumianowy o parametrach $r$ i $p$ opisuje rozkład zmiennej losowej $X$ równej liczbie porażek w doświadczeniach Bernoulliego powtarzanych do uzyskania pierwszych $r$ sukcesów, gdy prawdopodobieństwo pojedynczego sukcesu wynosi $p$. Jak z tego wynika, jest to rozkład postaci

$$P(X = x) = \begin{pmatrix} x+r-1 \\ r-1 \end{pmatrix} p^r (1-p)^x, \tag{3.75}$$

$x = 0, 1, 2, \ldots$. Zauważywszy, że

$$\begin{pmatrix} x+r-1 \\ r-1 \end{pmatrix} = \frac{(x+r-1)(x+r-2)\cdots(x+r-x)}{x!} \tag{3.76}$$

i przyjąwszy równość (3.76) za definicję współczynnika dwumianowego z dowolną, ustaloną dodatnią liczbą rzeczywistą $r$, można udowodnić, iż suma (3.75) wynosi 1. W ten sposób został zdefiniowany nowy rozkład prawdopodobieństwa, zwany uogólnionym ujemnym rozkładem dwumianowym z parametrami $r$ i $p$, gdzie $r$ jest dodatnią liczbą rzeczywistą i $p \in (0, 1)$. Rozkład ten nie ma żadnej prostej interpretacji, ale okazuje się, że dobrze opisuje np. pewne zjawiska geologiczne i biologiczne.
Jeśli wiadomo, że

$$EX = \frac{r(1-p)}{p} \text{ oraz } \sigma_X^2 = \frac{r(1-p)}{p^2},$$

oraz na podstawie równości (3.9) wykazać, że estymatory MM parametrów $r$ i $p$, oparte na $n$-elementowej próbie losowej z tego rozkładu mają postać, odpowiednio,

$$\hat{r} = \frac{\bar{X}^2}{\frac{1}{n}\sum X_i^2 - \bar{X}^2 - \bar{X}}$$

oraz

$$\hat{p} = \frac{\bar{X}}{\frac{1}{n}\sum X_i^2 - \bar{X}^2}.$$

**3.5.** Wykazać, że jednostronne przedziały ufności (3.23) i (3.24) mają poziom ufności $1 - \alpha$.

**3.6.** Zmierzono czas życia, czyli czas działania, próby losowej 16 żarówek o ustalonej mocy. Średni czas życia w próbie wyniósł 3000 godzin, natomiast odchylenie standardowe 20 godzin. Przy założeniu, że czas życia żarówki jest zmienną losową o rozkładzie normalnym, podać przedział ufności dla wartości średniej tego rozkładu na poziomie ufności 0,9.

**3.7.** Laboratorium zakładów chemicznych porównuje wydajność reakcji chemicznej przy zastosowaniu dwóch różnych katalizatorów. Wykonano 12 eksperymentów przy zastosowaniu pierwszego katalizatora, uzyskano średnią

wydajność reakcji w próbie równą 85 przy odchyleniu standardowym w próbie 4. Niezależnie wykonano 10 eksperymentów przy zastosowaniu drugiego katalizatora i uzyskano średnią wydajność 81 przy odchyleniu standardowym w próbie równym 5. Podać przedział ufności dla różnicy wartości średnich wydajności w obydwu populacjach na poziomie ufności 0,9. Założyć, że w obydwu przypadkach otrzymywane wydajności mają rozkład normalny o tym samym odchyleniu standardowym.

**3.8.** Siłownia reklamuje program odchudzający twierdząc, że ćwiczący zmniejsza swój obwód w talii w ciągu 5 dni ćwiczeń średnio o 2 cm. Zmierzono obwody w talii 6 mężczyzn biorących udział w programie przed rozpoczęciem ćwiczeń oraz po upływie 5 dni. W przypadku pierwszego mężczyzny uzyskano 95,5 cm przed i 93,9 cm po 5-dniowym cyklu ćwiczeń. W przypadku drugiego uzyskano 98,7 i 97,4 cm. W przypadku kolejnych uczestników badania uzyskano odpowiednio przed i po cyklu zajęć: 90,4 i 91,7 cm; 115,9 i 112,8 i 115,9 cm; 104,0 i 101,3 cm; 85,6 i 84,0 cm.
Założyć normalny rozkład różnic obwodów przed i po 5 dniach ćwiczeń, znaleźć przedział ufności dla średniego zmniejszenia obwodu na poziomie ufności 0,95. Czy otrzymany wynik świadczy, że twierdzenie siłowni jest uzasadnione?

**3.9.** Wygenerować próbę losową ze standardowego rozkładu normalnego o liczności 20. Skonstruować przedziały ufności dla wartości średniej tego rozkładu na poziomie ufności 0,95 i 0,99. Zastąpić dowolny z elementów próby wartością odstającą równą 3,1 i dla próby z rzekomym błędem zapisu skonstruować takie przedziały ufności jak wcześniej. Porównać odpowiednie przedziały z obydwu eksperymentów. Powtórzyć całe ćwiczenie 10 razy, ewentualnie zmieniając liczność próby, wartość odstającą itp.

**3.10.** (Uwaga: Próby losowe wygenerowane w tym zadaniu będą użyte w zad. 3.22.) Niekiedy czas życia urządzenia modeluje się za pomocą zmiennej losowej $X$ o rozkładzie logarytmiczno-normalnym danym gęstością

$$f(x) = \begin{cases} \frac{1}{\sigma x\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{\ln x-\mu}{\sigma}\right)^2\right), & \text{gdy } x > 0 \\ 0, & \text{gdy } x \leq 0, \end{cases}$$

gdzie $\mu$ jest dowolną stałą i $\sigma$ jest stałą dodatnią (logarytm naturalny zmiennej losowej o rozkładzie logarytmiczno-normalnym ma rozkład normalny o wartości średniej $\mu$ i odchyleniu standardowym $\sigma$). Wiadomo, że

$$EX = \exp\left(\mu + \frac{1}{2}\sigma^2\right) \text{ oraz } \sigma_X^2 = (\exp\sigma^2 - 1)\exp(2\mu + \sigma^2).$$

**a)** Wygenerować kilka prób losowych o liczności 500 z rozkładu logarytmiczno-normalnego o różnych parametrach, w tym próby z rozkładów o para-

metrach: $\mu = 0$, $\sigma = 1$; $\mu = 1, \sigma = 1$; $\mu = 0, \sigma = 1/2$. Zbadać kształty otrzymanych rozkładów.
**b)** Dla każdego z rozkładów zbadanych w punkcie a wygenerować parę niezależnych prób losowych o liczności 50. Dla każdej z otrzymanych par skonstruować przedziały ufności dla ilorazu wariancji na poziomie ufności 0,95 i 0,99.
**c)** Obliczyć wartości średnie i odchylenia standardowe każdego z rozkładów zbadanych poprzednio. Dla każdej otrzymanej pary (wartość srednia, odchylenie standardowe) wygenerować dwie niezależne próby losowe o liczności 50 z rozkładu normalnego o tych parametrach. Dla każdej z otrzymanych par prób z rozkładu normalnego skonstruować przedziały ufności dla ilorazu wariancji na poziomie ufności 0,95 i 0,99.
**d)** Powtórzyć punkt b i c kilkakrotnie. We wszystkich rozważonych przypadkach prawdziwa wartość ilorazu wariancji jest równa 1. Ocenić precyzję otrzymanych przedziałów ufności w zależności od rozkładu próby losowej.
**e)** Powtórzyć punkty b – d dla prób o liczności 100 i 500.

**3.11.** Uzasadnić, że przy zaobserwowanej częstości $\hat{p}$ przedział ufności dla nieznanej proporcji $p$ na poziomie ufności równym w przybliżeniu $1 - \alpha$ ma długość nie większą niż $l$, gdy liczność próby $n$ jest nie mniejsza niż

$$\frac{4z^2_{1-\alpha/2}\hat{p}(1-\hat{p})}{l^2},$$

i są spełnione warunki $np > 5$ oraz $n(1-p) > 5$. Pamiętając, że funkcja $y(x) = x(1-x)$ osiąga wartość maksymalną równą $1/4$ (dla $x = 1/2$), zauważyć, że otrzymany iloraz jest nie większy niż $z^2_{1-\alpha/2}/l^2$ (innymi słowy, jeżeli tylko $np > 5$ oraz $n(1-p) > 5$, to bez względu na wartość prawdziwej proporcji $p$ liczność próby przekraczająca wielkość $z^2_{1-\alpha/2}/l^2$ daje przedział ufności o długości nie większej niż $l$).

**3.12.** Podrzucano monetę 3 razy. Należy przetestować hipotezę $H_0$: $p = 1/2$ przy hipotezie alternatywnej $H_1$: $p = 2/3$, gdzie $p$ oznacza wypadnięcie orła. Procedura testowa polega na odrzuceniu hipotezy zerowej, gdy wypadły dwa lub więcej orłów. Wykazać, że moc tego testu wynosi $20/27$.

**3.13.** Na podstawie próby losowej z rozkładu normalnego o nieznanej wartości średniej $\theta$ i znanym odchyleniu standardowym 0,05, należy testować hipotezę (3.41) przy alternatywie (3.42). Skorzystać ze statystyki testowej (3.43). Wykazać, że liczność próby losowej musi wynosić przynajmniej 245, jeśli chcemy by na poziomie istotności 0,01 moc testu przekraczała 0,8, gdy prawdziwa wartość parametru $\theta$ różni się od przyjętej w hipotezie zerowej wartości $\theta_0$ o 0,01.

**3.14.** Na podstawie próby losowej o liczności $n$ z rozkładu normalnego o nieznanej wartości średniej $\theta$ i znanym odchyleniu standardowym 0,05, należy

testować hipotezę (3.41) przy alternatywie (3.50). Skorzystać ze statystyki testowej (3.43).
**a)** Wykazać, że przy liczności próby $n = 10$ moc testu na poziomie istotności 0,05 jest równa 0,599, gdy prawdziwa wartość parametru $\theta$ różni się od przyjętej w hipotezie zerowej wartości $\theta_0$ o 0,03.
**b)** Wykazać, że przy liczności próby $n = 5000$ moc testu na poziomie istotności 0,01 jest równa 0,691, gdy prawdziwa wartość parametru $\theta$ różni się od przyjętej w hipotezie zerowej wartości $\theta_0$ o 0,002.
**c)** Wykazać, że zmniejszenie w punkcie b liczności próby do 10 i zachowanie pozostałych warunków niezmienionych prowadzi do mocy 0,014.

**3.15.** Zauważyć, że ponieważ w zad. 3.13 i 3.14 jest znane odchylenie standardowe, za błąd standardowy średniej w próbie można uznać wielkość $\sigma/\sqrt{n}$, gdzie $\sigma$ oznacza odchylenie standardowe, a $n$ liczność próby. Dla wymienionych zadań obliczyć odchylenia standardowe średnich w próbie i na tej podstawie podać intuicyjne uzasadnienie wzrostu mocy testu o zadanym poziomie istotności, wynikające ze zwiększania liczności próby.

**3.16.** Dostawca drewnianych belek twiedzi, że ich średnia wytrzymałość jest równa $\mu = 40$ kg/cm$^2$. Wiadomo z doświadczenia, że wytrzymałość może być uznana za zmienną losową o rozkładzie normalnym i odchyleniu standardowym 2. Wybrano losowo 25 belek i zmierzono ich wytrzymałości. Otrzymano, że średnia wytrzymałość w próbie wynosi 39.

Celem sprawdzenia, czy twierdzenie dostawcy nie powinno być odrzucone, przetestować na poziomie istotności 0,1 hipotezę $H_0$: $\mu = 40$ przy alternatywie $H_1$: $\mu < 40$. Obliczyć moc testu dla $\mu = 39$.

**3.17.** Technolodzy proponują zmianę procesu produkcji przędzy bawełnianej. Nowa technologia ma przynieść zwiększenie średniej wytrzymałości przędzy na zerwanie. Pobrano dwie próby szpul przędzy. Każda próba liczy 10 szpul. Jedna próba zawiera szpule z przędzą produkowaną zgodnie ze starą technologią, druga zgodnie z nową. W przypadku starej technologii otrzymano średnią wytrzymałość w próbie równą 8,37 i wariancję w próbie 1,32. W przypadku nowej, odpowiednio, 9,62 i 1,18. Zakłada się, że wytrzymałość przędzy może być uznana za zmienną losową o rozkładzie normalnym. Przetestować na poziomie istotności 0,05 hipotezę o równości średnich wytrzymałości przy hipotezie alternatywnej, że po zastosowaniu nowej technologii otrzymuje się przędzę o większej średniej wytrzymałości na zerwanie.

**3.18.** Badany jest lek zatrzymujący rozwój białaczki. Wybrano próbę losową 9 myszy w zaawansowanym stadium choroby. Leczeniu poddano 5 myszy. Te, które poddano leczeniu, przeżyły (w latach): 2,1; 5,3; 1,4; 4,6; 0,9. Myszy, których nie poddano leczeniu przeżyły: 1,9; 0,5; 2,8; 3,1 lat.

Przy założeniu, że czas życia jest w obydwu przypadkach zmienną losową o rozkładzie normalnym z tym samym odchyleniem standardowym, przetestować hipotezę, że obydwa średnie czasy życia są sobie równe przy alternatywie, że średni czas życia myszy leczonych jest dłuższy. Na podstawie przeprowadzonego testu ocenić skuteczność leku.

**3.19.** Powtórzyć zad. 3.10 (punkty b – d) i zastąpić wszędzie konstrukcję przedziału ufności dla ilorazu wariancji konstrukcją testu F dla ilorazu wariancji. Za każdym razem przyjąć poziom istotności testu równy poziomowi ufności danego przedziału ufności.

**3.20.** Ocenia się, że w województwie $X$ korzystało bezprawnie z pewnej ulgi podatkowej 10% podatników. Istnieje obawa, że zmiana przepisów podatkowych mogła zwiększyć podany odsetek osób nieprawidłowo obliczających płacony przez nie podatek. Wylosowano 150 podatników i wykazano, że 21 z nich niesłusznie skorzystało ze wspomnianej ulgi. Skonstruować odpowiedni test i na tej podstawie ocenić zasadność istniejących obaw.

**3.21.** Powtórzyć zad. 1.5 i dodać wszędzie konstrukcję wykresu kwantylowego. Przedyskutować wyniki.

**3.22.** Sporządzić wykresy ramkowe, histogramy oraz wykresy kwantylowe dla prób losowych wygenerowanych w zad. 3.10.

**3.23.** Dla podanych dalej danych sporządzić: wykresy ramkowe, histogramy i wykresy kwantylowe oraz obliczyć współczynnik skośności i kurtozę. Zastosować testy kierunkowe oraz różne testy uniwersalne, w tym koniecznie test Shapiro–Wilka i Kołmogorowa i przetestować hipotezę o normalności rozkładu danej próby. Przedyskutować wyniki.
Dane są próby losowe latencji L3-N33 (patrz przykł. 1.4) i L5-P40 oraz składników losowych tej ostatniej R-P40 (patrz zad. 4.2); ponadto w poniższej tabeli są podane próby losowe latencji L3-P60 i L5-P60:

L3-P60

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 57,30 | 57,60 | 51,70 | 60,30 | 54,20 | 52,70 | 56,75 | 54,30 | 57,75 | 49,30 | 55,30 | 64,80 | 52,05 |
| 55,70 | 55,70 | 58,90 | 54,70 | 56,60 | 53,90 | 56,20 | 54,50 | 57,30 | 55,00 | 58,70 | 54,40 | 54,50 |
| 57,70 | 63,30 | 49,80 | 55,80 | 54,40 | 46,50 | 51,20 | 60,30 | 54,10 | 52,80 | 53,20 | 51,90 | 58,20 |
| 50,50 | 67,40 | 55,30 | 58,20 | 54,20 | 60,80 | 57,40 | 53,10 | 54,50 | 49,20 | 51,40 | 60,00 | 76,60 |
| 66,30 | 56,90 | 56,10 | 53,10 | 59,10 | 56,30 | 55,80 | 57,20 | 65,60 | 56,40 | | | |

L5-P60

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 77,55 | 77,25 | 73,40 | 69,90 | 73,50 | 71,35 | 82,65 | 85,75 | 84,25 | 78,30 | 71,40 | 85,60 | 87,50 |
| 75,95 | 69,65 | 77,75 | 76,05 | 80,50 | 74,40 | 73,35 | 70,10 | 84,15 | 80,30 | 89,65 | 77,00 | 73,30 |
| 68,00 | 82,00 | 73,85 | 73,45 | 75,50 | 67,80 | 63,85 | 79,10 | 62,85 | 83,40 | 73,15 | 67,80 | 71,45 |
| 67,95 | 79,60 | 73,25 | 78,15 | 73,00 | 83,35 | 65,15 | 85,60 | 70,35 | 76,30 | 75,10 | 77,35 | 76,60 |
| 69,15 | 87,25 | 72,45 | 75,10 | 74,80 | 76,45 | 71,75 | 78,50 | 73,15 | 79,40 | | | |

ROZDZIAŁ 4

# Analiza zależności zmiennych ilościowych

## 4.1. Wprowadzenie

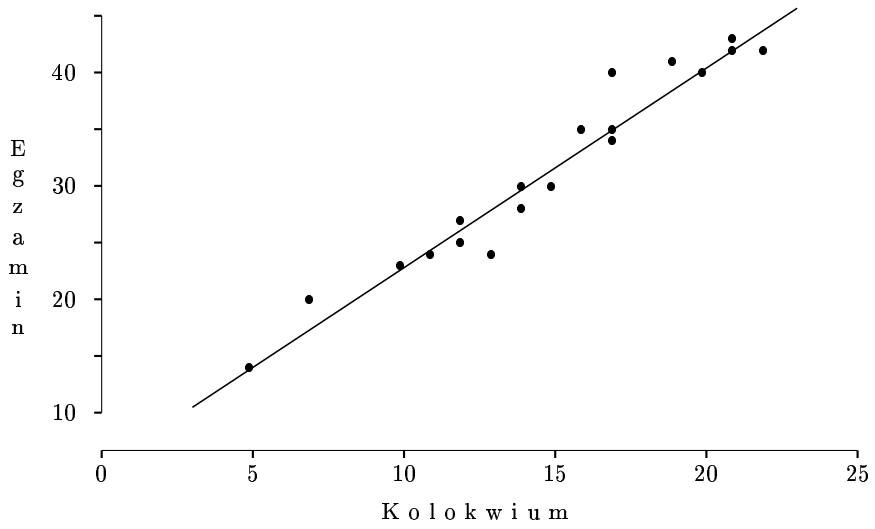
W typowych sytuacjach zbierania danych, nawet jeśli naszym celem jest jedynie analiza zachowania pewnej wielkości losowej $Y$, z reguły nie ograniczamy się tylko do obserwacji tej zmiennej, ale zbieramy również informacje towarzyszące, które mogą mieć znaczenie w analizie interesującej nas wielkości. W badaniach demograficznych, np. rozkładu wzrostu dorosłych Polaków, mamy z reguły do dyspozycji informacje dotyczące ich wieku, wagi, wykształcenia i stanu cywilnego. Analizując poziom inflacji w Polsce w kolejnych latach, bierzemy pod uwagę podstawowe wskaźniki gospodarcze takie, jak wysokość deficytu płatniczego, stopy procentowe czy wartość produktu narodowego przypadającego na jednego obywatela. Zbieranie informacji dodatkowej jest motywowane tym, że badana wielkość, choć losowa, w istotny sposób zależy od innych zmiennych i zrozumienie charakteru tej zależności może być pożyteczne w wielu zadaniach np. do przewidywania przyszłych wartości interesującej nas zmiennej. W tym rozdziale zajmiemy się analizą zależności między zmiennymi ilościowymi. Na początku rozważymy sytuację jedynie dwóch zmiennych $X$ i $Y$.

## 4.2. Analiza zależności dwóch zmiennych ilościowych

**Przykład 4.1.** Rozpatrzmy rezultaty kolokwium (na skali od 0 do 25 punktów) i egzaminu końcowego (na skali od 0 do 50 punktów) z rachunku prawdopodobieństwa i statystyki. W kolokwium i egzaminie brało udział 19 studentów pewnej szkoły technicznej (tab. 4.1).

**Tabela 4.1. Wyniki kolokwium i egzaminu końcowego (przykł. 4.1)**

| Numer studenta | Kolokwium | Egzamin końcowy | Numer studenta | Kolokwium | Egzamin końcowy |
|---|---|---|---|---|---|
| 1 | 7 | 20 | 11 | 5 | 14 |
| 2 | 11 | 24 | 12 | 12 | 27 |
| 3 | 12 | 25 | 13 | 16 | 35 |
| 4 | 14 | 30 | 14 | 14 | 28 |
| 5 | 17 | 35 | 15 | 21 | 42 |
| 6 | 15 | 30 | 16 | 20 | 40 |
| 7 | 21 | 43 | 17 | 17 | 34 |
| 8 | 22 | 42 | 18 | 10 | 23 |
| 9 | 19 | 41 | 19 | 17 | 40 |
| 10 | 13 | 24 | | | |

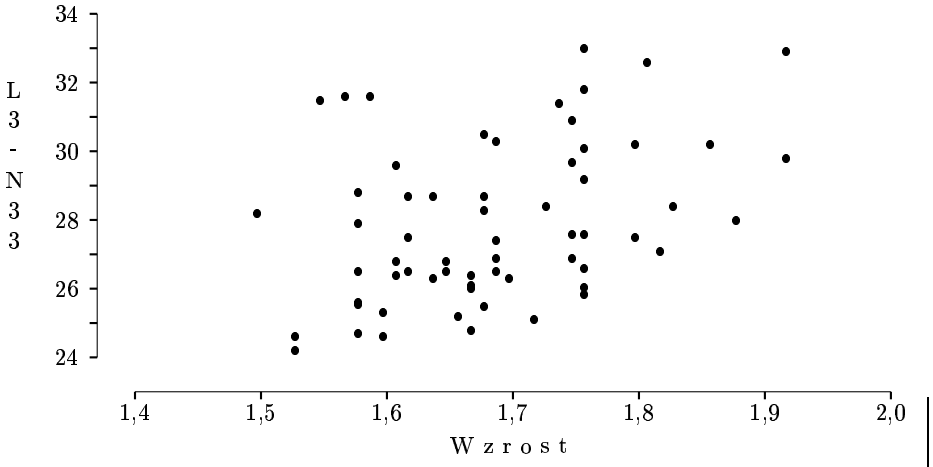



Rys. 4.1. Wykres rozproszenia dla danych z przykł. 4.1

Zależność między wynikiem egzaminu końcowego a kolokwium przedstawiono na rys. 4.1 w postaci wykresu par $(x, y)$, gdzie $x$ oznacza rezultat kolokwium, a $y$ rezultat egzaminu końcowego dla danego studenta. Wykres taki nosi nazwę **wykresu rozproszenia** i jest użytecznym graficznym przedstawieniem zależności między zmiennymi. Przykładowo, punkt $(21, 43)$ na wykresie odpowiada studentowi o numerze porządkowym 7, który uzyskał 21 punktów z kolokwium i 43 punkty z egzaminu końcowego. Zauważmy, że w naszym przykładzie dużym (odpowiednio małym) wartościom wyniku kolokwium odpowiadają z reguły duże (małe) wartości egzaminu końcowego. Mówimy w takim przypadku o zależności dodatniej między zmiennymi, w odróżnieniu od zależności ujemnej, gdy duże wartości jednej

zmiennej odpowiadają w większości przypadków małym wartościom drugiej zmiennej i podobnie małe wartości pierwszej zmiennej towarzyszą z reguły dużym wartościom drugiej zmiennej. W rozpatrywanym przykładzie chmura punktów ma w przybliżeniu charakter liniowy. Oznacza to, że przez chmurę punktów można przeprowadzić prostą, która dobrze oddaje charakter jej zmienności.

**Przykład 4.2.** Rozpatrzmy zależność latencji L3-N33 dla zdrowych osobników od ich wzrostu (porównaj przykł. 1.4 w rozdz. 1). Wykres rozproszenia przedstawiony jest na rys. 4.2. Chociaż zależność między zmiennymi jest podobnie jak w poprzednim przykładzie dodatnia, to siła zależności jest teraz znacznie słabsza niż poprzednio.




Rys. 4.2. Wykres rozproszenia dla danych z przykł. 4.2

W kontekście badania zależności będziemy używali następującej terminologii: zmienną $Y$ będącą wynikiem doświadczenia będziemy nazywali **zmienną objaśnianą** w odróżnieniu od **zmiennej objaśniającej** $X$, za pomocą zmian której chcemy objaśnić zmiany zmiennej $Y$. W przykładzie 4.1 zmienną objaśnianą jest rezultat egzaminu końcowego, zmienną objaśniającą jest rezultat kolokwium. Czasami zmienną $Y$ nazywa się również **zmienną zależną**, a zmienną $X$ **zmienną niezależną**. Przy konstrukcji wykresu rozproszenia wartości zmiennej objaśnianej umieszcza się na osi pionowej, a wartości zmiennej objaśniającej na osi poziomej. W pewnych sytuacjach nie jest jasne, która akurat zmienna ma być objaśniana, a która objaśniająca. W takiej sytuacji po prostu nie używamy tej terminologii. Na przykład możemy rozważyć rezultat egzaminu końcowego z rachunku prawdopodobieństwa i statystyki oraz egzaminu końcowego z inżynierii oprogramowania.

Jak można przypuszczać zmienne te są zależne, ale trudno w tym przypadku określić, która zmienna w tym przypadku jest zmienną objaśnianą, a która objaśniającą. W takiej sytuacji nie ma znaczenia, wartości której zmiennej umieścimy na osi pionowej, a które na osi poziomej.

W analizie wykresów rozproszenia najbardziej istotna jest ocena ogólnego charakteru zależności oraz odstępstw od niej. Ocena charakteru zależności zawiera ustalenie formy zależności, jej kierunku i siły. Typową formą zależności jest **przybliżona zależność funkcyjna**. Odpowiada ona sytuacji, kiedy chmura punktów tworzących wykres rozproszenia układa się dookoła wykresu pewnej funkcji. Gdy funkcja ta jest monotoniczna, mówimy o **zależności monotonicznej**, odpowiednio dodatniej i ujemnej. Rozproszenie punktów wokół wykresu decyduje o sile zależności: dla małego rozproszenia zależność jest silniejsza. Szczególne znaczenie ma sytuacja, gdy wykres funkcji, wokół której układają się punkty jest prostą. Jej rozpatrzeniu poświęcimy większą część materiału dwóch następnych punktów. Najbardziej istotnym odstępstwem od zależności jest występowanie pewnej liczby punktów, które odbiegają od podstawowej konfiguracji wykresu rozproszenia. Ich omówieniu poświęcimy część p. 4.2.4. Kłopot pojawia się wtedy, gdy wykres nie daje jednoznacznych przesłanek co do charakteru zależności. Stosuje się wtedy dodatkowe metody statystyczne umożliwiające wizualizację takich ukrytych zależności. Jedną z takich technik jest **średnia ruchoma**.

W celu skonstruowania średniej ruchomej dzieli się zakres zmienności zmiennej objaśnianej na pewną liczbę przedziałów $\{I_i\}$. Dla każdego przedziału $I_i$ identyfikuje się następnie wszystkie wartości $x_{j_1}, \ldots, x_{j_k}$ należące do tego przedziału i oblicza średnią z odpowiadających im wartości $y$: $\bar{y}_i = \frac{1}{k}\sum_{l=1}^{k} y_{j_l}$. Średnia ruchoma jest wykresem funkcji kawałkami stałej, która na przedziale $I_i$ przyjmuje wartość $\bar{y}_i$. Podobnie konstruuje się medianę ruchomą, zastępując średnią $\bar{y}_i$ przez medianę wartości $y_{j_1}, y_{j_2} \ldots, y_{j_k}$. Bardziej wyrafinowanym narzędziem od średniej ruchomej jest **lokalny estymator wielomianowy**, który przybliża lokalnie chmurę punktów na wykresie rozproszenia przez fragment wykresu wielomianu zadanego stopnia, dbając jednocześnie, aby otrzymane kawałki „sklejały się" w ciągłą funkcję. Nie będziemy go tutaj dokładniej omawiali.

### 4.2.1. Współczynnik korelacji próbkowej

Wprowadzimy obecnie pojęcie współczynnika korelacji próbkowej będącego estymatorem współczynnika korelacji, wprowadzonego w podrozdz. 2.3. Jego wartość obliczona dla konkretnych wartości próby ułatwia w wielu przypadkach określenie siły zależności. Współczynnik korelacji zmiennych losowych $X$ i $Y$ został zdefiniowany jako wartość średnia iloczynu standary-

zowanych zmiennych $(X - \mu_X)/\sigma_X$ i $(Y - \mu_Y)/\sigma_Y$. Współczynnik korelacji próbkowej jest odpowiednikiem tej definicji dla próby $(X_1, Y_1), \ldots, (X_n, Y_n)$.

**Definicja 4.1.** ***Współczynnikiem korelacji próbkowej*** *nazywamy zmienną losową*

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left( \frac{X_i - \bar{X}}{S_X} \right) \left( \frac{Y_i - \bar{Y}}{S_Y} \right),$$

*gdzie $\bar{X}$ i $S_X$ oznaczają średnią i odchylenie standardowe próby $X_1, X_2, \ldots, X_n$ i podobnie $\bar{Y}$ i $S_Y$ oznaczają średnią i odchylenie próby $Y_1, Y_2, \ldots, Y_n$.*

Ze względów tradycyjnych oznaczamy współczynnik korelacji w próbie małą literą $r$. Będziemy rozróżniać między współczynnikiem korelacji będącym zmienną losową, a jego wartością, oznaczaną również przez $r$, dla konkretnych wartości w próbie. Podstawowe własności próbkowego współczynnika korelacji są podobne do własności współczynnika korelacji $\rho$ dla pary zmiennych losowych. W szczególności, jeśli wykres rozproszenia ma charakter monotoniczny, wartość współczynnika korelacji próbkowej będzie dodatnia, gdy zależność jest dodatnia i ujemna, gdy zależność jest ujemna. *Sam fakt dodatniej wartości współczynnika korelacji nie może być interpretowany jako mówiący o dodatniej zależności zmiennych, gdy nie wiemy, czy zależność jest monotoniczna.* Zauważmy również, że ponieważ w próbkowym współczynniku korelacji wykorzystuje się jedynie wartości standaryzowane, nie zależy on od tego, w jakich jednostkach dokonujemy pomiaru wartości $X$ i $Y$.

Wymieńmy jeszcze dwie ważne własności próbkowego współczynnika korelacji, bedące odpowiednikami współczynnika korelacji między zmiennymi.

(1) *Próbkowy współczynnik korelacji jest zmienną losową ograniczoną przez liczby $-1$ i $1$: $-1 \leq r \leq 1$. Wartości $r$ bliskie $-1$ lub $1$ wskazują, że wykres rozproszenia jest skupiony wokół prostej.*

(2) *W przypadku liniowego charakteru wykresu rozproszenia próbkowy współczynnik korelacji mierzy siłę zależności między zmiennymi.*

Współczynnik korelacji między wynikiem egzaminu końcowego a wynikiem kolokwium w przykł. 4.1 wynosi $r = 0{,}973$. Wartość współczynnika korelacji bliska 1 wskazuje na przybliżoną liniowość wykresu rozproszenia i bardzo silną zależność dodatnią między zmiennymi. Podobnie, w przykł. 4.2 próbkowy współczynnik korelacji $r$ równa się 0,51. Z wykresu można wnosić, że zależność zmiennych jest monotoniczna, wartość współczynnika korelacji świadczy o umiarkowanie silnej zależności dodatniej między zmiennymi. Zauważmy, że w definicji próbkowego współczynnika korelacji nie jest istotne, która zmienna jest zmienną objaśnianą, a która objaśniającą. Role zmiennych $X$ i $Y$ można zamienić bez zmiany wartości współczynnika $r$. Wynika

z tego, że nie może być on użyty do ustalenia relacji przyczynowej między zmiennymi. Podkreślmy jeszcze, że ponieważ próbkowy współczynnik korelacji opiera się na średnich i odchyleniach standardowych, podobnie jak one nie jest odporny na obserwacje odstające.

### 4.2.2. Liniowa zależność między dwiema zmiennymi, prosta regresji

Rozpatrzmy sytuację, gdy wykres rozproszenia dla próby wartości $(x_1, y_1), \ldots, (x_n, y_n)$ wskazuje na wyraźną zależność liniową między zmiennymi $x$ i $y$. Tak jest w sytuacji przedstawionej na rys. 4.1, gdzie $x$ jest wynikiem kolokwium, a $y$ jest wynikiem egzaminu końcowego. Zastanówmy się nad metodami wyznaczenia prostej, która adekwatnie reprezentowałaby analizowaną chmurę punktów. Oczywiście, musimy zdecydować się na pewną miarę adekwatności dopasowania, która dobrze odpowiada naszym intuicjom i jednocześnie umożliwia jednoznaczne wyznaczenie poszukiwanej prostej. Rozpatrzmy dowolną, ale ustaloną prostą $y = b_0 + b_1 x$. Współczynnik $b_1$ nosi nazwę **współczynnika kierunkowego**, a $b_0$ **wyrazu wolnego**. Jeśli przyjmiemy, że prosta ma przybliżać daną chmurę punktów, wartość $\hat{y}_i = b_0 + b_1 x_i$ można interpretować jako **wartość $y$ przewidywaną na podstawie rozpatrywanej prostej** dla wartości zmiennej objaśniającej $x$ równej $x_i$. Błąd oszacowania, czyli tzw. **wartość resztowa** lub **rezyduum** wynosi $y_i - \hat{y}_i$. Chcielibyśmy, żeby dla prostej adekwatnie opisującej charakter zależności, wartości rezyduów były jak najmniejsze dla wszystkich $i = 1, \ldots, n$. Jednak jak łatwo zauważyć, przesunięcie prostej w kierunku ustalonego punktu próby, w celu zmniejszenia odstępstwa prostej od tego właśnie punktu powoduje z reguły jej odsunięcie od jakiegoś innego punktu. Tak więc postulat, że jednocześnie wszystkie rezydua są małe jest często trudny do zrealizowania. Podobnie jak przy definicji wariancji, za wskaźnik rozproszenia możemy przyjąć sumę kwadratów wartości wszystkich odchyłek, czyli sumę kwadratów rezyduów

$$S(b_0, b_1) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i))^2. \tag{4.1}$$

Różnica z definicją wariancji polega na tym, że estymowaną wartość $\hat{y}_i$ porównujemy teraz nie ze średnią wartości $\hat{y}_i$, ale z zaobserwowaną wartością $y_i$. Jako prostą reprezentującą chmurę punktów wybieramy prostą wyznaczoną przez współczynniki $b_0$ i $b_1$, dla których suma kwadratów błędów $S(b_0, b_1)$ traktowana jako funkcja swoich argumentów jest minimalna. Dopuszczamy zatem, że nie wszystkie rezydua są małe, ale w przypadku liniowego charakteru wykresu dużych rezyduów nie może być zbyt wiele. Metoda ta pochodzi

od F. Gaussa, który pierwszy wykorzystał ją przy analizie danych astronomicznych. Nosi ona nazwę **metody najmniejszych kwadratów**.

**Definicja 4.2.** ***Prostą regresji*** *opartą na metodzie najmniejszych kwadratów nazywamy prostą $b_0 + b_1 x$, dla której wartość sumy $S(b_0, b_1)$ w (4.1), traktowanej jako funkcja wszystkich możliwych wartości współczynnika kierunkowego i wyrazu wolnego, jest minimalna.*

Prostą regresji opartą na metodzie najmniejszych kwadratów nazywać będziemy w skrócie **prostą regresji MNK** lub **prostą MNK**. Wyznaczmy teraz jawną postać tej prostej. Po zróżniczkowaniu funkcji $S(b_0, b_1)$ względem $b_0$ i $b_1$ i przyrównaniu obu pochodnych cząstkowych do 0, otrzymamy

$$\frac{\partial S}{\partial b_0} = -2\sum_{i=1}^{n}(y_i - (b_0 + b_1 x_i)) = 0 \tag{4.2}$$

i

$$\frac{\partial S}{\partial b_1} = -2\sum_{i=1}^{n} x_i(y_i - (b_0 + b_1 x_i)) = 0. \tag{4.3}$$

Stąd

$$\sum_{i=1}^{n} y_i - nb_0 - b_1 \sum_{i=1}^{n} x_i = 0 \tag{4.4}$$

i

$$\sum_{i=1}^{n} x_i y_i - b_0 \sum_{i=1}^{n} x_i - b_1 \sum_{i=1}^{n} x_i^2 = 0. \tag{4.5}$$

Z równania (4.4) natychmiast otrzymujemy, że

$$b_0 = \frac{1}{n}(\sum_{i=1}^{n} y_i - b_1 \sum_{i=1}^{n} x_i) = \bar{y} - b_1 \bar{x}, \text{ gdzie } \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \ , \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \tag{4.6}$$

i po wstawieniu powyższego wyrażenia do równania (4.5) i po prostych przekształceniach mamy

$$\sum_{i=1}^{n} x_i y_i - (\bar{y} - b_1 \bar{x}) \sum_{i=1}^{n} x_i - b_1 \sum_{i=1}^{n} x_i^2 = 0$$

i

$$\sum_{i=1}^{n} x_i(y_i - \bar{y}) = b_1(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2).$$

Zatem, ponieważ $\sum_{i=1}^n x_i^2 - n\bar{x}^2 = \sum_{i=1}^n (x_i - \bar{x})^2$,

$$b_1 = \frac{\sum_{i=1}^n x_i(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}. \quad (4.7)$$

Dwie ostatnie równości wynikają z faktu, że $\sum_{i=1}^n (x_i - \bar{x}) = 0$ i $\sum_{i=1}^n (y_i - \bar{y}) = 0$. Tak więc prosta regresji oparta na metodzie najmniejszych kwadratów jest jednoznacznie wyznaczona przez współczynniki $b_0$ i $b_1$ określone w równaniach (4.6) i (4.7). Wartość współczynnika $b_1$ jest obliczana najpierw i wstawiana do równania (4.6) w celu obliczenia wartości wyrazu wolnego $b_0$. Wartość $y = b_0 + b_1 x$ nazywamy **wartością przewidywaną** zmiennej objaśnianej na podstawie prostej MNK dla wartości zmiennej objaśniającej równej $x$.

**Przykład 4.1 cd.** Tabela 4.2 zawiera część wydruku SAS dotyczącego analizy regresji dla przykł. 4.1. Prosta regresji MNK ma postać $y = 5{,}2 + 1{,}76 \times x$, zatem $b_0 = 5{,}2$ i $b_1 = 1{,}76$. Dokładność obliczeń powinna być o rząd większa od dokładności danych, dlatego wyniki podawane przez SAS zaokrąglamy do pierwszego miejsca po przecinku. Przykładowo, wartość przewidywana dla studenta o numerze 2, który uzyskał 24 punkty na kolokwium, wynosi $5{,}2 + 1{,}76 \times 11 = 24{,}6$, a wartość rezyduum wynosi $24 - 24{,}6 = -0{,}6$. Z wykresu rozproszenia z naniesioną prostą regresji MNK widzimy, że prosta MNK bardzo dobrze odzwierciedla charakter chmury punktów. Potwierdza to naszą intuicję, że charakter zależności jest w przybliżeniu liniowy.

Przyjrzyjmy się teraz własnościom współczynników $b_0$ i $b_1$, wynikającym bezpośrednio z definicji. Ciąg par $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ jest otrzymany jako wynik eksperymentu. Oczywiście, dysponując narysowaną prostą regresji MNK, możemy na wykresie łatwo wyznaczyć wartość $b_0$: jest to wartość prostej regresji dla $x = 0$. Co więcej, nachylenie $b_1$ łatwo odczytać z wykresu jako przyrost wartości prostej przy jednostkowym przyroście argumentu $x$

$$b_1 = \frac{b_0 + b_1(x+1) - (b_0 + b_1 x)}{(x+1) - x}.$$

Przykładowo, różnica wartości rezultatu kolokwium dla siódmego i ósmego studenta wynosi 1, a odpowiednia różnica wartości przewidywanych $43{,}9 - -42{,}2$ jest równa 1,7. Różnica między 1,7 a wartością $b_1$ wynika z błędu zaokrąglenia. Zauważmy jeszcze, że używając def. 4.1 i określenia współ-

czynnika nachylenia $b_1$, łatwo możemy otrzymać związek tego współczynnika z wartością współczynnika korelacji próbkowej

$$b_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x^2} = \\ = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} \frac{s_y}{s_x} = r\frac{s_y}{s_x}. \tag{4.8}$$

(Przypomnijmy, że oznaczenia $s_y$ używamy, gdy rozpatrujemy odchylenie standardowe ustalonych wartości $y_1, y_2, \ldots, y_n$ zmiennych $Y_1, Y_2, \ldots, Y_n$.)

**Tabela 4.2. Wydruk z pakietu SAS dotyczący analizy regresji dla przykł. 4.1**

```
The SAS System       18:22 Monday, August 21, 2000   1

Model Equation
KONCOWY   =   5.1999   +   1.7604  KOLOKW

                              Parametric Regression Fit
                                      Model              Error
Curve   Degree(Polynomial)    DF  Mean Square   DF  Mean Square   R-Square
                         1     1    1251.3860   17       4.0733     0.9476

                              Parametric Regression Fit

                                 F Stat    Prob > F
                               307.2191      0.0001

                     Summary of Fit
Mean of Response           31.4211  R-Square        0.9476
Root MSE                    2.0182  Adj R-Sq        0.9445

                     Analysis of Variance
Source      DF  Sum of Squares  Mean Square      F Stat    Prob > F
Model        1       1251.3860    1251.3860    307.2191      0.0001
Error       17         69.2456       4.0733           .           .
C Total     18       1320.6316            .           .           .

                     Type III Tests
Source      DF  Sum of Squares  Mean Square      F Stat    Prob > F
KOLOKW       1       1251.3860    1251.3860    307.2191      0.0001

                     Parameter Estimates                              Var
Variable    DF  Estimate  Std Error   T Stat  Prob >|T|  Tolerance Inflation
INTERCEPT    1    5.1999     1.5660   3.3205     0.0040          .   0.0000
KOLOKW       1    1.7604     0.1004  17.5277     0.0001     1.0000   1.0000
```

Ponadto,

*prosta regresji MNK przechodzi przez punkt* $(\bar{x}, \bar{y})$.

Wynika to bezpośrednio z równania (4.6):

$$b_0 + b_1\bar{x} = \bar{y} - b_1\bar{x} + b_1\bar{x} = \bar{y},$$

a zatem wartością prostej MNK dla argumentu $\bar{x}$ jest $\bar{y}$.

Rozważmy teraz pewne własności rezyduów $e_i = y_i - \hat{y}_i, i = 1, \ldots, n$. Z faktu, że prosta MNK minimalizuje sumę kwadratów rezyduów wynika, że

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (y_i - \hat{y}_i) = 0. \tag{4.9}$$

Rzeczywiście, jeśli skorzystamy z równości $\hat{y}_i = b_0 + b_1 x_i$, to otrzymamy

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i)) = 0$$

na mocy równania (4.2). Widzimy zatem, że wartości rezyduów nie mogą być zupełnie dowolne. W szczególności niemożliwe jest, aby wszystkie rezydua były jednocześnie dodatnie albo wszystkie jednocześnie ujemne. Z równania (4.9) po podzieleniu obu stron przez $n$ wynika również, że średnia $\bar{y}$ jest równa średniej $\bar{\hat{y}}$ wartości przewidywanych $\hat{y}_i$; $\bar{\hat{y}} = n^{-1} \sum_{i=1}^{n} \hat{y}_i$. Zauważmy, że równość $e_i = 0$ dla pewnego $i$ oznacza, że prosta MNK przechodzi przez punkt $(x_i, y_i)$, a spełnienie tej równości dla wszystkich $i$ oznacza, że wszystkie punkty na wykresie rozproszenia są współliniowe i leżą na prostej MNK. Podobnie, z równania (4.3) wynika natychmiast, że $\sum_{i=1}^{n} x_i e_i = 0$. Na podstawie tej równości w połączeniu z równością (4.9) i faktem, że $\hat{y}_i = b_0 + b_1 x_i$, otrzymujemy

$$\sum_{i=1}^{n} \hat{y}_i e_i = 0. \tag{4.10}$$

Równości (4.9) i (4.10) wykorzystamy w dalszym ciągu tego punktu.

**Rozkład całkowitej zmienności zmiennej objaśnianej.** Zastanówmy się teraz nad oceną dobroci dopasowania prostej MNK zbudowanej na podstawie zmiennej objaśniającej $x$. W przypadku, gdybyśmy dysponowali tylko wartościami zmiennej objaśnianej $y_1, \ldots, y_n$, zmienność moglibyśmy ocenić

za pomocą ich wariancji $s_y^2$ lub, pomijając czynnik $(n-1)^{-1}$ w definicji wariancji, za pomocą **całkowitej sumy kwadratów** $SST$

$$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

(ang. *total sum of squares*); będziemy używali tego oznaczenia ze względu na powszechne użycie tego terminu w pakietach. Okazuje się, że zmienność opisana przez $SST$ jest sumą dwóch składników, które też mają interpretację zmienności

$$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = SSE + SSR. \quad (4.11)$$

Pierwszy ze składników nosi nazwę **sumy kwadratów błędów** (ang. *error sum of squares*), a drugi **regresyjnej (lub modelowej) sumy kwadratów** (ang. *regression (model) sum of squares*). Regresyjną sumę kwadratów $SSR$ możemy interpretować jako indeks zmienności wartości przewidywanych $\hat{y}_i$ wokół swojej średniej $\bar{\hat{y}}$ (równej wartości średniej $\bar{y}$), a sumę kwadratów błędów $SSE$ jako indeks zmienności rezyduów wokół swojej średniej równej 0.

Równość (4.11) wynika z oczywistej równości $y_i - \bar{y} = y_i - \hat{y}_i + \hat{y}_i - \bar{y}$, która po podniesieniu do kwadratu i zsumowaniu po $i = 1, \ldots, n$, prowadzi do równości

$$SST = SSE + SSR + 2\sum_{i=1}^{n}(\hat{y}_i - \bar{y})e_i.$$

Ostatni czynnik po prawej stronie jest równy 0 na mocy równości (4.10) i (4.9). Podkreślmy, że równość (4.11) zachodzi w sytuacji, gdy $\hat{y}_i$ oznaczają wartości przewidywane, obliczane na podstawie prostej MNK i nie musi być koniecznie spełniony dla żadnej innej prostej przybliżającej chmurę punktów.

Zauważmy, że regresyjną sumę kwadratów $SSR$ możemy interpretować jako wskaźnik zmienności wartości przewidywanych na podstawie prostej MNK, a więc tej części indeksu zmienności $SST$ wartości $y_i$, którą można objaśnić za pomocą zależności liniowej między zmienną objaśnianą a objaśniającą. Analogicznie, sumę kwadratów błędów SSE możemy traktować jako indeks zmienności wartości $y_i$ nie wyjaśnionej przez zależność liniową. W sytuacji, gdy chmura punktów na wykresie rozproszenia jest silnie skupiona wokół prostej MNK, składnik $SSE$ jest mały w porównaniu ze składnikiem $SST$.

*Zatem stosunek* $SSR/SST = 1 - SSE/SST$ *zwany* **współczynnikiem determinacji** *określa stopień, w jakim zależność liniowa między zmienną objaśnianą a objaśniającą tłumaczy zmienność wykresu rozproszenia.*

**Przykład 4.1 cd.** Z tabeli 4.2 odczytujemy, że całkowita suma kwadratów jest równa $SST = 1320{,}63$, suma kwadratów błędów wynosi $SSE = 69{,}25$, a regresyjna suma kwadratów jest różnicą tych wielkości $SSR = 1251{,}39 = 1320{,}63 - 69{,}25$. Współczynnik determinacji wynosi w tym przypadku $SSR/SST = 1251{,}39/1320{,}63 = 0{,}95$.

Okazuje się, że wartość współczynnika determinacji jest ściśle związana z wartością współczynnika korelacji próbkowej dla próby $(x_1, y_1), \ldots, (x_n, y_n)$.

**Stwierdzenie 4.1.** *Zachodzi równość*

$$r^2 = SSR/SST = \frac{\textit{zmienność wyjaśniona przez model}}{\textit{zmienność całkowita}}.$$

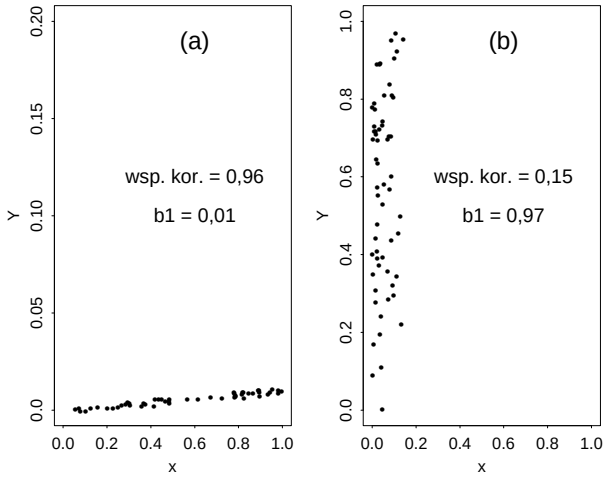
Dowód stwierdzenia jest prostym wnioskiem z równości (4.8) i definicji współczynnika $b_0$. Mianowicie, prawa strona ostatniej równości wynosi

$$\frac{\sum_{i=1}^n (b_1 x_i + b_0 - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = \frac{\sum_{i=1}^n (b_1 x_i - b_1 \bar{x})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = b_1^2 \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} =$$

$$= \frac{\frac{r^2 s_y^2}{s_x^2} \sum_{i=1}^n (x_i - \bar{x})^2}{(n-1) s_y^2} = r^2.$$

Zauważmy, że stwierdzenie 4.1 jest zgodne z własnością (1) współczynnika korelacji. Ścisłe dopasowanie prostej do wykresu odpowiada wartościom współczynnika determinacji bliskim 1, co pociąga za sobą, na mocy stwierdzenia, fakt, że wartość współczynnika korelacji jest bliska 1 lub $-1$. Odnotujmy jednocześnie, że ze stwierdzenia wynika, iż **wartość $r^2$ jest bardziej adekwatnym wskaźnikiem stopnia zależności liniowej niż sama wartość współczynnika korelacji próbkowej** $r$. Trudno jest podać granicę wartości współczynnika $r^2$, od której stopień zależności liniowej uważa się za silny. Jest to bardzo mocno związane ze specyfiką i pochodzeniem danych. Archeolog, badający zależność występowania określonych artefaktów na różnych stanowiskach uzna wartość $r^2 = 0{,}7$ za zadowalającą, gdy dla fizyka badającego zależność pewnych parametrów w ściśle zaplanowanym eksperymencie dolną granicą wartości $r^2$ może być wartość 0,9.

Odnotujmy na zakończenie tego punktu, że powyższa interpretacja współczynnika korelacji oraz równość (4.8) umożliwia właściwą ocenę funkcji jaką pełnią współczynnik kierunkowy $b_1$ i współczynnik korelacji $r$ w analizie zależności. Współczynnik kierunkowy $b_1$ określa, jak „stroma" jest prosta




Rys. 4.3. Wykresy rozproszenia dla przypadku dużej wartości współczynnika korelacji i małej wartości współczynnika kierunkowego $b_1$ i odwrotnie

regresji MNK, czyli umożliwia ocenę, jak istotna jest zmienna niezależna dla wyjaśnienia zmian zmiennej zależnej. Współczynnik korelacji określa, w jakim stopniu chmura punktów na wykresie jest rozproszona wokół prostej MNK. Znajomość wartości jednego ze współczynników nie wyznacza wartości drugiego. Co więcej, można łatwo sobie wyobrazić przybliżony związek liniowy, dla którego współczynnik korelacji jest bliski 1, a prosta MNK jest prawie pozioma (czyli współczynnik kierunkowy jest bliski 0; por. rys. 4.3a), jak również sytuację odwrotną, gdy współczynnik kierunkowy $b_1$ jest duży, a współczynnik korelacji mały (por. rys. 4.3b). W analizie zależności liniowej konieczna jest znajomość *obu* tych współczynników jednocześnie – współczynnik $b_1$ jest odpowiedzialny za **szybkość zmiany zmiennej objaśnianej**, a współczynnik korelacji $r$ za **stopień skupienia wykresu wokół prostej MNK**. Najmniej istotny przy ocenie zależności jest wyraz $b_0$: odgrywa on rolę cechowania i ma interpretację wartości zmiennej objaśnianej przy niewystępowaniu czynnika objaśniającego. Dlatego znacznie większą wagę będzie miało dla nas ustalenie, czy współczynnik kierunkowy $b_1$ jest różny od zera niż czy wyraz wolny $b_0$ przyjmuje pewną konkretną wartość.

### 4.2.3. Model zależności liniowej

Zauważmy, że przedstawione podejście opiera się jedynie na analizie danych w postaci wykresu rozproszenia i nie wymaga żadnych założeń dotyczących zależności między zmienną objaśnianą a objaśniającą. Możemy jed-

nak, jak robiliśmy to już poprzednio, abstrahować od konkretnych wartości $(x_1, y_1), \ldots, (x_n, y_n)$ i potraktować $b_0$ i $b_1$ jako zmienne losowe zdefiniowane na podstawie próby losowej $(x_1, Y_1), \ldots, (x_n, Y_n)$, gdzie $Y_i$ jest objaśnianą zmienną losową odpowiadającą wartości $x_i$ zmiennej objaśniającej $x$. Formalnie odpowiada to podstawieniu we wzorach (4.6) i (4.7) zmiennych losowych $Y_i$ w miejsce wartości $y_i$. Podkreślmy, że wartości $x_i$ traktujemy jako wartości **deterministycznej** zmiennej $x$, które wybieramy w celu obserwacji odpowiadającej zmiennej **losowej** $Y_i$. Badanie własności probabilistycznych współczynników $b_0$ i $b_1$ będzie wymagało sformułowania modelu zależności między wartościami $x_i$ a zmiennymi $Y_i$, $i = 1, 2, \ldots, n$. Ogólne założenie dotyczące powiązania tych zmiennych wygląda następująco:

*Dla pewnych stałych $\beta_0$ i $\beta_1$ zachodzi*

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad i = 1, \ldots, n, \tag{4.12}$$

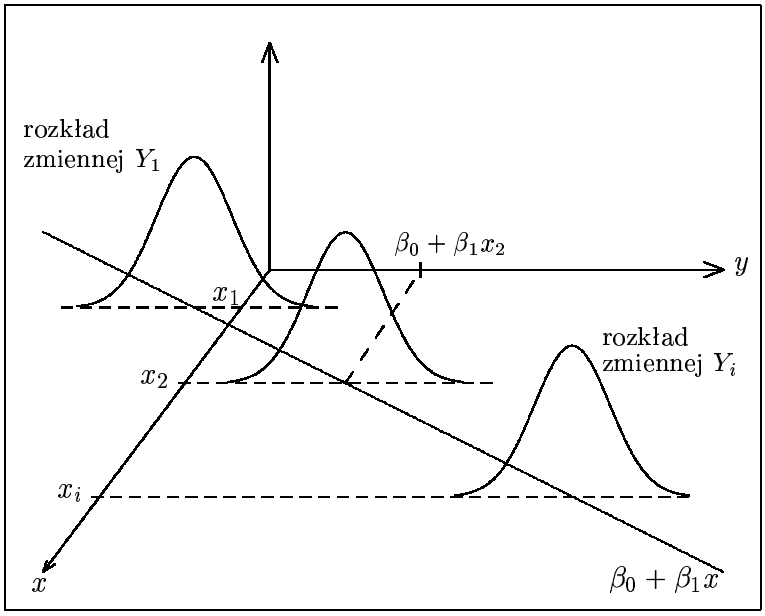
*gdzie $\varepsilon_i$ są niezależnymi zmiennymi losowymi o tym samym rozkładzie ze średnią 0 i wariancją $\sigma^2$. Przyjmujemy ponadto, że wartości $x_i$ nie są wszystkie równe jednej liczbie.*

Prosta $y = \beta_0 + \beta_1 x$ nosi nazwę **prostej regresji**, a współczynniki $\beta_0$ i $\beta_1$ są odpowiednio jej **wyrazem wolnym** i **współczynnikiem kierunkowym**. W tym momencie nie przyjmujemy żadnych założeń dotyczących wspólnego rozkładu zmiennych $\varepsilon_i$, poza tym, że ma on wartość średnią równą 0 i nieznaną wariancję $\sigma^2$. Model zależy więc od **trzech** nieznanych parametrów $\beta_0, \beta_1$ i $\sigma^2$. Zmienne $\varepsilon_i$ noszą nazwę **losowych błędów w modelu regresji**, a ich wariancja $\sigma^2$ wariancji błędów w modelu. Wymieńmy podstawowe własności rozkładów zmiennych $Y_i$ wynikające z założenia (4.12). Zauważmy najpierw, że wartość średnia zmiennej $Y_i$ jest równa

$$\mu_{Y_i} = E(\beta_0 + \beta_1 x_i) + E(\varepsilon_i) = \beta_0 + \beta_1 x_i. \tag{4.13}$$

Odpowiada to następującej intuicji: wartość obserwowana w eksperymencie jest równa funkcji liniowej (ściślej afinicznej) zmiennej objaśniającej plus pewien błąd poczyniony podczas obserwacji, którego wartość średnia jest 0. Uśrednienie wielu obserwacji $y$ poczynionych dla ustalonej wartości $x$ powinno dać nam zatem, na mocy prawa wielkich liczb, wartość bliską $\beta_0 + \beta_1 x$. Powód, dla którego rozpatrywany model nazywamy modelem zależności liniowej, mimo że zależność wartości średniej $\mu_{Y_i}$ od zmiennej objaśniającej $x_i$ jest afiniczna[1] , a nie liniowa, omówimy w p. 4.2.5.

[1] Zależność $y = ax + b$ dla $b \neq 0$ nazywamy zależnością afiniczną.




Rys. 4.4. Związek między rozkładami zmiennych $Y_i$

Analogicznie, rozpatrując wariancję zmiennej $Y_i$, otrzymujemy

$$\sigma^2_{Y_i} = \mathrm{Var}(\beta_0 + \beta_1 x_i) + \mathrm{Var}(\varepsilon_i) = \sigma^2.$$

Wariancja zmiennych losowych $Y_i$ jest zatem stała i równa wariancji błędów $\sigma^2$. Związek między rozkładami zmiennych $Y_i$ przedstawiono na rys. 4.4, wszystkie z nich mają rozkład o dokładnie takim samym kształcie, jedyną różnicą między nimi jest tylko zmienna wartość średnia, która dla rozkładu $Y_i$ wynosi $\beta_0 + \beta_1 x_i$. Zatem w przypadku, gdy $\beta_1 \neq 0$, czyli gdy mamy do czynienia z rzeczywistą zależnością zmiennej objaśnianej od objaśniającej, zmienne $Y_i$ **nie** mają tego samego rozkładu. Oznacza to, że próba $Y_i, i = 1, \ldots, n$ nie jest prostą próbą losową, ale staje się nią, gdy od każdej ze zmiennych odjąć jej wartość średnią.

W eksperymencie obserwujemy wartości zmiennej objaśniającej $x$ równe $x_i, i = 1, \ldots, n$ i odpowiadające jej zmienne objaśniane $Y_i$: $(x_1, Y_1), \ldots$ $\ldots, (x_n, Y_n)$. Przy przyjęciu, że jest spełniony warunek (4.12), naszym celem jest wnioskowanie o nieznanych współczynnikach $\beta_0$ i $\beta_1$. Oczywiście, naturalnymi estymatorami tych współczynników są zmienne losowe

$$b_0 = \bar{Y} - b_1 \bar{x},$$

$$b_1 = \sum_{i=1}^{n} (x_i - \bar{x})(Y_i - \bar{Y}) / \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} (x_i - \bar{x}) Y_i / \sum_{i=1}^{n} (x_i - \bar{x})^2$$

określone w równaniach (4.6) i (4.7); $\bar{Y} = n^{-1}\sum_{i=1}^{n} Y_i$. Wyznaczona przez nie prosta MNK, równa $b_0+b_1x$, jest oszacowaniem nieznanej prostej regresji $\beta_0+\beta_1 x$. Zatem adekwatność przybliżenia wykresu rozproszenia przez prostą MNK zależy od własności współczynników $b_0$ i $b_1$. Zanim zajmiemy się zbadaniem własności rozkładów estymatorów $b_0$ i $b_1$, udowodnimy pewien potrzebny nam fakt.

**Stwierdzenie 4.2.** $\mathrm{Cov}(\bar{Y}, b_1) = 0$.

W celu udowodnienia stwierdzenia, zauważmy, że korzystając z równości $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$, definicji $b_1$ i definicji (4.12), mamy

$$b_1 = \beta_1 \frac{\sum_{i=1}^{n}(x_i-\bar{x})x_i}{\sum_{i=1}^{n}(x_i-\bar{x})^2} + \frac{\sum_{i=1}^{n}(x_i-\bar{x})\varepsilon_i}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \beta_1 + \frac{\sum_{i=1}^{n}(x_i-\bar{x})\varepsilon_i}{\sum_{i=1}^{n}(x_i-\bar{x})^2}. \quad (4.14)$$

Zatem oznaczając przez $\bar{\varepsilon}$ średnią błędów $\varepsilon_i$, korzystając z niezależności błędów i własności $\mathrm{Cov}(X+a, Y+b) = \mathrm{Cov}(X,Y)$ otrzymamy, że kowariancja $\mathrm{Cov}(\bar{Y}, b_1)$ jest równa

$$\mathrm{Cov}\Big(\bar{\varepsilon}, \frac{\sum_{i=1}^{n}(x_i-\bar{x})\varepsilon_i}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\Big) = \frac{1}{n}\frac{\sum_{i=1}^{n}\mathrm{Var}(\varepsilon_i,\varepsilon_i)(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} =$$

$$= \frac{\sigma^2}{n}\frac{\sum_{i=1}^{n}(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = 0,$$

gdyż

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n}(\beta_1 x_i + \beta_0) + \bar{\varepsilon}. \quad (4.15)$$

Wartości średnie i wariancje estymatorów $b_0$ i $b_1$ są podane w następnym stwierdzeniu.

**Stwierdzenie 4.3.** *Zmienne losowe $b_0$ i $b_1$ są nieobciążonymi estymatorami współczynników $\beta_0$ i $\beta_1$: $\mu_{b_0} = \beta_0$ i $\mu_{b_1} = \beta_1$. Ich wariancje wynoszą*

$$\sigma^2_{b_0} = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right) \quad (4.16)$$

i

$$\sigma^2_{b_1} = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}. \quad (4.17)$$

Udowodnijmy powyższe własności. Zauważmy, że ponieważ wartości średnie błędów są równe 0, z równości (4.15) wynika, że

$$\mu_{\bar{Y}} = \frac{1}{n}\Big(\sum_{i=1}^{n} \beta_0 + \beta_1 x_i\Big) = \beta_0 + \beta_1 \bar{x},$$

a zatem na mocy równości (4.6)

$$\mu_{b_0} = \mu_{\bar{Y}} - \mu_{b_1}\bar{x} = \beta_0 + \beta_1\bar{x} - \mu_{b_1}\bar{x} = \beta_0 + \bar{x}(\beta_1 - \mu_{b_1}).$$

Zatem, aby udowodnić nieobciążoność estymatora $b_0$, wystarczy udowodnić nieobciążoność estymatora $b_1$. Ale to prosto wynika z równości (4.14) mówiącej, że estymator $b_1$ różni się od współczynnika $\beta_1$ o kombinację liniową zmiennych $\varepsilon_i$ o wartości średniej 0.

W celu udowodnienia własności (4.17), zauważmy, że zmienną $b_1$ można zapisać równoważnie jako $b_1 = \sum_{i=1}^{n}(x_i - \bar{x})Y_i / \sum_{i=1}^{n}(x_i - \bar{x})^2$, a zatem

$$\sigma^2_{b_1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2\sigma^2_{Y_i}}{\Big(\sum_{i=1}^{n}(x_i - \bar{x})^2\Big)^2} = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Równość (4.16) wynika prosto ze stwierdzenia 2.9 oraz równości (4.15) na podstawie równości $b_0 = \bar{y} - b_1\bar{x}$. Mianowicie, ponieważ średnia $\bar{Y}$ jest nieskorelowana z estymatorem $b_1$, więc z równości (4.15) wynika, że $\sigma^2_{\bar{Y}} = \sigma^2/n$, skąd

$$\sigma^2_{b_0} = \sigma^2_{\bar{Y}} + \sigma^2_{\bar{x}b_1} - 2\mathrm{Cov}(\bar{Y}, \bar{x}b_1) = \sigma^2_{\bar{Y}} + (\bar{x})^2\sigma^2_{b_1} = \frac{\sigma^2}{n} + \frac{\sigma^2\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

### 4.2.4. Wnioskowanie w modelu zależności liniowej

Zajmowaliśmy się dotąd tylko oszacowaniem współczynników prostej regresji $\beta_0$ i $\beta_1$. Zauważmy, że nie mniej ważne jest oszacowanie trzeciego nieznanego parametru modelu: wariancji błędów $\sigma^2$. Analizując wzory (4.16) i (4.17), dostrzeżemy, że bez oszacowania wariancji $\sigma^2$ niemożliwe jest oszacowanie wariancji $\sigma^2_{b_0}$ i $\sigma^2_{b_1}$, a bez tego nie mamy informacji na temat zmienności estymatorów $b_0$ i $b_1$. Estymator wariancji błędów $\sigma^2$ jest oparty na następującej intuicji. Rezydua $Y_i - \hat{Y}_i$ będąc odchyłkami obserwacji $Y_i$ od nieznanej wartości prostej regresji MNK są empirycznymi odpowiednikami odchyłek obserwacji $Y_i$ od nieznanej prostej regresji $\beta_0 + \beta_1 x_i$. Zatem wariancja rezyduów w próbie powinna być naturalnym oszacowaniem wariancji

błędów $\sigma^2$. Pamiętając o tym, że średnia rezyduów jest równa 0 i zastępując w definicji wariancji próbkowej czynnik $(n-1)^{-1}$ czynnikiem $(n-2)^{-1}$, otrzymamy następujący estymator wariancji $\sigma^2$.

**Definicja 4.3.** ***Błędem średniokwadratowym*** $S^2$ *nazywamy estymator wariancji* $\sigma^2$ *określony jako wariancja próbkowa rezyduów pomnożona przez czynnik* $(n-1)/(n-2)$

$$S^2 = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2} = \frac{SSE}{n-2}.$$

Oznaczenie błędu średniokwadratowego jest takie samo jak wariancji; w podrozdz. 4.2 przez symbol $S^2$ będziemy zawsze rozumieli błąd średniokwadratowy. Czynnik $(n-2)$ występujący w mianowniku wyrażenia określającego $S^2$ nosi nazwę liczby stopni swobody rezyduów. Okazuje się, że właśnie dla takiego mnożnika *estymator* $S^2$ *jest nieobciążonym estymatorem* $\sigma^2$: $\mu_{S^2} = \sigma^2$. Fakt występowania w definicji czynnika $(n-2)^{-1}$ często jest tłumaczony następująco: aby otrzymać nieobciążony estymator wariancji $\sigma^2$ dzielimy sumę kwadratów rezyduów przez ich liczbę $n$ pomniejszoną o liczbę niezależnych ograniczeń, które spełniają wartości resztowe. Ponieważ mamy dwa takie ograniczenia zadane przez warunki $\sum_{i=1}^{n} e_i = 0$ i $\sum_{i=1}^{n} e_i x_i = 0$, liczba stopni swobody przypisana $SSE$ wynosi $(n-2)$. Analogicznie, liczba stopni swobody przypisana $SST$ wynosi $n-1$, ponieważ $Y_1, \ldots, Y_n$ spełniają warunek $n^{-1}\sum_{i=1}^{n} Y_i = \bar{Y}$; a liczba stopni swobody dla $SSR$ wynosi 1 (wartości $\hat{Y}_i$ są w pełni określone przez dwa parametry i spełniają jeden warunek $n^{-1}\sum_{i=1}^{n} \hat{Y}_i = \bar{Y}$). Równanie (4.11) jest często uzupełniane o równanie dla liczby stopni swobody $n-1 = n-2+1$ zapisywane jako

*Całkowita liczba stopni swobody =*

*= Liczba stopni swobody rezyduów + Liczba stopni swobody modelu.*

Stwierdzenie 4.3 na podstawie def. 4.3 prowadzi do następującej definicji błędów standardowych estymatorów $b_0$ i $b_1$.

**Definicja 4.4.** *Błędy standardowe* $SE_{b_0}$ *i* $SE_{b_1}$ *otrzymujemy jako dodatnie pierwiastki kwadratowe następujących wyrażeń:*

$$SE_{b_0}^2 = S^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)$$

*i*

$$SE_{b_1}^2 = \frac{S^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Oczywiście, ponieważ $S^2$ jest nieobciążonym estymatorem $\sigma^2$, wyrażenia $SE^2_{b_0}$ i $SE^2_{b_1}$ są nieobciążonymi estymatorami wariancji $\sigma^2_{b_0}$ i $\sigma^2_{b_1}$.

**Przykład 4.1 cd.** Z tabeli 4.2 odczytujemy, że estymator wariancji $S^2$ wynosi 4,07, a błędy standardowe dla estymatorów $b_1$ i $b_0$ wynoszą odpowiednio 0,10 i 1,57.

### Testy dotyczące współczynników $\beta_0$ i $\beta_1$

W poprzednim punkcie wyprowadziliśmy ogólne wzory na wartości średnie i wariancje wyrazu wolnego $b_0$ i nachylenie $b_1$ prostej MNK. Bez wprowadzenia dodatkowych założeń nie jesteśmy w stanie ustalić dokładnego rozkładu tych estymatorów. Aby to zrobić, musimy wiedzieć, jaki jest wspólny rozkład błędów $\varepsilon_i$. Poza podstawowym założeniem (4.12) przyjmijmy od tej pory dodatkowe założenie

*Rozkład każdego z błędów $\varepsilon_i$ jest normalny.*

Tak więc przyjmujemy, że jest spełnione równanie (4.12), a błędy $\varepsilon_i$, $i = 1, \ldots, n$ tworzą prostą próbę losową z rozkładu normalnego $N(0, \sigma)$. W tym przypadku mówimy, że są spełnione założenia liniowego modelu regresji jednokrotnej. Podkreślmy, że podobnie jak założenia (4.12), założenia o rozkładzie normalnym błędów nie możemy również przyjąć w sposób automatyczny. Analiza rezyduów służąca jego weryfikacji zostanie omówiona w następnym punkcie. Z założeń modelu możemy wyprowadzić rozkład estymatorów współczynników prostej regresji jak również ich wersji studentyzowanych.

**Twierdzenie 4.1.** (1) *Rozkład estymatora $b_1$ jest rozkładem normalnym $N(\beta_1, \frac{\sigma}{\{\sum_{i=1}^n (x_i - \bar{x})^2\}^{1/2}})$. Ponadto, dla studentyzowanego estymatora $b_1$ zachodzi*

$$\frac{b_1 - \beta_1}{SE_{b_1}} = \frac{(b_1 - \beta_1)\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}{S} \sim t_{n-2}. \tag{4.18}$$

(2) *Rozkład estymatora $b_0$ jest rozkładem normalnym*

$$N(\beta_0, \sigma(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2})^{1/2}).$$

*Ponadto dla studentyzowanego estymatora $b_0$ zachodzi*

$$\frac{b_0 - \beta_0}{SE_{b_0}} = \frac{(b_0 - \beta_0)}{S\sqrt{\frac{1}{n} + \bar{x}^2 / \sum_{i=1}^{n}(x_i - \bar{x})^2}} \sim t_{n-2}. \tag{4.19}$$

Część twierdzenia dotycząca normalności estymatorów jest wnioskiem ze stwierdzenia mówiącego, że kombinacja liniowa zmiennych losowych o rozkładzie normalnym ma rozkład normalny. Ponieważ z równości (4.14) wynika, że estymator $b_1$ jest kombinacją liniową błędów $\varepsilon_i$, a więc estymator $b_1$ ma rozkład normalny. Jego wartość średnia i wariancja została ustalona w stwierdzeniu 4.3. Ponadto, ponieważ $b_0 = \bar{y} - \bar{x}b_1$, estymator wyrazu wolnego jest również kombinacją zmiennych losowych o rozkładzie normalnym. Zgodnie ze stwierdzeniem 4.3 jego wartość średnia wynosi $\beta_0$, a wariancja $\sigma^2(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2})$. Przyjęcie założenia o normalności błędów umożliwia sformułowanie jeszcze jednej ważnej własności estymatorów $b_0$ i $b_1$. Okazuje się, że w tej sytuacji są one nie tylko estymatorami MNK, ale również estymatorami największej wiarogodności parametrów $\beta_0$ i $\beta_1$ odpowiednio. Z własności tej wynika m.in., że są one również estymatorami NMW (nieobciążonymi o minimalnej wariancji).

Użycie metodologii przedstawionej w rozdz. 3 umożliwia wykorzystanie równości (4.18) i (4.19) do konstrukcji przedziałów ufności dla współczynników $\beta_1$ i $\beta_0$ prostej regresji na podstawie estymatorów $b_1$ i $b_0$ oraz odpowiednich statystyk testowych. *Podkreślmy raz jeszcze, że konstrukcja ta jest prawidłowa, to znaczy przedziały mają postulowany poziom ufności, gdy jest spełnione równanie (4.12) i błędy mają rozkład normalny $N(0, \sigma)$.*

**Stwierdzenie 4.4.** *Przedział ufności na poziomie ufności $1 - \alpha$ dla współczynnika $\beta_1$ ma postać*

$$b_1 \pm t_{1-\alpha/2,n-2} \times SE_{b_1} = b_1 \pm t_{1-\alpha/2,n-2} \frac{S}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}.$$

*Przedział ufności na poziomie ufności $1-\alpha$ dla wyrazu wolnego $\beta_0$ ma postać*

$$b_0 \pm t_{1-\alpha/2,n-2} \times SE_{b_0} = b_0 \pm t_{1-\alpha/2,n-2} S \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}.$$

Na podstawie powyższych wzorów można skonstruować przedział ufności dla współczynnika kierunkowego $\beta_1$ dla danych z przykł. 4.1, przy założeniu, że spełniają one założenia modelu liniowego. Przedział ufności na poziomie 95% jest równy $(1{,}5, 1{,}8)$ i daje nam precyzyjne oszacowanie najbardziej prawdopodobnych wartości współczynnika kierunkowego. Zauważmy,

że długość przedziału ufności jest związana ze zmiennością danych wyjaśnianą przez model liniowy. Dla przykładu 4.2 współczynnik determinacji $R^2 = 0{,}26$, zatem model regresji jednokrotnej wyjaśnia jedynie 26% zmienności danych. Jednocześnie przedział ufności dla współczynnika kierunkowego wynosi $(10{,}5, 26{,}7)$ i jest zbyt szeroki z praktycznego punktu widzenia.

Równości (4.18) i (4.19) można również wykorzystać do skonstruowania statystyk testowych dla testowania na poziomie istotności $\alpha$ hipotezy

$$H_0\colon\ \beta_1 = \beta_{1,0} \qquad \text{przeciwko alternatywie} \qquad H_1\colon\ \beta_1 \neq \beta_{1,0},$$

gdzie $\beta_{1,0}$ jest pewną ustaloną liczbą. Statystyka testowa mająca postać

$$t = \frac{b_1 - \beta_{1,0}}{S}\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{4.20}$$

ma **przy prawdziwości hipotezy** $H_0$ rozkład $t_{n-2}$. Zatem obszar krytyczny testu hipotezy $H_0$ przeciwko hipotezie $H_1$ na poziomie istotności $\alpha$ ma postać

$$\{t\colon\ |t| \geq t_{1-\alpha/2,n-2}\}.$$

Analogicznie, przy zachowaniu tej samej hipotezy zerowej i zmianie hipotezy alternatywnej na jednostronną postaci $H_1\colon\ \beta_1 > \beta_{1,0}$ obszar krytyczny ma postać

$$\{t\colon\ t \geq t_{1-\alpha,n-2}\},$$

a dla hipotezy alternatywnej $H_1\colon\ \beta_1 < \beta_{1,0}$ ma postać

$$\{t\colon\ t \leq -t_{1-\alpha,n-2}\}.$$

Pakiety statystyczne podają z reguły $p$–wartość dla odpowiednich testów dwustronnych. Przykładowo, w przypadku testowania hipotezy $H_0\colon\ \beta_0 = 0$ przeciwko hipotezie alternatywnej $H_1\colon\ \beta_0 \neq 0$ w przykł. 4.1 statystyka $t$ wynosi 3,32, a odpowiednia $p$–wartość (dla 17 stopni swobody) wynosi 0,004. Jednakże $p$–wartość dla testu tej samej hipotezy zerowej przeciwko jednostronnej hipotezie alternatywnej $H_1\colon\ \beta_0 > 0$ wynosi 0,002.

Inne podejście do testowania hipotezy $H_0\colon\ \beta_1 = 0$ przeciwko alternatywie $H_1\colon\ \beta_1 \neq 0$ można zaproponować na podstawie rozkładu zmienności zmiennej objaśnianej omówionego w p. 4.2.2. Okazuje się, że przy spełnieniu hipotezy $H_0$, regresyjna suma kwadratów $SSR$ ma rozkład $\chi^2$ z 1 stopniem swobody, a suma kwadratów błędów ma rozkład $\chi^2$ z $n-2$ stopniami swobody. Ponadto, zmienne $SSE$ i $SSR$ są niezależne. Zatem zgodnie z rozważaniami w p. 3.3.2 statystyka

$$F = \frac{SSR/1}{SSE/(n-2)} \quad \text{ma rozkład } F_{1,n-2},$$

gdzie $F_{1,n-2}$ oznacza rozkład $F$ Snedecora z parametrami $(1, n-2)$. Stąd test o zbiorze krytycznym $\{F\colon F \geq f_{1-\alpha,1,n-2}\}$ jest testem na poziomie istotności $\alpha$. Zauważmy, że jest to zgodne z intuicją, gdyż dla $\beta_1 \neq 0$ regresyjna suma kwadratów powinna przyjmować duże wartości w stosunku do sumy kwadratów błędów. Zatem duże wartości statystyki $F$ powinny wskazywać na niespełnienie hipotezy $H_0$. Okazuje się jednak, że w wyniku tego rozumowania nie otrzymujemy żadnego nowego testu. Mianowicie, w sytuacji jednej zmiennej objaśniającej $F = t^2$, gdzie $t$ jest statystyką zdefiniowaną w (4.20) dla $\beta_{1,0} = 0$, a zatem wynik testu opartego na statystyce $F$ jest taki sam jak wynik testu opartego na statystyce $t$. W przykładzie 1 wartość statystyki $F$ wynosi 307,22 i jest równa wartości statystyki $t = 17{,}528$ podniesionej do kwadratu. Również $p$–wartości dla testowania $H_0$: $\beta_1 = 0$ przeciwko $H_1$: $\beta_1 \neq 0$ są takie same i wynoszą 0,0001. Zauważmy jednocześnie, że statystyki $F$, w odróżnieniu od statystyki $t$ nie możemy użyć w przypadku testowania jednostronnej hipotezy alternatywnej.

**Problem prognozy.** Dotąd dopasowanie modelu regresji liniowej służyło nam do analizy charakteru zależności między rozpatrywanymi zmiennymi. Przypuśćmy, że dopasowanie modelu do danych uznaliśmy za zadowalające: wartość współczynnika determinacji jest duża, a rezydua nie wskazują na istnienie wyraźnych odstępstw od założeń modelowych (problem ten zostanie dokładniej omówiony w następnym punkcie). Wtedy postulowany model może być użyty również do innych celów. Możemy starać się przewidzieć, jak będzie się zachowywać zmienna objaśniana, gdy zmienna objaśniająca przyjmie nową wartość $x_0$ różną od dotychczasowych wartości zmiennych objaśniających $x_1, \ldots, x_n$. W przykładzie pierwszym może interesować nas na przykład jaka jest średnia wartość wyniku egzaminu końcowego dla osób, które uzyskały z kolokwium 18 punktów. Żeby móc rozwiązać tak postawione zadanie prognozy, wartość $x_0$ nie może znacząco odstawać od „centralnej" części zbioru wartości zmiennej objaśniającej. W przeciwnym przypadku dokonamy nieuzasadnionej **ekstrapolacji** na zakres wartości $x$, o którym na podstawie danych nie mamy żadnych, albo bardzo mało informacji. Załóżmy, że powyższy warunek jest spełniony. Wtedy może interesować nas odpowiedź na dwa *różne* pytania. Możemy chcieć ocenić

**wartość średnią** *zmiennej objaśnianej w sytuacji, gdy zmienna objaśniająca $x$ jest równa $x_0$,*

lub chcieć przewidzieć

**przyszłą wartość** *zmiennej objaśnianej przy tym samym warunku $x = x_0$.*

Zauważmy, że nie zetknęliśmy się dotychczas z problemem podobnym do drugiego problemu: dotąd zawsze staraliśmy się estymować pewien nieznany, ale stały (na mocy definicji!) parametr rozkładu. Omówimy po kolei te

dwa problemy, zauważając najpierw fakt istotny dla analizy obydwu z nich. Mianowicie, z analizy modelu wynika, że możemy przyjąć, iż obserwacja $Y(x_0)$ dla $x = x_0$ będzie spełniała następujące równanie:

$$Y(x_0) = \beta_0 + \beta_1 x_0 + \varepsilon_0, \tag{4.21}$$

gdzie $\varepsilon_0$ jest zmienną losową o rozkładzie normalnym $N(0, \sigma)$, niezależną od zmiennych $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$. To ostatnie założenie jest naturalne w świetle założenia, że wartość $x_0$ jest różna od wartości $x_1, x_2, \ldots, x_n$.

**Przewidywanie (prognoza) wartości średniej zmiennej $Y(x_0)$.** Z równania (4.21) po obliczeniu i porównaniu wartości średnich jego obydwu stron wynika, że

$$\mu_{Y(x_0)} = \beta_0 + \beta_1 x_0.$$

Oczywistym oszacowaniem tej wartości będzie wartość prostej MNK dla argumentu $x = x_0$

$$\hat{Y}(x_0) = b_0 + b_1 x_0. \tag{4.22}$$

Z nieobciążoności estymatorów $b_0$ i $b_1$ wynika, że estymator $\hat{Y}(x_0)$ jest nieobciążonym estymatorem wartości $\mu_{Y(x_0)}$, mianowicie

$$\mu_{\hat{Y}(x_0)} = E(b_0 + x_0 b_1) = \beta_0 + \beta_1 x_0.$$

Podobnie możemy obliczyć wariancję estymatora, pamiętając, że estymator $b_1$ jest nieskorelowany ze średnią $\bar{Y}$. Mamy

$$\begin{aligned} \sigma^2_{\hat{Y}(x_0)} &= \mathrm{Var}(b_0 + b_1 x_0) = \mathrm{Var}(\bar{Y} + b_1(x_0 - \bar{x})) = \\ &= \sigma^2_{\bar{Y}} + (x_0 - \bar{x})^2 \sigma^2_{b_1} = \sigma^2 \Big( \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \Big). \end{aligned} \tag{4.23}$$

Możemy zatem, podobnie jak w poprzednim punkcie, zdefiniować błąd standardowy estymatora $\hat{Y}(x_0)$ następująco:

$$SE_{\hat{Y}(x_0)} = S \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

i udowodnić następujące stwierdzenie.

**Stwierdzenie 4.5.** *Estymator $\hat{Y}(x_0)$ wartości średniej $\mu_{Y(x_0)}$ zmiennej objaśnianej $Y$ dla wartości zmiennej objaśniającej $x$ równej $x_0$ ma rozkład normalny $N(\mu_{Y(x_0)}, \sigma_{\hat{Y}(x_0)})$. Ponadto,*

$$\frac{\hat{Y}(x_0) - \mu_{Y(x_0)}}{SE_{\hat{Y}(x_0)}} \sim t_{n-2}.$$

Stwierdzenie 4.5 prowadzi do następującej postaci przedziału ufności na poziomie ufności $1-\alpha$ dla wartości średniej $\mu_{Y(x_0)} = \beta_0 + \beta_1 x_0$:

$$\hat{Y}(x_0) \pm t_{1-\alpha/2,n-2} SE_{\hat{Y}(x_0)}. \tag{4.24}$$

Zauważmy, że długość przedziału ufności nie jest stała dla wszystkich $x_0$: im punkt $x_0$ jest bardziej oddalony od średniej $\bar{x}$, tym przedział jest dłuższy. Im dalej znajdujemy się od centralnej części wartości zmiennej objaśniającej, tym bardziej prognoza staje się niepewna.

**Przewidywanie (prognoza) przyszłej wartości $Y(x_0)$.** Podobnie jak w przypadku prognozy wartości średniej, jako estymator przyszłej wartości $Y(x_0)$ służy nam $\hat{Y}(x_0)$ w (4.22). Zauważmy, że w tym przypadku mamy trudniejsze zadanie: staramy się estymować nie wartość stałą, ale zmienną losową. Oceńmy, jak duża jest zmienność różnicy między estymatorem $\hat{Y}(x_0)$ a zmienną $Y(x_0)$. Ponieważ zmienne $\hat{Y}(x_0)$ i $Y(x_0)$ są niezależne, więc na mocy (4.23)

$$\sigma^2_{\hat{Y}(x_0)-Y(x_0)} = \sigma^2_{\hat{Y}(x_0)} + \sigma^2_{Y(x_0)} = \sigma^2_{\hat{Y}(x_0)} + \sigma^2 = \sigma^2\Big(1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{\sum_{i=1}^n (x-x_i)^2}\Big).$$

Zauważmy, że wartość $\sigma^2_{\hat{Y}(x_0)-Y(x_0)}$ różni się o wielkość $\sigma^2$ od wartości $\sigma^2_{\hat{Y}(x_0)}$. Jest to zrozumiałe w świetle definicji (4.21) i (4.22). Ostatnie równanie prowadzi do naturalnej definicji błędu standardowego estymatora różnicy $\hat{Y}(x_0) - Y(x_0)$

$$SE_{\hat{Y}(x_0)-Y(x_0)} = S\sqrt{1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{\sum_{i=1}^n (x_i-\bar{x})^2}}.$$

Można teraz sformułować stwierdzenie analogiczne do stwierdzenia 4.5.

**Stwierdzenie 4.6.** *Różnica $\hat{Y}(x_0) - Y(x_0)$ ma rozkład normalny $N(0, \sigma_{\hat{Y}(x_0)-Y(x_0)})$. Ponadto,*

$$\frac{\hat{Y}(x_0) - Y(x_0)}{SE_{\hat{Y}(x_0)-Y(x_0)}} \sim t_{n-2}.$$

Oczywiście na mocy stwierdzenia 4.6 przedział ufności dla zmiennej $Y(x_0)$ wynosi

$$\hat{Y}(x_0) \pm t_{1-\alpha/2,n-2} SE_{\hat{Y}(x_0)-Y(x_0)}. \tag{4.25}$$

**Przykład 4.1 cd.** Dla przykładu 4.1 wartość przewidywana $\hat{Y}(18)$ wynosi $5{,}2+1{,}76\times 18 = 36{,}9$ i jest zarówno estymatorem wartości średniej zmiennej objaśnianej dla wyniku kolokwium równym 18 punktów, jak i estymatorem losowego wyniku egzaminu studenta, który uzyskał tyle punktów na kolokwium. Przedział ufności na poziomie 95% dla średniej $\mu_{Y(18)}$ wynosi $36{,}9 \pm 1{,}1$. Przedział ufności na tym samym poziomie ufności dla zmiennej $\hat{Y}(x_0)$ wynosi $36{,}9 \pm 4{,}4$.

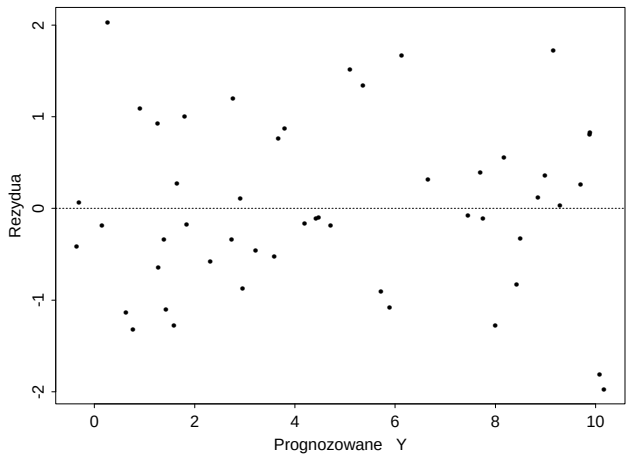
### 4.2.5. Analiza wartości resztowych

Podkreśliliśmy kilkakrotnie w p. 4.2.2–4.2.4, że poprawność testów dotyczących parametrów modelu oraz prognozy przyszłych wartości zależy w istotny sposób od poprawności postulowanego modelu. Dlatego jest niezwykle ważne, aby ocenić, czy dane nie wskazują na istotne odstępstwa od przyjętych założeń. Służy temu analiza rezyduów, która opiera się na tym samym uzasadnieniu, które poprzednio doprowadziło nas do definicji estymatora wariancji błędów $\sigma^2$. Wartość resztową $e_i = Y_i - \hat{Y}_i = Y_i - (b_0 x_i + b_1)$ można bowiem traktować jako przybliżenie błędu $\varepsilon_i = Y_i - (\beta_0 + \beta_1 x_i)$.

Zastanówmy się najpierw, jak zachowują się rezydua, gdy model regresji liniowej jest poprawny, czyli załóżmy, że jest spełnione równanie (4.12), przy czym błędy $\varepsilon_i$ mają rozkład normalny $N(0, \sigma)$. W tym przypadku ciąg rezyduów powinien zachowywać się w przybliżeniu tak, jak ciąg niezależnych zmiennych losowych o rozkładzie normalnym. W szczególności, wykres rezyduów względem numeru porządkowego lub odpowiedniej wartości zmiennej objaśniającej powinien przedstawiać chmurę punktów skupioną dookoła osi $x$ i nie mającą żadnej wyraźnej struktury czy tendencji, podobnie jak chmura punktów przedstawiona na rys. 4.5. W istocie, ciąg rezyduów obserwacji względem prostej MNK w modelu zależności liniowej (4.12) stanowi *w przybliżeniu* ciąg niezależnych zmiennych losowych o tym samym rozkładzie. Pewną trudność może powodować fakt, że rezydua nie są całkowicie niezależne (przypomnijmy, że suma rezyduów wynosi 0!) i nie mają tej samej wariancji, nawet w przypadku adekwatności liniowego modelu regresji. W celu stwierdzenia, jak bardzo może się różnić zmienność poszczególnych rezyduów, obliczymy, ile wynosi wariancja $i$-tego z nich.

**Stwierdzenie 4.7.** *Wariancja rezyduum $e_i = Y_i - \hat{Y}_i$ ma postać*

$$\sigma^2_{e_i} = \sigma^2\Big(1 - \Big(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\Big)\Big).$$




Rys. 4.5. Typowy wykres rezyduów dla liniowego modelu regresji

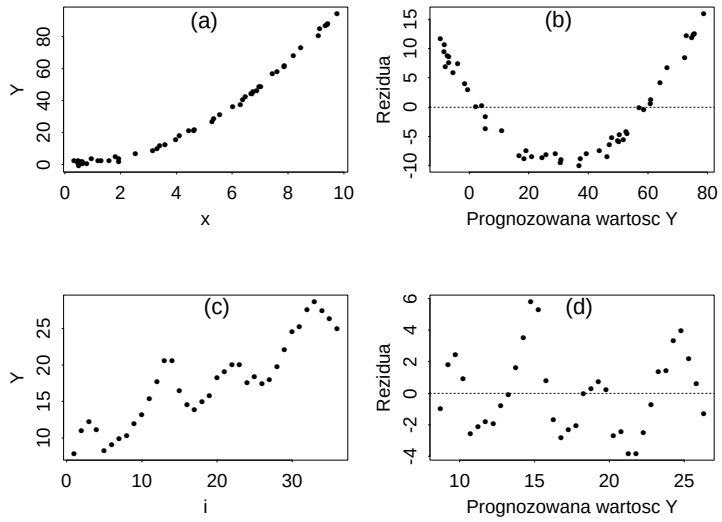
Dowód stwierdzenia wynika z definicji rezyduów i równości (4.7)

$$\sigma^2_{e_i} = \sigma^2_{Y_i} + \sigma^2_{\hat{Y}_i} - 2\mathrm{Cov}(Y_i, \hat{Y}_i) =$$
$$= \sigma^2 + \sigma^2(\tfrac{1}{n} + \tfrac{(x_i-\bar{x})^2}{\sum_{i=1}^n (x_i-\bar{x})^2}) - 2\mathrm{Cov}(Y_i, \bar{Y} + b_1(x_i - \bar{x})) =$$
$$= \sigma^2 + \sigma^2(\tfrac{1}{n} + \tfrac{(x_i-\bar{x})^2}{\sum_{i=1}^n (x_i-\bar{x})^2}) - 2(\tfrac{\sigma^2}{n} + \tfrac{\sigma^2(x_i-\bar{x})^2}{\sum_{i=1}^n (x_i-\bar{x})^2}) =$$
$$= \sigma^2(1 - (\tfrac{1}{n} + \tfrac{(x_i-\bar{x})^2}{\sum_{i=1}^n (x_i-\bar{x})^2})).$$

Powyższe stwierdzenie prowadzi do następującej definicji **błędu standardowego rezyduum** $e_i$

$$SE_{e_i} = S\sqrt{1 - \Big(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\Big)}$$

oraz **studentyzowanego rezyduum** $r_i = e_i/SE_{e_i}$. W przypadku dużej liczności próby wariancja $\sigma^2_{e_i}$ jest w przybliżeniu równa wariancji błędów $\sigma^2$ i podobnie błąd $SE_{e_i}$ jest w przybliżeniu równy błędowi $S$. W takim przypadku nie ma znaczenia, czy rozpatrujemy wykres rezyduów, czy wykres rezyduów studentyzowanych. Jednakże w przypadku małej liczności danych, dla których wartości zmiennej objaśniającej nie są w miarę równomiernie rozłożone, niektóre błędy $SE_{e_i}$ mogą znacznie odbiegać od błędu $S$. W takim przypadku warto w analizie rezyduów zastąpić rezydua rezyduami




Rys. 4.6. Charakter zależności zmiennej objaśnianej od objaśniającej

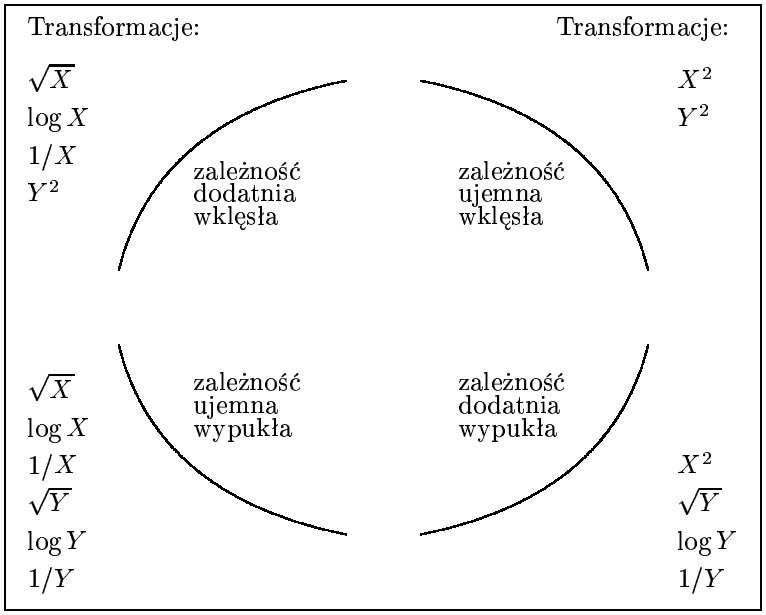
studentyzowanymi. Problem ten ma dużo większe znaczenie w przypadku występowania wielu zmiennych objaśniających i powrócimy do niego w następnym podrozdziale.

Spróbujmy teraz odpowiedzieć na pytanie, jak można stwierdzić, że któreś z założeń modelowych nie jest spełnione? Rozpatrzmy najpierw przypadek, gdy wprawdzie zachodzi równość (4.12), ale wspólny rozkład błędów $\varepsilon_i$ różni się znacznie od rozkładu normalnego. Sytuację taką można wykryć, jeśli przeanalizuje się histogram bądź wykres kwantylowy rezyduów (lub studentyzowanych rezyduów) omówiony w rozdz. 3. Przypomnijmy, że w sytuacji, gdy rezydua mają w przybliżeniu rozkład normalny, punkty na wykresie kwantylowym powinny skupiać się wokół pewnej prostej. Ponadto, na podstawie wykresu kwantylowego możemy stwierdzić, czy rozkład błędów jest istotnie skośny lub ma ogony malejące szybciej lub wolniej od ogonów rozkładu normalnego.

Innnym przypadkiem odstępstwa od modelu jest błędna specyfikacja funkcji regresji w równości (4.12). Oznacza to, że jest spełniona równość $Y_i = f(x_i) + \varepsilon_i, i = 1, \ldots, n$, ale funkcja regresji $f(x)$ nie jest postaci $\beta_0 + \beta_1 x$. Odstępstwo tego typu daje się z reguły zinterpretować na podstawie wykresu rezyduów. Z rysunku 4.6a i b można wnosić o nieliniowym charakterze zależności zmiennej objaśnianej od objaśniającej. Przedstawione rezydua pochodzą z modelu $Y = x^2 + \varepsilon$. Dobrze jest jednak uświadomić sobie, że czasami można podać więcej niż jeden możliwy powód takiego charakteru wykresu rezyduów. Na rys. 4.6.c i d przedstawiono wykres przebiegu zmien-

nych związanych zależnością $Y_i = 10 + 0{,}5 \times i + \varepsilon_i$, gdzie $\varepsilon_i$ są wartościami takich zmiennych, że $\varepsilon_i$ jest ujemnie zależne od $\varepsilon_{i-1}$. W tym przypadku wykres przebiegu można zinterpretować jako wskazujący na zależność krzywoliniową, gdy tymczasem w rzeczywistości zależność jest liniowa, a charakter wykresu jest spowodowany zależnością błędów.

Z niektórymi przypadkami błędnej specyfikacji modelu łatwo sobie poradzić, nie rezygnując z modelu liniowego. Jeśli np. zamiast równości (4.12) jest spełnione równanie $Y_i = \beta_0 + \beta_1 x_i^2 + \varepsilon_i$ dla $i = 1, 2, \ldots, n$, to wystarczy dokonać zamiany zmiennej objaśniającej $x_i$ na zmienną $x_i' = x_i^2$. Dalszej analizy dokonujemy dla par $(x_i', Y_i)$ $i = 1, \ldots, n$, które spełniają założenia modelu regresji liniowej. Jedynym ograniczeniem przy stosowaniu tej ogólnej metody podstawienia jest tylko liniowy charakter równania wyjściowego ze względu na współczynniki $\beta_0$ i $\beta_1$. Jak wybrać odpowiednią transformację zmiennej objaśniającej? Niekiedy z wykresu rozproszenia można odczytać jaka jest przybliżona zależność funkcyjna $y$ od $x$ i wtedy właśnie taką transformację stosujemy. Z reguły w przypadku przybliżonej zależności dodatniej i zależności opisanej przez funkcję wklęsłą próbuje się zastosować funkcje $f(x) = \sqrt{x}$ lub $f(x) = \log(x)$. Przykładowe transformacje dla zależności dodatniej wypukłej i ujemnej wypukłej są podane na rys. 4.7.




Rys. 4.7. Przykładowe transformacje dla różnych typów zależności

Inny, częsty w praktyce przypadek odstępstwa przedstawiono na rys. 4.8. Wskazuje on na zwiększanie zmienności błędów wraz ze wzrostem wartości

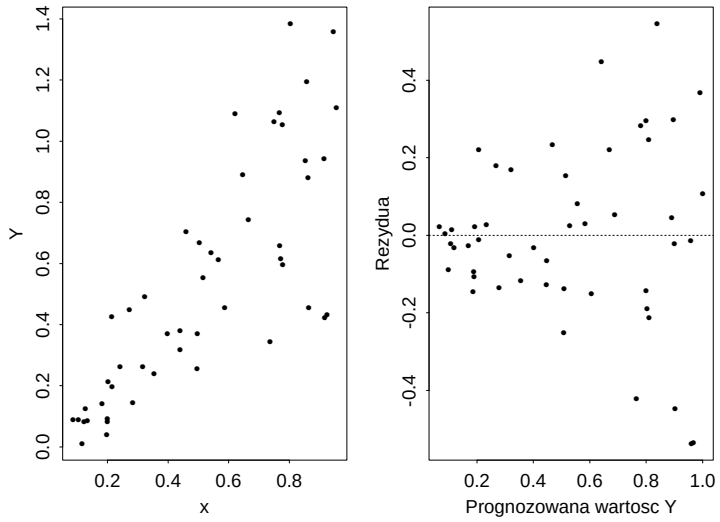
zmiennej objaśniającej. Łatwo można sobie wyobrazić, jak będzie wyglądał wykres rezyduów, gdy zmienność jest rosnącą, a następnie malejącą funkcją wielkości zmiennej objaśniającej. Co robimy w sytuacji, gdy analiza rezyduów prowadzi nas do konkluzji, że błędy są wprawdzie przypuszczalnie niezależne i mają w przybliżeniu rozkład normalny, ale ich zmienność istotnie zależy od numeru obserwacji? Przyjmijmy, że zachodzi równanie (4.12), ale $\mathrm{Var}(\varepsilon_i) = \sigma_i^2$. Chcemy w tym przypadku zmodyfikować kryterium, na podstawie którego wybraliśmy nasz estymator parametrów modelu. Rozumowanie jest następujące: jeśli przypuszczamy, że zmienność $i$-tego błędu jest większa niż zmienność pozostałych błędów, to powinniśmy dopuścić, że wyraz $(y_i - (b_0 + b_1 x_i))^2$ będzie większy od analogicznych wyrazów dla pozostałych obserwacji, gdzie $b_0$ i $b_1$ oznaczają teraz zmodyfikowane estymatory parametrów. Jednakże estymator MNK został wybrany na podstawie minimalizacji sumy $\sum_{j=1}^{n}(y_j - (b_0 + b_1 x_j))^2$, na wartość której wszystkie składniki mają taki sam wpływ. Żeby większa wartość czynnika $(y_i - (b_0 + b_1 x_i))^2$ nie miała zbyt dużego wpływu na rozwiązanie, zmniejszamy znaczenie tego wyrazu w sumie, zastępując go wyrazem $w_i(y_i - (b_0 + b_1 x_i))^2$, gdzie $w_i$ jest pewną wagą, której wartość jest mniejsza niż wagi wybrane dla innych obserwacji. Prowadzi to do problemu minimalizacji sumy

$$\sum_{j=1}^{n} w_j (y_j - (b_0 + b_1 x_j))^2 = \sum_{j=1}^{n} w_i e_i^2, \tag{4.26}$$

gdzie waga $w_j$ powinna być tym *mniejsza*, im *większa* jest wariancja $\sigma_i^2$.

Estymator otrzymany w wyniku minimalizacji sumy (4.26) nazywa się **estymatorem otrzymanym metodą najmniejszych ważonych kwadratów** lub w skrócie estymatorem MNWK. Rozsądnym kandydatem na wagę $w_i$ jest wartość $\sigma_i^{-2}$. Wynika to stąd, że w takiej sytuacji przybliżając rezydua przez odpowiadające im błędy, otrzymamy przybliżoną równość $\mathrm{E}(e_i^2 w_i) \approx 1$, a zatem wartości oczekiwane wszystkich wyrazów w sumie (4.25) są w tym przypadku w przybliżeniu równe. Ponieważ prawie nigdy nie znamy wartości wariancji poszczególnych błędów, wagę $w_i$ przyjmuje się równą $\hat{\sigma}_i^{-2}$, gdzie $\hat{\sigma}_i$ jest pewnym estymatorem odchylenia standardowego $\sigma_i$. Często rozpatrywanym estymatorem wartości $\sigma_i$ jest wartość przewidywana dla $i-$tej obserwacji w modelu regresji z tą samą zmienną objaśniającą, gdy za wartości zmiennej objaśnianej przyjmuje się wartości rezyduów. Ponieważ estymator taki ma dużą zmienność, z reguły stosuje się go jako punkt początkowy dla $i$-tej wagi w procedurze iteracyjnej, służącej do wyznaczenia estymatora MNWK.

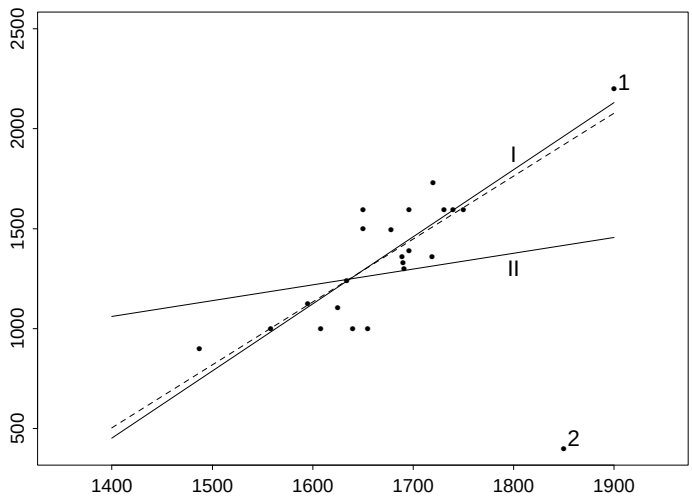
Następnym problemem, z którym możemy się zetknąć, jest nieadekwatność modelu spowodowana tym, że zapomnieliśmy uwzględnić jedną lub więcej




Rys. 4.8. Zwiększanie się zmienności błędów wraz ze wzrostem wartości zmiennej objaśniającej

istotnych zmiennych objaśniających. Jeśli istnieją dodatkowe zmienne mające liniowy wpływ na zmienną objaśniającą oznacza to, że powinniśmy rozważyć ogólniejszy model liniowy, uwzględniający te zmienne. Sytuacja staje się trudna, gdy pewne z potencjalnych zmiennych objaśniających są od siebie silnie zależne i prowadzi to do pytania, które zmienne objaśniające powinniśmy uwzględnić. Problem ten jest związany ściśle z problemem selekcji zmiennych w modelu regresji i zostanie omówiony dokładniej w p. 4.3.3. Ograniczymy się tutaj do stwierdzenia, że w tym przypadku pożyteczne może być sporządzenie wykresu rezyduów (dla modelu z jedną zmienną objaśniającą) względem wartości potencjalnie istotnych zmiennych nieuwzględnionych w modelu regresji. Jeśli taki wykres rozproszenia wyraźnie wskazuje na zależność, to analiza regresji powinna być powtórzona przy uwzględnieniu pominiętej zmiennej.

Na zakończenie tego punktu omówmy krótko kwestię wpływu wartości odstających na estymację prostej regresji. **Obserwacjami odstającymi w modelu regresji liniowej** nazywamy obserwacje nie spełniające równości (4.12). Identyfikacja wartości odstających ma w liniowej analizie regresji podstawowe znaczenie, między innymi dlatego, że ich występowanie może mieć duży wpływ na postać prostej MNK. Rozpatrzmy rys. 4.9, na którym przedstawiono diagram rozproszenia składający się z zasadniczej chmury punktów oraz dwóch obserwacji 1 i 2. Jak widać z wykresu rozproszenia, obserwacja 2 jest obserwacją odstającą. Linia przerywana oznacza prostą regresji dla zasadniczej grupy punktów, proste I i II to proste MNK po dołącze-




Rys. 4.9. Wykres rozproszenia składający się z zasadniczej chmury punktów oraz dwóch obserwacji 1 i 2

niu odpowiednio punktów 1 albo 2. Uwzględnienie obserwacji 2 powoduje dużą zmianę współczynnika kierunkowego prostej; punkt ten „przyciąga" do siebie prostą regresji MNK. O takiej obserwacji mówimy, że jest **obserwacją wpływową**, gdyż jej dołączenie do zbioru danych ma duży wpływ na przebieg prostej MNK. *Obserwacja, dla której wartość zmiennej objaśniającej znacząco odbiega od typowych wartości tej zmiennej, jest potencjalną obserwacją wpływową.* Zaznaczmy na marginesie, że to właśnie dla takich punktów odchylenie standardowe rezyduum odbiega znacznie od wartości $S$. Zauważmy ponadto, że gdyby punkt 2 przesunąć wzdłuż osi $x$ tak, aby wartość zmiennej objaśniającej wyniosła np. 1650, to byłby on nadal obserwacją odstającą, ale najpewniej nie byłby obserwacją wpływową. Inaczej jest z wpływem punktu 1 na postać prostej MNK. Punkt ten jest położony blisko prostej MNK wyznaczonej na podstawie zasadniczej grupy punktów i dlatego jego dołączenie do zbioru danych nie wpływa w sposób zasadniczy na postać tej prostej. Obserwacja 1 **nie** jest obserwacją wpływową ani obserwacją odstającą, mimo że podobnie jak dla obserwacji 2 jej wartość zmiennej objaśniającej różni się znacznie od typowej wartości tej zmiennej. W przypadku jednej zmiennej objaśniającej obserwacje odstające można identyfikować optycznie, analizując wykres rozproszenia. Będzie to jednak operacja heurystyczna, gdyż nie wiemy, na ile jedna ze współrzędnych obserwacji musi różnić się od pozostałych, aby uznać ją za obserwację odstającą. Przypomnijmy, że próba $\{Y_i\}$ nie jest prostą próbą losową i dlatego nie możemy zastosować metod identyfikacji wartości odstających omówionych

w rozdz. 1. Możemy jednak analizować rezydua lub rezydua studentyzowane, które powinny zachowywać się podobnie jak prosta próba losowa $\{\varepsilon_i\}$, i szukać wśród nich wartości odstających (pewne ulepszenie tego postępowania omówimy w następnym podrozdziale). Załóżmy np. na moment, że w przykł. 4.1 błędnie zarejestrowano wynik egzaminu końcowego dla czternastego studenta: zamiast 28 punktów zapisano 8. Równanie prostej MNK dla zmienionego zbioru danych wynosi $y = 3{,}5 + 1{,}8 \times x$ (zauważmy, że w porównaniu z poprzednią sytuacją współczynnik kierunkowy prostej nie zmienił się, natomiast wyraz wolny zmniejszył się o 1,7). Wartość rezyduum dla czternastego studenta wynosi teraz $8 - (3{,}5 + 1{,}8 \times 14) = -20{,}8$ i jest wartością odstającą od typowych wartości rezyduów. Czternasta obserwacja jest obserwacją odstającą i wpływową. Kłopot w tym, że w ten sposób nie wykryjemy wpływowych obserwacji odstających, dla których rezydua są małe. Metody identyfikacji takich obserwacji oraz formalne metody testowania, czy dana obserwacja jest obserwacją odstającą bądź wpływową poznamy w następnym punkcie.

## 4.3. Analiza zależności wielu zmiennych ilościowych

Sytuacja, w której zmienna objaśniana zależy tylko od jednej zmiennej objaśniającej, zdarza się rzadko. Z reguły jesteśmy w stanie wyróżnić kilka zmiennych, które mogą mieć wpływ na zmienną objaśnianą. Na przykład rozpatrując liczbę sprzedanych komputerów na pewnym obszarze w danym roku, możemy przypuszczać, że na jej wielkość może mieć wpływ nie tylko liczba mieszkańców tego obszaru, ale również liczba firm mających tam swoją siedzibę oraz procent ogólnej liczby mieszkańców z podstawowym wykształceniem informatycznym. W sytuacji, gdy dysponujemy $p-1$ zmiennymi objaśniającymi, $p \geq 2$, nasze dane dla $n$ obiektów możemy przedstawić następująco:

$$\begin{array}{cccc} y_1 & x_{11} & \dots & x_{1,p-1} \\ \vdots & \vdots & & \vdots \\ y_n & x_{n1} & \dots & x_{n,p-1}, \end{array}$$

gdzie $x_{ik}$ jest wartością $k$-tej zmiennej objaśniającej dla $i$-tego obiektu, a $y_i$ jest wartością zmiennej objaśnianej dla tego obiektu. We wspomnianym przykładzie indeks $i$ odpowiadałby numerowi kolejnego obszaru, na którym analizuje się sprzedaż komputerów. Rozpatrzmy jeszcze jeden przykład, który omówimy dokładnie po przedstawieniu części teoretycznej.

**Przykład 4.3.** Analizie poddano następujące parametry 24 samochodów znajdujących się na polskim rynku:

$Y$ – średnie zużycie paliwa na 100 km (zmienna objaśniana);
$x_1$ – pojemność silnika ($cm^3$);
$x_2$ – moc silnika (KM);
$x_3$ – ładowność (l);
$x_4$ – masa (kG);
$x_5$ – długość (cm);
$x_6$ – szerokość (cm).

Wartości zmiennych dla poszczególnych marek są zawarte w tab. 4.3. Zgodnie z wprowadzoną notacją $x_{3,1} = 1498$ jest wartością pojemności silnika (w $cm^3$) dla trzeciego obiektu, czyli samochodu Daewoo Lanos.

**Tabela 4.3. Parametry wybranych samochodów (przykł 4.3)**

| | Marka samochodu | $Y$ | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ | $x_6$ |
|---|---|---|---|---|---|---|---|---|
| 1 | Citroen Saxo3D | 6.3 | 1124 | 60 | 505 | 825 | 3178 | 1595 |
| 2 | Citroen ZX | 8.0 | 1360 | 75 | 655 | 1125 | 4108 | 1719 |
| 3 | Daewoo Lanos | 8.5 | 1498 | 86 | 480 | 1096 | 4047 | 1678 |
| 4 | Polonez Caro | 9.3 | 1598 | 76 | 440 | 1120 | 4369 | 1650 |
| 5 | Maluch | 5.5 | 652 | 24 | 320 | 600 | 3109 | 1377 |
| 6 | Fiat Uno Fire | 5.9 | 999 | 45 | 430 | 805 | 3689 | 1558 |
| 7 | Fiat CC Happy | 5.9 | 899 | 41 | 440 | 710 | 3227 | 1487 |
| 8 | Fiat Punto | 6.5 | 1108 | 54 | 475 | 875 | 3760 | 1625 |
| 9 | Ford Fiesta | 6.4 | 1242 | 75 | 435 | 940 | 3828 | 1634 |
| 10 | Ford Escort | 6.6 | 1299 | 60 | 435 | 1065 | 4104 | 1691 |
| 11 | Ford Mondeo | 8.2 | 1597 | 90 | 675 | 1275 | 4670 | 1750 |
| 12 | Mercedes C280T | 10.1 | 2799 | 197 | 480 | 1430 | 4516 | 1720 |
| 13 | Opel Corsa | 6.3 | 998 | 54 | 420 | 930 | 3729 | 1608 |
| 14 | Opel Astra | 6.8 | 1388 | 60 | 445 | 1010 | 4051 | 1696 |
| 15 | Peugeot 306 | 7.6 | 1360 | 75 | 450 | 1010 | 3995 | 1689 |
| 16 | Renault Megane | 6.7 | 1598 | 90 | 485 | 1010 | 3931 | 1696 |
| 17 | Skoda Octavia | 7.3 | 1598 | 75 | 510 | 1160 | 4511 | 1731 |
| 18 | Łada Samara | 7.1 | 1499 | 70 | 425 | 945 | 4006 | 1650 |
| 19 | Lancia | 9.2 | 1998 | 155 | 550 | 1450 | 4687 | 1822 |
| 20 | Toyota Corolla | 6.9 | 1332 | 85 | 470 | 1110 | 4100 | 1690 |
| 21 | VW Polo | 5.9 | 999 | 50 | 495 | 880 | 3715 | 1655 |
| 22 | VW Passat | 7.5 | 1595 | 100 | 625 | 1125 | 4675 | 1740 |
| 23 | Volvo S40 | 8.6 | 1731 | 115 | 434 | 1286 | 4480 | 1720 |
| 24 | Seat Ibiza | 6.0 | 1000 | 50 | 460 | 935 | 3853 | 1640 |

W tej książce ograniczamy się do modeli liniowych; gdy chodzi o modele ogólniejsze, p. (J. Koronacki, J. Ćwik (2005)).

### 4.3.1. Model liniowy regresji wielokrotnej

Przypuśćmy, że wpływ każdej rozpatrywanej zmiennej objaśniającej na zmienną objaśnianą jest liniowy i nie zależy on od wartości innych zmiennych. Wielkość $y_i$ traktujemy jako wartość zmiennej losowej $Y_i$ i rozpatrujemy równość będącą uogólnieniem równości (4.12)

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_{p-1} x_{i,p-1} + \varepsilon_i, \quad i = 1, 2, \ldots, n, \tag{4.27}$$

gdzie $\varepsilon_i$ dla $i = 1, \ldots, n$ są błędami losowymi o takim samym rozkładzie mającym wartość średnią 0 i nieznaną wariancję $\sigma^2$, a $\beta_0, \ldots, \beta_{p-1}$ są nieznanymi parametrami. W dalszym ciągu będziemy stosowali zapis kolumnowy wektorów, oznaczając przez $\mathbf{z}'$ transpozycję wektora $\mathbf{z}$. Tak więc wektor $(x_{i1}, x_{i2}, \ldots, x_{i,p-1})'$ jest wektorem wartości zmiennych objaśniających dla $i$-tej zmiennej objaśnianej $Y_i$. Aby uwzględnić wyraz wolny $\beta_0$ powiększmy ostatni wektor o dodatkową współrzędną $x_{i0}$ równą 1 i zdefiniujmy wektor $\mathbf{x}_i' = (x_{i0}, x_{i1}, \ldots, x_{i,p-1})$. Rozpatrzmy następnie macierz $\mathbf{X}$ wymiaru $n \times p$, której $i$-tym wierszem jest wektor $\mathbf{x}_i'$

$$\mathbf{X} = (x_{ij})_{\substack{i=1,\ldots,n \\ j=0,\ldots,p-1}} = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1,p-1} \\ 1 & x_{21} & \ldots & x_{2,p-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \ldots & x_{n,p-1} \end{pmatrix}.$$

Niech $\mathbf{Y}' = (Y_1, Y_2, \ldots, Y_n)$ będzie wektorem zmiennych objaśnianych, a $\varepsilon' = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)$ wektorem błędów. Wówczas $n$ równości w (4.27) można zwarcie przedstawić w postaci macierzowej

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon, \tag{4.28}$$

gdzie $\beta' = (\beta_0, \beta_1, \ldots, \beta_{p-1})$ jest wektorem nieznanych współczynników. W sytuacji, gdy jest spełniona równość (4.28) oraz współrzędne wektora $\varepsilon$ są niezależnymi zmiennymi losowymi, mówimy, że są spełnione założenia **liniowego modelu regresji wielokrotnej**. Określenie „wielokrotny" odnosi się do kilku (więcej niż jedna) zmiennych objaśniających występujących w równości (4.28). Kolejne zmienne objaśniające będziemy oznaczać symbolami $x_1, \ldots, x_{p-1}$. Dla $p - 1 = 1$ rozpatrujemy tylko jedną zmienną objaśniającą i wtedy model regresji wielokrotnej sprowadza się do rozpatrywanego poprzednio modelu regresji jednokrotnej. *W dalszym ciągu ograniczymy się do analizy sytuacji, gdy wspólnym rozkładem błędów jest rozkład normalny* $N(0, \sigma)$. Zauważmy, że w tym przypadku równanie (4.28) zawiera $p + 1$ nieznanych parametrów: $p$ współrzędnych wektora $\beta$ i nieznaną wariancję

błędu $\sigma^2$. Podkreślmy, że przymiotnik liniowy oznacza, że *parametry* są liniowymi współczynnikami równania (4.27), zmienne objaśniające nie muszą być liniowe. Na przykład równanie

$$y = \beta_0 + \beta_1 \sin x_1 + \beta_2 \ln x_2 + \varepsilon$$

jest równaniem liniowym regresji, gdy dokonamy zamiany zmiennych objaśniających na $\tilde{x}_1 = \sin x_1$ i $\tilde{x}_2 = \ln x_2$. Natomiast równanie

$$y = \beta_0 + \sin(\beta_1 x_1)$$

nie jest liniowe. Powyższa możliwość przekształcania zmiennych objaśniających w równaniach pozornie nieliniowych powoduje, że za pomocą równań regresji liniowej można adekwatnie modelować różnorakie problemy zależności. Na przykład równanie regresji wielomianowej $y = \beta_0 + \beta_1 x + \ldots + \beta_{p-1} x^{p-1} + \varepsilon$ po podstawieniu $x_i = x^i$ dla $i = 1, \ldots, p-1$ staje się szczególnym przypadkiem równania regresji liniowej z $p-1$ zmiennymi objaśniającymi.

Zaznaczmy na marginesie, że próbę $Y_1, Y_2, \ldots, Y_n$ z rozkładu o średniej $\mu$ i wariancji $\sigma^2$ można zapisać następująco:

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \mu + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix},$$

gdzie $\varepsilon_i$ są niezależnymi „błędami" o średniej 0. Zatem problem estymacji wartości średniej i wariancji rozkładu omówiony w p. 2.2.4 jest szczególnym przypadkiem estymacji parametrów modelu liniowego. Dla regresji jednokrotnej mamy $p = 2$ (gdyż $p - 1 = 1$) i równanie (4.28) redukuje się do równania

$$\mathbf{Y} = \begin{pmatrix} 1 & x_{1,1} \\ \vdots & \vdots \\ 1 & x_{n,1} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix},$$

identycznego z równaniem (4.12). Z postaci równania (4.28) jest również oczywiste, dlaczego rozpatrywany model nazywamy modelem zależności liniowej, a nie afinicznej, jak sugerowałoby to równanie (4.27). Wyraz wolny $\beta_0$ w równaniu (4.27) można potraktować jako współczynnik odpowiadający dodatkowej zmiennej objaśniającej $x_0$ stale równej 1. W ten sposób, zmieniając liczbę zmiennych objaśniających z $p-1$ na $p$, możemy mówić o liniowym charakterze zależności.

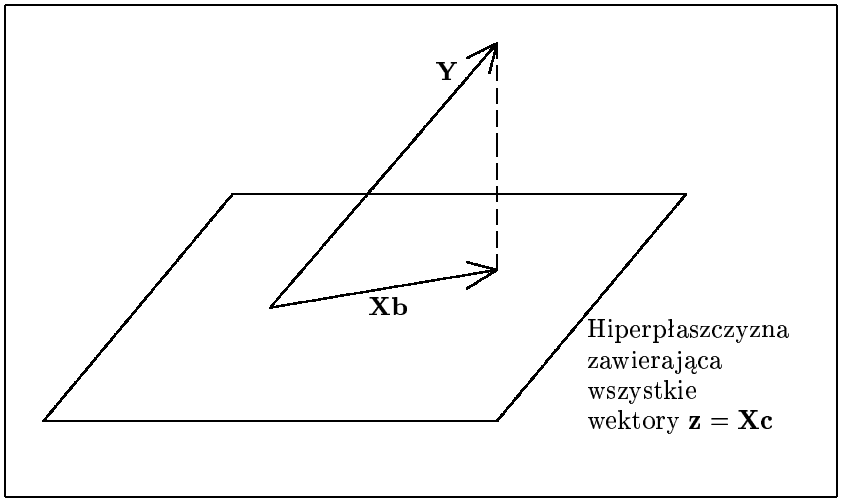
Zauważmy, że zgodnie z założeniem modelowym (4.27), zmienna $Y_i$ jest zmienną losową o wartości średniej

$$\mu_{Y_i} = \beta_0 + \sum_{j=1}^{p-1} \beta_i x_{ij} = \beta' \mathbf{x}_i,$$

gdzie $\beta' = (\beta_0, \ldots, \beta_{p-1})$. Oznaczając przez $\mu_{\mathbf{Y}} = (\mu_{Y_1}, \ldots, \mu_{Y_n})'$ wektor wartości oczekiwanych możemy ostatnią równość zapisać w sposób bardziej zwarty jako $\mu_{\mathbf{Y}} = \mathbf{X}\beta$. Biorąc pod uwagę fakt, że $\mathbf{x}_i$ jest wektorem deterministycznych wartości zmiennych objaśniających dla $i$-tego obiektu i stosując równanie (4.27), otrzymujemy $\sigma^2_{Y_i} = \sigma^2$ oraz, ponieważ zmienne $Y_i$ i $Y_j$ są nieskorelowane dla $i \neq j$, mamy $\mathrm{Cov}(Y_i, Y_j) = 0$. Jeśli przez $\mathbf{\Sigma_Y}$ oznaczymy macierz kowariancji wektora $\mathbf{Y}$

$$\mathbf{\Sigma_Y} = \begin{pmatrix} \mathrm{Cov}(Y_1, Y_1) & \ldots & \mathrm{Cov}(Y_1, Y_n) \\ \mathrm{Cov}(Y_2, Y_1) & \ldots & \mathrm{Cov}(Y_2, Y_n) \\ \vdots & \vdots & \ddots \\ \mathrm{Cov}(Y_n, Y_1) & \ldots & \mathrm{Cov}(Y_n, Y_n) \end{pmatrix},$$

to otrzymujemy, że $\mathbf{\Sigma_Y} = \sigma^2\mathbf{I}$, gdzie $\mathbf{I}$ jest macierzą jednostkową wymiaru $n \times n$, to znaczy macierzą mającą jedynki na przekątnej i zera poza nią. Podobnie jak w przypadku regresji jednokrotnej zmienne $Y_1, Y_2, \ldots, Y_n$ nie tworzą z reguły prostej próby losowej, ale stają się nią po odjęciu odpowiadających wartości średnich.




Rys. 4.10. Wektor **Xb** jako rzut prostopadły wektora **Y**

**Metoda najmniejszych kwadratów.** Podobnie jak w przypadku jednej zmiennej objaśniającej, gdy staraliśmy się tak wybrać estymatory $b_0$ i $b_1$, aby suma kwadratów odległości $Y_i$ od prostej $b_0 + b_1 x$ była minimalna, chcemy teraz minimalizować wyrażenie

$$Q(\mathbf{b}) = \sum_{i=1}^{n}(Y_i - (b_0 + b_1 x_{i1} + \ldots + b_{p-1}x_{i,p-1}))^2 = (\mathbf{Y} - \mathbf{Xb})'(\mathbf{Y} - \mathbf{Xb}),$$

gdzie $\mathbf{b}' = (b_0, b_1, \ldots, b_{p-1})$. Odpowiada to poszukiwaniu takiego wektora $\mathbf{b}$, że odległość wektora $\mathbf{Y}$ od zbioru wektorów $\{\mathbf{z} = \mathbf{Xc}$ dla pewnego $\mathbf{c}\}$ jest równa odległości wektora $\mathbf{Y}$ od wektora $\mathbf{Xb}$. Wektor $\mathbf{Xb}$ jest niczym innym jak rzutem prostopadłym wektora $\mathbf{Y}$ na ten zbiór (porównaj rys. 4.10).

**Definicja 4.5.** *Estymatorem wyznaczonym metodą najmniejszych kwadratów (MNK) nazywamy wektor* $\mathbf{b}$ *minimalizujący funkcję* $Q(\cdot)$.

Wyznaczmy postać estymatora MNK. W wyniku rozpisania kwadratów w definicji $Q(\mathbf{b})$ otrzymujemy

$$Q(\mathbf{b}) = \sum_{i=1}^{n} Y_i^2 - 2\sum_{j=0}^{p-1}\sum_{i=1}^{n} b_j x_{ij} Y_i + \sum_{i,j=0}^{p-1}\sum_{k=1}^{n} b_i b_j x_{ki} x_{kj}.$$

W celu zidentyfikowania minimum funkcji $Q$ znajdźmy punkt $\mathbf{b}$, w którym zerują się pochodne cząstkowe $\frac{\partial Q}{\partial b_i}(\mathbf{b})$ dla dowolnego $i = 0, 1, \ldots, p-1$. Po zróżniczkowaniu względem $b_{j_0}$ otrzymujemy

$$\frac{\partial Q}{\partial b_{j_0}}(\mathbf{b}) = -2\sum_{i=1}^{n} x_{ij_0} Y_i + 2\sum_{i=0}^{p-1}\sum_{k=1}^{n} b_i x_{ki} x_{kj_0}, \quad j_0 = 0, 1, \ldots, p-1. \tag{4.29}$$

Zdefiniujmy wektor $\frac{\mathbf{d}}{\mathbf{db}}(Q)$, którego współrzędne są kolejnymi pochodnymi cząstkowymi funkcji $Q$

$$\frac{\mathbf{d}}{\mathbf{db}}(Q) = \begin{pmatrix} \frac{\partial Q}{\partial b_0} \\ \vdots \\ \frac{\partial Q}{\partial b_{p-1}} \end{pmatrix}$$

i zauważmy, że równość (4.29) można zapisać następująco:

$$\frac{\mathbf{d}}{\mathbf{db}}(Q) = -2\mathbf{X}'\mathbf{Y} + 2\mathbf{X}'\mathbf{Xb}.$$

Zatem wektor $\mathbf{b}$ minimalizujący funkcję $Q$ spełnia równanie (zwane czasami równaniem normalnym)

$$\mathbf{X}'\mathbf{Xb} = \mathbf{X}'\mathbf{Y}.$$

Jeśli kolumny macierzy $\mathbf{X}'\mathbf{X}$ są liniowo niezależne, to jest ona odwracalna i otrzymujemy jawną postać estymatora MNK

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}. \tag{4.30}$$

W dalszym ciągu będziemy zakładali, że macierz $\mathbf{X}'\mathbf{X}$ można odwrócić i estymator MNK jest dany wzorem (4.30). Aby przekonać się, że po rozwiązaniu (4.30) otrzymujemy postać estymatorów wyprowadzonych poprzednio dla regresji jednokrotnej, zapiszmy równanie (4.28) w nieco innej postaci

$$Y_i = \beta_0 + \beta_1\bar{x} + \beta_1(x_i - \bar{x}) + \varepsilon_i = \tilde{\beta}_0 + \beta_1(x_i - \bar{x}) + \varepsilon_i.$$

Dla tego przedstawienia otrzymujemy na podstawie $\sum_{i=1}^n (x_i - \bar{x}) = 0$

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 - \bar{x} \\ \vdots & \vdots \\ 1 & x_n - \bar{x} \end{pmatrix} \quad \text{i} \quad \mathbf{X}'\mathbf{X} = \begin{pmatrix} n & 0 \\ 0 & \sum_{i=1}^n (x_i - \bar{x})^2 \end{pmatrix}.$$

Ponadto

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} n^{-1} & 0 \\ 0 & (\sum_{i=1}^n (x_i - \bar{x})^2)^{-1} \end{pmatrix}, \qquad \mathbf{X}'\mathbf{Y} = \begin{pmatrix} \sum_{i=1}^n Y_i \\ \sum_{i=1}^n (x_i - \bar{x})Y_i \end{pmatrix}.$$

Zgodnie ze wzorem (4.30) otrzymujemy znaną już postać estymatora współczynnika nachylenia $b_1 = \sum_{i=1}^n (x_i - \bar{x})Y_i / \sum_{i=1}^n (x_i - \bar{x})^2$. Z reguły w przypadku więcej niż jednej zmiennej nie można jednak uzyskać jawnej postaci wektora estymatorów $\mathbf{b}$.

### 4.3.2. Własności estymatorów MNK

Omówimy tylko podstawowe własności estymatorów MNK: ich nieobciążoność oraz postać odchyleń standardowych. Sformułujmy najpierw pożyteczną własność

**Stwierdzenie 4.8.** *Niech $\mathbf{X}$ będzie $r$-wymiarowym wektorem losowym o wartości średniej $\mu_{\mathbf{X}}$ i macierzy kowariancji $\Sigma_{\mathbf{X}}$ oraz niech $\mathbf{A}$ będzie macierzą rozmiaru $s \times r$. Wówczas dla $s$-wymiarowego wektora losowego $\mathbf{Y} = \mathbf{A}\mathbf{X}$ mamy*

$$\mu_{\mathbf{Y}} = \mathbf{A}\mu_{\mathbf{X}} \quad \text{i} \quad \Sigma_{\mathbf{Y}} = \mathbf{A}\Sigma_{\mathbf{X}}\mathbf{A}'.$$

Zauważmy, że ponieważ mamy $Y_i = \sum_{j=1}^r a_{ij}X_j$, dla $i = 1, \ldots, s$, zatem $\mu_{Y_i} = \sum_{j=1}^r a_{ij}\mu_{X_j}$, z czego wynika pierwsza z równości w stwierdzeniu. Ponadto, analogicznie otrzymujemy

$$E(Y_iY_h) = \sum_{j,k=1}^r a_{ij}a_{hk}E(X_jX_k)$$

oraz

$$EY_iEY_h = \sum_{j,k=1}^{r} a_{ij}a_{hk}EX_jEX_k.$$

Zatem

$$\mathrm{Cov}(Y_i, Y_h) = E(Y_iY_h) - EY_iEY_h = \sum_{j,k=1}^{r} a_{ij}a_{hk}\mathrm{Cov}(X_j, X_k),$$

skąd wynika druga z równości w stwierdzeniu.

Na podstawie powyższej własności łatwo udowodnić następujące stwierdzenie.

**Stwierdzenie 4.9.** *Estymator* $\mathbf{b}$ *jest nieobciążonym estymatorem* $\beta$: $\mu_{\mathbf{b}} = \beta$ *oraz* $\Sigma_{\mathbf{b}} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$.

Pierwsza część stwierdzenia wynika z pierwszej własności w stwierdzeniu 4.8, zastosowanej do macierzy $\mathbf{A} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ i faktu, że $\mu_{\mathbf{Y}} = \mathbf{X}\beta$. Jeśli skorzystamy z drugiej własności, równości $(\mathbf{AB})' = \mathbf{B}'\mathbf{A}'$ i $\Sigma_{\mathbf{Y}} = \sigma^2\mathbf{I}$, to mamy

$$\begin{aligned}\Sigma_{\mathbf{b}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\sigma^2\mathbf{I})((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')' = \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\sigma^2\mathbf{I})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1},\end{aligned} \tag{4.31}$$

gdzie użyliśmy faktu, że macierz $(\mathbf{X}'\mathbf{X})^{-1}$ jest symetryczna. Wynika to z symetrii macierzy $\mathbf{X}'\mathbf{X}$, dla której element o współrzędnych $(i,j)$ ma postać $\sum_{k=1}^{n} x_{ki}x_{kj}$.

W szczególności wariancja estymatora $b_i$ jest z dokładnością do czynnika $\sigma^2$ równa $i$-temu wyrazowi na diagonalnej macierzy $(\mathbf{X}'\mathbf{X})^{-1}$: $\sigma^2_{b_i} = \mathrm{Cov}(b_i, b_i) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}_{ii}$, $i = 0, \ldots, p-1$. W przypadku regresji jednokrotnej otrzymujemy np. $\sigma^2_{b_1} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}_{11} = \sigma^2 / \sum_{i=1}^{n}(x_i - \bar{x})^2$.

Analogicznie jak w przypadku regresji jednokrotnej wartością przewidywaną dla $i$-tej obserwacji nazwiemy wartość

$$\hat{Y}_i = b_0 + b_1x_{i1} + \ldots + b_{p-1}x_{i,p-1} = \mathbf{x}_i'\mathbf{b}.$$

Wektor wszystkich wartości przewidywanych będzie miał postać $\hat{\mathbf{Y}} = (\hat{Y}_1, \hat{Y}_2, \ldots, \hat{Y}_n)' = \mathbf{Xb}$ i na mocy równania (4.30) może być zapisany następująco:

$$\hat{\mathbf{Y}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{HY}, \tag{4.32}$$

gdzie $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. Macierz $\mathbf{H}$ jest symetryczna i ma własność $\mathbf{H}^2\mathbf{y} = \mathbf{H}\mathbf{y}$ dla dowolnego wektora $\mathbf{y}$. Rzeczywiście,

$$\mathbf{H}^2 = (\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' =$$

$$= \mathbf{X}\mathbf{I}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{H}.$$

Macierz $\mathbf{H}$ pojawia się często przy okazji diagnostyki modelu liniowego.

**Wartości resztowe (rezydua).** Wartość resztową $e_i$ dla $i$-tej obserwacji definiujemy jako różnicę między wartością zmiennej objaśniającej $Y$ dla $i$-tej obserwacji i odpowiednią wartością przewidywaną $\hat{Y}_i = \mathbf{x}_i'\mathbf{b}$. Zatem wektor wartości resztowych można zapisać następująco:

$$\mathbf{e} = (e_1, e_2, \ldots, e_n)' = \mathbf{Y} - \hat{\mathbf{Y}} = (\mathbf{I} - \mathbf{H})\mathbf{Y}.$$

Ponieważ na mocy stwierdzenia 4.9 wektor $\mathbf{b}$ jest nieobciążonym estymatorem wektora współczynników $\beta$, więc wartość oczekiwana wektora rezyduów $E\mathbf{e}$ jest równa wektorowi zerowemu. Ponadto macierz kowariancji wektora $\mathbf{e}$ jest równa

$$\Sigma_{\mathbf{e}} = \sigma^2(\mathbf{I} - \mathbf{H}). \tag{4.33}$$

Powyższą równość wynikającą prosto ze stwierdzenia 4.8 i faktu, że $\mathbf{H}^2 = \mathbf{H}$, pozostawiamy jako zadanie. Podobnie jak w przypadku regresji jednokrotnej, rezydua w problemie regresji wielokrotnej służą do diagnostyki modelu oraz do estymacji nieznanej wariancji $\sigma^2$. Do tego ostatniego celu analogicznie jak poprzednio wykorzystamy sumę kwadratów wartości resztowych $SSE = \sum_{i=1}^{n} e_i^2$ (którą można zapisać jako $\mathbf{e}'\mathbf{e}$), pomnożoną przez odpowiednią stałą. Stałą chcielibyśmy wybrać tak, aby zapewniała nam nieobciążoność estymatora wariancji. Można udowodnić, że w modelu liniowej regresji wielokrotnej $E\mathbf{e}'\mathbf{e} = (n-p)\sigma^2$. Zatem, definiując

$$S^2 = \frac{1}{(n-p)}\sum_{i=1}^{n} e_i^2 = \frac{1}{(n-p)}\mathbf{e}'\mathbf{e} = \frac{1}{(n-p)}SSE,$$

uzyskamy nieobciążony estymator nieznanej wariancji $\sigma^2$. Powyższa definicja jest zgodna z def. 4.3 błędu średniokwadratowego dla regresji jednokrotnej. Liczba $n-p$ nosi nazwę liczby stopni swobody sumy kwadratów błędów i może być interpretowana jako liczba stopni swobody obserwacji, równa $n$, pomniejszona o liczbę więzów nakładanych na $\hat{y}_i$, $i = 1, \ldots, n$, równą $p$. Teraz zgodnie z równaniem (4.31) prosto otrzymujemy wyrażenia na odchylenia standardowe estymatorów $b_i$ współczynników $\beta_i$ jako dodatnie pierwiastki z wyrażeń

$$SE_{b_i}^2 = S^2(\mathbf{X}'\mathbf{X})_{ii}^{-1},\ i = 0, 1, \ldots, p-1. \tag{4.34}$$

Spora część narzędzi wnioskowania, które poznaliśmy w przypadku regresji jednokrotnej, przenosi się bez dużych zmian na przypadek wielu zmiennych objaśniających. W szczególności identycznie możemy zdefiniować całkowitą sumę kwadratów $SST = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$ i regresyjną sumę kwadratów $SSR = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$ i udowodnić, że zmienność całkowita opisana przez $SST$ jest sumą $SSE$ i $SSR$

$$SST = SSE + SSR. \tag{4.35}$$

Modyfikacji ulega jedynie równość dla liczby stopni swobody związana z odpowiednimi sumami kwadratów. Liczba stopni swobody związana z $SSE$ wynosi $n - p$, a związana z $SSR$ wynosi $p - 1$ i jest równa liczbie faktycznych zmiennych objaśniających w modelu regresji. Ponieważ liczba stopni swobody związana z $SST$ nie zależy od liczby zmiennych objaśniających i wynosi $n - 1$, więc zachodzi równanie $n - 1 = n - p + p - 1$, czyli

$$\begin{gathered}\text{całkowita liczba stopni swobody} = \\ = \text{liczba stopni swobody błędów} + \text{regresyjna liczba stopni swobody.}\end{gathered}$$

Jak omawialiśmy to poprzednio, regresyjną sumę kwadratów możemy interpretować jako część zmienności całkowitej tłumaczoną przez model, zatem stosunek $SSR/SST$ można przyjąć jako wskaźnik adekwatności modelu. Nosi on nazwę **współczynnika determinacji wielokrotnej** i jest oznaczany przez $R^2$. Na mocy równości (4.35) otrzymujemy

$$R^2 = SSR/SST = 1 - SSE/SST.$$

Dla sytuacji jednej zmiennej objaśniającej współczynnik determinacji wielokrotnej $R^2$ pokrywa się ze współczynnikiem determinacji (jednokrotnej) $r^2$ zdefiniowanym w podrozdz. 4.2.

**Testy dla wektora współczynników $\beta$.** Przypuśćmy, że wybraliśmy pewien zbiór zmiennych objaśniających i możemy przyjąć, że zachodzi dla nich liniowy model regresji (4.28). Bardzo ważne jest w tej sytuacji przekonanie się, czy faktycznie którakolwiek ze zmiennych objaśniających $x_1, \ldots, x_{p-1}$ ma wpływ na zmienną objaśnianą $Y$. W sytuacji adekwatności modelu (4.28) odpowiada to stwierdzeniu, czy któryś ze współczynników $\beta_1, \ldots, \beta_{p-1}$ jest różny od 0. Współczynnik $\beta_0$ pomijamy, ponieważ odpowiada on stałej w modelu regresji. Zgodnie z metodologią przedstawioną poprzednio interesujący nas fakt formułujemy jako hipotezę *alternatywną* dla sytuacji testowania

$$H_0\colon\ \beta_1 = \beta_2 = \ldots = \beta_{p-1} = 0$$

przeciwko

$$H_1\text{: który\'s ze współczynników } \beta_1, \ldots, \beta_{p-1} \text{ jest różny od 0.}$$

Do testowania hipotezy $H_0$ wykorzystamy porównanie regresyjnej sumy kwadratów $SSR$ z sumą kwadratów błędów $SSE$. Przy niespełnieniu $H_0$, regresyjna suma kwadratów powinna przyjmować relatywnie duże wartości w stosunku do sumy kwadratów błędów. Żeby sprecyzować dokładnie, co oznacza stwierdzenie „relatywnie duże wartości", musimy wiedzieć, jakie wartości stosunku $SSR/SSE$ są typowe przy spełnieniu $H_0$. Wykorzystamy w tym celu fakt, z którego wynika, że jeśli hipoteza $H_0$ jest prawdziwa, to

$$\sigma^{-2}SSR \text{ ma rozkład } \chi^2 \text{ z } p-1 \text{ stopniami swobody}$$

i

$$\sigma^{-2}SSE \text{ ma rozkład } \chi^2 \text{ z } n-p \text{ stopniami swobody,}$$

oraz wyrażenia $SSR$ i $SSE$ są niezależne. Zatem, gdy jest prawdziwa hipoteza $H_0$, statystyka

$$F = \frac{\frac{1}{\sigma^2}\frac{SSR}{(p-1)}}{\frac{1}{\sigma^2}\frac{SSE}{(n-p)}} = \frac{\frac{SSR}{(p-1)}}{\frac{SSE}{(n-p)}} \sim F_{p-1,n-p},$$

gdzie $F_{p-1,n-p}$ oznacza rozkład $F$ Snedecora z parametrami $p-1$ i $n-p$. Tak więc wartości przyjmowane przez statystykę $F$ powinny być typowymi wartościami z rozkładu $F_{p-1,n-p}$. W tej sytuacji test o zbiorze krytycznym

$$\{F\text{: } F \geq f_{1-\alpha,p-1,n-p}\}$$

jest testem na poziomie istotności $\alpha$ dla testowania hipotezy $H_0$.

Podkreślmy, że jeśli odrzucimy hipotezę zerową, to nie możemy stwierdzić, które współczynniki $\beta_j$ są istotnie różne od 0. Rozpatrzmy zatem nieco inaczej postawiony problem. Przy zachodzeniu założenia (4.28) chcemy testować hipotezę

$$H_0\text{: } \beta_{i_0} = 0 \text{ przeciwko alternatywie } H_1\text{: } \beta_{i_0} \neq 0,$$

gdzie $i_0$ jest pewnym ustalonym indeksem zmiennej objaśniającej. Statystykę testową do testowania hipotezy $H_0$ przeciwko alternatywie $H_1$ łatwo skonstruować, rozpatrując studentyzowaną wartość $t = (b_{i_0} - \beta_{i_0})/SE_{b_{i_0}}$ równą, przy zachodzeniu $H_0$, $b_{i_0}/SE_{b_{i_0}}$. Zauważmy, że duże co do wartości bezwzględnej wartości $t$ wskazują na możliwość zachodzenia hipotezy alternatywnej. Okazuje się ponadto, że przy spełnieniu hipotezy $H_0$

$$t \sim t_{n-p},$$

gdzie $t_{n-p}$ oznacza statystykę $t$ Studenta z $n-p$ stopniami swobody. Zatem zbiór krytyczny

$$\{t\colon\ |t| \geq t_{1-\alpha/2,n-p}\}$$

określa test na poziomie istotności $\alpha$. Analogicznie można zbudować zbiór krytyczny dla testowania $H_0$ przeciwko hipotezie jednostronnej. Zastanówmy się jeszcze przez chwilę, co oznacza ostatnia rozpatrywana hipoteza zerowa. Pamiętajmy, że modelem, którego prawdziwość zakładamy, jest model (4.28), czyli oprócz zmiennej $x_{i_0}$ rozpatrujemy jeszcze wpływ $p-2$ innych zmiennych objaśniających. Hipoteza $H_0$ mówi, że zmienna $x_{i_0}$ nie ma istotnego wpływu na zmienną objaśnianą, *gdy w modelu uwzględniliśmy już wpływ zmiennych* $x_1,\ldots,x_{i_0-1},x_{i_0+1},\ldots,x_{p-1}$. W tym sensie powiększenie zbioru zmiennych objaśniających $I_{i_0}=\{x_1,\ldots,x_{i_0-1},x_{i_0+1},\ldots,x_{p-1}\}$ o zmienną $x_{i_0}$ nie jest celowe. Nie musi oznaczać to jednak, że zmienna $x_{i_0}$ po pominięciu pozostałych zmiennych nie ma wpływu na zmienną $Y$, a raczej, że funkcja objaśniająca tej zmiennej jest zrealizowana przez pozostałe zmienne modelu. Do problemu tego wrócimy w następnym punkcie przy okazji omawiania problemu selekcji zmiennych.

**Częściowe wykresy regresji.** W celu wizualizacji wpływu dowolnej zmiennej objaśniającej $x_{i_0}$ na zmienną objaśnianą po pominięciu liniowego wpływu pozostałych zmiennych objaśniających ze zbioru $I_{i_0}$ używa się często częściowych wykresów regresji.[2]. Rozpatrzmy regresję zmiennej $Y$ względem zmiennych objaśniających ze zbioru $I_{i_0}$. Niech $\mathbf{r}_{Y,i_0}$ będzie odpowiadającym wektorem rezyduów. Analogicznie niech $\mathbf{r}_{x_{i_0}}$ będzie wektorem rezyduów zmiennej $x_{i_0}$ otrzymanych po wykonaniu regresji zmiennej $x_{i_0}$ względem zmiennych ze zbioru $I_{i_0}$. Częściowy wykres regresji jest wykresem rozproszenia wektora $\mathbf{r}_{Y,i_0}$ względem wektora $\mathbf{r}_{x_{i_0}}$ ($i$-ty punkt na wykresie ma za pierwszą współrzędną $i$-ty punkt pierwszego wektora, a za drugą współrzędną $i$-ty punkt drugiego) uzupełnionym o prostą regresji $\mathbf{r}_{Y,i_0}$ od $\mathbf{r}_{x_{i_0}}$. Grupowanie się punktów na wykresie wokół prostej regresji o bliskim zeru współczynniku nachylenia świadczy o niewielkim wpływie zmiennej $x_{i_0}$, po uwzględnieniu wpływu pozostałych zmiennych, na zmienną $Y$. Gdy prosta ta ma istotnie niezerowy współczynnik, wpływ zmiennej $x_{i_0}$ jest znaczący. Jednocześnie nieliniowy charakter wykresu rozproszenia może świadczyć o nieliniowym wpływie zmiennej $x_{i_0}$ na zmienną $Y$ i sugerować odpowiednią transformację tej pierwszej. Wykres ten jest również często używany do wykrywania obserwacji wpływowych i odstających. Przykład zastosowania częściowego wykresu regresji omówiono w zad. 4.9.

**Prognoza.** Cel prognozy i metody rozwiązania tego problemu w przypadku regresji wielokrotnej są bardzo podobne do przypadku regresji jednokrotnej i dlatego omówimy tylko pokrótce konieczne adaptacje poprzednich rozwią-

[2]Nazwa angielska: **partial regression** lub **partial leverage plots.**

zań. Jak pamiętamy, możemy tutaj rozpatrzeć dwa różne problemy: przewidywanie wartości średniej zmiennej objaśniającej i przewidywanie przyszłej wartości tej zmiennej dla ustalonego wektora $\mathbf{x}_0$, różnego od wektorów $\mathbf{x}_i, i = 1, \ldots, n$. W obu przypadkach estymator poszukiwanej wielkości jest równy prognozie $\hat{Y}(\mathbf{x}_0) = \mathbf{x}_0'\mathbf{b}$. Dla pierwszego problemu, estymator $\hat{Y}(\mathbf{x}_0)$ jest nieobciążonym estymatorem wartości średniej $\mu_{Y(\mathbf{x}_0)}$

$$\mu_{\hat{Y}(\mathbf{x}_0)} = E\mathbf{x}_0'\mathbf{b} = \mathbf{x}_0'\beta = \mu_{Y(\mathbf{x}_0)}.$$

Ponadto, korzystając z równości (4.31) i stwierdzenia 4.8, łatwo otrzymujemy

$$\sigma^2_{\hat{Y}(\mathbf{x}_0)} = \mathbf{x}_0'\mathbf{\Sigma}_b\mathbf{x}_0 = \sigma^2\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0. \tag{4.36}$$

Ostatnia równość prowadzi do następującej definicji błędu standardowego estymatora $\hat{Y}(\mathbf{x}_0)$

$$SE_{\hat{Y}(\mathbf{x}_0)} = S\sqrt{\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0}$$

i rutynowej konstrukcji przedziału ufności dla $\mu_{Y(\mathbf{x}_0)}$ na poziomie istotności $1 - \alpha$, postaci $\hat{Y}(\mathbf{x}_0) \pm t_{1-\alpha/2,n-p}SE_{\hat{Y}(\mathbf{x}_0)}$.

W przypadku przewidywania przyszłej wartości $Y(\mathbf{x}_0)$, analogicznie jak w p. 4.2.3 otrzymujemy

$$\sigma^2_{\hat{Y}(\mathbf{x}_0)-Y(\mathbf{x}_0)} = \sigma^2(1 + \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0)$$

i stąd

$$SE^2_{\hat{Y}(\mathbf{x}_0)-Y(\mathbf{x}_0)} = S^2(1 + \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0).$$

Przedział ufności dla $Y(\mathbf{x}_0)$ ma postać $\hat{Y}(\mathbf{x}_0) \pm t_{1-\alpha/2,n-p}SE_{\hat{Y}(\mathbf{x}_0)-Y(\mathbf{x}_0)}$. Wzór (4.36) jest również przydatny do stwierdzenia, czy przy prognozie nie dokonujemy nieuzasadnionej ekstrapolacji danych. W takim przypadku możemy spodziewać się, że zmienność prognozy wartości $Y(\mathbf{x}_0)$ będzie duża. Zauważmy, że w przypadku więcej niż dwóch zmiennych objaśniających nie jesteśmy w stanie ocenić tego wizualnie. Jak przekonamy się przy omawianiu obserwacji wpływowych pewną miarą odstępstwa $\mathbf{x}_0$ od wartości średniej obserwacji jest wartość $h_0 = \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0$. Jeśli $h_0$ jest typową wartością z zakresu wyznaczonego przez wartości $h_i = \mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i$ dla obserwacji $X_1, X_2, \ldots, X_n$, nie zachodzi niebezpieczeństwo ekstrapolacji danych. W przeciwnym przypadku należy zaniechać prognozy; zauważmy, że w tej sytuacji przedział ufności byłby bardzo szeroki, wskazując na małą użyteczność takiego postępowania.

### 4.3.3. Diagnostyka modelu regresji

Najważniejszym narzędziem diagnostycznym jest tak samo jak w przypadku regresji jednokrotnej wykres rezyduów. Nie ulega zmianie identyfikacja takich odstępstw modelowych jak możliwa nieliniowość równania regresji, skorelowanie i niejednakowa zmienność poszczególnych błędów, jak również odstępstwo ich rozkładu od rozkładu normalnego. Poniżej omówimy kilka problemów diagnostycznych, których bądź dotąd nie omawialiśmy szczegółowo, bądź takich, które pojawiają się przy rozpatrzeniu kilku zmiennych objaśniających. Zanim jednak przejdziemy do omówienia, jak przejawiają się specyficzne odstępstwa modelowe, zastanówmy się nad możliwością zastosowania pewnego ogólnego podejścia do sprawdzania adekwatności modelu. Wiemy, że w przypadku, gdy model adekwatnie opisuje dane, estymator $S^2$ jest z reguły rozsądnym estymatorem wariancji błędów $\sigma^2$. W przeciwnym przypadku spodziewamy się, że $S^2$ przeszacowywuje $\sigma^2$. Nasuwa to pomysł, żeby sprawdzać jak bardzo wartość $S^2/\sigma^2$ odbiega od 1. Problem w zastosowaniu tej metody polega na tym, że z reguły nie znamy wartości wariancji $\sigma^2$. Możemy dysponować dobrym oszacowaniem $\hat{\sigma}^2$, otrzymanym np. na podstawie poprzednich eksperymentów i wtedy porównujemy wartość $S^2$ z wartością $\hat{\sigma}^2$. Gdy tak nie jest, potrzebujemy niezależnego od modelu estymatora wariancji. Z reguły możemy otrzymać go tylko wtedy, gdy dysponujemy tzw. **replikacjami**, czyli niezależnymi obserwacjami o takich samych wartościach wszystkich zmiennych objaśniających. Poza sytuacją eksperymentalną takich danych nie mamy praktycznie nigdy. Wtedy zmuszeni jesteśmy poszukiwać specyficznych odstępstw od założeń modelowych.

**Identyfikacja obserwacji odstających.** Obserwacje odstające w modelu regresji liniowej są to obserwacje nie spełniające równania (4.28). Jako takie wymagają identyfikacji i interpretacji. Chcemy wiedzieć, czy obserwacja odstająca jest wynikiem błędu w zapisie danych, czy być może wskazuje na to, że równanie (4.28) jest spełnione tylko dla pewnego zakresu wartości zmiennych objaśniających. Z tego powodu poświęcimy więcej miejsca identyfikacji takich obserwacji. Podstawowym narzędziem do tego celu jest wykres studentyzowanych rezyduów i jego modyfikacje. Korzystając z równości (4.33), błąd standardowy $i$-tego rezyduum można zdefiniować jako $SE_{e_i} = S\sqrt{1-h_i}$, gdzie $h_i = h_{ii}$ jest $i$-tym elementem diagonalnym macierzy $\mathbf{H}$. **Studentyzowana wartość resztowa** ma zatem postać

$$r_i = e_i / SE_{e_i}.$$

Rozpatrując wartości $r_i$ zamiast $e_i$ uwzględniamy różną zmienność rozkładów rezyduów, która może powodować, że niektóre wartości $e_i$ są pozornie odstające. Sporządzenie wykresu studentyzowanych rezyduów względem ich indeksu umożliwia zidentyfikowanie dużych wartości, które przypuszczalnie

odpowiadają obserwacjom odstającym. Metoda ta zawodzi jednak w przypadku wpływowej obserwacji odstającej $Y_i$, dla której różnica $Y_i - \mathbf{x}_i'\mathbf{b}$ jest mała. W celu poprawnej identyfikacji również tych obserwacji odstających rozpatruje się następującą modyfikację $i$-tego rezyduum

$$d_i = Y_i - \hat{Y}_{i(i)},$$

gdzie $\hat{Y}_{i(i)}$ jest wartością przewidywaną zmiennej objaśnianej w modelu regresji (4.28) dla $\mathbf{x} = \mathbf{x}_i$ na podstawie zbioru danych $\mathbf{J}_i$, który powstaje z całego zbioru przez pominięcie obserwacji $i$-tej:

$$\mathbf{J}_i = \{(Y_1, \mathbf{x}_1), \ldots, (Y_{i-1}, \mathbf{x}_{i-1}), (Y_{i+1}, \mathbf{x}_{i+1}), \ldots (Y_n, \mathbf{x}_n)\}.$$

Zauważmy, że dla wpływowej obserwacji odstającej wartość $d_i$ w odróżnieniu od wartości $e_i$ nie będzie bliska 0. Wielkość $d_i$ nazywamy **rezyduum modyfikowanym**, a jego studentyzowaną wersję

$$t_i = d_i / SE_{d_i}$$

**studentyzowanym rezyduum modyfikowanym**. Wartość błędu standardowego $SE_{d_i}$ możemy łatwo znaleźć, zauważając, że jest to estymator zmienności estymatora wartości oczekiwanej zmiennej objaśnianej dla $\mathbf{x} = \mathbf{x}_i$ na podstawie danych ze zbioru $\mathbf{J}_i$ i dokonując prostej modyfikacji jego błędu standardowego (por. wzór (4.36)). Okazuje się jednak, że szczęśliwie nie trzeba obliczać postaci estymatorów modyfikowanych w celu obliczenia wartości $t_i$, gdyż jest ona prosto związana z wartością $e_i$ i wartością $SSE$. Mianowicie, zachodzi równość

$$t_i = e_i \left( \frac{n-p-1}{SSE(1-h_{ii}) - e_i^2} \right)^{1/2}.$$

Ponadto zachodzi następująca własność umożliwiająca formalne stwierdzenie, czy dana obserwacja jest czy nie jest obserwacją odstającą na poziomie istotności $\alpha$, mianowicie

$$t_i \sim t_{n-p-1}. \tag{4.37}$$

Zmiana liczby stopni swobody z $n-p$ na $n-p-1$ jest związana z faktem, że estymacja błędu standardowego jest dokonywana na podstawie zbioru $\mathbf{J}_i$, zawierającego $n-1$, a nie $n$ elementów. Jednakże, gdy chcemy testować hipotezę, że żadna z obserwacji nie jest obserwacją odstającą, musimy odpowiednio zmodyfikować poziom istotności testu. Problem polega na tym, że nawet, jeśli testujemy hipotezę tylko dla indeksu $i_0$ odpowiadającego największemu $d_{i_0}$, to w sposób milczący testujemy ją dla wszystkich obserwacji, czyli zamiast hipotezy $H_0$: obserwacja $i_0$ nie jest obserwacją odstającą,

testujemy jednocześnie $n$ hipotez $H_{0i}$: obserwacja $i$ nie jest obserwacją odstającą, $i = 1, 2, \ldots, n$. Przypuśćmy, że chcemy przeprowadzić na pewnym poziomie istotności test hipotezy, że żadna obserwacja nie jest obserwacją odstającą i odrzucimy ją, gdy test hipotezy $H_{0i}$ oparty na własności (4.37) dla przynajmniej jednej obserwacji ją odrzuci. Wtedy

$$\begin{gathered} P(\text{żaden z testów nie odrzuca przy braku obserwacji odstających}) = \\ = 1 - P(\text{przynajmniej jeden test odrzuca}) \geq \\ \geq 1 - \sum_{i=1}^{n} P(i\text{-ty test odrzuca, gdy } H_{0i} \text{ prawdziwa}) = 1 - n\alpha, \end{gathered}$$

gdzie $\alpha$ oznacza poziom istotności indywidualnego testu. Z powyższej nierówności wynika, że jeśli chcemy, aby test dla wszystkich obserwacji miał poziom istotności $\alpha$, testy dla poszczególnych obserwacji powinny mieć poziomy istotności $\alpha/n$. Takie ustalanie poziomu indywidualnych testów jest znane jako **procedura Bonferroniego** i ma zastosowanie nie tylko do identyfikacji obserwacji odstających. Wadą tego podejścia jest to, że taki test dla wszystkich obserwacji ma w rzeczywistości poziom istotności znacznie mniejszy od $\alpha$, co powoduje, że znajduje on mniej obserwacji odstających niż test **dokładnie** na poziomie istotności $\alpha$.

Zauważmy na koniec, że występowanie dużych wartości rezyduów (a więc obserwacji potencjalnie odstających) może oznaczać, że wprawdzie jest spełnione równanie (4.28), ale rozkład błędów ma ogony znacznie wolniej malejące niż ogony rozkładu normalnego. W takim przypadku możemy skonstruować M-estymator wektora $\beta$ podobnie do M-estymatorów parametru położenia omówionych w p. 3.2.3. Odpowiednikiem minimalizacji funkcji (3.14) jest w takiej sytuacji minimalizacja funkcji

$$\sum_{i=1}^{n} \rho((y_i - \mathbf{x}_i'\beta)/\sigma),$$

gdzie funkcja $\rho$ jest jedną z funkcji rozpatrywanych w p. 3.2.3.

**Identyfikacja obserwacji wpływowych.** Obserwacją wpływową nazywamy taką obserwację, której usunięcie ze zbioru danych powoduje dużą zmianę wektora estymatorów MNK. Obserwacje odstające mogą, ale nie muszą być takimi obserwacjami. Inną grupę punktów, wśród których mogą znajdować się obserwacje wpływowe, stanowią te, dla których wektor wartości zmiennych objaśniających jest znacznie oddalony od typowego wektora wartości objaśniających ze zbioru danych. Zauważmy, że nie nazywamy takich obserwacji odstającymi, gdyż fakt przyjmowania nietypowych wartości dla zmiennych objaśniających nie ma nic wspólnego z zachodzeniem (lub nie) równania (4.28). W przypadku regresji jednokrotnej wyróżnienie tych punktów było proste, gdyż można było je wizualnie zidentyfikować

np. na podstawie histogramu wartości zmiennej objaśniającej. W przypadku wielu zmiennych znaczne odbieganie wektora $\mathbf{x}$ od wektora średnich $\bar{\mathbf{x}} = (1, \bar{x}_1, \ldots, \bar{x}_{p-1})$, gdzie $\bar{x}_k = \frac{1}{n}\sum_{i=1}^{n} x_{i,k}$, nie musi wcale oznaczać, że któraś ze współrzędnych wektora $\mathbf{x}$ będzie znacznie odstawać od odpowiadającej współrzędnej wektora średnich. Okazuje się jednak, że pewna globalna miara odstępstwa obserwacji $\mathbf{x}_i$ od $\bar{\mathbf{x}}$ jest zadana przez $i$-ty wyraz diagonalny $h_i = h_{ii}$ macierzy $\mathbf{H}$ (czasami zwany **wpływem** tej obserwacji). Ponieważ wiadomo, że

$$\sum_{i=1}^{n} h_i = p \text{ oraz dla każdego } i \quad 1 \geq h_i \geq \frac{1}{n},$$

możemy zatem przyjąć, że typowa wartość $h_i$ nie przekracza znacznie wartości $p/n$. Dlatego często przyjmuje się, że powinniśmy przyjrzeć się bliżej obserwacji $(\mathbf{x}_i, Y_i)$, dla której

$$h_i \geq \frac{2p}{n}; \tag{4.38}$$

jest to potencjalna obserwacja wpływowa. Oczywiście, próg $2p/n$ w tej regule jest arbitralny, co czyni ją całkowicie heurystyczną. Dodatkowym uzasadnieniem wnikliwego rozpatrzenia obserwacji o dużym wpływie $h_i$ jest fakt wynikający z równania (4.33), a mianowicie że $\mathrm{Var}(e_i) = \sigma^2(1 - h_i)$. Zatem, dla obserwacji o dużych wpływach, wariancja odpowiadającego rezyduum jest mała. Inaczej mówiąc, w takiej sytuacji wartość przewidywana jest „zmuszana" do pozostawania blisko zaobserwowanej wartości. Każdą potencjalnie wpływową obserwację spełniającą (4.38) usuwamy kolejno ze zbioru danych i sprawdzamy na ile zmienił się nowy wektor współczynników $\mathbf{b}$ w porównaniu z wektorem $\mathbf{b}$ obliczonym na podstawie całego zbioru danych lub na ile zmieniły się prognozy zaobserwowanych wartości.

Alternatywnie, proces identyfikacji obserwacji wpływowych można oprzeć na obliczeniu tzw. **odległości Cooke'a** $D_i$. Jest ona oparta na koncepcji użytej przy wprowadzaniu modyfikowanych rezyduów:

$$D_i = \frac{\sum_{j=1}^{n}(\hat{Y}_j - \hat{Y}_{j(i)})^2}{pS^2} = \frac{e_i^2}{pS^2}\frac{h_i}{(1-h_i)^2}, \tag{4.39}$$

gdzie $\hat{Y}_{j(i)}$ jest wartością przewidywaną dla $j$-tej obserwacji obliczoną na podstawie danych z usuniętą obserwacją $i$-tą. Druga z równości jest własnością odległości Cooke'a ułatwiającą jej obliczenie. Tak więc wartość $D_i$ odpowiada wpływowi, jaki na prognozę znanych wartości zmiennej objaśnianej ma usunięcie ze zbioru danych $i$-tej obserwacji. Duża wartość $D_i$ wskazuje na

znaczny wpływ usunięcia, czyli $i$-ta obserwacja jest obserwacją wpływową. **Diagram Cooke'a**, będący wykresem odległości Cooke'a w funkcji indeksu obserwacji, służy do identyfikacji obserwacji wpływowych. Zauważmy, że duża wartość $h_i$ jest tylko jedną z przyczyn dużej wartości $D_i$. Obserwacja może być wpływowa (mieć dużą wartość $D_i$) przy umiarkowanej wartości $h_i$, ale mając dużą wartość resztową. Ostatnia uwaga tłumaczy, dlaczego często preferuje się analizę odległości Cooke'a zamiast analizy wartości wpływów.

**Współliniowość.** Rozważmy teraz problem współliniowości zmiennych. Problem ten nie występował w przypadku regresji jednokrotnej. Rozpatrzmy przykładowe równanie regresji

$$Y = 5x_1 + 2x_2 + \varepsilon \tag{4.40}$$

w sytuacji, gdy zmienne $x_1$ i $x_2$ są związane deterministyczną zależnością liniową $x_2 = 2x_1$. Wtedy można łatwo zauważyć, że zmienna $Y$ jest opisana nie tylko równaniem (4.40), ale również np. równaniami $Y = 7x_1 + x_2 + \varepsilon$ oraz $Y = 11x_1 - x_2 + \varepsilon$. Tak więc w przypadku liniowej zależności zmiennych ten sam model regresji liniowej jest opisany wieloma różnymi równaniami. Możemy teraz łatwo wyobrazić sobie, co będzie się działo z estymatorami parametrów dla zależności (4.40). Ponieważ dane są opisywane wieloma „konkurencyjnymi" równaniami, estymator $(b_1, b_2)$ może być równie dobrze bliski wektorowi $(5, 2)$, jak i np. wektorowi $(11, -1)$. Oczywiście, w takiej sytuacji trudno poprawnie interpretować otrzymane wartości estymatorów: np. estymator $b_2 = 2$ sugeruje dodatnią zależność zmiennych $x_2$ i $Y$ przy ustalonym poziomie zmiennej $x_1$; odwrotny wniosek wyciągnęlibyśmy po otrzymaniu wartości $b_2 = -1$. Jednocześnie jest jasne, że w takiej sytuacji zmienność estymatorów może być bardzo duża. Sytuacja taka zachodzi nie tylko w przypadku ścisłej, ale również przybliżonej liniowej zależności (**współliniowości**) zmiennych objaśniających. Odpowiada ona warunkowi, że macierz $\mathbf{X'X}$ jest bliska macierzy nieodwracalnej. Postać estymatorów MNK podana w (4.30) tłumaczy, dlaczego w tym przypadku możemy spodziewać się ich niestabilnego zachowania.

Jak wykryć fakt występowania współliniowości zmiennych objaśniających? Jedną z możliwości jest analiza korelacji tych zmiennych. Jeśli wartość bezwzględna współczynnika korelacji $|r_{x_1,x_2}|$ jest bliska 1, to wskazuje ona na przybliżoną liniową zależność zmiennych $x_1$ i $x_2$. Używając tej metody nie wykryjemy jednak związków liniowych wiążących więcej niż dwie zmienne jednocześnie. W tym celu często używa się współczynników determinacji wielokrotnej $R_i^2$ dla hipotetycznego modelu liniowego, w którym $x_i$ jest zmienną objaśnianą, a $x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_{p-1}$ są zmiennymi objaśniającymi. Wartość $R_i^2$ bliska 1 wskazuje na współliniowość zmiennej $x_i$ i pozostałych zmiennych objaśniających. Alternatywnie, analizuje

się tzw. współczynniki podbicia wariancji (ang. *variance inflation factor*) $VIF_i = (1 - R_i^2)^{-1}$. Duża wartość $VIF_i$ dla pewnego $i$ wskazuje na potencjalną liniową zależność $i$-tej zmiennej objaśniającej od zmiennych pozostałych. Uzasadnieniem użycia współczynników podbicia wariancji jest związek między wariancją estymatora $b_i$ a $VIF_i$. Można udowodnić, że

$$\sigma_{b_i}^2 = \sigma^2 VIF_i / (\sum_{k=1}^{n} (x_{ki} - \bar{x}_i)^2),$$

zatem duża wartość współczynnika $VIF_i$ z reguły pociąga za sobą dużą zmienność estymatora $b_i$, co może być spowodowane współliniowością w danych.

Jednym ze sposobów radzenia sobie ze zjawiskiem współliniowości jest użycie zamiast estymatorów MNK estymatorów otrzymanych metodą tzw. **regresji grzbietowej** (p. (J. Koronacki, J. Ćwik (2005))). Tej ostatniej metody nie będziemy tutaj dokładnie omawiać. Stwierdźmy jedynie, że opiera się ona na nałożeniu górnego ograniczenia na wartość sumy kwadratów współrzędnych wektora $\beta$ (równej $\beta'\beta$) i minimalizacji funkcji $Q$ z def. 4.5 przy tym ograniczeniu.

**Wybór zmiennych objaśniających w liniowym modelu regresji.** Celem selekcji zmiennych w regresji jest wybór „najlepszego" podzbioru zmiennych objaśniających. Jednym z podstawowych powodów, dla których staramy się ograniczyć zbiór zmiennych mogących mieć wpływ na zmienną objaśnianą, jest zasada metodologiczna, z której wynika, że spośród grona modeli adekwatnie opisujących zbiór danych, najlepszym modelem jest model najprostszy. Nie wdając się tutaj w szczegółowe omówienie tej zasady, zauważmy tylko, że w analizie regresji prosty model daje największą możliwość zrozumienia istoty zależności zmiennej objaśnianej od zmiennych objaśniających. Ponadto, powracając do roli statystyków zaobserwujmy, że estymacja dużej liczby parametrów, wśród których są przypuszczalnie parametry zbyteczne, z reguły prowadzi do dużej zmienności estymatorów. Estymatory te stają się tym samym mniej użyteczne, niż w przypadku prostego modelu z prawidłowo wybranymi zmiennymi. Ponadto, nie starając ograniczyć się zbioru zmiennych, narażamy się na to, że część z nich może być współliniowa, co jak wiemy ma negatywny wpływ na stabilność i możliwości interpretacji estymatorów.

W przypadku małej, nie większej niż 4 liczby potencjalnych zmiennych objaśniających jesteśmy w stanie przeanalizować szczegółowo adekwatność modelu regresji zmiennej objaśnianej względem dowolnego podzbioru tych zmiennych. Dla większej liczby zmiennych jesteśmy zmuszeni uciec się do metod automatycznej selekcji.

Omówimy tylko dwie metody selekcji sekwencyjnej: **metodę eliminacji** i **metodę dołączania**. **Metodę eliminacji** inicjuje się w modelu, w którym uwzględniono wszystkie potencjalnie interesujące nas zmienne (krok 1). Następnie, zakładając prawdziwość tego modelu, testujemy indywidualne hipotezy o istotności poszczególnych zmiennych i usuwamy tę zmienną, dla której $p$–wartość odpowiadającego testu $t$ jest *największą* $p$–wartością przekraczającą ustalony poziom $\alpha$ (krok 2). Później dopasowujemy mniejszy model z usuniętą zmienną i powracamy do kroku 2. Procedurę przerywamy, gdy w pewnym kroku wszystkie $p$–wartości są mniejsze od $\alpha$. **Metoda dołączania** startuje od modelu zawierającego tylko stałą (krok 1), następnie wybiera się tę spośród możliwych zmiennych, dla których $p$–wartość odpowiadającego jej testu $t$ jest *najmniejszą* $p$–wartością mniejszą od $\alpha$ (krok 2). Rozpatrując wszystkie możliwe zmienne nie znajdujące się w modelu, powtarzamy krok 2. Procedurę zatrzymujemy, kiedy żadnemu z potencjalnych kandydatów na włączenie do modelu nie odpowiada $p$–wartość mniejsza od $\alpha$.

Metody sekwencyjne nie są czasochłonne obliczeniowo, mają jednak kilka wad. Najpoważniejszą z nich jest fakt, że pechowy wybór zmiennej dokonany na pewnym etapie selekcji nie może być już później skorygowany. Jeśli np. faktycznymi zmiennymi objaśniającymi są zmienne $\{x_1, x_2\}$, a na pierwszym etapie procedury dołączania wybierzemy zmienną $x_3$, to nigdy już nie otrzymamy właściwego zbioru zmiennych objaśniających. Wady tej nie ma metoda **selekcji (regresji) krokowej**, będąca połączeniem metody eliminacji i metody dołączania, w której na każdym kroku można odrzucić lub dodać zmienną. Przypuśćmy przykładowo, że na pewnym etapie selekcji wybraliśmy zmienne $x_3, x_5$ ze zbioru zmiennych $\{x_1, \ldots, x_7\}$. Postępując teraz tak jak w metodzie dołączania, staramy się powiększyć zbiór $\{x_3, x_5\}$ o jedną z pozostałych zmiennych. Załóżmy, że dołączamy zmienną $x_1$, tzn. $p$–wartość odpowiadającego jej testu $t$ jest najmniejszą $p$–wartością mniejszą od $\alpha$. Przeglądamy teraz zmienne uprzednio włączone do modelu, w tym przypadku zmienne $x_3$ i $x_5$, i stosując metodę eliminacji sprawdzamy, czy któraś z nich nie jest zbyteczna (przy włączonej do modelu zmiennej $x_1$). Postępowanie to powtarzamy na każdym kroku selekcji krokowej.

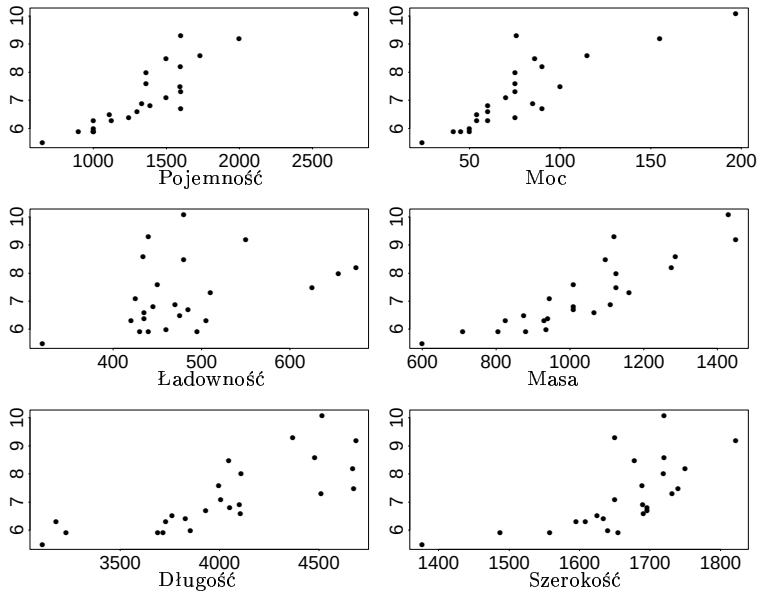
Problemem metod selekcji jest wybór progu $\alpha$, który musi być wybierany arbitralnie (najczęściej przyjmuje się $\alpha = 0,05 \div 0,1$ dla procedury eliminacji i $\alpha = 0,1 \div 0,2$ dla metody dołączania) i nie można nadać mu konkretnej interpretacji. Związane to jest z tym, że w opisanych procedurach selekcji mamy do czynienia z testowaniem wielu hipotez, których wyniki zależą od siebie. Nie opisujemy tu często stosowanych procedur selekcji opartych na minimalizacji kryteriów z jednej strony „promujących" jakość dopasowania postulowanego modelu do danych, a z drugiej strony „karzących" za dużą liczbę parametrów w modelu. Są one omówione w specjalistycznych książkach poświęconych analizie regresji.

### 4.3.4. Analiza zależności parametrów samochodów

Rozpatrzmy dane omówione w przykł. 4.3. Diagramy rozproszenia zużycia paliwa względem pojedynczych zmiennych objaśniających przedstawiono na rys. 4.11. Stwierdzamy istnienie silnej zależności zużycia paliwa od masy, pojemności i mocy, umiarkowanie silnej od szerokości i długości samochodu i praktyczny brak zależności od jego ładowności. Przykładowe wartości $R^2$ wynoszą 0,77 dla pary zużycie i pojemność, 0,76 dla pary zużycie i moc, i $R^2 = 0{,}60$ dla pary zużycie i szerokość. Rozpatrzmy najpierw model regresji liniowej uwzględniający wszystkie sześć zmiennych objaśniających. Wartość $R^2$ dla takiego modelu wynosi 0,87, czyli 87% zmienności zużycia paliwa jest tłumaczone przez ten model zależności. Test $F$ odrzuca hipotezę $H_0$: $\beta_1 = \beta_2 = ... = \beta_6 = 0$ (liczby stopni swobody wynoszą $p - 1 = 6$ i $n - p = 17$), odpowiednia $p$–wartość jest mniejsza od 0,001. Rozpatrzmy teraz indywidualne testy istotności dla współczynników. Stwierdzamy, że na poziomie istotności 0,05, tylko współczynniki $\beta_4$ i $\beta_6$ odpowiadające masie i szerokości, okazują się statystycznie różne od 0. O ile na podstawie wykresów rozproszenia spodziewamy się, że ładowność nie ma istotnego liniowego wpływu na zużycie paliwa, o tyle na pierwszy rzut oka wydaje się dziwne, że dotyczy to również pojemności, mocy i długości. Wyjaśnieniem tego faktu jest bardzo silna zależność między pojemnością a mocą (współczynnik korelacji między tymi zmiennymi wynosi 0,95). Dlatego nie odrzuciliśmy hipotezy $H_0$: $\beta_1 = 0$ jak również hipotezy $H_0$: $\beta_2 = 0$. Gdy w modelu uwzględniono pojemność, moc silnika nie ma już praktycznie wartości objaśniającej. Podobnie ma się rzecz z pojemnością, gdy w modelu uwzględnimy moc silnika. Również długość samochodu jest bardzo silnie zależna od masy ($R^2 = 0{,}84$) i szerokości ($R^2 = 0{,}76$). Zauważmy, że bardzo duże wartości współczynnika podbicia wariancji dla mocy ($VIF_2 = 17{,}45$), masy ($VIF_4 = 23{,}71$) i pojemności ($VIF_1 = 14{,}69$) świadczą o występowaniu współliniowości wśród zmiennych objaśniających.

Rozpatrzmy teraz mniejszy model uwzględniający jako zmienne objaśniające tylko masę i szerokość. W tabeli 4.4 przedstawiono wyniki analizy regresji dla tego modelu. Zauważmy, że dla tego modelu $R^2 = 0{,}83$, a zatem procent objaśnianej zmienności jest praktycznie taki sam co dla rozpatrywanego poprzednio pełnego modelu. Czytelnikowi pozostawiamy stwierdzenie, że indywidualne dołączenie każdej ze zmiennych: długość, pojemność i ładowność nie prowadzi do istotnego poprawienia modelu (test $t$ odrzuca hipotezy o istotności dołączonych zmiennych), a powiększenie modelu o zmienną mocy prowadzi do problemu współliniowości.

Na podstawie wykresu kwantylowego stwierdzamy, że rozkład rezyduów dla mniejszego modelu umiarkowanie odstaje od rozkładu normalnego. Ponieważ w analizie posługujemy się głównie testem $t$, który jest odporny na nie-




Rys. 4.11. Wykresy rozproszenia zużycia paliwa od pojemności silnika, mocy, ładowności, masy, długości i szerokości samochodu (przykł. 4.3)

duże odstępstwa od normalności, odstępstwo takie jest dopuszczalne. Analiza studentyzowanych i modyfikowanych studentyzowanych rezyduów umożliwia stwierdzenie, że wartości odpowiadające Polonezowi Caro są wyraźnie większe od innych: $r_4 = 2{,}56$ i $t_4 = 3{,}02$. Jest to potencjalnie obserwacja odstająca. Ma ona również (obok Malucha) o rząd większą od innych obserwacji wartość odległości Cooke'a $D = 0{,}21$. Usunięcie jej ze zbioru danych powoduje wiecej niż dziesięcioprocentową zmianę wartości współczynników modelu i wzrost współczynnika determinacji do $R^2 = 0{,}87$. Usunięcie Malucha (przy pozostawieniu Poloneza Caro) nie jest tak znaczące, choć powoduje również około dziesięcioprocentową zmianę współczynnika odpowiadającego szerokości. Usuńmy obie te obserwacje ze zbioru danych. Przeglądając wartości odległości Cooke'a znajdujemy jedną obserwację potencjalnie wpływową. Jest to Mercedes C280T, dla którego wartość odległości Cooke'a jest równa $D = 0{,}33$, czyli o rząd przekracza inne wartości odległości. Z analizy tabeli danych wynika natychmiast, że Mercedes ma największe zużycie paliwa (10,1 litra). Po usunięciu również tej obserwacji równanie regresji ma postać

$$\text{Zużycie} = 7{,}066 + 0{,}0064 \times \text{Masa} + 0{,}0040 \times \text{Szerokość}$$

i współczynnik determinacji zmniejsza wartość z $R^2 = 0{,}86$ (dla danych bez Poloneza i Malucha) do $R^2 = 0,80$. Jest to związane z bardzo dużą zmianą $SST$ i $SSR$ po usunięciu tej obserwacji. Taka sytuacja może się zdarzyć

przy pominięciu obserwacji wpływowej. Całkowita suma kwadratów SST zmniejsza się bowiem z 28,09 do 19,25, a regresyjna suma kwadratów SSR zmniejsza się z 24,06 do 15,42.

**Tabela 4.4. Wydruk z pakietu SAS dotyczący analizy regresji dla przykł. 4.3**

```
The SAS System      20:43 Friday, December 1, 2000   1

Model Equation
ZUZYCIE   =    10.8078   +   0.0081  MASA    -    0.0072   SZEROKOS

                      Summary of Fit
Mean of Response            7.2125  R-Square        0.8335
Root MSE                    0.5297  Adj R-Sq        0.8176

                    Analysis of Variance
Source      DF  Sum of Squares  Mean Square      F Stat    Prob > F
Model        2         29.4939      14.7470     52.5578      0.0001
Error       21          5.8923       0.2806           .           .
C Total     23         35.3862            .           .           .

                    Type III Tests
Source      DF  Sum of Squares  Mean Square      F Stat    Prob > F
MASA         1         13.9873      13.9873     49.8502      0.0001
SZEROKOS     1          2.1632       2.1632      7.7097      0.0113

                    Parameter Estimates                          Var
Variable   DF  Estimate  Std Error   T Stat  Prob>|T|  Tolerance  Inflation
INTERCEPT   1   10.8078     3.3027   3.2724    0.0036          .     0.0000
MASA        1    0.0081     0.0011   7.0605    0.0001     0.2177     4.5942
SZEROKOS    1   -0.0072     0.0026  -2.7766    0.0113     0.2177     4.5942
```

## 4.4. Zadania

**4.1.** Rozpatrzmy próbę dwuwymiarową postaci

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $x$: | $-5$ | $-4$ | $-3$ | $-2$ | $-1$ | $0$ | $1$ | $2$ | $3$ | $4$ | $5$ |
| $y$: | $15$ | $6$ | $-1$ | $-6$ | $-9$ | $-10$ | $-9$ | $-6$ | $-1$ | $6$ | $15$ |

Narysować wykres rozproszenia danych i obliczyć próbkowy współczynnik korelacji. Czy na podstawie wykresu można stwierdzić, że między zmiennymi $x$ i $y$ istnieje zależność? Pominąć dwie ostatnie obserwacje i obliczyć jeszcze raz współczynnik korelacji. Czy zmienne $x$ i $y$ są ujemnie zależne? Zinterpretować wyniki w świetle własności (2) próbkowego współczynnika korelacji.

**4.2.** W tabeli 4.5 podano 62 wyniki pomiaru wzrostu osoby, jej latencji L5-P40 oraz wartości rezyduum R-P40, otrzymane w wyniku rozwiązania zadania regresji liniowej ze wzrostem jako zmienną objaśniającą i latencją jako zmienną objaśnianą. Przeprowadzić pełną analizę regresji latencji L5-P40 względem wzrostu i ocenić czy rozwiązanie można uznać za zadowalające.

**Tabela 4.5. Dane do zad. 4.2**

| Wzrost | L5-P40 | R-P40 | Wzrost | L5-P40 | R-P40 | Wzrost | L5-P40 | R-P40 |
|---|---|---|---|---|---|---|---|---|
| 1,67 | 50,15 | 1,129 | 1,86 | 58,45 | 5,176 | 1,73 | 53,35 | 2,986 |
| 1,59 | 44,50 | −2,730 | 1,76 | 50,60 | −0,436 | 1,82 | 50,80 | −1,579 |
| 1,61 | 48,60 | 0,922 | 1,83 | 54,10 | 1,497 | 1,75 | 54,50 | 3,688 |
| 1,50 | 46,90 | 1,685 | 1,75 | 52,25 | 1,438 | 1,69 | 43,25 | −6,219 |
| 1,67 | 51,25 | 2,229 | 1,68 | 50,50 | 1,255 | 1,76 | 50,75 | −0,286 |
| 1,62 | 46,50 | −1,402 | 1,61 | 49,45 | 1,772 | 1,64 | 44,55 | −3,799 |
| 1,76 | 49,10 | −1,936 | 1,76 | 47,95 | −3,086 | 1,76 | 47,15 | −3,886 |
| 1,67 | 48,40 | −0,621 | 1,66 | 48,70 | −0,097 | 1,75 | 47,65 | −3,162 |
| 1,69 | 50,30 | 0,831 | 1,76 | 51,05 | 0,014 | 1,62 | 51,30 | 3,398 |
| 1,76 | 53,95 | 2,914 | 1,80 | 58,60 | 6,669 | 1,92 | 54,50 | −0,117 |
| 1,74 | 46,05 | −4,538 | 1,72 | 47,25 | −2,890 | 1,70 | 45,45 | −4,243 |
| 1,88 | 57,20 | 3,478 | 1,53 | 44,50 | −1,387 | 1,64 | 53,85 | 5,501 |
| 1,58 | 48,45 | 1,444 | 1,76 | 48,65 | −2,386 | 1,69 | 50,15 | 0,681 |
| 1,75 | 51,25 | 0,438 | 1,92 | 48,60 | −6,017 | 1,65 | 48,45 | −0,123 |
| 1,61 | 45,00 | −2,678 | 1,58 | 46,80 | −0,206 | 1,53 | 46,60 | 0,713 |
| 1,58 | 46,95 | −0,056 | 1,80 | 51,80 | −0,131 | 1,68 | 52,45 | 3,205 |
| 1,58 | 45,75 | −1,256 | 1,69 | 47,60 | −1,869 | 1,55 | 45,70 | −0,635 |
| 1,68 | 47,05 | −2,195 | 1,57 | 49,25 | 2,468 | 1,67 | 54,00 | 4,979 |
| 1,68 | 48,80 | −0,445 | 1,58 | 49,20 | 2,194 | 1,81 | 54,15 | 1,995 |
| 1,58 | 45,10 | −1,906 | 1,65 | 47,55 | −1,023 | 1,60 | 47,90 | 0,446 |
| 1,60 | 44,00 | −3,454 | 1,62 | 49,55 | 1,648 | | | |

**4.3.** Na podstawie tab. 4.2 zawierającej wydruk SAS analizy regresji dla przykł. 4.1:

**a)** Obliczyć korelację między wynikiem kolokwium a wynikiem egzaminu końcowego i wariancję wyniku egzaminu końcowego.

**b)** Przeprowadzić test hipotezy $H_0$: $\beta_1 = 1{,}7$ przeciwko hipotezie $H_1$: $\beta_1 \neq$ $\neq 1{,}7$ i przeciwko hipotezie $H_1$: $\beta_1 > 1{,}7$.

**4.4.** W poniższej tabeli przedstawiono 20 wyników dotyczących ciężaru odlewu przed (zmienna $x$) i po obróbce skrawaniem (zmienna $y$); wartości obu zmiennych są podane w kilogramach.

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $x$ | 2,570 | 2,565 | 2,550 | 2,555 | 2,555 | 2,590 | 2,580 | 2,545 |
| $y$ | 1,935 | 1,920 | 1,905 | 1,910 | 1,900 | 1,950 | 1,905 | 1,915 |
| $x$ | 2,535 | 2,530 | 2,525 | 2,520 | 2,515 | 2,515 | 2,490 | 2,490 |
| $y$ | 1,890 | 1,880 | 1,885 | 1,870 | 1,885 | 1,890 | 1,875 | 1,895 |
| $x$ | 2,520 | 2,575 | 2,600 | 2,645 | | | | |
| $y$ | 1,870 | 1,835 | 1,945 | 2,015 | | | | |

**a)** Sporządzić wykres rozproszenia zmiennej $y$ względem zmiennej $x$ i nanieść na niego prostą regresji. Na podstawie wykresu zidentyfikować dwie obserwacje odstające. Sprawdzić, czy odpowiadają im największe wartości studentyzowanych i modyfikowanych studentyzowanych rezyduów[3].

**b)** Usunąć jedną z obserwacji odstających ze zbioru danych, zachowując drugą i odwrotnie. Narysować odpowiednie proste regresji. Usunięcie której z obserwacji ma większy wpływ na nachylenie prostej regresji? Jaka jest wartość odległości Cooke'a dla tej obserwacji?

**c)** Usunąć obie obserwacje odstające ze zbioru danych i nanieść prostą regresji na nowy wykres rozproszenia. Sporządzić wykres rezyduów w funkcji wartości przewidywanych i przeprowadzić diagnostykę modelu. Czy model regresji jednokrotnej można uznać za adekwatny? Przeprowadzić test hipotezy $\beta_1 = 0$ przeciwko hipotezie $\beta_1 \neq 0$ i test analogicznej hipotezy dla wyrazu wolnego. Skonstruować przedział ufności dla współczynnika nachylenia na poziomie ufności 0,95 i sprawdzić, czy jego wartość zgadza się ze stwierdzeniem 4.4 i wartością $S$ obliczoną w pakiecie.

**d)** Porównać i zinterpretować wartości współczynnika determinacji w sytuacji a i c.

**e)** Na podstawie obliczeń z punktu c obliczyć średnią wartość odlewu po skrawaniu dla wartości odlewu nieobrobionego równego 2,57 kg.

**4.5.** Rozpatrzmy model regresji jak w równaniu (4.12), ale z pominięciem wyrazu stałego

$$Y_i = \beta_1 x_i + \varepsilon_i \qquad i = 1, 2, \ldots, n,$$

gdzie $\varepsilon_i, i = 1, 2, \ldots, n$ są niezależnymi błędami o rozkładzie normalnym $N(0, \sigma)$.

**a)** Postępując analogicznie jak w p. 4.2.2 wykazać, że postać estymatora współczynnika kierunkowego $\beta_1$ metodą najmniejszych kwadratów w tym modelu ma postać

$$b_1 = \frac{\sum_{i=1}^{n} Y_i x_i}{\sum_{i=1}^{n} x_i^2}$$

(można ją formalnie otrzymać po podstawieniu $\bar{x} = 0$ w (4.7)). Ponieważ estymuje się w tym modelu jeden, a nie dwa współczynniki równania regresji, więc liczba stopni swobody związana z sumą kwadratów błędów $SSE = \sum_{i=1}^{n}(Y_i - b_1 x_i)^2$ wynosi $n-1$ i nieobciążonym estymatorem wariancji błędów w tym modelu $\sigma^2$ jest $S^2 = (n-1)^{-1} SSE$. Statystyka $(b_1 - \beta_1)/S$ służy

[3] W niektórych pakietach rezydua studentyzowane nazywa się rezyduami standaryzowanymi, a modyfikowane rezydua studentyzowane zwane są rezyduami studentyzowanymi. Warto sprawdzić, jaka terminologia jest stosowana!

do testowania hipotezy, że w uproszczonym modelu wartość współczynnika nachylenia faktycznie wynosi $\beta_1$.

**b)** W zadaniu 4.4 c test hipotezy $H_0$: $\beta_0 = 0$ przeciwko $H_1$: $\beta_0 \neq 0$ nie został odrzucony, co nasuwa podejrzenie, że być może w tym przypadku wyraz stały w modelu regresji można pominąć. Dopasować prostą regresji, zakładając prawdziwość uproszczonego modelu, przeprowadzić jego diagnostykę i porównać wyniki z wynikami zad. 4.3 c. Który model bardziej adekwatnie opisuje dane? Przeprowadzić Test hipotezy $H_0$: $\beta_1 = 0$ przeciwko $H_1$: $\beta_1 \neq 0$.

**4.6. a)** Na podstawie danych dotyczących wielkości opadów i temperatury w lipcu w Warszawie (zad. 1.3) sporządzić wykres rozproszenia temperatury od wielkości opadów i nanieść na niego prostą regresji. Na podstawie wykresu rezyduów przeprowadzić diagnostykę modelu. Jaki procent zmienności temperatury jest tłumaczony przez zaproponowaną zależność od ilości opadów, a jaki nie jest? Czy wykres rezyduów umożliwia wyjaśnienie bardzo niskiej wartości współczynnika determinacji? Wykonać testy dotyczące wyrazu wolnego i współczynnika nachylenia prostej regresji.

**b)** Zauważyć, że jeśli pominie się jednostki przy analizie rozproszenia wartości $x$ i $y$, to można stwierdzić, że zakres zmienności poziomu opadów jest dziesięciokrotnie większy od zakresu zmienności temperatury. Jak zmieniłaby się wartość estymatora $b_1$, gdyby temperaturę wyrażać w nowych jednostkach równych 0,1 stopnia (tzn. 20°C byłoby równe 200 jednostkom w nowej skali)? Czy taka sama zmiana nastąpiłaby, gdyby poziom opadów rejestrować w centymetrach, a nie w milimetrach?

**4.7.** Dane omówione szczegółowo w zad. 4.7 zawierają wielkości zawartości tkanki tłuszczowej (zmienna $x_2$) i wagi (zmienna $x_4$) dla grupy 252 mężczyzn.

**a)** Narysować diagram rozproszenia zmiennej $x_2$ od zmiennej $x_4$, nanieść na niego prostą regresji i przeprowadzić diagnostykę modelu. Obliczyć współczynnik korelacji między zmiennymi. Zinterpretować wyniki.

**b)** W przypadku, gdy próba dwuwymiarowa jest prostą próbą losową z dwuwymiarowego rozkładu normalnego ze współczynnikiem korelacji $\rho = 0$ i $r$ oznacza współczynnik korelacji próbkowej, statystyka $r\sqrt{n-2}/(1-r^2)$ ma rozkład Studenta $t_{n-2}$. Wyznaczyć zbiór krytyczny dla testu hipotezy $H_0$: $\rho = 0$ przeciwko $H_1$: $\rho > 0$ w tej sytuacji. Sprawdzić na podstawie wykresów kwantylowych, że próby wartości zawartości tkanki tłuszczowej i wagi można uznać za realizację prób z rozkładu bliskiego rozkładowi normalnemu i przetestować hipotezę, że odpowiednie zmienne są niezależne (to jest $\rho = 0$) przeciwko hipotezie, że są dodatnio zależne.

**4.8.** Dane zebrane przez dr. G. Fischera i udostępnione do ogólngo użytku w bibliotece StatLib (zbiór *bodyfat* w `http://lib.stat.cmu.edu/datasets`)

dotyczą zależności między procentową zawartością tkanki tłuszczowej w ciele ludzkim i jego rozmaitych parametrów antropometrycznych (między innymi obwodu szyi, klatki piersiowej, bioder, brzucha, bicepsu, wagi itp.). Ilość tkanki tłuszczowej oblicza się na podstawie zależności empirycznej zwanej równaniem Siriego, wiążącej tę ilość z gęstością ciała. Do obliczenia gęstości ciała niezbędna jest znajomość jego objętości, której wartość uzyskuje się na podstawie zanurzenia go w wodzie. Ze względu na kłopotliwość tej procedury ważne jest wyznaczenie adekwatnego zbioru zmiennych objaśniających dla zawartości tłuszczu $x_2$. We wstępnej analizie stwierdzono, że największe współczynniki determinacji jednokrotnej zmienna $x_2$ osiąga z następującymi zmiennymi: obwodem brzucha $x_8$ ($R^2 = 0{,}66$), obwodem klatki piersiowej $x_7$ ($R^2 = 0{,}49$), ciężarem $x_4$ ($R^2 = 0{,}37$) i obwodem bioder $x_{10}$ ($R^2 = 0{,}32$).

**a)** Równanie Siriego stwierdza, że procentowa zawartość tkanki tłuszczowej $x_2$ spełnia równanie $x_2 = \beta_0 + \beta_1/x_1$, gdzie $x_1$ jest gęstością ciała. Odpowiednio przekształcając zmienną $x_1$ obliczyć przybliżone wartości $\beta_0$ i $\beta_1$ na podstawie zbioru *bodyfat*. Przeanalizować wykres rezyduów. Czy zależność ma raczej charakter deterministyczy czy losowy? Uzasadnić.

**b)** Sporządzić wykresy rozproszenia zmiennej objaśnianej $x_2$ względem zmiennych objaśniających $x_4, x_7, x_8$ i $x_{10}$ i zmiennych objaśniających względem siebie (łącznie 10 wykresów). Przeanalizować wykresy. Obliczyć odpowiadające współczynniki korelacji i sprawdzić, czy adekwatnie mierzą one siłę zależności badanych zmiennych.

**c)** Dokonać analizy regresji wielokrotnej zmiennej $x_2$ względem zmiennych $x_4, x_7, x_8$ i $x_{10}$. Ocenić na podstawie wykresu kwantylowego rezyduów czy można przyjąć założenie o normalności błędów.

**d)** Przeprowadzić testy dla poszczególnych współczynników równania regresji. Wyjaśnić pozorną sprzeczność między nieodrzucaniem testu o zerowości współczynnika dla obwodu klatki piersiowej $x_7$ i stosunkowo dużą wartością współczynnika determinacji jednokrotnej między zmiennymi $x_2$ i $x_7$.

**e)** Usunąć zmienną obwodu klatki piersiowej $x_7$ ze zbioru danych i powtórzyć analizę regresji zmiennej $x_2$ względem $x_4, x_8$ i $x_{10}$. Porównać wartości współczynników determinacji i współczynników przy odpowiednich zmiennych objaśniających obliczone teraz i w punkcie c. Czy usunięcie zmiennej $x_7$ ze zbioru zmiennych $\{x_4, x_7, x_8, x_{10}\}$ wydaje się uzasadnione?

**f)** Przyjmuje się z reguły, że wartość współczynnika podbicia wariancji pewnej zmiennej większa od 10 wskazuje na silną zależność tej zmiennej od pozostałych zmiennych objaśniających. W analizie dokonanej w punkcie c współczynnik $VIF_4$ dla zmiennej wagi ciała $x_4$ jest równy 10,358. Zamiast zmiennej $x_7$ ze zbioru zmiennych objaśniających $\{x_4, x_7, x_8, x_{10}\}$ usunąć zmienną wagi ciała $x_4$. Jaki jest rezultat testu hipotezy $H_0$: $\beta_7 = 0$

przeciwko $H_1$: $\beta_7 \neq 0$, gdzie $\beta_7$ jest współczynnikiem równania regresji odpowiadającym zmiennej $x_7$? Porównać wynik z wynikiem punktu d oraz e i zinterpretować go.

**4.9.** Rozpatrzyć model regresji liniowej zmiennej $x_2$ w zbiorze *bodyfat* względem zmiennych $x_7$ i $x_8$. Sporządzić wykresy regresji częściowej dla obu zmiennych w tym modelu. Która zmienna wydaje się mieć większy wpływ na zmienną $x_2$ po usunięciu wpływu drugiej zmiennej? Czy jest to zgodne z wynikami testu $t$ o istotności poszczególnych zmiennych?

**4.10.** Dokonać doboru zmiennych metodą selekcji krokowej dla zbioru *bodyfat*.

**4.11.** Dane przedstawione poniżej są wynikami egzaminu końcowego (zmienna Final) oraz dwóch testów międzysemestralnych (Test1 i Test2) ze wstępu do rachunku prawdopodobieństwa dla 33 studentów pewnej wyższej szkoły inżynierskiej.

| Test1 | Test2 | Final | Test1 | Test2 | Final | Test1 | Test2 | Final |
|---|---|---|---|---|---|---|---|---|
| 19 | 29 | 25 | 5 | 24 | 18 | 14,5 | 31 | 29 |
| 12 | 27 | 17,5 | 16 | 30 | 33 | 20 | 37 | 38 |
| 16 | 35 | 32,5 | 14 | 31 | 22,5 | 20 | 31,5 | 40 |
| 12,5 | 37 | 21 | 20 | 33 | 29,5 | 18 | 32,5 | 18 |
| 14,5 | 25,5 | 18,5 | 18,5 | 34,5 | 32,5 | 10 | 25 | 21 |
| 16,5 | 35 | 40,5 | 15,5 | 28 | 14 | 17,5 | 27,5 | 14,5 |
| 12 | 25 | 20,5 | 9 | 31,5 | 10,5 | 15 | 32,5 | 27,5 |
| 6 | 22 | 25,5 | 13 | 30,5 | 25,5 | 15,5 | 34 | 26,5 |
| 13,5 | 34 | 25,5 | 13,5 | 23,5 | 23,5 | 13 | 30 | 15 |
| 16,5 | 30,5 | 34,5 | 14 | 12 | 11 | 12,5 | 28,5 | 12 |
| 14,5 | 35,5 | 19,5 | 19,5 | 28,5 | 22 | 17 | 41,5 | 32,5 |

**a)** Dokonać analizy regresji wyniku egzaminu końcowego względem wyników obu testów i zinterpretować wyniki. Jaki procent zmienności jest tłumaczony przez model? Przeprowadzić test $t$ istotności indywidualnych zmiennych w modelu.

**b)** Określić wartości całkowitej sumy kwadratów, regresyjnej sumy kwadratów i sumy kwadratów błędów w tym modelu oraz związane z nimi liczby stopni swobody. Na podstawie tych wartości obliczyć z definicji wartość statystyki $F$ dla testowania hipotezy $H_0$: $\beta_1 = \beta_2 = 0$ przeciwko hipotezie $H_1$ mówiącej, że któryś z tych współczynników jest różny od zera. Przetestować tę hipotezę. Czy wynik testu $F$ jest zrozumiały w świetle wyników testów $t$ dla indywidualnych zmiennych w modelu?

**c)** Wyróżnić obserwacje spełniające kryterium potencjalnych obserwacji wpływowych (4.38) na podstawie wyrazów diagonalnych macierzy $\mathbf{H}$. Zidentyfikować je na wykresie rozproszenia rezyduów względem wartości przewidywanych. Sprawdzić, czy odpowiadają im duże wartości odległości Cooke'a.

ROZDZIAŁ 5

# Analiza wariancji

## 5.1. Wprowadzenie

Analiza regresji, jak wiemy, ma na celu ustalenie przybliżonej zależności funkcyjnej wiążącej wartości zmiennej zależnej, czyli objaśnianej, z wartościami zmiennych niezależnych, czyli zmiennych objaśniających. Łatwo sobie wyobrazić nieomal dowolną liczbę przykładów z dziedziny nauk technicznych, przyrodniczych, ekonomicznych, psychologicznych i innych, w których chodzi o ustalenie takiej właśnie zależności. Łatwo jednak przedstawić również takie sytuacje, w których wartości zmiennej objaśnianej możemy poznać dla niewielkiej tylko liczby różnych wartości jednej lub wielu zmiennych objaśniających. Wyobraźmy sobie, że przedmiotem naszego zainteresowania jest poprawa praktycznie dowolnego procesu technologicznego realizowanego na skalę przemysłową, którego nie można zbadać w warunkach laboratoryjnych i o którym wiedzę czerpiemy, zmieniając nastawy odpowiednich urządzeń w trakcie produkcji. Nietrudno zrozumieć, że w takiej sytuacji badania będziemy zmuszeni ograniczyć do jak najmniejszej liczby eksperymentów. Poprawę procesu technologicznego poprzedza się wówczas bardzo dokładną analizą jego dotychczasowego przebiegu, umożliwiającą wybranie możliwie niewielu zmiennych objaśniających, z których każda może w naszych doświadczeniach przyjąć tylko kilka (najlepiej dwie) wartości. Na przykład z analizy pewnego procesu produkcji wykładziny z tworzywa poliuretanowego, pokrywającej tablice rozdzielcze samochodów, wynikło, że poprawę jakości wykładziny można być może uzyskać po odpowiedniej zmianie temperatury izocyjanianu, ciśnienia wtłaczania izocyjanianu do formy i ewentualnie średnicy otworu, którym izocyjanian jest wtłaczany. W takiej sytuacji nie ma sensu badanie dokładnego kształtu funkcji wiążącej jakość wykładziny ze zmiennymi objaśniającymi. Wymagałoby to zbyt wielu eksperymentów, co byłoby zbyt czasochłonne, sparaliżowałoby produkcję i w rezultacie kosztowałoby zbyt wiele. Technolodzy mogą natomiast ustalić po dwie możliwe wartości każdej z trzech zmiennych i zlecić sprawdzenie dla

jakiego zestawu wartości temperatury, ciśnienia oraz średnicy można oczekiwać najwyższej jakości wykładziny (w tak zaplanowanym doświadczeniu mamy tylko 8 możliwych zestawów wartości zmiennych objaśniających).

W niniejszym rozdziale zmienną objaśnianą będziemy nazywać **zmienną odpowiedzi**, natomiast zmienne objaśniające **czynnikami** (z tak rozumianym pojęciem czynnika spotkaliśmy się już w podrozdz. 2.5). Możliwe wartości czynnika będziemy nazywać jego **poziomami**.

Pomijając to, że czynniki mogą występować na niewielu poziomach, często dokładna postać zależności między zmienną odpowiedzi a czynnikami nas po prostu nie interesuje. Wystarczy nam ogólna odpowiedź na pytanie, czy zmiana poziomu danego czynnika ma wpływ na średnią wartość zmiennej odpowiedzi. Tak jest np. wtedy, gdy pytamy czy poziom zanieczyszczenia gleby daną substancją ma wpływ na średnią wagę suchego ziarna pszenicy.

Nic zatem dziwnego, że obok analizy regresji istnieje metoda szczególnie dobrze dostosowana do badania problemów podobnych do opisanych. Metoda ta, jak zobaczymy w p. 5.2.2 blisko zresztą związana z analizą regresji, jest znana jako analiza wariancji (w p. 5.2.1 dowiemy się skąd wzięła się jej nieco zaskakująca nazwa). Jest to metoda tym popularniejsza i tym ważniejsza, że doskonale nadaje się do stosowania w przypadku występowania czynników, które nie są zmiennymi ilościowymi, lecz jakościowymi.

Czynniki jakościowe rzeczywiście łatwo można spotkać w praktyce. Stale np. są badane różne rodzaje margaryny na zawartość tłuszczów nasyconych. W badaniach takich margaryna odgrywa rolę czynnika, jej zaś różne rodzaje odpowiadają różnym poziomom tego czynnika. Zmienną odpowiedzi jest zawartość tłuszczów nasyconych. Naturalne jest pytanie czy średnia wartość zawartości tłuszczów jest taka sama dla każdego poziomu czynnika, czyli każdego rodzaju margaryny, czy też któraś z margaryn ma średnio mniej lub więcej tłuszczów nasyconych (generalnie, ze względu na konkurencję na rynku, oczekuje się, że zawartość tłuszczów jest taka sama dla różnych rodzajów margaryny). Z innymi przykładami czynników jakościowych spotkaliśmy się już w przykł. 2.29 i 2.30.

Dokładne sformułowanie zadania analizy wariancji przedstawimy w kolejnych podrozdziałach tego rozdziału, gdzie oddzielnie rozważymy analizę **jednoczynnikową**, gdy zmienna odpowiedzi może zależeć od tylko jednego czynnika oraz analizę **dwuczynnikową**, gdy zmienna odpowiedzi może zależeć od dwóch czynników. Obydwa wymienione problemy są jakościowo różne, w drugim bowiem przypadku poza niezależnymi od siebie wpływami każdego z czynników oddzielnie, może występować ich **interakcja**, czyli łączne oddziaływanie na zmienną odpowiedzi. Znając rozwiązanie zagadnie-

nia analizy dwuczynnikowej łatwo już sobie uzmysłowić, jak powinno wyglądać jego uogólnienie na problem z wieloma czynnikami, którego omówienie pomijamy.

## 5.2. Analiza jednoczynnikowa

### 5.2.1. Test $F$ analizy wariancji

Analiza wariancji ma szczególnie długą tradycję jej stosowania w naukach rolniczych. Na przykład może nas interesować czy zastosowanie ustalonego nawozu w czterech różnych ilościach ma wpływ na średnią wielkość zbioru pszenicy ozimej danego gatunku. Odpowiedź na takie pytanie wymaga przeprowadzenia eksperymentu na odpowiednio wielu poletkach, o których założymy, że nie różnią się niczym poza użytą ilością nawozu. W takim przypadku zmienną odpowiedzi jest wielkość zbioru, czynnikiem zaś ilość nawozu. Jest to oczywiście czynnik typu ilościowego, występujący na czterech poziomach.

Jeżeli interesuje nas wytrzymałość pewnego stopu na rozerwanie i dysponujemy próbkami trzech różnych odmian tego stopu, to mamy do czynienia z czynnikiem występującym na trzech poziomach, przy czym tym razem czynnik ma charakter jakościowy. Rzeczywiście, dany poziom czynnika to po prostu dana odmiana stopu. Wytrzymałość stopu jest zmienną odpowiedzi.

Ogólnie rzecz ujmując załóżmy, że jest dany jeden czynnik, który może mieć wpływ na interesującą nas zmienną odpowiedzi (niekiedy taki czynnik będziemy oznaczać dużą literą A). Załóżmy dalej, że czynnik (ilościowy lub jakościowy) może występować na $k$ różnych poziomach. Zadanie **jednoczynnikowej analizy wariancji**, zwanej też **analizą jednokierunkową**, polega na teście hipotezy o równości wartości średnich zmiennej odpowiedzi dla wszystkich $k$ poziomów czynnika,

$$H_0\colon\ \mu_1 = \mu_2 = \ldots = \mu_k, \tag{5.1}$$

gdzie $\mu_i$, $i = 1, 2, \ldots, k$, oznacza średnią wartość zmiennej odpowiedzi dla $i$-tego poziomu czynnika, przy alternatywie

$$H_1\colon \text{ przynajmniej dwie spośród średnich } \mu_1, \mu_2, \ldots, \mu_k \text{ są różne.} \tag{5.2}$$

Równoważne sformułowanie hipotezy alternatywnej brzmi: nie wszystkie średnie $\mu_1, \mu_2, \ldots, \mu_k$ są równe. W przypadku odrzucenia hipotezy zerowej naturalne jest zbadanie, jak średnie $\mu_1, \mu_2, \ldots, \mu_k$ różnią się między sobą. Do tego problemu wrócimy w następnym punkcie tego podrozdziału. Przedtem

należy zająć się zagadnieniem weryfikacji (czyli przetestownia prawdziwości) hipotezy $H_0$.

Wracając do przykładu rolniczego z początku tego punktu łatwo zaproponować naturalne podejście do weryfikacji hipotezy $H_0$. Po pierwsze musimy dysponować pewną populacją poletek do obsiania pszenicą. Jak pamiętamy z podrozdz. 2.5, poletka takie nazywamy **jednostkami eksperymentalnymi**. Każdej jednostce przypisujemy następnie poziom czynnika, jaki będzie zastosowany do tej jednostki. W naszym przykładzie sprowadza się to do przypisania każdemu poletku ilości nawozu, która będzie na nim wysypana. W ten sposób otrzymujemy $n_1$ jednostek poddanych działaniu czynnika na poziomie pierwszym, $n_2$ jednostek poddanych działaniu czynnika na poziomie drugim i ostatecznie $n_k$ jednostek poddanych działaniu czynnika na poziomie $k$-tym. Łatwo się domyślić, że życzymy sobie, by działaniu czynnika na każdym poziomie była poddana więcej niż jedna jednostka eksperymentalna. Rzecz w tym, iż umożliwi to w rozsądny i naturalny sposób oszacowanie średniej wartości zmiennej odpowiedzi dla każdego poziomu czynnika. Takim naturalnym estymatorem średniej wartości zmiennej odpowiedzi jest oczywiście średnia arytmetyczna, czyli średnia w próbie $n_i$ wartości zmiennej odpowiedzi, otrzymanych w wyniku zastosowania czynnika na $i$-tym poziomie (w naszym przykładzie wartościami zmiennej odpowiedzi są wielkości plonów zebranych z poletek). Jak doskonale wiemy, im większa jest liczność $n_i$, tym mniejsza jest wariancja średniej, a więc tym mniejsza zmienność naszego estymatora. Ostatnim i już nieoczywistym krokiem jest opracowanie procedury testowej opartej na średnich próbkowych uzyskanych dla każdego poziomu czynnika.

Zanim przejdziemy do omówienia samej procedury testowej, zatrzymajmy się jeszcze przez chwilę przy czynności przypisywania jednostek eksperymentalnych poziomom. Czynność tę nazywamy **planem doświadczenia**. Dla uproszczenia opisu, w całym rozdziale będziemy rozważać tylko plany **zrównoważone**, czyli takie, w których $n_1 = n_2 = \ldots = n_k = n$. Zakładać też będziemy, że $n > 1$ i będziemy mówić (por. podrozdz. 2.5), że mamy do czynienia z **replikacjami** doświadczenia, ponieważ każde doświadczenie – czyli zastosowanie czynnika na danym poziomie – jest powtarzane, a zatem „replikowane".

Istotne znaczenie przypisuje się (zrównoważonemu) planowi **całkowicie zrandomizowanemu**, w którym jednostki eksperymentalne przypisuje się poziomom czynnika w sposób losowy. Jeśli dysponujemy całą populacją jednostek eksperymentalnych, to losowo wybieramy po jednej jednostce dla każdego poziomu czynnika i procedurę tę powtarzamy (replikujemy) jeszcze $n-1$ razy, losując spośród nie wylosowanych dotąd jednostek.

Randomizacja planu, jak na to zwróciliśmy uwagę w podrozdz. 2.5, ma na

celu możliwie skuteczne przeciwdziałanie wpływowi na wynik procedury testowej czynników, które nie są kontrolowane, a które mogą oddziaływać na prowadzony eksperyment. W naszym przykładzie czynnikiem kontrolowanym jest oczywiście ilość użytego na poletku nawozu. Poza tym zakłada się, że poletka się między sobą nie różnią, czyli że w eksperymencie nie występują żadne inne czynniki. Z taką idealną sytuacją mamy jednak do czynienia nadzwyczaj rzadko. W rzeczywistości, mimo jak największej staranności, z jaką wybiera się jednostki eksperymentalne, na wynik doświadczenia najczęściej wpływają i inne czynniki, kórych możemy nie znać, które zwykle uznajemy za losowe i których na pewno nie kontrolujemy, czyli nie mamy na nie wpływu. Wilgotność i skład fizykochemiczny gleby na różnych poletkach są z pewnością nieidentyczne, chociaż w sformułowaniu zadania od tego faktu abstrahujemy. Decyzja o nieuwzględnieniu w analizie innych czynników niż wybrany i kontrolowany przez eksperymentatora sugeruje, że poletka o nieco różnych wilgotnościach i składach gleby powinny mieć taką samą szansę znalezienia się w każdej z $k$ grup poletek. I stąd właśnie wynika randomizacja planu. W przeciwnym przypadku mielibyśmy większą szansę popełnienia błędu „systematycznego": np. zastosowania nawozu na poziomie pierwszym głównie na poletkach o większej od przeciętnej wilgotności oraz nawozu na poziomie $k$-tym głównie na poletkach o najniższej wilgotności. Odrzucenie wówczas hipotezy zerowej może być wyłącznie konsekwencją takiego błędu systematycznego, a nie wpływu ilości nawozu na wielkość plonu.

Zauważmy jeszcze, że potrzeba przypisania poziomów czynnika jednostkom eksperymentalnym nie zawsze jest tak oczywista, jak w naszym przykładzie rolniczym. Na początku tego punktu wspomnieliśmy także o przykładzie, w którym dysponujemy próbkami trzech różnych odmian stopu. Jeżeli owe trzy typy stopu pochodzą od trzech producentów, oferujących nam swój wyrób, etap wyboru jednostek eksperymentalnych powinien polegać na odpowiednim, losowym przypisaniu do konkretnych zadań zarówno aparatury pomiarowej, jak i osób wykonujących pomiary. Jeżeli to my w swoich laboratoriach pracujemy nad owym stopem, etap wyboru jednostek eksperymentalnych jest jeszcze bardziej skomplikowany. Proces wytworzenia próbki stopu jest realizowany w konkretnym laboratorium z użyciem konkretnej aparatury przez konkretnych ludzi. W takich sytuacjach randomizacja eksperymentu jest bardzo istotnym zagadnieniem, którego właściwego rozwiązania nie będziemy tu wprawdzie rozważać, ale które musi zapobiec powstaniu błędów systematycznych skutkiem np. niewłaściwego rozdziału zadań na różne aparaty użyte w procesie produkcji różnych odmian stopu.

Wróćmy do ogólnego problemu analizy wariancji. Ponieważ dla każdego poziomu kontrolowanego czynnika wykonujemy więcej niż jedno doświadczenie, możemy mówić o wariancji zmiennej odpowiedzi dla tego poziomu (wariancję dla $i$-tego poziomu będziemy oznaczać $\sigma_i^2$).

Wszystkie dalsze rozważania będziemy prowadzić przyjmując, że jest spełnione następujące podstawowe założenie analizy wariancji:

**Podstawowe założenie analizy wariancji:** *Dla każdego poziomu (kontrolowanego) czynnika rozkład zmiennej odpowiedzi jest normalny z taką samą wariancją* $\sigma^2$, $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2 = \sigma^2$.

Analizę zawsze rozpoczynamy od sprawdzenia czy jest spełnione podane założenie. Sposób postępowania podajemy w dalszym ciągu tego punktu w kontekście przykł. 5.1.

Formalnie obserwacje zmiennej odpowiedzi możemy zapisać następująco:

$$Y_{ij} = \mu_i + \varepsilon_{ij}, \tag{5.3}$$

gdzie $i = 1, 2, \ldots, k$, $j = 1, 2, \ldots, n$, $Y_{ij}$ oznacza $j$-tą obserwację na $i$-tym poziomie czynnika, $\mu_i$ oznacza jak poprzednio średnią wartość zmiennej odpowiedzi dla $i$-tego poziomu czynnika oraz $\varepsilon_{ij}$ są niezależnymi zmiennymi losowymi o rozkładzie normalnym $N(0, \sigma)$. Dla każdej ustalonej wartości wskaźnika $i$, a zatem dla każdego ustalonego poziomu czynnika mówimy, że obserwacje $Y_{ij}$, $j = 1, 2, \ldots, n$, tworzą **grupę**. Wartości średnie $\mu_i$ można zapisać następująco:

$$\mu_i = \mu + \beta_i, \tag{5.4}$$

gdzie $i = 1, 2, \ldots, k$, $\mu$ jest ogólną wartością średnią, natomiast $\beta_i$ jest efektem $i$-tego poziomu czynnika; zakładamy przy tym, że

$$\beta_1 + \beta_2 + \cdots + \beta_k = 0. \tag{5.5}$$

Bez takiego założenia wielkości $\mu$ i $\beta_i$ nie są określone jednoznacznie; zwiększenie $\mu$ o dowolną ustaloną stałą $a$ oraz zmniejszenie wszystkich efektów $\beta_i$, $i = 1, 2, \ldots, k$ o tę samą stałą $a$ nie zmienia wartości $\mu_i$.

Odnotujmy, że w świetle warunku (5.5) hipotezie zerowej możemy nadać postać

$$H_0\colon\ \beta_1 = \beta_2 = \cdots = \beta_k = 0.$$

Zauważmy też, że milcząco przyjmujemy, iż efekty $\beta_i$ są stałymi, a nie zmiennymi losowymi. Założenie o stałości wszystkich efektów czynnikowych będzie obowiązywać we wszystkich naszych rozważaniach dotyczących analizy wariancji. (Pomijane przez nas uogólnienie analizy wariancji na przypadek losowych efektów czynnikowych jest nazywane **analizą komponentów wariancyjnych**, wariancja zmiennej odpowiedzi jest bowiem wówczas pewną kombinacją wariancji składowych: w modelu (5.3)–(5.4) jest kombinacją wariancji zmiennej $\varepsilon_{ij}$ oraz wariancji efektu czynnikowego $\beta_i$).

Jeśli przyjmiemy założenie o równości wariancji zmiennej odpowiedzi dla wszystkich poziomów czynnika, to możemy oprzeć test hipotezy (5.1) przy

alternatywie (5.2) na porównaniu dwóch estymatorów tej samej wariancji $\sigma^2$. Stąd właśnie bierze się nazwa całej procedury – wprawdzie interesuje nas równość (lub nie) wartości średnich, ale testowanie sprowadza się do analizy wariancji. Dokonajmy najpierw rozkładu całkowitej zmienności zmiennej odpowiedzi, podobnego do rozkładu zmienności zmiennej objaśnianej w analizie regresji. Określmy w tym celu następujące wielkości: średnią wartości zmiennej odpowiedzi dla $i$-tego poziomu czynnika

$$\bar{y}_{i\cdot} = \frac{1}{n}\sum_{j=1}^{n} y_{ij};$$

średnią ogólną

$$\bar{y}_{\cdot\cdot} = \frac{1}{kn}\sum_{i=1}^{k}\sum_{j=1}^{n} y_{ij};$$

oraz **całkowitą sumę kwadratów**

$$SST = \sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{\cdot\cdot})^2,$$

która opisuje całkowitą zmienność zmiennej odpowiedzi i która – jak można udowodnić – jest sumą zmienności międzygrupowej $SSA$ i zmienności wewnątrzgrupowej $SSE$, gdzie

$$SSA = n\sum_{i=1}^{k}(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2$$

oraz

$$SSE = \sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i\cdot})^2.$$

Zmienność międzygrupowa opisuje zmienność średnich charakteryzujących grupy względem średniej ogólnej (przypomnijmy, że litera A w symbolu $SSA$ odnosi się do czynnika, który tak właśnie oznaczamy). Łatwo też zauważyć, że zmienność wewnątrzgrupowa $SSE$ charakteryzuje zmienność wewnątrz grup i że może jednocześnie być traktowana jako suma kwadratów reszt (czyli różnic między wartościami zaobserwowanymi i ich średnimi próbkowymi). Jak wspomnieliśmy, prawdziwa jest równość

$$SST = SSA + SSE.$$

Rozumowanie podobne do przeprowadzonego w rodziale 4 pozwala orzec, że liczba stopni swobody związanych z $SST$ wynosi $kn - 1$, liczba stopni

swobody związanych z $SSA$ jest równa $k-1$, zmienności $SSE$ zaś odpowiada liczba stopni swobody $kn-k = k(n-1)$. Bez trudu możemy także zauważyć, że błąd średniokwadratowy

$$s^2 = \frac{SSE}{k(n-1)} = \frac{s_1^2 + s_2^2 + \cdots + s_k^2}{k}, \tag{5.6}$$

gdzie $s_i^2$, $i = 1, 2, \ldots, k$ są wariancjami w próbach, które tu nazywamy grupami, jest nieobciążonym estymatorem wariancji $\sigma^2$. Jest to estymator tym lepszy, że oparty na uśrednieniu wszystkich estymatorów wariancji w $k$ grupach. Co więcej, ponieważ estymator ten jest oparty na zmienności w grupach, jest nieobciążony bez względu na to, czy hipoteza (5.1) jest prawdziwa. Nietrudno też zauważyć, że – podobnie jak regresyjna suma kwadratów, $SSR$, w rozdz. 4 – zmienność międzygrupowa $SSA$ może być podstawą konstrukcji innego estymatora wariancji $\sigma^2$, ale wtedy tylko, gdy jest prawdziwa hipoteza (5.1). Można udowodnić, że jeśli hipoteza (5.1) jest spełniona, to wielkość

$$\frac{SSA}{k-1}$$

jest nieobciążonym estymatorem wariancji $\sigma^2$. Jeżeli natomiast jest prawdziwa hipoteza (5.2), estymator $SSA/(k-1)$ ma tendencję do przyjmowania dużych wartości, czyli większych od $\sigma^2$. Jak z tego wynika, statystyka

$$F = \frac{SSA/(k-1)}{SSE/[k(n-1)]} \tag{5.7}$$

może być uznana za naturalną statystykę testową, służącą do testownia hipotezy (5.1) przy alternatywie (5.2). Okazuje się, że jeżeli jest spełniona hipoteza zerowa (5.1), to statystyka (5.7) ma rozkład $F$ Snedecora z $k-1$ oraz $k(n-1)$ stopniami swobody. Zgodnie z wcześniejszymi uwagami, zbiorem krytycznym testu hipotezy (5.1) na poziomie istotności $\alpha$ jest zbiór

$$\{F\colon\ F \geq f_{1-\alpha,k-1,k(n-1)}\},$$

gdzie $f_{1-\alpha,k-1,k(n-1)}$ jest kwantylem rzędu $1-\alpha$ rozkładu $F$ z $k-1$ i $k(n-1)$ stopniami swobody.

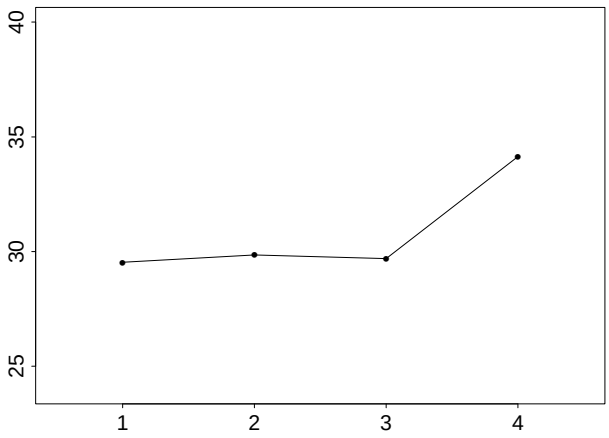
Przedstawiony test nazywamy **testem $F$ analizy wariancji**. Podkreślenia wymaga to, że jest to inny test niż test $F$ rozważony w rodz. 3. Tam porównywaliśmy wariancje dwóch niezależnych prób losowych, natomiast tym razem licznik i mianownik statystyki testowej są oparte na tych samych obserwacjach. Różnicę między obydwoma testami szczególnie dobrze widać w przypadku analizy jednoczynnikowej, gdy kontrolowany czynnik przyjmuje tylko dwie wartości oraz $n_1 = n_2 = n$. Nietrudno wówczas wykazać

(por. zad. 5.1), że – po oczywistej zmianie we wzorze (3.60) wielkości $\bar{X}_1$ i $\bar{X}_2$ na $\bar{Y}_{1\cdot}$ i $\bar{Y}_{2\cdot}$ – kwadrat statystyki danej tym wzorem jest równy statystyce (5.7) z estymatorem wariancji $S_p$ danym wzorem (3.29). Innymi słowy test $F$ analizy wariancji jest wówczas równoważny testowi $t$ na równość wartości średnich dwóch prób i, tak samo jak ten ostatni, jest odporny na nieduże odchylenia od normalności rozkładu obserwacji oraz na istnienie małej różnicy między wariancjami obydwu grup. Okazuje się, że test $F$ analizy wariancji zachowuje wymienione własności odporności także wtedy, gdy $k$ jest większe od 2 oraz liczności w grupach nie różnią się zbytnio między sobą. Czytelnik na pewno pamięta, że test $F$ omówiony w rodz. 3 jest nieodporny na odstępstwa od normalności rozkładu obu prób.

**Przykład 5.1.** Pewna uczelnia zdecydowała się wprowadzić semestralne kursy wyrównawcze z analizy matematycznej i algebry dla studentów II semestru, którzy z trudem przebrnęli przez I semestr zajęć matematycznych. Opracowano cztery różne programy kursu. Czterdziestu czterech zagrożonych studentów skierowano w sposób losowy na kursy wyrównawcze, po 11 na każdy kurs. Równolegle studenci uczęszczali na zajęcia II semestru analizy matematycznej. Po upływie semestru postanowiono zbadać czy średnie wyniki egzaminu z analizy matematycznej II są takie same dla wszystkich czterech kursów. Oto wyniki końcowego egzaminu w czterech grupach (na egzaminie można było zdobyć najwyżej 50 punktów):

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Kurs nr 1 | 28,2 | 36,1 | 26,8 | 28,0 | 25,1 | 27,8 | 33,3 | 26,5 | 31,6 |
| | 32,4 | 29,0 | | | | | | | |
| Kurs nr 2 | 29,6 | 27,1 | 34,7 | 24,3 | 29,6 | 33,9 | 26,9 | 33,2 | 32,3 |
| | 24,9 | 31,8 | | | | | | | |
| Kurs nr 3 | 30,9 | 24,5 | 37,4 | 29,6 | 27,4 | 29,1 | 33,0 | 28,6 | 31,1 |
| | 27,4 | 27,7 | | | | | | | |
| Kurs nr 4 | 30,8 | 29,6 | 35,4 | 34,9 | 40,0 | 28,5 | 32,5 | 30,6 | 37,5 |
| | 35,8 | 39,9 | | | | | | | |

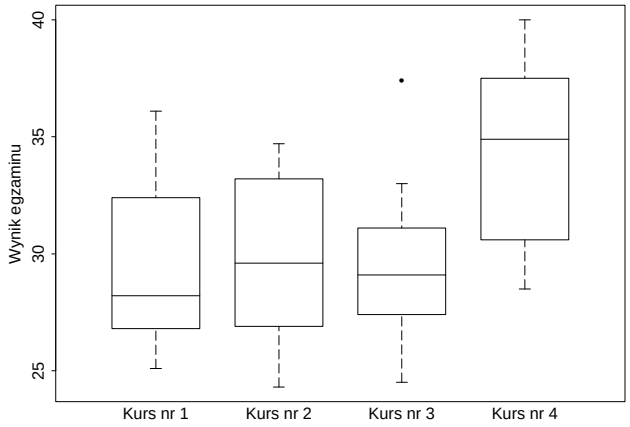
W przykładzie mamy zatem do czynienia z jednym czynnikiem jakościowym – kursem wyrównawczym – o czterech poziomach, czyli czterech różnych programach kursu. Na wykresie średnich w grupach (rys. 5.1) obserwujemy wyraźną różnicę między średnią wyników dla czwartego kursu (średnia równa 34,14) i pierwszych trzech kursów (odpowiednio 29,53; 29,85; 29,70). Oczywiście to czy dostrzeżona różnica jest statystycznie istotna zależy od zmienności wewnątrz grup. Rysunek 5.2, na którym są pokazane wspólnie wykresy ramkowe dla wszystkich grup, sugeruje, że jest wielce prawdopodobne odrzucenie hipotezy (5.1) na rzecz przyjęcia hipotezy alternatywnej (5.2).



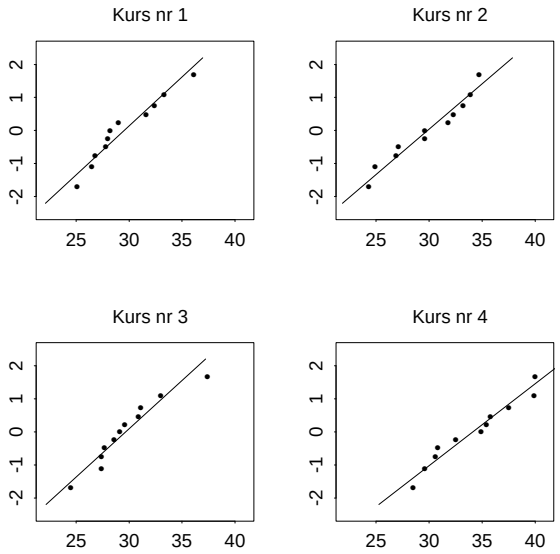

Rys. 5.1. Wykres średnich w grupach dla danych z przykł. 5.1

Zastosowanie testu $F$ analizy wariancji wymaga co najmniej przybliżonego spełnienia podstawowego założenia tej analizy: wariancje w grupach powinny być przynajmniej w przybliżeniu równe oraz zmienne odpowiedzi odpowiadające danej grupie powinny mieć ten sam rozkład normalny lub bliski normalnemu. Ze względów, o których była już mowa, nie zalecamy oparcia sprawdzenia jednorodności (czyli równości) wariancji na teście $F$ z rozdz. 3. Można natomiast oprzeć się na teście Levene'a, o którym wspomnieliśmy w tamtym rozdziale i który w rzeczywistości został zbudowany właśnie jako test jednorodności wariancji dla problemu z dwiema lub więcej grupami. Testu Levene'a nie będziemy dokładniej opisywać i zadowolimy się stwierdzeniem, że każdy nowoczesny pakiet statystyczny umożliwia jego użycie. Innym bardzo popularnym testem jednorodności wariancji, którego omówienie pomijamy i który jest zawarty w pakietach statystycznych, jest test Bartletta. Pozostawiamy Czytelnikowi sprawdzenie, że żaden z tych dwóch testów nie pozwala odrzucić hipotezy o równości wariancji w czterech grupach z przykł. 5.1 (por. zad. 5.2).

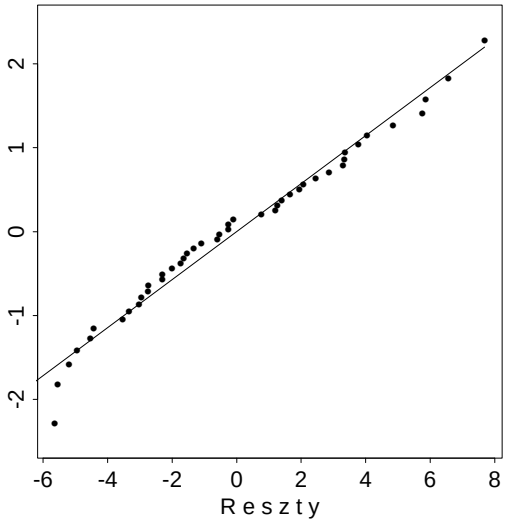
Normalność rozkładu zmiennej odpowiedzi sprawdzamy w sposób przybliżony, korzystając z wykresu kwantylowego. Pokazane na rys. 5.3 wykresy dla czterech grup z przykł. 5.1 świadczą o tym, że założenie normalności jest najprawdopodobniej przynajmniej w przybliżeniu spełnione. Ze względu na małą liczność grup trudno tu oczekiwać zupełnie jednoznacznych konkluzji. Ponieważ liczność $n$ nie jest „dramatycznie" mała, można też posłużyć się testem Shapiro–Wilka. Zważywszy jednak, że wariancje w grupach możemy uznać za równe, lepiej jest sporządzić wykres kwantylowy (oraz wykonać test Shapiro–Wilka) dla wszystkich reszt $Y_{ij} - \bar{Y}_{i\cdot}$, $i = 1, 2, \ldots, k$, $j = 1, 2, \ldots, n$.




Rys. 5.2. Wykresy ramkowe dla danych z przykł. 5.1




Rys. 5.3. Wykresy kwantylowe dla 4 grup z przykł. 5.1




Rys. 5.4. Wykres kwantylowy dla reszt $Y_{ij} - \bar{Y}_{i\cdot}$, $i = 1, 2, 3, 4$, $j = 1, 2, \ldots, 11$ dla danych z przykł. 5.1

Jak wynika z rys. 5.4, rozkład reszt można uznać za normalny. Z testu Shapiro–Wilka otrzymujemy $p$–wartość równą 0,37, co potwierdza trafność naszej konkluzji.

Założenia o normalności rozkładów oraz jednorodności wariancji mogą być nie spełnione skutkiem istnienia obserwacji odstających. Obserwacje takie możemy znaleźć, korzystając np. z wykresów ramkowych i należy je usunąć. Jeżeli założenie o normalności rozkładów, albo też założenie o jednorodności wariancji, musi być zdecydowanie odrzucone (analiza istnienia obserwacji odstających może nie dać pożądanych rezultatów), postuluje się znalezienie takiej transformacji danych, która umożliwi wymienione założenia uznać za spełnione w przypadku przekształconej zmiennej odpowiedzi. Postępowanie jest w takim przypadku podobne do postępowania krótko opisanego w rozdz. 4.

Z testu $F$ analizy wariancji otrzymujemy w przykł. 5.1 $p$–wartość mniejszą niż 0,012, co zgodnie z oczekiwaniem uzasadnia odrzucenie hipotezy zerowej (5.1) na korzyść hipotezy alternatywnej (5.2).

Na koniec tego punktu należy jeszcze wspomnieć o przynajmniej dwóch zagadnieniach. Po pierwsze, w ramach wstępnej analizy danych w grupach warto zbadać korelację między odchyleniem standardowym a średnią w grupach. Zdarza się, że mimo iż test jednorodności nie odrzuca hipotezy o wariancjach w grupach, to jednocześnie odnotowuje się wyraźną dodatnią korelację między odchyleniem standardowym a średnią – im większa średnia

w grupie tym też większe okazuje się być odchylenie standardowe. Przypadki takie wymagają ostrożności przy wyciąganiu wniosków. Jeżeli w grupie charakteryzującej się szczególnie dużą średnią i wariancją występują obserwacje odstające, to należy je usunąć. Jeżeli obserwacji takich nie ma i test $F$ analizy wariancji odrzucił hipotezę zerową o równości średnich, należy zdawać sobie sprawę, że mogło to być skutkiem istnienia tej szczególnie dużej średniej, która to szczególnie duża wartość mogła się z kolei zdarzyć skutkiem istnienia większej niż w innych grupach wariancji obserwacji. Innymi słowy, otrzymany wynik testu $F$ nie musi być w pełni wiarygodny i ostateczne wnioski należy – jak zawsze – wyciągnąć po wielostronnej i krytycznej analizie danego problemu. W następnym punkcie wspominamy o skorzystaniu z pewnych znanych nam już sposobów do wykrywania obserwacji odstających oraz prowadzenia analizy wariancji, gdy wariancje w grupach nie są równe.

Po drugie, chociaż test $F$ analizy wariancji jest odporny na umiarkowane odstępstwa od normalności rozkładów obserwacji, to jednak duże odstępstwa mogą prowadzić do błędnych wniosków. W przypadku zachodzenia takich dużych odstępstw zaleca się odwołanie do metod omówionych w podrozdz. 9.5.

### 5.2.2. Związki z analizą regresji

Pokrótce omówimy teraz związek jednoczynnikowej analizy wariancji z analizą regresji. Zauważyliśmy już analogię łączącą rozkład całkowitej zmienności zmiennej odpowiedzi w analizie wariancji z rozkładem zmienności zmiennej objaśnianej w analizie regresji. W rzeczywistości model analizy wariancji (5.3)–(5.5) można przedstawić jako szczególny przypadek modelu regresyjnego z nieznanymi parametrami $\mu$ oraz $\beta_1, \ldots, \beta_k$. W modelu regresji wystarczy przy tym uwzględnić tylko parametry $\mu$ oraz (na przykład) $\beta_1, \ldots, \beta_{k-1}$, ponieważ na mocy związku (5.5)

$$\beta_k = -\beta_1 - \beta_2 - \cdots - \beta_{k-1}.$$

Konstrukcję właściwego modelu regresji zilustrujemy na przykładzie problemu z czynnikiem występującym na czterech poziomach ($k = 4$), z trzema replikacjami doświadczenia dla każdego poziomu czynnika ($n = 3$). Łatwo sprawdzić, że w tym przypadku równania (5.3) możemy zapisać łącznie jako równanie macierzowe postaci

$$\begin{bmatrix} Y_{11} \\ Y_{12} \\ Y_{13} \\ Y_{21} \\ Y_{22} \\ Y_{23} \\ Y_{31} \\ Y_{32} \\ Y_{33} \\ Y_{41} \\ Y_{42} \\ Y_{43} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & -1 & -1 & -1 \\ 1 & -1 & -1 & -1 \\ 1 & -1 & -1 & -1 \end{bmatrix} \begin{bmatrix} \mu \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{13} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{23} \\ \varepsilon_{31} \\ \varepsilon_{32} \\ \varepsilon_{33} \\ \varepsilon_{41} \\ \varepsilon_{42} \\ \varepsilon_{43} \end{bmatrix}.$$

Po oznaczeniu

$$\mathbf{Y} = \begin{bmatrix} Y_{11} \\ Y_{12} \\ Y_{13} \\ Y_{21} \\ Y_{22} \\ Y_{23} \\ Y_{31} \\ Y_{32} \\ Y_{33} \\ Y_{41} \\ Y_{42} \\ Y_{43} \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & -1 & -1 & -1 \\ 1 & -1 & -1 & -1 \\ 1 & -1 & -1 & -1 \end{bmatrix}, \quad \beta = \begin{bmatrix} \mu \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}, \quad \varepsilon = \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{13} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{23} \\ \varepsilon_{31} \\ \varepsilon_{32} \\ \varepsilon_{33} \\ \varepsilon_{41} \\ \varepsilon_{42} \\ \varepsilon_{43} \end{bmatrix},$$

otrzymujemy liniowy model regresji wielokrotnej (4.28),

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon,$$

z bardzo prostą macierzą $\mathbf{X}$. Mając problem jednoczynnikowej analizy wariancji z dowolną zadaną liczbą $k$ poziomów czynnika oraz dowolną zadaną liczbą $n$ replikacji, Czytelnik nie będzie miał kłopotów z napisaniem odpowiadającego mu modelu regresji wielokrotnej.

W języku modelu regresji test hipotezy (5.1) przy hipotezie alternatywnej (5.2) możemy zapisać w postaci

$$H_0\colon\ \beta_1 = \beta_2 = \cdots = \beta_{k-1} = 0$$

przy alternatywie

$$H_1\text{: } \text{któryś ze współczynników } \beta_1, \ldots, \beta_{k-1} \text{ jest różny od 0.}$$

Jest to ten sam problem testowania, z którym zetknęliśmy się w p. 4.3.2 omawiając testy dla wektora współczynników $\beta$ w liniowym modelu regresji. Także sposób przeprowadzenia powyższego testu przedstawiony w p. 4.3.2 jest równoważny testowi $F$ analizy wariancji z poprzedniego punktu. Musimy wszakże pamiętać, że wymieniona równoważność testów $F$ z p. 4.3.2 i 5.2.1 dotyczy tylko przypadków ze szczególną postacią macierzy $\mathbf{X}$, odpowiadającą oczywiście problemowi analizy wariancji.

Fakt, że jednoczynnikowa analiza wariancji może być traktowana jako szczególny przypadek wielokrotnej analizy regresji ma oczywistą i ważną konsekwencję – wszędzie gdzie może to przynieść korzyść warto przy rozwiązywaniu problemu analizy wariancji sięgnąć do odpowiednich metod opracowanych w ramach analizy regresji. Takie odwołanie się do metod tej ostatniej ma na przykład sens, gdy pytamy o obserwacje odstające. Odwołanie się do odpowiedniego fragmentu p. 4.3.3 każe spytać o wariancję rezyduów $e_{ij}$, czyli rezyduów odpowiadających zaobserwowanym wartościom zmiennej odpowiedzi $Y_{ij}$, $i = 1, \ldots, k$, $j = 1, \ldots, n$. Na mocy własności (4.33) musimy zatem odpowiedzieć na pytanie o elementy przekątnej głównej macierzy $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. Można wykazać, że w przypadku planu zrównoważonego z $n$ replikacjami każdego doświadczenia, elementy przekątnej głównej macierzy $\mathbf{H}$ są wszystkie równe $1/n$. Mamy zatem

$$\sigma^2_{e_{ij}} = \sigma^2(1 - \frac{1}{n}) = \frac{\sigma^2(n-1)}{n}.$$

Dalej poszukiwanie obserwacji odstających możemy już prowadzić tak, jak to opisano w p. 4.3.3 (należy jedynie pamiętać, że parametrowi $p$ z rozdz. 4 odpowiada liczba $k$ poziomów czynnika w analizie wariancji oraz parametrowi $n$ z rozdz. 4 odpowiada iloczyn $kn$). Zauważmy jeszcze w tym miejscu, że równość wariancji wszystkich rezyduów jest konsekwencją zrównoważenia planu – gdyby liczba replikacji doświadczenia z $i$-tym poziomem czynnika była równa $n_i$, $i = 1, \ldots, k$, i nie wszystkie liczby $n_i$ byłyby sobie równe, wariancja rezyduum $e_{ij}$ wynosiłaby

$$\sigma^2_{e_{ij}} = \sigma^2(1 - \frac{1}{n_i}) = \frac{\sigma^2(n_i - 1)}{n_i}.$$

Jak wynika z dwóch ostatnich wzorów, posługiwanie się rezyduami studentyzowanymi (zamiast zwykłymi) nie jest potrzebne, gdy plan jest zrównoważony. Jest natomiast koniecznością, gdy liczby replikacji doświadczeń $n_i$, $i = 1, \ldots, k$, istotnie się różnią między sobą.

Omawiając problem regresji linowiej zwróciliśmy w p. 4.2.5 uwagę na możliwość skorzystania z metody najmniejszych ważonych kwadratów (MNWK), gdy wariancje błędów nie są równe. Jeżeli w zadaniu analizy wariancji mamy do czynienia z nierównością wariancji w grupach, możemy odwołać się do modelu regresyjnego i tak jak tam oprzeć się w naszym zadaniu na MNWK.

W podrozdziale 5.3 nie będziemy już wracać do zagadnienia związku między analizą wariancji a analizą regresji. Stwierdźmy tylko, że przedstawienie modeli rozważanych w podrozdz. 5.3 jako szczególnych przypadków modelu regresyjnego jest tak samo łatwe jak w przypadku analizy jednoczynnikowej. Dodajmy też, że dołączeniem do modeli analizy wariancji ilościowych zmiennych objaśniających zajmuje się **analiza kowariancji**.

### 5.2.3. Porównania wielokrotne

Odrzucenie hipotezy (5.1) rodzi oczywiste pytanie o różnice między wartościami średnimi. Średnie mogą się różnić między sobą na wiele sposobów. Nawiązując do użytego wcześniej sformułowania hipotezy alternatywnej, orzekającego, że *co najmniej dwie średnie różnią się między sobą*, poszukiwanie tych różnic oprzemy na porównaniach wartości średnich parami. Jeżeli, tak jak w przykł. 5.1 interesują nas 4 wartości średnie, porównaniem obejmujemy pary $\bar{y}_1$ i $\bar{y}_2$, $\bar{y}_1$ i $\bar{y}_3$, $\bar{y}_1$ i $\bar{y}_4$, $\bar{y}_2$ i $\bar{y}_3$, $\bar{y}_2$ i $\bar{y}_4$, oraz $\bar{y}_3$ i $\bar{y}_4$. Odpowiednie procedury prowadzenia takich porównań nazywamy **porównaniami wielokrotnymi**.

Jest jasne, że poczynione założenia o problemie analizy wariancji umożliwiają porównania wielokrotne oprzeć na statystyce typu (3.60) oraz na wielokrotnym zastosowaniu testów równości dwóch średnich.

Równość dwóch średnich jest w oczywisty sposób równoważna zerowaniu się ich różnicy. Z kolei różnica dwóch średnich jest szczególnym przypadkiem tzw. kontrastu między średnimi, czyli takiej kombinacji liniowej średnich $\sum_{i=1}^{k} c_i\mu_i$, w której ustalone stałe $c_i$ spełniają warunek $\sum_{i=1}^{k} c_i = 0$ (niektóre z tych stałych mogą być równe zeru). Niekiedy badacza rzeczywiście interesują kontrasty o więcej niż dwóch niezerowych stałych $c_i$. Tak jest np. wtedy, gdy badamy dwie możliwe terapie i interesuje nas średnia z ich wyników w porównaniu ze średnią (nieleczonej) grupy kontrolnej. Naturalne jest wówczas zbadanie kontrastu przypisującego taki sam współczynnik 0,5 obu średnim odpowiadającym terapiom oraz współczynnik $-1$ średniej odpowiadającej grupie kontrolnej. Zerowanie się takiego kontrastu oznacza, że średnio biorąc proponowane terapie nie dają żadnej zmiany wyniku badania lekarskiego. Zagadnieniem badania kontrastów nie będziemy się dalej zajmować.

Obecna sytuacja różni się od rozważonego w p. 3.4.1 problemu porównania średnich dwóch niezależnych populacji normalnych tym, że tym razem mamy

w ogólności do czynienia z więcej niż dwiema grupami danych. Po pierwsze zatem, stosowany obecnie estymator wspólnej dla wszystkich grup wariancji ma postać (5.6) i, po drugie, porównaniem należy objąć więcej niż jedną parę średnich. Ten drugi aspekt wymaga szczególnej uwagi. Mamy mianowicie do czynienia z koniecznością jednoczesnego wykonania całej serii testów.

Przyjrzyjmy się temu problemowi dokładniej. Procedura porównań wielokrotnych najpierw wymaga obliczenia statystyk $t_{ij}$ na podstawie zaobserwowanych wartości średnich $\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_k$, gdzie

$$t_{ij} = \frac{\bar{y}_i - \bar{y}_j}{s\sqrt{2/n}}, \tag{5.8}$$

przy czym $i < j$, $i, j = 1, 2, \ldots, k$ oraz (patrz (5.6))

$$s^2 = \frac{SSE}{k(n-1)}.$$

Otrzymujemy w ten sposób $K = \frac{k(k-1)}{2}$ statystyk testowych, umożliwiających przeprowadzenie tyluż testów hipotezy zerowej

$$H_{0ij}\colon\ \mu_i = \mu_j$$

przy hipotezie alternatywnej

$$H_{1ij}\colon\ \mu_i \neq \mu_j.$$

Problemem jest właściwe wyznaczenie wspólnej dla wszystkich testów wartości krytycznej na zadanym poziomie istotności procedury. Zauważmy, że musimy tu wprowadzić rozróżnienie między poziomem istotności pojedynczego testu i poziomem istotności całej procedury, ta ostatnia bowiem opiera się na zastosowaniu serii testów. Poziom istotności pojedynczego testu mówi o prawdopodobieństwie błędnego odrzucenia hipotezy o równości tej konkretnej pary średnich, której test dotyczy. Poziom istotności procedury powinien podawać prawdopodobieństwo błędnego odrzucenia którejkolwiek z hipotez $H_{0ij}$, $i < j$, $i, j = 1, 2, \ldots, k$. Z problemem łącznego poziomu istotności serii testów zetknęliśmy się już w p. 4.3.3 przy okazji testowania hipotezy, że żadna obserwacja w zagadnieniu regresji liniowej nie jest obserwacją odstającą.

Rozumując tak jak w punkcie 4.3.3 raz jeszcze dochodzimy do procedury Bonferroniego: chcąc by poziom istotności procedury, czyli łączny poziom istotności serii testów wynosił co najwyżej $\alpha$, poziom istotności każdego indywidualnego testu powinien wynosić $\alpha/K$. Ostatecznie zatem procedura Bonferroniego ma następującą postać. Dla każdej z $K$ par średnich obliczamy wartość (5.8) statystyki $t$ z $kn - k = k(n-1)$ stopniami swobody

(taka liczba stopni swobody wynika z użycia estymatora (5.6) nieznanej wariancji $\sigma^2$). Korzystając z tablic lub z pakietu statystycznego znajdujemy kwantyl rzędu $1-\alpha/(2K)$ rozkładu $t$ Studenta z $k(n-1)$ stopniami swobody. Oznaczmy podany kwantyl symbolem $t^\star$. Jeżeli dla którejkolwiek pary średnich $\bar{y}_i, \bar{y}_j$ jest spełniony warunek

$$t^\star \leq |t_{ij}|, \tag{5.9}$$

gdzie $t_{ij}$ jest wartością statystyki testowej dla pary $\bar{y}_i, \bar{y}_j$, to odrzucamy hipotezę zerową o równości średnich $\bar{y}_i$ oraz $\bar{y}_j$. Stosując pojedyncze testy do wszystkich par średnich, dowiadujemy się na poziomie istotności równym co najwyżej $\alpha$ (w praktyce mniejszym od $\alpha$) czy, i jeśli tak, to które pary mają nierówne wartości średnie.

Dobrze znaną wadą procedury Bonferroniego jest konieczność stosowania indywidualnych testów na bardzo niskim poziomie istotności. Jeżeli np. jest 5 poziomów czynnika, $K = 10$ i chcąc otrzymać $\alpha = 0{,}05$, musimy dla pojedynczych testów przyjąć poziom istotności równy 0,005. W rezultacie odrzucenie hipotezy zerowej staje się niezwykle trudne.

Pamiętając o dualności problemów testowania hipotez i konstruowania przedziałów ufności, testową procedurę Bonferroniego możemy z łatwością zamienić na metodę budowy $K$ przedziałów ufności postaci

$$[\bar{y}_i - \bar{y}_j - t^\star s_p\sqrt{2/n}, \bar{y}_i - \bar{y}_j + t^\star s_p\sqrt{2/n}], \tag{5.10}$$

$i < j$, $i, j = 1, 2, \ldots, k$. Pojedynczy przedział ufności charakteryzuje się poziomem ufności $1-\alpha/K$. Łączny poziom ufności wszystkich przedziałów jest nie mniejszy (w praktyce większy) niż $1-\alpha$. Innymi słowy, mamy szansę nie mniejszą niż $1-\alpha$, że prawdziwe są nasze wnioski dotyczące równości (lub nie) średnich w badanych $K$ parach. Jeżeli dla danych średnich $\bar{y}_i$ oraz $\bar{y}_j$ przedział ufności zawiera 0, orzekamy równość średnich. Jeżeli przedział ufności dla średnich $\bar{y}_i$ oraz $\bar{y}_j$ nie zawiera zera, orzekamy, że średnie te są różne.

Wadą testowej procedury Bonferroniego jest to, że jej poziom istotności jest w rzeczywistości mniejszy od założonego. Podobnie, rzeczywisty poziom ufności łącznej analizy przedziałów ufności jest większy od założonego poziomu $1-\alpha$. Wad tych nie ma procedura zaproponowana przez Scheffégo. Nie ma ich także, ale tylko wtedy, gdy $n_1 = n_2 = \ldots = n_k = n$, procedura zaproponowana przez Tukeya. Obydwie procedury opierają się na statystyce (5.8), ale różnią się od procedury Bonferroniego – i między sobą – sposobem obliczania wartości krytycznej $t^\star$. Procedury testowe Tukeya i Scheffégo mają zatem nadal postać (5.8)–(5.9) i ich przedziały ufności mają postać (5.10), ale inne niż w przypadku procedur Bonferroniego są wartości krytyczne $t^\star$.

Nie będziemy tu wnikać w sposoby obliczania wartości $t^\star$ w przypadku procedur Tukeya i Scheffégo, zadowalając się stwierdzeniem, że znaleźć je można w tablicach oraz, że są zawarte w statystycznych pakietach komputerowych. W przypadku wielokrotnych porównań par wartości średnich w grupach o tej samej liczności szczególnie poleca się procedurę Tukeya. Procedura ta daje nie tylko dokładną wartość poziomu istotności lub ufności, ale otrzymywana wartość krytyczna $t^\star$ jest mniejsza niż uzyskiwana metodą Scheffégo. W rezultacie np. przedziały ufności otrzymane metodą Tukeya są węższe od odpowiadających im przedziałów ufności uzyskanych metodą Scheffégo (ta ostatnia procedura jest polecana zwłaszcza wtedy, gdy analizuje się nie tylko różnice średnich w parach, ale także tzw. kontrasty między więcej niż dwiema średnimi).

**Przykład 5.1 cd.** W tabelach 5.1 i 5.2 podano $p$–wartości dla testów porównań wielokrotnych otrzymane metodą Tukeya i Scheffégo. Na poziomie istotności 0,05 procedura Tukeya odrzuca hipotezę o równości średnich $\bar{y}_4$ i $\bar{y}_1$, $\bar{y}_4$ i $\bar{y}_2$ oraz $\bar{y}_4$ i $\bar{y}_3$. Procedura każe oczywiście przyjąć hipotezę o równości średnich w pozostałych parach. Wynik ten jest zgodny z wnioskami, jakie płyną z przyjrzenia się rys. 5.2. Procedura Scheffégo nie prowadzi do tak zdecydowanych konkluzji.

**Tabela 5.1 $p$–wartości dla testu porównań wielokrotnych otrzymane metodą Tukeya**

| | $\bar{y}_1$ | $\bar{y}_2$ | $\bar{y}_3$ | $\bar{y}_4$ |
|---|---|---|---|---|
| $\bar{y}_1$ | – | 0,996847 | 0,999571 | 0,024213 |
| $\bar{y}_2$ | 0,996847 | – | 0,999698 | 0,040251 |
| $\bar{y}_3$ | 0,999571 | 0,999698 | – | 0,031784 |
| $\bar{y}_4$ | 0,024213 | 0,040251 | 0,031784 | – |

**Tabela 5.2 $p$–wartości dla testu porównań wielokrotnych otrzymane metodą Scheffégo**

| | $\bar{y}_1$ | $\bar{y}_2$ | $\bar{y}_3$ | $\bar{y}_4$ |
|---|---|---|---|---|
| $\bar{y}_1$ | – | 0,997605 | 0,999641 | 0,043509 |
| $\bar{y}_2$ | 0,997605 | – | 0,999748 | 0,068104 |
| $\bar{y}_3$ | 0,999641 | 0,999748 | – | 0,055341 |
| $\bar{y}_4$ | 0,043509 | 0,068104 | 0,055341 | – |

### 5.2.4. Zrandomizowany plan blokowy

Potrzeba dokonania randomizacji planu eksperymentu, jak na to zwróciliśmy uwagę w p. 5.2.1, bierze się stąd, że w praktyce jednostki ekspery-

mentalne różnią się między sobą ze względu na takie czy inne efekty niekontrolowane. Niekiedy złym skutkom istnienia niekontrolowanej zmienności, zwanej **zmiennością uboczną**, można przeciwdziałać w inny, bardziej efektywny sposób niż stosując plan całkowicie zrandomizowany. Sposób ten, zasygnalizowany już w podrozdz. 2.5, omówimy teraz nieco szerzej.

Wyobraźmy sobie, że chcemy porównać skuteczność trzech leków przeciw tej samej chorobie. Czynnikiem kontrolowanym jest w tym przypadku lek, występujący na trzech poziomach, natomiast zmienną odpowiedzi jest jego odpowiednio mierzona skuteczność, na przykład czas do wyzdrowienia. Wiemy, że wiek i ogólny stan zdrowia pacjenta mają charakter zmiennych ukrytych (por. podrozdz. 2.5) i mogą mieć wpływ na interesującą nas zmienną odpowiedzi. Możemy zatem zdecydować, że wszystkie jednostki eksperymentalne będą np. osobami w wieku od 30 do 40 lat o dobrym ogólnym stanie zdrowia.

W ten sposób zredukujemy zmienność uboczną, ale ceną za to będzie ograniczenie zakresu ważności otrzymanych wniosków. Po dokonaniu analizy wariancji dowiemy się, czy leki są równie skuteczne, czy też któryś z nich (oraz który) jest lepszy w leczeniu generalnie zdrowych osób w wieku od 30 do 40 lat. Nic nie będziemy wiedzieć o skuteczności leków wśród młodszych i starszych, ogólnie zdrowych pacjentów oraz wśród osób w dowolnym wieku, ale o ogólnie nie najlepszym stanie zdrowia.

Jednakże zgodzenie się na to, by populacja jednostek eksperymentalnych zawierała osoby w różnym wieku i niekoniecznie ogólnie zdrowe, nie tylko wprowadza zmienność uboczną, ale – zwłaszcza przy pechowym podziale populacji jednostek na grupy – może doprowadzić do fałszywych wniosków. Po pierwsze zmienność uboczna może „przesłonić" faktyczną różnicę w skuteczności obydwu leków. Mianowicie, na skutek istnienia wspomnianej zmienności rozproszenie rozkładów zmiennej odpowiedzi w grupach może okazać się zbyt duże, by móc uznać za istotną otrzymaną różnicę między średnimi. Ponadto może się zdarzyć, że podział jednostek na grupy jest korzystny dla jednej i niekorzystny dla dwóch pozostałych grup. Na przykład wyleczenie osób z pierwszej grupy może być łatwiejsze, ponieważ zrządzeniem losu znalazły się w niej osoby młodsze i ogólnie zdrowsze. W rezultacie, mimo że wszystkie trzy leki są tak samo skuteczne, analiza wariancji może doprowadzić do fałszywego wniosku, iż skuteczniejszy jest lek zaaplikowany pierwszej grupie osób.

W opisanym problemie możliwym i najlepszym wyjściem jest dokonanie podziału pacjentów na grupy opierając się na **zrandomizowanym planie blokowym**. Z planem takim zetknęliśmy się już w podrozdz. 2.5. Przyjmijmy, że interesujący nas czynnik występuje na $k$ poziomach. Zasadą planu jest utworzenie najpierw $l$ **bloków**, czyli grup jednostek eksperymentalnych o liczności $k$ w taki sposób, by jednostki w bloku były tak jednorodne (po-

dobne do siebie), jak to tylko jest możliwe. Następnie jednostkom w bloku przypisuje się losowo poziomy czynnika, tak by każda jednostka w bloku była poddana działaniu czynnika na innym poziomie. Takie postępowanie umożliwia zredukowanie zmienności ubocznej w blokach.

W naszym przykładzie bloki powinny być trzyelementowe i w każdym powinniśmy umieścić pacjentów w możliwie podobnym wieku, mających możliwie podobne wyniki badań medycznych, najlepiej też tej samej płci. Warto tu odnotować, że przykład ten bardzo przypomina sytuację testowania hipotezy (3.61) – różnica polega na dysponowaniu nie parami, a trójkami obserwacji.

Nierzadko cały blok eksperymentów można przeprowadzić na jednej jednostce eksperymentalnej. W przypadku badania wytrzymałości różnych stopów w kilku laboratoriach naturalne może być przeprowadzenie w każdym laboratorium jednego bloku eksperymentów. Gdy badamy różne ulepszenia benzyny samochodowej, każdy z bloków eksperymentów można przeprowadzić, korzystając z jednego samochodu. Wiadomo, że na wielkich lotniskach pasażerskich praca kontrolerów lotów jest wyjątkowo stresująca. Dlatego co jakiś czas sprawdza się różne sposoby polepszenia warunków ich pracy. Zawsze trzeba jednak pamiętać, że reakcja kontrolera na daną zmianę warunków jego pracy jest sprawą indywidualną. Dlatego, jeśli np. zapada decyzja o porównaniu trzech nowych trybów pracy kontrolera, za bloki uznaje się losowo wybranych do eksperymentu kontrolerów i na każdym sprawdza działanie wszystkich trzech poziomów badanego czynnika. Czytelnik z łatwością może podać przykłady innych problemów, w których rozsądne jest zastosowanie planu blokowego.

Ponieważ zarysowany schemat postępowania jest inny niż opisany w p. 5.2.1 oraz wyraźnie uwzględnia się zjawisko zmienności ubocznej, niezbędne jest stosowne przeformułowanie zadania analizy wariancji. I tak podstawowe założenie analizy wariancji przyjmuje postać:

**Podstawowe założenie analizy wariancji:** *Dla każdego poziomu (kontrolowanego) czynnika w każdym bloku rozkład zmiennej odpowiedzi jest normalny z taką samą wariancją* $\sigma^2$, $\sigma_{ib}^2 = \sigma^2$, *gdzie* $\sigma_{ib}^2$ *oznacza wariancję zmiennej odpowiedzi dla* $i$*-tego poziomu czynnika w* $b$*-tym bloku,* $i = 1, 2, \ldots, k$, $b = 1, 2, \ldots, l$.

Obserwacje zmiennej odpowiedzi możemy zapisać następująco:

$$Y_{ib} = \mu + \beta_i + \rho_b + \varepsilon_{ib}, \tag{5.11}$$

gdzie $i = 1, 2, \ldots, k$, $b = 1, 2, \ldots, l$, $Y_{ib}$ oznacza obserwację na $i$-tym poziomie czynnika w $b$-tym bloku, $\mu$ jest ogólną wartością średnią, $\beta_i$ jest efektem $i$-tego poziomu czynnika, $\rho_b$ jest efektem $b$-tego bloku oraz $\varepsilon_{ib}$ są niezależnymi zmiennymi losowymi o rozkładzie normalnym $N(0, \sigma)$. Nadal przy

tym obowiązuje warunek (5.5) i z tych samych względów nakładamy także analogiczny warunek na efekty blokowe,

$$\rho_1 + \rho_2 + \cdots + \rho_l = 0. \tag{5.12}$$

Obserwacje zmiennej odpowiedzi na zadanym poziomie czynnika są indeksowane numerem bloku, ponieważ w każdym bloku mamy po jednej obserwacji na tym poziomie. Na każdym poziomie czynnika mamy $l$ obserwacji zmiennej odpowiedzi.

Testowana hipoteza zerowa przyjmuje postać

$$H_0\colon\ \beta_1 = \beta_2 = \cdots = \beta_k = 0 \tag{5.13}$$

przy alternatywie

$$H_1\colon \text{ nie wszystkie efekty } \beta_i \text{ są równe zeru.} \tag{5.14}$$

Zauważmy, że wartości średnie obserwacji zmiennej odpowiedzi na zadanym poziomie czynnika mogą zależeć od efektu blokowego i mogą przeto być różne w różnych blokach. Zauważmy też, że nie interesuje nas tu pytanie czy i ewentualnie jak różnią się wartości $\rho_b$, $b = 1, 2, \ldots, l$.

Stosowny test hipotezy (5.13) przy alternatywie (5.14) konstruuje się całkowicie analogicznie do testu analizy wariancji omówionego w p. 5.2.1. Różnica polega na konieczności uwzględnienia zmienności między blokami. Niech zatem, podobnie jak poprzednio,

$$\bar{y}_{i\cdot} = \frac{1}{l}\sum_{b=1}^{l} y_{ib}$$

będzie średnią wartością zmiennej odpowiedzi dla $i$-tego poziomu czynnika,

$$\bar{y}_{\cdot b} = \frac{1}{k}\sum_{i=1}^{k} y_{ib}$$

niech będzie średnią wartością zmiennej odpowiedzi w $b$-tym bloku oraz niech

$$\bar{y}_{\cdot\cdot} = \frac{1}{kl}\sum_{i=1}^{k}\sum_{b=1}^{l} y_{ib}$$

będzie średnią ogólną. Można wykazać, że całkowita suma kwadratów

$$SST = \sum_{i=1}^{k}\sum_{b=1}^{l}(y_{ib} - \bar{y}_{\cdot\cdot})^2,$$

jest sumą zmienności $SSA$ między poziomami czynnika, zmienności $SSBL$ między blokami oraz zmienności losowej $SSE$, czyli sumy kwadratów reszt,

$$SST = SSA + SSBL + SSE,$$

gdzie

$$SSA = l\sum_{i=1}^{k}(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2,$$

$$SSBL = k\sum_{b=1}^{l}(\bar{y}_{\cdot b} - \bar{y}_{\cdot\cdot})^2$$

oraz

$$SSE = \sum_{i=1}^{k}\sum_{b=1}^{l}(y_{ib} - \bar{y}_{i\cdot} - \bar{y}_{\cdot b} + \bar{y}_{\cdot\cdot})^2.$$

Wyjaśnienia wymaga postać sumy kwadratów reszt. Przyjęty model (5.11) wskazuje, że reszty mają postać

$$y_{ib} - \hat{\mu} - \hat{\beta}_i - \hat{\rho}_b, \tag{5.15}$$

gdzie $\hat{\mu}$, $\hat{\beta}_i$ i $\hat{\rho}_b$ są odpowiednimi estymatorami wielkości $\mu$, $\beta_i$ i $\rho_b$. W naszym przypadku

$$\hat{\mu} = \bar{y}_{\cdot\cdot},$$

$$\hat{\beta}_i = \bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot},$$

oraz

$$\hat{\rho}_b = \bar{y}_{\cdot b} - \bar{y}_{\cdot\cdot}.$$

Odejmowanie średniej ogólnej w dwóch ostatnich estymatorach wynika stąd, że wartość średnia zmiennej odpowiedzi dla $i$-tego czynnika jest równa sumie $\mu+\beta_i$, natomiast wartość średnia zmiennej odpowiedzi w $b$-tym bloku równa się sumie $\mu + \rho_b$. Po wstawieniu powyższych estymatorów do wzoru (5.15) otrzymuje się wcześniej podaną postać sumy kwadratów reszt, $SSE$.

Rozumując podobnie jak we wcześniejszych przypadkach otrzymuje się, że liczba stopni swobody związanych z $SST$ wynosi $kl - 1$, liczba stopni swobody związanych z $SSA$ jest równa $k - 1$, zmienności $SSBL$ odpowiada $l - 1$ stopni swobody i ostatecznie sumie kwadratów reszt $SSE$ odpowiada $(k - 1)(l - 1)$ stopni swobody. Statystyka testowa $F$ analizy wariancji, której możemy użyć do testownia hipotezy (5.13) przy alternatywie (5.14), ma postać

$$F = \frac{SSA/(k-1)}{SSE/[(k-1)(l-1)]}. \tag{5.16}$$

Jeżeli jest spełniona hipoteza zerowa, statystyka (5.16) ma rozkład $F$ Snedecora z $k-1$ oraz $(k-1)(l-1)$ stopniami swobody. Nietrudno zauważyć, że estymatory występujące w liczniku i mianowniku statystyki testowej (5.16) mają własności analogiczne do estymatorów tworzących statystykę (5.7) i w rezultacie zbiorem krytycznym testu hipotezy (5.16) na poziomie istotności $\alpha$ jest zbiór

$$\{F\colon\ F \geq f_{1-\alpha,k-1,(k-1)(l-1)}\},$$

gdzie $f_{1-\alpha,k-1,(k-1)(l-1)}$ jest kwantylem rzędu $1-\alpha$ rozkładu $F$ z $k-1$ i $(k-1)(l-1)$ stopniami swobody.

Jak zobaczymy, analiza jednoczynnikowa z planem blokowym jest bardzo podobna do szczególnego przypadku analizy dwuczynnikowej.

## 5.3. Analiza dwuczynnikowa

Oczywiste jest, że w wielu sytuacjach praktycznych możemy kontrolować poziom więcej niż jednego czynnika. Na początku tego rozdziału wspomnieliśmy o autentycznym przykładzie, w którym były kontrolowane trzy czynniki związane z izocyjanianem, głównym obok alkoholu wielowodorotlenowego składnikiem tworzywa pokrywającego tablicę rozdzielczą samochodu. W tym podrozdziale ograniczymy się do analizy dwuczynnikowej. Ustalenie przez technologów jednej tylko możliwej temperatury wtłaczanego do formy izocyjanianu uczyniłoby z podanego przykładu taki właśnie problem. Z kolei podrozdział 5.2 rozpoczęliśmy przykładem problemu jednoczynnikowego, w którym interesowało nas czy zastosowanie ustalonego nawozu w czterech różnych ilościach ma wpływ na wielkość zbioru pszenicy ozimej zadanego gatunku. W pewnych warunkach niezbędne jest uwzględnienie przynajmniej jednego jeszcze czynnika, a mianowicie metody nawadniania poletek, gdy np. mamy do dyspozycji dwie takie metody (por. zad. 5.3 i 5.7).

Innym przykładem może być zbadanie wpływu na wielkość plonu użytego rodzaju kukurydzy oraz zastosowanej gęstości zasiewu (mierzonej ilością ziarna wysianego na jednostce powierzchni). Mamy przy tym do dyspozycji ustaloną liczbę, powiedzmy $k$, rodzajów ziarna oraz możemy stosować $l$ gęstości zasiewu. W ogólności, niech nasze badanie dotyczy dwóch czynników, A i B (np. rodzaju kukurydzy i gęstości zasiewu). Problem **dwuczynnikowej** (inaczej **dwukierunkowej**) analizy wariancji polega na zbadaniu czy na wartość zmiennej odpowiedzi ma wpływ wybór poziomu czynnika A, wybór poziomu czynnika B oraz, jak zobaczymy w pewnym sensie w pierwszym rzędzie, czy istnieje interakcja między czynnikami A i B, czyli czy czynniki

współdziałają między sobą, wspólnie oddziałując na zmienną odpowiedzi (zagadnienie interakcji zostanie dokładnie wyjaśnione w dalszym ciągu podrozdziału).

Jak się łatwo zorientować, czynniki A i B mogą wystąpić w $kl$ różnych kombinacjach, które niekiedy nazywamy $kl$ sposobami oddziaływania na jednostki eksperymentalne. Aż do odwołania będziemy przyjmować, że oddziaływaniu każdej z $kl$ kombinacji została poddana więcej niż jedna jednostka eksperymentalna, czyli że mamy do czynienia z replikacjami każdego doświadczenia. Dla prostoty będziemy przyjmować, że plan (całkowicie zrandomizowanego) doświadczenia jest zrównoważony, czyli że liczba replikacji każdej z $kl$ kombinacji jest taka sama (równa $n$).

**Przykład 5.2.** Narodowa Fundacja Nauki w USA (w skrócie NSF) bada mediany zarobków rocznych brutto wśród naukowców i inżynierów ze stopniem naukowym doktora w różnych grupach zawodowych i różnych regionach kraju. W tabeli 5.3 są podane takie mediany (z roku 1995) dla trzech stanów regionu środkowo-atlantyckiego (śa) oraz dla trzech stanów regionu północno-wschodnio-środkowego (pwś), w podziale na następujące (szeroko rozumiane) grupy zawodowe: matematycy i informatycy (mat.-inf.); biolodzy i naukowcy pokrewnych specjalności (biol.); fizycy i naukowcy pokrewnych specjalności (fiz.); socjolodzy i naukowcy pokrewnych specjalności (soc.); inżynierowie (inż.). Mamy zatem do czynienia z dwoma czynnikami, regionem USA i zawodem, przy czym pierwszy czynnik występuje na $k = 2$ poziomach, a drugi na $l = 5$ poziomach. Dla każdej z 10 kombinacji mamy 3 replikacje doświadczenia (zarobki są podane w tysiącach dolarów).

**Tabela 5.3**

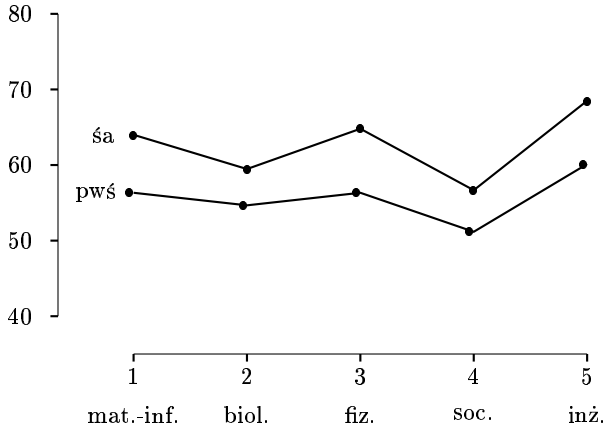
| | mat.-inf. | biol. | fiz. | soc. | inż. |
|---|---|---|---|---|---|
| śa | 74 63 55 | 67,4 58 53 | 71,4 63 60 | 60 55 55 | 72,5 68 65 |
| pwś | 56 57 56 | 55 60 49 | 55 60 54 | 48,5 55 50 | 56,4 68 56 |

Zanim dokładnie sformułujemy problem dwuczynnikowej analizy wariancji, przyjrzyjmy się wykresom średnich w grupach zawodów, obliczanych oddzielnie dla każdego z dwóch regionów (rys. 5.5). Średnie te są podane także w tab. 5.4.

**Tabela 5.4**

| | mat.-inf. | biol. | fiz. | soc. | inż. |
|---|---|---|---|---|---|
| śa | 64,00 | 59,47 | 64,80 | 56,67 | 68,50 |
| pwś | 56,33 | 54,67 | 56,33 | 51,17 | 60,13 |

Oczywiście w naszym przypadku mamy na rysunku dwie łamane – gdyby regionów było więcej, odpowiednio też więcej byłoby wykresów obrazujących średnie dla różnych zawodów w tych regionach. W przypadku każdej z grup zawodowych reprezentanci regionu środkowo-atlantyckiego zarabiają (w rozważanym przez nas sensie) lepiej niż ich koledzy i koleżanki z regionu północno-wschodnio-środkowego. Międzyregionalne różnice tych średnich są podobne w grupie matematyków i informatyków, fizyków oraz inżynierów, gdzie oscylują (w zaokrągleniu) między 7,7 a 8,5 tysięcy dolarów. Międzyregionalne różnice w grupie biologów i socjologów są równe, odpowiednio, 4,8 i 5,5 tysięcy dolarów.

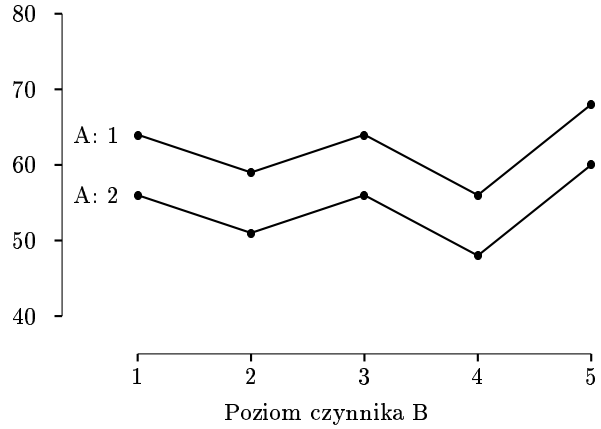



Rys. 5.5. Wykres średnich w grupach zawodów dla dwóch regionów (patrz tab. 5.4)

Wprawdzie obydwie łamane nie są do siebie równoległe, ale wyobraźmy sobie przez chwilę, że tak właśnie jest. Taka sytuacja jest w pewnym sensie prostsza i dlatego warto od niej zacząć interpretację wyników. Tabela 5.5. oraz rys. 5.6 opisują fikcyjny przykład, w którym czynnik A występuje na jednym z dwóch poziomów (1 lub 2), natomiast czynnik B na pięciu poziomach, $1, 2, \ldots, 5$. W tabeli i na rysunku są podane średnie wewnątrzgrupowe zmiennej odpowiedzi.

**Tabela 5.5**

| | | B | | | | | Średnia |
|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | dla poziomu A |
| A | 1 | 64 | 59 | 64 | 56 | 68 | 62,20 |
| | 2 | 56 | 51 | 56 | 48 | 60 | 54,20 |
| Średnia dla poziomu B | | 60 | 55 | 60 | 52 | 64 | |




Rys. 5.6. Wykres średnich wewnątrzgrupowych (patrz tab. 5.5)

Gdy czynnik A zmienia poziom z 1 na 2, natomiast czynnik B pozostaje na nie zmienionym poziomie (1 lub 2, lub ... 5), wartość zmiennej odpowiedzi zmienia się o $-8$. Zauważmy, że zmiana ta jest – i musi być – równa różnicy między średnimi wartościami zmiennej odpowiedzi dla poziomu 1 i 2 czynnika A, uśrednionymi względem wszystkich poziomów czynnika B (średnie te nazywamy często brzegowymi średnimi czynnika A). Gdy czynnik B zmienia poziom np. z 4 na 5, natomiast czynnik A pozostaje na nie zmienionym poziomie (1 lub 2), wartość zmiennej odpowiedzi zmienia się o 12. Tym razem zmiana ta musi być równa różnicy między średnimi wartościami zmiennej odpowiedzi dla poziomu 4 i 5 czynnika B, uśrednionymi względem wszystkich poziomów czynnika A (średnie te nazywamy często brzegowymi średnimi czynnika B). Jak widać, w opisywanej sytuacji zmiana średniej wartości zmiennej odpowiedzi, odpowiadająca zmianie poziomu jednego czynnika, **nie zależy** od tego, jaki jest poziom drugiego czynnika. Mówimy wówczas o **braku interakcji** między czynnikami. W takim przypadku jest uzasadnione przyjęcie, że poziomy obydwu czynników wpływają na średnią wartość zmiennej odpowiedzi w sposób **addytywny**:

$$\mu_{ij} = \mu + \alpha_i + \beta_j, \tag{5.17}$$

gdzie $i = 1, 2, \ldots, k$, $j = 1, 2, \ldots, l$, $\mu_{ij}$ oznacza średnią na $i$-tym poziomie czynnika A i $j$-tym poziomie czynnika B, $\mu$ jest ogólną wartością średnią, $\alpha_i$ jest efektem $i$-tego poziomu czynnika A, $\beta_j$ jest efektem $j$-tego poziomu czynnika B. Efekty $\alpha_i$, $i = 1, 2, \ldots, k$, oraz $\beta_j$, $j = 1, 2, \ldots, l$, często nazywamy efektami głównymi, odpowiednio, czynnika A i B. Oczywiście w praktyce – gdy o całości zjawiska współdecydują elementy losowe – wartość średnia $\mu_{ij}$ jest nieobserwowalna. Obserwujemy natomiast zmienną odpowiedzi, czyli

$$Y_{ijm} = \mu + \alpha_i + \beta_j + \varepsilon_{ijm}, \tag{5.18}$$

gdzie $i = 1, 2, \ldots, k$, $j = 1, 2, \ldots, l$, $m = 1, 2, \ldots, n$, $Y_{ijm}$ oznacza $m$-tą obserwację na $i$-tym poziomie czynnika A i $j$-tym poziomie czynnika B, $\mu$ jest ogólną wartością średnią, $\alpha_i$ jest efektem $i$-tego poziomu czynnika A, $\beta_j$ jest efektem $j$-tego poziomu czynnika B oraz $\varepsilon_{ijm}$ są niezależnymi zmiennymi losowymi o rozkładzie normalnym $N(0, \sigma)$.

Zamiast średnich $\mu_{ij}$ na wykresy nanosi się średnie próbkowe

$$\bar{y}_{ij\cdot} = \frac{1}{n} \sum_{m=1}^{n} y_{ijm}.$$

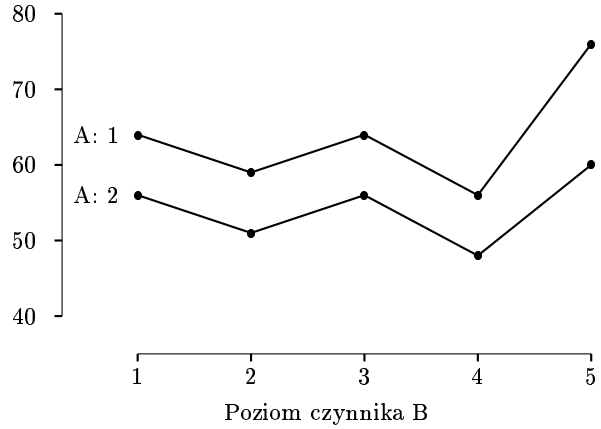
Ze względu na obecność zmiennych losowych $\varepsilon_{ijm}$, nawet wtedy gdy model (5.17)–(5.18) adekwatnie opisuje badane zjawisko, trudno w praktyce spotkać idealnie równoległe wykresy średnich. Tylko testy statystyczne mogą odpowiedzieć na pytanie, czy należy odrzucić hipotezę o adekwatności modelu addytywnego (5.17)–(5.18), czyli o braku interakcji. Zanim na to pytanie odpowiemy w kontekście przykł. 5.2, przyjrzyjmy się jeszcze innej sytuacji idealnej, gdy wykresy średnich nie są równoległe i automatycznie możemy stąd wnosić, że model addytywny nie jest adekwatny.

Przyjmijmy, że taki właśnie przypadek jest opisany w tab. 5.6, w której są podane odpowiednie średnie wartości zmiennej odpowiedzi oraz na rys. 5.7.

**Tabela 5.6**

| | | B | | | | | Średnia |
|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | dla poziomu A |
| A | 1 | 64 | 59 | 64 | 56 | 76 | 63,80 |
| | 2 | 56 | 51 | 56 | 48 | 60 | 54,20 |
| Średnia dla poziomu B | | 60 | 55 | 60 | 52 | 68 | |

Tym razem, jeżeli poziom czynnika B jest ustalony, zmiana wartości zmiennej odpowiedzi, gdy czynnik A zmienia poziom z 1 na 2, zależy od poziomu czynnika B (jest taka sama, gdy poziom czynnika B wynosi 1, 2, 3 lub 4 i inna, gdy wynosi 5). Podobnie, zmiana wartości zmiennej odpowiedzi, gdy czynnik B zmienia poziom z 4 na 5, zależy od poziomu czynnika A. A zatem zmiana średniej wartości zmiennej odpowiedzi, odpowiadająca zmianie poziomu jednego czynnika, **zależy tym razem od tego, jaki jest poziom drugiego czynnika**. Mówimy wówczas o **istnieniu interakcji** między czynnikami. W naszym przykładzie interakcja między czynnikami prowadzi np. do tego, że zmiana poziomu czynnika B z 4 na 5 daje zmianę średniej wartości zmiennej odpowiedzi o 20, gdy A = 1, oraz o 12, gdy A = 2.




Rys. 5.7. Wykres średnich wewnątrzgrupowych (patrz tab. 5.6)

Odpowiednik modelu (5.17), uwzględniający obecność interakcji, przyjmuje postać

$$\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}, \tag{5.19}$$

gdzie $i = 1, 2, \ldots, k$, $j = 1, 2, \ldots, l$, $\gamma_{ij}$ jest interakcją między $i$-tym poziomem czynnika A i $j$-tym poziomem czynnika B oraz nie jest zmienione znaczenie innych wielkości występujących w równaniu (5.19). Analogicznie, obserwacje zmiennej odpowiedzi są opisywane równaniem

$$Y_{ijm} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijm}. \tag{5.20}$$

Jak w przypadku analizy jednoczynnikowej przyjmujemy, że efekty $\alpha_i$ spełniają odpowiednik warunku (5.5). Ponadto z tych samych względów co poprzednio nakładamy warunki

$$\beta_1 + \beta_2 + \cdots + \beta_l = 0, \tag{5.21}$$

$$\gamma_{1j} + \gamma_{2j} + \cdots + \gamma_{kj} = 0 \tag{5.22}$$

dla $j = 1, 2, \ldots, l$ oraz

$$\gamma_{i1} + \gamma_{i2} + \cdots + \gamma_{il} = 0 \tag{5.23}$$

dla $i = 1, 2, \ldots, k$. Nadal też obowiązuje podstawowe założenie analizy wariancji, któremu obecnie możemy nadać następującą postać.

**Podstawowe założenie analizy wariancji:** *Dla każdej z kl możliwych kombinacji poziomów czynników A i B, rozkład zmiennej odpowiedzi jest normalny z taką samą wariancją* $\sigma^2$, $\sigma^2_{ijm} = \sigma^2$, *gdzie* $\sigma^2_{ijm}$ *oznacza wariancję zmiennej odpowiedzi dla m-tej obserwacji na i-tym poziomie czynnika A i j-tym poziomie czynnika B,* $i = 1, 2, \ldots, k$, $j = 1, 2, \ldots, l$.

Podobnie do przypadku analizy jednoczynnikowej, interesuje nas test hipotezy o braku istnienia efektu głównego pochodzącego od czynnika A, czyli test hipotezy

$$H_{0A}\colon\ \alpha_1 = \alpha_2 = \cdots = \alpha_k = 0 \tag{5.24}$$

przy alternatywie

$$H_{1A}\colon \text{ nie wszystkie efekty } \alpha_i \text{ są równe zeru} \tag{5.25}$$

oraz test hipotezy o braku istnienia efektu głównego pochodzącego od czynnika B, czyli test hipotezy

$$H_{0B}\colon\ \beta_1 = \beta_2 = \cdots = \beta_l = 0 \tag{5.26}$$

przy alternatywie

$$H_{1B}\colon \text{ nie wszystkie efekty } \beta_j \text{ są równe zeru.} \tag{5.27}$$

Problem dwuczynnikowy jest przy tym bogatszy od jednoczynnikowego nie tylko dlatego, że efekty główne mogą pochodzić od dwóch, a nie jednego czynnika. Dodatkowo należy jeszcze rozważyć hipotezę o istnieniu – lub nie – interakcji między obydwoma czynnikami. Interesuje nas zatem test hipotezy o nieistnieniu interakcji,

$$H_{0AB}\colon\ \gamma_{ij} = 0 \text{ dla wszystkich par } i, j,\ i = 1, 2, \ldots, k,\ j = 1, 2, \ldots, l \tag{5.28}$$

przy alternatywie

$$H_{1AB}\colon \text{ nie wszystkie interakcje } \gamma_{ij} \text{ są równe zeru.} \tag{5.29}$$

Zauważmy, że testy (5.24) i (5.26) jest sens przeprowadzać wtedy tylko, gdy nie ma podstaw do odrzucenia hipotezy (5.28). Rzeczywiście, jeżeli interakcje są obecne, nie ma sensu badanie efektu głównego pochodzącego od jednego czynnika bez równoczesnego odwołania się do drugiego czynnika. A zatem test hipotezy (5.28) powinien być przeprowadzony pierwszy.

Jak to czyniliśmy dotąd, tak i tym razem wszystkie testy oprzemy na konstrukcji statystyk $F$ mających postać ilorazów odpowiednich składników całkowitej sumy kwadratów

$$SST = \sum_{i=1}^{k}\sum_{j=1}^{l}\sum_{m=1}^{n}(y_{ijm} - \bar{y}_{\ldots})^2,$$

gdzie

$$\bar{y}_{\ldots} = \frac{1}{kln}\sum_{i=1}^{k}\sum_{j=1}^{l}\sum_{m=1}^{n} y_{ijm}.$$

Chcąc dokonać stosownego rozkładu całkowitej sumy kwadratów $SST$, wprowadźmy jeszcze następujące średnie:

$$\bar{y}_{ij\cdot} = \frac{1}{n}\sum_{m=1}^{n} y_{ijm},$$

$$\bar{y}_{\cdot j\cdot} = \frac{1}{kn}\sum_{i=1}^{k}\sum_{m=1}^{n} y_{ijm}$$

oraz

$$\bar{y}_{i\cdot\cdot} = \frac{1}{ln}\sum_{j=1}^{l}\sum_{m=1}^{n} y_{ijm}.$$

Wiadomo, że

$$SST = SSA + SSB + SSAB + SSE,$$

gdzie

$$SSA = ln\sum_{i=1}^{k}(\bar{y}_{i\cdot\cdot} - \bar{y}_{\cdots})^2,$$

$$SSB = kn\sum_{j=1}^{l}(\bar{y}_{\cdot j\cdot} - \bar{y}_{\cdots})^2,$$

$$SSAB = n\sum_{i=1}^{k}\sum_{j=1}^{l}(\bar{y}_{ij\cdot} - \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot j\cdot} + \bar{y}_{\cdots})^2$$

oraz

$$SSE = \sum_{i=1}^{k}\sum_{j=1}^{l}\sum_{m=1}^{n}(y_{ijm} - \bar{y}_{ij\cdot})^2.$$

$SSA$ jest zmiennością między poziomami czynnika A, $SSB$ jest zmiennością między poziomami czynnika B, $SSAB$ jest zmiennością wynikającą z interakcji między czynnikami oraz $SSE$ jest zmiennością wewnątrzgrupową i zarazem sumą kwadratów reszt.

Liczba stopni swobody związanych z $SST$ wynosi $kln-1$, liczba stopni swobody związanych z $SSA$ jest równa $k-1$, liczba stopni swobody związanych z $SSB$ jest równa $l-1$, zmienności $SSAB$ odpowiada $(k-1)(l-1)$ stopni swobody i ostatecznie sumie kwadratów reszt $SSE$ odpowiada $kl(n-1)$ stopni swobody. Statystyka testowa $F$ analizy wariancji, której możemy użyć do testowania hipotezy (5.28) przy alternatywie (5.29), ma postać

$$F_{AB} = \frac{SSAB/(k-1)(l-1)}{SSE/[kl(n-1)]}. \tag{5.30}$$

Jeżeli jest spełniona hipoteza zerowa, statystyka (5.30) ma rozkład $F$ Snedecora z $(k-1)(l-1)$ oraz $kl(n-1)$ stopniami swobody. Statystyka, której możemy użyć do testowania hipotezy (5.24) przy alternatywie (5.25), ma postać

$$F_A = \frac{SSA/(k-1)}{SSE/[kl(n-1)]}. \tag{5.31}$$

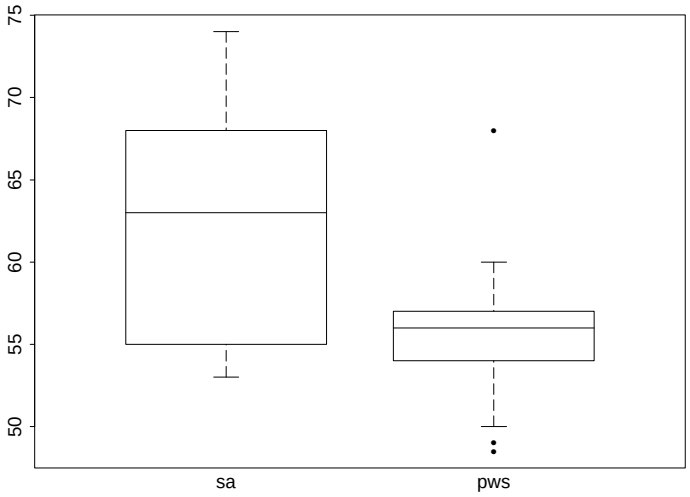
Jeżeli jest spełniona hipoteza zerowa, statystyka (5.31) ma rozkład $F$ Snedecora z $k-1$ oraz $kl(n-1)$ stopniami swobody. Analogicznie, statystyka do testowania hipotezy (5.26) przy alternatywie (5.27), ma postać

$$F_B = \frac{SSB/(l-1)}{SSE/[kl(n-1)]} \tag{5.32}$$

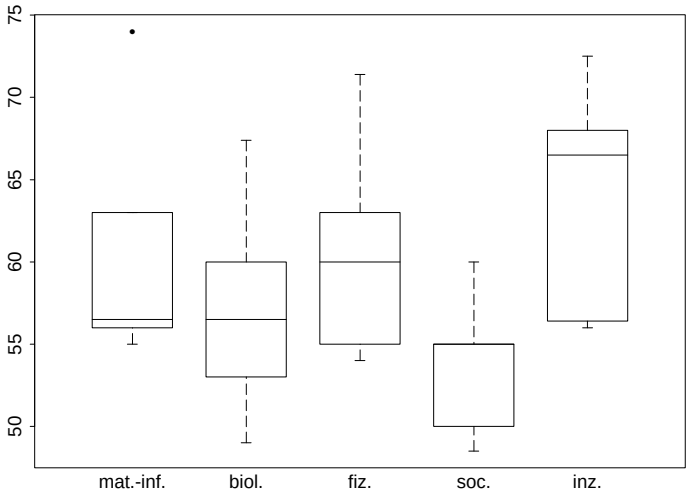
i jeżeli jest spełniona hipoteza zerowa, statystyka ta ma rozkład $F$ Snedecora z $l-1$ oraz $kl(n-1)$ stopniami swobody. Postaci zbiorów krytycznych powyższych testów są w pełni analogiczne do rozważanych wcześniej, więc je pomijamy.

**Przykład 5.2 cd.** Z wykresu kwantylowego oraz testu Shapiro–Wilka wynika, że rozkład reszt wewnątrzgrupowych można uznać za normalny; także testy jednorodności wariancji umożliwiają wykorzystanie analizy wariancji do sprawdzenia istnienia interakcji oraz - przy ich braku – efektów głównych (por. zad. 5.5). Test (5.30) uniemożliwia odrzucenie hipotezy o braku interakcji. Z kolei testy (5.31) i (5.32) wykazują obecność efektów głównych. W przypadku czynnika pierwszego (regionu) dokonywanie porównań wielokrotnych jest oczywiście niepotrzebne. W przypadku drugiego czynnika (grupy zawodów) tylko metoda Tukeya prowadzi do wykrycia różnicy między średnimi zarobkami w grupie inżynierów i socjologów. Różnice między innymi grupami są statystycznie nieistotne. (Porównaj zad. 5.6). Otrzymane wyniki dotyczące efektów głównych dobrze ilustrują wykresy ramkowe (rys. 5.8 i 5.9).

Aż do tej chwili zakładaliśmy, że mamy do czynienia z $n > 1$ replikacjami każdego doświadczenia. Jeżeli $n = 1$, znika oczywiście zmienność wewnątrzgrupowa $SSE$. Jeżeli jednak można z góry założyć, że w badanym problemie nie ma interakcji między czynnikami A i B, rolę sumy kwadratów reszt może odgrywać składnik $SSAB$. Rzeczywiście, przemianowanie bloków z p. 5.2.4 na czynnik B czyni obydwa problemy – analizy wariancji z zastosowaniem zrandomizowanego planu blokowego oraz analizy dwuczynnikowej z $n = 1$ – równoważnymi. W szczególności, suma kwadratów reszt w przypadku planu blokowego ma identyczną wówczas postać co zmienność $SSAB$ w analizie




Rys. 5.8. Wykresy ramkowe zmiennej odpowiedzi dla 2 poziomów czynnika pierwszego (region)




Rys. 5.9. Wykresy ramkowe zmiennej odpowiedzi dla 5 poziomów czynnika drugiego (grupa zawodowa)

dwuczynnikowej. Zatem, gdy $n = 1$ i przy założeniu braku interakcji, statystyka (5.16) umożliwia testowanie braku efektu głównego pochodzącego od czynnika A w problemie z dwoma czynnikami. Na tej samej zasadzie, statystyka (5.16) z licznikiem postaci $SSB/(l-1)$ umożliwia testowanie braku efektu głównego od czynnika B.

Inna sprawa, że łatwość dokonania takiej analizy nie może przesłonić potrzeby weryfikacji założenia o braku interakcji.

**Przykład 5.3.** Fundacja NSF porównuje zarobki w różnych grupach ludzi i zawodów. W tabeli 5.7 (z roku 1995) są podane mediany wynagrodzeń brutto wśród inżynierów rasy białej i żółtej w USA, z podziałem na zawody: inż. chemik, inż. lądowy, inż. elektryk lub informatyk, inż. mechanik.

**Tabela 5.7**

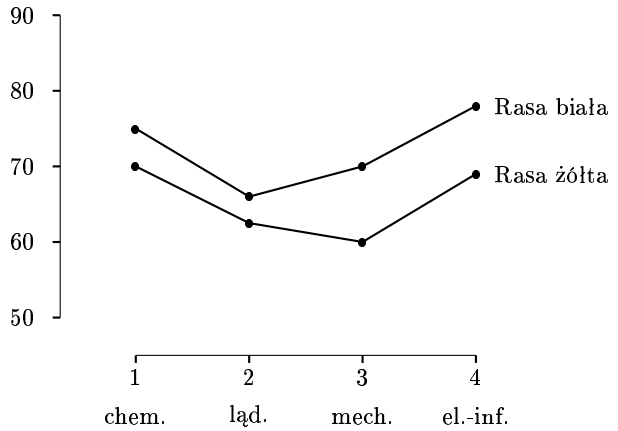
| | chem. | ląd. | el-inf. | mech. |
|---|---|---|---|---|
| Rasa biała | 75,00 | 66,00 | 78,00 | 70,00 |
| Rasa żółta | 70,00 | 62,50 | 69,00 | 60,00 |

Mamy zatem do czynienia z dwoma czynnikami: rasą (gdzie $k = 2$) i zawodem (gdzie $l = 4$). Ponieważ dla każdej z $kl = 8$ kombinacji mamy tylko po jednej obserwacji, można by się pokusić o wykonanie analizy dwuczynnikowej, jednak wyłącznie przy założeniu nieistnienia interakcji między czynnikami. Wykresy zarobków dla obydwu ras są podane na rys. 5.10[1]. Na podstawie danych NSF nie możemy ocenić zmienności wewnątrzgrupowej, nie możemy więc autorytatywnie orzec czy rzeczywiście – tak jak to sugeruje rys. 5.10 – mamy do czynienia z interakcją między rasą i zawodem (dla inżynierów pochodzenia europejskiego oraz inżynierów pochodzenia azjatyckiego mediany zarobków zachowują się wyraźnie odmiennie: wśród osób rasy białej inżynier lądowy jest gorzej wynagradzany niż inżynier mechanik, natomiast wśród Azjatów jest odwrotnie). A zatem bez uzyskania dodatkowej i wiarygodnej informacji, która uzasadniałaby (raczej zaskakujące w tym przypadku) założenie braku interakcji, nie możemy tym razem skorzystać z wyników p. 5.2.3.

W przypadku przykł. 5.3 najprawdopodobniej mamy do czynienia z interakcją między czynnikami. Przyjąwszy, że tak rzeczywiście jest, warto porównać rys. 5.7 i 5.10, by na koniec zwrócić uwagę na możliwość występowania dwóch różnych rodzajów interakcji. Zmiana wartości zmiennej odpowiedzi

[1] Pokazana na rysunku kolejność zawodów nie odpowiada kolejności kolumn w tab. 5.7. Zmiany tego typu umożliwiają łatwiejsze zauważenie różnic między efektami głównymi czynników. Nowoczesne oprogramowania wykonują takie zmiany automatycznie.

przy zmianie czynnika B z poziomu 4 na 5 (por. rys. 5.7) zależy od poziomu czynnika A. Niemniej jednak zmiany te, chociaż zależą od poziomu czynnika A, mają ten sam znak: w obydwu przypadkach wartość zmiennej odpowiedzi rośnie (nachylenia odpowiednich odcinków na rys. 5.7, odpowiadające zmianie czynnika B z poziomu 4 na 5, mają ten sam znak dla obydwu poziomów czynnika A). Można powiedzieć, że dla obydwu poziomów czynnika A jest zachowana tendencja zmiany wartości zmiennej odpowiedzi. Inaczej jest w przypadku rys. 5.10, czyli w przykł. 5.3. Tym razem zarobki niektórych kategorii inżynierów rasy białej i żółtej mają przeciwstawne tendencje: wśród białych inżynier lądowy jest gorzej wynagradzany od inżyniera mechanika, natomiast wśród Azjatów inżynier lądowy zarabia lepiej niż inżynier mechanik. Nachylenia odpowienich odcinków łamanych są przeciwnych znaków. Ten ostatni efekt jest jeszcze bardziej widoczny, gdy dane z tab. 5.7 uzupełnić o mediany wynagrodzeń tzw. inżynierów przemysłowych (również zaczerpnięte z bazy danych NSF z roku 1995). Mediana ta wynosi 59 tysięcy dolarów dla inżynierów rasy białej i 65 tysięcy dla inżynierów rasy żółtej. Mediany zarobków dla obydwu ras i pięciu grup zawodowych są zilustrowane na rys. 5.11. Na przykład tendencje zmian są wyraźnie różne, gdy porównać kategorię inżynierów lądowych z kategorią inżynierów przemysłowych.

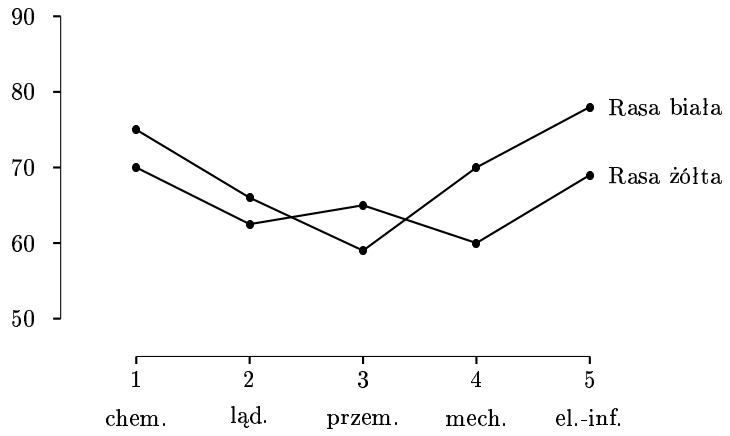



Rys. 5.10. Wykresy median zarobków inżynierów rasy białej i żółtej (tab. 5.7)

Jak wynika z przedstawionej analizy rys. 5.7 i 5.10 oraz z wcześniejszych analiz wykresów średnich, wykresy te są wartościową – chociaż nieformalną – graficzną metodą badania problemów dwuczynnikowych.

Na zakończenie niniejszego podrozdziału wróćmy jeszcze na chwilę do przykładów rozpoczynających go, dotyczących kolejno jakości tworzywa poliuretanowego oraz wielkości zbioru pszenicy. W przykładach tych chodziło

o taki dobór wartości czynników determinujących wartości zmiennych odpowiedzi, by otrzymać poprawę jakości tworzywa oraz możliwie duży plon pszenicy. Były to zatem zadania optymalizacji, w których liczyliśmy na ich przybliżone rozwiązanie. Otóż należy w tym miejscu podkreślić, że jeżeli w danym zadaniu są obecne interakcje, to niewłaściwe jest optymalizowanie (np. maksymalizacja, jak w naszych przykładach) wartości zmiennej odpowiedzi najpierw względem jednego czynnika i następnie względem czynnika drugiego. Skutkiem istnienia interakcji, czyli nieadekwatności modelu addytywnego, taka procedura może dać wynik gorszy od osiągalnego dzięki jednoczesnej optymalizacji względem obydwu czynników.




Rys. 5.11. Wykresy median zarobków inżynierów rasy białej i żółtej dla pięciu grup zawodowych

## 5.4. Zadania

**5.1.** Wykazać, że w przypadku problemu jednoczynnikowej analizy wariancji ze zrównoważonym całkowicie zrandomizowanym planem doświadczenia test $F$ oparty na statystyce (5.7) jest równoważny testowi $t$ opartemu na statystyce (3.60).
**Wskazówka:** Przyjąć w statystyce (3.60) $\bar{X}_i = \bar{Y}_{i\cdot}$, $i = 1, 2$, i zauważyć, że $\bar{\bar{Y}}_{\cdot\cdot} = (\bar{Y}_{1\cdot} + \bar{Y}_{2\cdot})/2$ oraz, że są równoważne wzory (3.29) i (5.6).

**5.2.** Korzystając z pakietu statystycznego sprawdzić, czy dla danych z przykł. 5.1 można przyjąć hipotezę o równości wariancji w grupach. Badanie oprzeć na testach Levene'a i Bartletta.

**5.3.** W tabeli są podane wielkości zbioru pszenicy ozimej otrzymane przy zastosowaniu czterech możliwych dawek azotu jako nawozu. Każdą z dawek

azotu zastosowano na 8 poletkach (dawki są zakodowane za pomocą liczb: $1, \ldots, 4$). Dla każdego z 32 poletek jest podana wartość zebranego plonu (w kwintalach na hektar).

| | Dawka 1 azotu | Dawka 2 azotu | Dawka 3 azotu | Dawka 4 azotu |
|---|---|---|---|---|
| Plon | 64,5 66,3 69,3 67,0 74,0 75,8 72,0 72,5 | 64,8 66,5 66,8 67,3 77,3 71,5 74,0 74,5 | 69,3 70,3 70,0 69,0 76,3 72,0 72,5 76,8 | 69,0 71,5 71,3 72,0 77,0 74,5 79,0 79,8 |

Zbadać, czy średnia wielkość plonu zależy od użytej ilości nawozu. Analizę wariancji poprzedzić sprawdzeniem, czy przynajmniej w przybliżeniu są spełnione odpowiednie założenia. Powtórzyć analizę opierając się na odpowiednim modelu regresji wielokrotnej. (Przykład pochodzi z podręcznika: A. Dąbrowski, S. Gnot, Michalski A., Strzednicka J. (1997): *Statystyka: 15 godzin z pakietem Statgraphics.* Wydawnictwo Akademii Rolniczej we Wrocławiu, Wrocław, gdzie stanowi ćwiczenie 5.)

**5.4.** Anderson D.R., Sweeney D.J. i Williams T.A. ((1986): *Statistics: Concepts and Applications.* West Publishing Company, St. Paul) opisują doświadczenie przeprowadzone na jednym z lotnisk w Cleveland w stanie Ohio. Doświadczenie było oparte na zrandomizowanym planie blokowym i jego celem była weryfikacja hipotezy, że trzy różne odmiany stanowiska pracy kontrolera lotów wywołują średnio taki sam stres u kontrolera (stanowisko pracy jest czynnikiem występującym na trzech poziomach). Rolę sześciu bloków pełniło sześciu losowo wybranych kontrolerów. Każdy musiał przepracować na każdym z trzech stanowisk ustaloną liczbę godzin (kolejność przypisania kontrolerowi stanowisk pracy była określona losowo). Wyniki doświadczenia, czyli wartość stresu obliczona na podstawie wywiadu i badań medycznych, odpowiadająca danemu kontrolerowi (blokowi) i pracy na danym stanowisku, są podane w poniższej tabeli.

| | Blok 1 | Blok 2 | Blok 3 | Blok 4 | Blok 5 | Blok 6 |
|---|---|---|---|---|---|---|
| Stanowisko 1 | 15 | 14 | 10 | 13 | 16 | 13 |
| Stanowisko 2 | 15 | 14 | 11 | 12 | 13 | 13 |
| Stanowisko 3 | 18 | 14 | 15 | 17 | 16 | 13 |

Sprawdzić, czy na poziomie istotności 0,05 hipoteza o braku efektu poziomu czynnika powinna zostać odrzucona.

**5.5.** Sprawdzić, czy przykład 5.2 można rozwiązać metodą dwuczynnikowej analizy wariancji.

**5.6.** Sprawdzić, czy w przykładzie 5.2 nie ma podstaw do odrzucenia hipotezy o braku interakcji między czynnikami, zbadać obecność efektów głównych oraz do drugiego czynnika (grupy zawodów) zastosować metodę porównań wielokrotnych Tukeya i Scheffégo. Skomentować otrzymane wyniki.

**5.7.** W zadaniu 5.3 pominęliśmy informację, że każde poletko było nawadniane jedną z dwóch metod: 16 poletek było nawodnione metodą 1 oraz 16 metodą 2. Przy zadanej metodzie nawadniania, daną ilością azotu nawożono 4 poletka. Pełny opis rzeczywiście przeprowadzonego doświadczenia oraz uzyskane wyniki zawiera poniższa tabela.

| | Dawka 1 azotu | Dawka 2 azotu | Dawka 3 azotu | Dawka 4 azotu |
|---|---|---|---|---|
| Metoda 1 nawadniania | 64,5 66,3<br>69,3 67,0 | 64,8 66,5<br>66,8 67,3 | 69,3 70,3<br>70,0 69,0 | 69,0 71,5<br>71,3 72,0 |
| Metoda 2 nawadniania | 74,0 75,8<br>72,0 72,5 | 77,3 71,5<br>74,0 74,5 | 76,3 72,0<br>72,5 76,8 | 77,0 74,5<br>79,0 79,8 |

Sprawdzić, czy przynajmniej w przybliżeniu są spełnione odpowiednie założenia analizy wariancji. Jeśli można, przeprowadzić dwuczynnikową analizę wariancji, by stwierdzić czy interakcje między obydwoma czynnikami (nawozem i metodą nawadniania) są istotne. Jeżeli interakcje nie są istotne, sprawdzić, czy obecne są efekty główne. Porównać wyniki dotyczące efektu głównego związanego z nawozem z wynikami z zad. 5.3 i wyjaśnić ewentualne różnice. (Por. ćwiczenie 6 w podręczniku: Dąbrowski A., Gnot S., Michalski A., Strzednicka J. (1997): *Statystyka: 15 godzin z pakietem Statgraphics.*)

**5.8.** W pewnej uczelni technicznej postanowiono sprawdzić celowość wprowadzenia nowego kursu analizy matematycznej II, mającego ewentualnie zastąpić kurs tradycyjny. Nowy kurs opiera się na wychodzeniu od przykładów rzeczywistych problemów technicznych i wprowadzaniu nowych pojęć i metod rozwiązywania problemów zależnie od wynikłych potrzeb. Zajęcia postanowiono poprowadzić w dwóch tokach, odpowiadających dwóm typom kursów: tradycyjnemu (T) i nowemu (N) . Studentów podzielono na podstawie wyników egzaminu z analizy matematycznej I na sześć grup: studentów bardzo zdolnych, zdolnych, przeciętnych, prawie przeciętnych, słabych i bardzo słabych. Następnie z grup studentów bardzo zdolnych (B), przeciętnych (P) i słabych (S) wylosowano po 24 studentów i losowo przydzielono ich do 8 grup (w każdej grupie studenckiej było przynajmniej 18 studentów, przy czym starano się by w każdej grupie studenckiej było po 3 studentów o danym poziomie zdolności). W losowy sposób po 4 grupy znalazły się w toku T i N nauczania analizy. Studenci wiedzieli tylko, że zostali arbitralnie podzieleni na dwa toki studiów, nie wiedzieli zaś o prowadzeniu eksperymentu. W poniższej tabeli są podane wyniki egzaminu końcowego z analizy II dla dwóch kursów (toku) analizy (T i N), przy czym dla każdego kursu wyniki obejmują po 12 studentów o zdolnościach B, P i S.

| | Grupa B | Grupa P | Grupa S |
|---|---|---|---|
| Kurs T | 46,0 43,0 50,0 47,5<br>44,5 47,0 45,5 43,5<br>45,0 45,5 45,0 47,5 | 34,0 36,5 36,0 37,5<br>32,0 38,0 35,0 36,5<br>37,0 34,5 35,5 35,0 | 29,0 27,5 25,5 30,0<br>27,0 26,5 23,0 28,5<br>27,5 27,0 26,5 26,5 |
| Kurs N | 44,0 46,5 43,0 44,5<br>42,5 45,5 45,0 43,5<br>44,5 47,0 45,5 44,0 | 39,5 37,0 37,5 37,0<br>37,5 40,5 38,5 37,5<br>33,0 34,0 37,5 41,5 | 32,5 31,0 33,0 29,0<br>31,0 28,5 31,5 34,5<br>34,0 33,0 33,0 30,0 |

Dokonać analizy wariancji z kursem analizy (T i N) i zdolnościami (B, P i S) jako czynnikami. Dokładnie opisać istniejące interakcje, wykorzystując wykresy: średniego wyniku w funkcji poziomu zdolności przy ustalonym kursie (2 łamane na wspólnym rysunku, jedna odpowiadająca kursowi T i jedna kursowi N); średniego wyniku w funkcji kursu przy ustalonym poziomie zdolności (3 odcinki na wspólnym rysunku, po jednym dla każdego poziomu zdolności). W szczególności odpowiedzieć na pytania: Czy wyniki studentów B istotnie zależą od typu kursu? Jakich odpowiedzi należy udzielić na podobne pytania w przypadku studentów P i S? Czy różnice między wynikami studentów B, P i S w przypadku kursu T są podobne do odpowiadających im różnic w przypadku kursu N?

**5.9.** Poniższa tabela zawiera wyniki zmiennej $SDNN$ rytmu serca dziewcząt i chłopców w kategoriach wiekowych 11–12 lat, 13–14 lat oraz 15–16 lat (jest to mały fragment bardzo obszernego studium dr J. Rękawek z Kliniki Kardiologii Centrum Zdrowia Dziecka w Warszawie, które obejmowało wiele charakterystyk rytmu serca nie tylko dziewcząt i chłopców, ale także młodszych dzieci, niemowląt i noworodków, przy czym w każdej grupie było przynajmniej 30 obserwacji).

| | 11–12 | 13–14 | 15–16 |
|---|---|---|---|
| DZ | 162,7 154,1 137,4<br>128,7 183,3 128,7<br>114,6 109,7 126,7<br>129,4 171,4 | 159,4 188,8 148,2<br>153,7 185,3 145,4<br>158,1 193,3 136,0<br>124,4 161,0 | 206,0 135,4 238,3<br>191,4 162,7 179,8<br>165,3 136,0 200,1<br>193,3 163,7 |
| CH | 192,7 126,4 188,2<br>168,1 132,4 182,3<br>151,3 171,7 159,6<br>179,2 158,2 | 183,4 180,6 164,7<br>226,6 208,7 144,9<br>200,0 156,2 130,4<br>193,6 186,5 | 203,0 181,8 180,2<br>142,6 140,9 194,5<br>176,2 147,6 157,6<br>202,8 232,3 |

Dokonać analizy wariancji z płcią jako jednym czynnikiem (na dwóch poziomach zakodowanych jako DZ i CH) oraz wiekiem jako drugim czynnikiem (na trzech poziomach zakodowanych jako 11–12, 13–14 i 15–16). Jak wyjaśnić brak podstaw do odrzucenia hipotezy o istnieniu interakcji na podstawie odpowiedniego testu $F$ analizy wariancji, mimo że wykresy średnich wewnątrzgrupowych sugerują ich istnienie? Czy jest to raczej wynik dużej zmienności wewnątrzgrupowej, czy też odstępstwa rozkładu reszt od normalności? Zinterpretować wyniki dla efektów głównych, w tym wyniki porównań wielokrotnych.

**5.10.** W poniższej tabeli (por. przykł. 5.2) są podane mediany zarobków rocznych brutto osób ze stopniem naukowym doktora w dwóch regionach USA: w regionie środkowo-atlantyckim (śa) oraz zachodnio-południowo-środkowym (zpś). W każdym regionie są podane mediany z trzech stanów dla następujących (szeroko rozumianych) grup zawodowych: matematycy i informatycy (mat.-inf.); biolodzy i naukowcy pokrewnych specjalności (biol.); fizycy i naukowcy pokrewnych specjalności (fiz.); socjolodzy i naukowcy pokrewnych specjalności (soc.); inżynierowie (inż.). Wszystkie dane pochodzą z roku 1995 i zostały zebrane przez fundację NSF.

| | mat.-inf. | biol. | fiz. | soc. | inż. |
|---|---|---|---|---|---|
| Region śa | 74 63 55 | 67,4 58 53 | 71,4 63 60 | 60 55 55 | 72,5 68 65 |
| Region zpś | 45 45 58,7 | 46 48,9 52 | 58 50 59 | 50 58 48 | 62 64 70 |

Sformułować i rozwiązać problem dwuczynnikowej analizy wariancji z regionem USA i grupą zawodową jako czynnikami.

**5.11.** W poniższej tabeli są zawarte czasy przeżycia (w jednostkach dziesięciogodzinnych) zwierząt poddanych leczeniu po podaniu im pewnego preparatu. Preparat był podawany w jednej z trzech możliwych dawek (zakodowanych jako dawki m, ś i d). Zwierzę było poddawane jednej z czterech możliwych terapii (zakodowanych jako terapie 1, 2, 3 i 4). Dla każdej z 12 kombinacji dawki preparatu i leczenia doświadczeniu poddawano 4 losowo wybrane zwierzęta.

| | Terapia 1 | Terapia 2 | Terapia 3 | Terapia 4 |
|---|---|---|---|---|
| Dawka m | 0,31 0,45 | 0,82 1,10 | 0,43 0,45 | 0,45 0,71 |
| | 0,46 0,43 | 0,88 0,72 | 0,63 0,76 | 0,66 0,62 |
| Dawka ś | 0,36 0,29 | 0,92 0,61 | 0,44 0,35 | 0,56 1,02 |
| | 0,40 0,23 | 0,49 1,24 | 0,31 0,40 | 0,71 0,38 |
| Dawka d | 0,22 0,21 | 0,30 0,37 | 0,23 0,25 | 0,30 0,36 |
| | 0,18 0,23 | 0,38 0,29 | 0,24 0,22 | 0,31 0,33 |

Stosując analizę wariancji, w tym uwzględniając możliwość istnienia interakcji, odpowiedzieć na pytanie czy (i ewentualnie jak) średni czas przeżycia zależy od zastosowanych dawek preparatu oraz sposobu leczenia. (Dane pochodzą z pracy: Box G.E.P, Cox D.R.: *Journal of the Royal Statistical Society* A, vol. 143, s. 383-340[2].)

**Wskazówka:** Zauważyć najpierw, że rozkłady reszt – jak zwykle w przypadku czasów przeżycia – są prawostronnie skośne i zamiast tychże czasów badać ich odwrotności (czasy przeżycia najczęściej poddaje się przekształceniu logarytmicznemu, ale Box i Cox wykazali, że w tym przypadku lepsze dopasowanie do rozkładu normalnego uzyskuje się, stosując przekształcenie $1/x$).

[2]Dane są powtórzone w zbiorze Hand D. i in. (1994): *A handbook of small data sets*. Londyn, Chapman & Hall.

ROZDZIAŁ 6

# Analiza danych jakościowych

## 6.1. Wprowadzenie

Niewiele miejsca poświęciliśmy dotąd statystycznej analizie danych jakościowych. W podrozdziale 1.2 przedstawiliśmy wykresy słupkowe i kołowe oraz w rozdz. 5 stosunkowo wyczerpująco omówiliśmy analizę wariancji, gdzie co prawda zmienna odpowiedzi jest ilościowa, ale czynniki mogą być – i często są – jakościowe. By wspomniany brak uzupełnić, analizie danych jakościowych poświęcamy niniejszy rozdział.

Zacząć wypada od zwrócenia uwagi na to, że dane jakościowe mogą być dwojakiego typu. Mogą mianowicie opisywać cechy nominalne lub uporządkowane i (odpowiednio) nosić nazwę danych **nominalnych** lub danych o **wartościach uporządkowanych**.

Niech, tak jak w przykł. 1.1, interesującą nas cechą będzie wyznanie religijne, populacją zaś, w której chcemy poznać rozkład tej cechy, niech będą mieszkańcy Warszawy w zadanym roku. Cechę tę opisują następujące kategorie: katolik, prawosławny, ewangelik, żyd oraz kategoria „inne wyznania", z których wymienienia rezygnujemy. Zgodnie z wcześniej przyjętą konwencją możemy powiedzieć, że dane pochodzące z rejestracji wyznania kolejnych warszawiaków przyjmują **wartości** lub **poziomy** o podanych właśnie nazwach. Wyznanie religijne jest cechą nominalną, nie ma bowiem żadnej relacji jakoś porządkującej kategorie wyznaniowe – są to po prostu różne kategorie i tyle. Tak samo rzecz się ma z takimi cechami jak kolor włosów, nazwa miejscowości, w której może być zameldowany mieszkaniec województwa pomorskiego, czy zawód.

Inaczej jest, gdy interesującą nas cechą jest stopień sympatii do pewnej partii politycznej, przy czym cecha ta może występować na pięciu poziomach: nienawidzę, nie lubię, jest mi obojętna, lubię, bardzo lubię. Ów stopień sympatii jest ewidentnie uporządkowany, ponieważ możliwe poziomy sympatii

potrafimy uporządkować od najgłębszej antypatii do wielkiej sympatii. Podobnie mamy do czynienia z cechą uporządkowaną, gdy pytamy nie o kolor włosów, a o to czy ktoś jest łysy, ma mało włosów, ma typową liczbę włosów na głowie, czy też ma gęstą czuprynę. Cechę uporządkowaną otrzymamy także wówczas, gdy mieszkańców Pomorza podzielimy na mieszkających na wsi, w małym mieście, średnim mieście oraz dużym mieście.

Kierując się względami językowymi często zamiast mówić o poziomach, na jakich może występować dana cecha, będziemy zamiennie posługiwać się pojęciem **kategorii**. Samą cechę nazywać też będziemy zmienną (często, chociaż tego terminu nie będziemy używać w niniejszej książce, zamiast o cechach mówi się o **atrybutach** obiektu).

Przyglądając się uważniej podanym przykładom cech jakościowych musimy dostrzec, iż różnią się co do swego statusu nie tylko tym, że są albo nominalne, albo uporządkowane. Cechy nominalne są cechami fundamentalnie jakościowymi w tym sensie, że nie mają żadnych związków z cechami ilościowymi. Przy tym cecha taka pozostaje fundamentalnie jakościową także wtedy, gdy jej kategorie zakodujemy z jakiegoś względu jako liczby – liczba odgrywa wówczas jedynie rolę nazwy i nie ma żadnego sensu np. odejmowanie jednej liczby od drugiej (co miałoby oznaczać odejmowanie ewangelika od katolika?).

W przeciwieństwie do cech nominalnych, cechy uporządkowane mają – w mniejszym lub większym stopniu – związek z pewnymi cechami ilościowymi. W przypadku sympatii do określonej partii politycznej można zauważyć, że niejako w tle mamy na myśli jakąś cechę mierzalną o nieskończenie wielu wartościach. Możemy powiedzieć, że między nienawiścią a miłością do partii politycznej (nie ma tu nic do rzeczy, że takie skrajne uczucia w odniesieniu do partii politycznej nie są racjonalne) istnieje *continuum* różnych stopni sympatii. Ponieważ jednak owej intuicji mierzalności sympatii na skali o nieskończenie wielu wartościach nie umiemy sprecyzować, nie tylko mamy rację cechę tę uznając za jakościową, tyle że uporządkowaną, ale też uznając jej związek z jakąś cechą ilościową za bardzo niejasny.

Z kolei taka cecha jak fakt mieszkania na wsi lub w małym, średnim czy dużym mieście może mieć mocny i dobrze określony związek z cechą ilościową, a mianowicie z liczbą mieszkańców miast Pomorza. Otóż interesująca nas tu cecha jakościowa mogła powstać przez ustalenie, iż małymi miastami nazywamy miasta o liczebności mieszkańców do 50000, natomiast miastami średnimi są miasta zamieszkiwane przez więcej niż 50000, ale nie więcej niż 200000 osób. A zatem cechy ewidentnie jakościowe mogą powstawać przez **dyskretyzację** jakiejś cechy ilościowej (w naszym przykładzie cechy opisującej liczbę mieszkańców miast).

W niniejszym wprowadzeniu musimy jeszcze poruszyć dwie kwestie interpretacyjne – traktowania danych jakościowych jako ilościowych i odwrotnie, traktowania cech ilościowych tak jakby były jakościowymi.

Oceniamy studentów, stosując stopnie z pozoru liczbowe, 2, 3, 4, 5, ewentualnie jeszcze z połówkami, 2,5, itd. Dziwi nas, że np. Anglosasi stosują wyłącznie oceny literowe, D zamiast naszej dwójki, C i B oraz A zamiast piątki. A przecież oceny opisują uporządkowaną cechę jakościową! Student, który ma trójkę nie jest o 1 lepszy od studenta z dwóją i o 2 gorszy od studenta z piątką. Jeszcze śmieszniej brzmiałoby stwierdzenie, że student czwórkowy jest 2 razy lepszy od dwójkowego. Liczby 2, 3, 4, 5 są po prostu kodami ocen jakościowych, mówiących, że student może być niedostatecznie przygotowany, słabo, dobrze lub bardzo dobrze przygotowany. Gdy zatem mamy dokonać analizy danych z pozoru liczbowych, trzeba zacząć od sprawdzenia, czy nie są to w istocie kody kategorii jakościowych i jeśli tak, to – gdy jest to tylko możliwe – postępować zgodnie z procedurami właściwymi dla danych jakościowych.

Niekiedy usprawiedliwia się pogwałcenie podanej właśnie reguły, trzeba jednak zawsze zachować krytycyzm wobec takiego, metodologicznie niepoprawnego postępowania. Na przykład, jeżeli ocenom z kolokwiów, testów i ustnych odpowiedzi w ciągu semestru trudno jest nadać obiektywny, ilościowy charakter, wypada pozostać przy ocenach porządkowych. Jeżeli są to oceny zakodowane liczbowo, trudno na koniec semestru postąpić inaczej niż wystawić ocenę równą średniej z ocen uzyskanych w ciągu semestru, mimo, że jest to krok arbitralny, a jego ścisłe uzasadnienie nie istnieje.

W tym miejscu konieczna jest ważna dygresja. Zwróciliśmy uwagę, że stwierdzenie iż student czwórkowy jest dwa razy lepszy od dwójkowego jest jeszcze bardziej rażąco absurdalne niż opinia, że trójkę i dwójkę dzieli odległość równa 1. Rzecz w tym, że dane ilościowe mogą być dwojakiego typu – niektóre można mierzyć na tzw. skali ilorazowej, inne zaś tylko na tzw. skali przedziałowej. O skali **ilorazowej** mówimy wtedy, gdy wyniki obserwacji są określone w sposób jednoznaczny z dokładnością do użytej jednostki – tak jest, gdy mierzymy np. długość albo objętość. O skali **przedziałowej** mówimy wtedy, gdy mamy możliwość wyboru nie tylko jednostki pomiaru, ale i położenia zera na osi pomiarowej – tak jest np. w przypadku określania temperatury na skali innej niż bezwględna. Zauważmy, że w przypadku temperatury w skali Fahrenheita zero jest przesunięte względem temperatury w skali Celsjusza o 32°C. Skutkiem względności położenia zera, w przypadku skali przedziałowej nie ma sensu mówienie, że np. temperatura 36°C jest dwa razy wyższa od temperatury 18°C. Natomiast można powiedzieć, że pierwsza temperatura jest o 18°C wyższa od drugiej. Z kolei, w przypadku skali bezwzględnej lorda Kelvina możemy powiedzieć, że temperatura 305 K jest dwa razy wyższa od temperatury 152,5 K i że temperatury te różnią się o 152,5 K – temperatura

w kelwinach jest mierzona na skali ilorazowej. Ocena z przedmiotu jest mierzona na skali porządkowej i stąd błędem jest formułowanie sądów właściwych dla pomiarów na skali przedziałowej, ale szczególnie absurdalnie brzmi ferowanie wyroków uzsadnionych tylko w przypadku pomiarów na skali ilorazowej. Inna sprawa, że gdy studentom zadaje się w teście 100 pytań, za trafną odpowiedź egzaminowana osoba otrzymuje 1 punkt, za odpowiedź błędną 0 punktów i wynikowa ocena jest równa sumie zdobytych punktów, to ocena ta jest mierzona na skali ilorazowej.

Przejdźmy teraz do drugiej z wymienionych kwestii interpretacyjnych, czyli do traktowania zmiennych ilościowych jak jakościowych. Część Czytelników pamięta zapewne tragedię amerykańskiego wahadłowca Challenger, który spłonął w roku 1986 kilkanaście sekund po starcie z Przylądka Canaveral. Tragedii uniknięto by, gdyby wcześniej dostrzeżono, że szansa uszkodzenia podczas startu wahadłowca pewnych pierścieni uszczelniających zależy od temperatury otoczenia na ziemi. Krytyczny start odbywał się w wyjątkowo niskiej temperaturze i gdyby wiedziano, iż szansa powstania uszkodzeń rosła wraz z obniżaniem się temperatury, lot Challengera odłożono by. Można było w tamtym krytycznym dniu oprzeć się na 24 wcześniejszych startach wahadłowców, w trakcie których żadne uszkodzenie nie pojawiło się 17 razy, jedno uszkodzenie powstało 5 razy, dwa uszkodzenia 1 raz i również raz wystąpiły trzy uszkodzenia (temperatury otoczenia przed startem obejmowały przedział od 50 do 85 stopni Fahrenheita; dane cytujemy za: Foster D.P., Stine R.A., Waterman R.P. (1998): *Business analysis using regression.* Springer, New York). Liczba uszkodzeń jest oczywiście zmienną ilościową. Jednak wobec tego, że zmienna ta przyjmuje zaledwie 4 różne wartości odległe jedna od drugiej przynajmniej o 1, posłużenie się modelem regresyjnym z liczbą uszkodzeń jako zmienną objaśnianą, jest bardzo kontrowersyjne. Jak zobaczymy w podrozdz. 6.4, wystarczy przeprowadzenie bardzo prostej dyskretyzacji temperatury (np. podzielenie jej na zaledwie dwa przedziały, do 65°F i powyżej 65°F) oraz potraktowanie otrzymanej zmiennej zdyskretyzowanej jako jednej cechy uporządkowanej, zaś faktu wystąpienia lub nie uszkodzeń jako drugiej takiej cechy. Dla podanych cech można łatwo ustalić, czy istnieje między nimi interesująca nas zależność.

Niekiedy, gdy zmienna ilościowa przyjmuje bardzo mało wartości, nie jest błędem metodologicznym potraktowanie takiej zmiennej jako jakościowej. Można powiedzieć, iż nieraz uzasadnione jest potraktowanie możliwych wartości zmiennej ilościowej jako różnych kategorii zmiennej jakościowej o uporządkowanych wartościach. Co więcej, dyskretyzację zmiennej ilościowej (w naszym przykładzie temperatury) można również potraktować jako przekształcenie tej zmiennej w uporządkowaną cechę jakościową.

W literaturze anglosaskiej takie zmienne powstałe ze zmiennych ilościowych oraz rzeczywiste zmienne jakościowe opatruje się wspólną nazwą zmiennych

lub danych **skategoryzowanych**. Czerpiąc z terminologii anglojęzycznej niniejszy rozdział można by zatytułować „Analiza danych skategoryzowanych". My jednak pozostaniemy przy mówieniu o danych jakościowych oraz ilościowych traktowanych jako jakościowe.

Podkreślmy jeszcze, że traktowanie oryginalnych zmiennych ilościowych jako uporządkowanych cech jakościowych zawsze wymaga usprawiedliwienia przez cel, jakiemu ma służyć. W podanym wyżej przykładzie usprawiedliwieniem jest uzyskanie adekwatnej odpowiedzi na pytanie o istnienie zależności między zmiennymi.

Trzeba jednak zawsze pamiętać, że potraktowanie zmiennej ilościowej jako jakościowej uniemożliwia uzyskanie odpowiedzi na wszelkie pytania wymagające ilościowego charakteru zmiennej.

Z traktowaniem danych ilościowych jako danych o uporządkowanych wartościach zetkniemy się jeszcze w rozdz. 9, poświęconym tzw. metodom rangowym. Sytuacja tam rozważana istotnie się różni od aktualnie omawianej. Ujmując rzecz najkrócej i nieprecyzyjnie: gdy obecnie chcemy mieć do czynienia z kilkudziesięcioma lub więcej danymi, które należą do kilku uporządkowanych kategorii, w rozdz. 9 zajmować się będziemy problemami, w których na kilkadziesiąt danych najwyżej kilka razy liczność danych należących do jednej kategorii może nieznacznie przekraczać 1.

Następny podrozdział jest poświęcony statystycznej analizie rozkładu jednej zmiennej. W podrozdziale 6.4 przedmiotem naszego zainteresowania jest para zmiennych jakościowych. Rozważamy zagadnienia niezależności zmiennych lub ich zależności. Próbę losową pary zmiennych jakościowych, na którą składa się $n$ par obserwacji, będziemy zapisywać w formie **tablicy kontyngencji**, zwanej także **tablicą wielodzielczą** (w przypadku analizowania par zmiennych losowych, **tablicą dwudzielczą**). Załóżmy, że jedna zmienna jakościowa, $X$, ma $k$ kategorii, zakodowanych jako $x_1, x_2, \ldots, x_k$, natomiast druga zmienna, $Y$, ma $l$ kategorii, $y_1, y_2, \ldots, y_l$. Załóżmy dalej, że w próbie $n$ par $(x, y)$, gdzie $x$ jest zaobserwowaną kategorią zmiennej $X$ oraz $y$ jest zaobserwowaną kategorią zmiennej $Y$, mamy $n_{11}$ par $(x_1, y_1)$, $n_{12}$ par $(x_1, y_2)$, ..., $n_{kl}$ par $(x_k, y_l)$. Ogólnie możemy napisać, że $n_{ij}$ jest liczbą wystąpienia w próbie par obserwacji $(x_i, y_j)$. Oczywiście

$$\sum_{i=1}^{k} \sum_{j=1}^{l} n_{ij} = n.$$

Używając języka algebry liniowej możemy powiedzieć, że tablica kontyngencji jest macierzą o $k$ wierszach i $l$ kolumnach z elementami $n_{ij}$ na przecięciu $i$-tego wiersza oraz $j$-tej kolumny, $i = 1, 2, \ldots, k$, $j = 1, 2, \ldots, l$:

| | $y_1$ | $y_2$ | $\dots$ | $y_j$ | $\dots$ | $y_l$ |
|---|---|---|---|---|---|---|
| $x_1$ | $n_{11}$ | $n_{12}$ | $\dots$ | $n_{1j}$ | $\dots$ | $n_{1l}$ |
| $x_2$ | $n_{21}$ | $n_{22}$ | $\dots$ | $n_{2j}$ | $\dots$ | $n_{2l}$ |
| $\vdots$ | | | | | | |
| $x_i$ | $n_{i1}$ | $n_{i2}$ | $\dots$ | $n_{ij}$ | $\dots$ | $n_{il}$ |
| $\vdots$ | | | | | | |
| $x_k$ | $n_{k1}$ | $n_{k2}$ | $\dots$ | $n_{kj}$ | $\dots$ | $n_{kl}$ |

Mówimy, że podana tablica ma wymiar $k \times l$. Plan eksperymentu, który polega na wylosowaniu próby $n$ jednostek, gdzie $n$ jest ustaloną liczbą, i przypisaniu kategorii jakie w przypadku każdej jednostki odpowiadają zmiennym $X$ oraz $Y$, nazywamy **planem krzyżowym**.

Z tablicami kontyngencji zetknęliśmy się już w podrozdz. 3.4, ale tam jej elementami były prawdopodobieństwa zdarzeń, $P(X = x_i, Y = y_j)$. Zakładaliśmy wówczas znajomość łącznego rozkładu prawdopodobieństwa zmiennych $(X, Y)$. Teraz nasze zadanie polega na wnioskowaniu o tym rozkładzie na podstawie danych. Można przy tym wykazać, że iloraz

$$\frac{n_{ij}}{n}$$

jest estymatorem największej wiarogodności prawdopodobieństwa $P(X = x_i, Y = y_j)$. Jak z tego wynika, zapisana obecnie tablica kontyngencji pełni funkcję próbkowego odpowiednika tablic z rozdz. 3.

**Przykład 6.1.** Odpowiedź na pytanie czy można było uniknąć tragedii Challengera w roku 1986 wymaga zbadania danych z 24 wcześniejszych startów wahadłowców w przestrzeń kosmiczną. Tak jak to już zasugerowaliśmy, opiszmy każdy start dwiema cechami: temperaturą otoczenia zdyskretyzowaną do dwóch przedziałów oraz uszkodzeniami krytycznych pierścieni uszczelniających (ta druga cecha ma także dwie kategorie: brak uszkodzeń i obecność przynajmniej 1 uszkodzenia). Dane możemy przedstawić w następującej tablicy kontyngencji o wymiarach $2 \times 2$:

| | Brak | Obecność |
|---|---|---|
| $\leq 65°\mathrm{F}$ | 0 | 4 |
| $> 65°\mathrm{F}$ | 17 | 3 |

Sytuacja była rzeczywiście zadziwiająca. Podczas gdy wszystkim czterem startom w temperaturze nie przekraczającej 65°F towarzyszyły

uszkodzenia pierścieni, w aż 17 z 20 startów w wyższych temperaturach wszystko odbyło się bez żadnych kłopotów. Do przykładu 6.1 wrócimy w podrozdz. 6.4.

Podrozdział 6.3 jest poświęcony weryfikacji hipotezy, że dwie lub więcej populacji ma ten sam rozkład prawdopodobieństwa. Hipotezę taką nazywamy **hipotezą o jednorodności rozkładów**. Oczywiście zakładamy zatem, że każda z populacji jest opisana jakościową zmienną losową mogącą przyjmować takie same wartości.

**Przykład 6.2.** Pewna uczelnia prowadzi wśród studentów ankiety, mające na celu ocenę zarówno programów nauczania, jak i kadry nauczającej. Ćwiczenia z analizy matematycznej oraz algebry prowadzą trzej asystenci. Każdy jest oceniany przez każdego studenta. Student może uznać sposób prowadzenia zajęć za zdecydowanie niezadowalający, niezadowalający, mierny, dobry i bardzo dobry. Ocena jest obserwowaną zmienną losową o kategoriach zakodowanych odpowiednio jako: bnzd, nzd, mrn, db, bdb. Każdemu z trzech asystentów można przypisać rozkład tej zmiennej. Interesuje nas, czy trzy rozkłady są takie same. Odpowiedzi musimy udzielić na podstawie zebranych danych, czyli ocen zawartych w ankietach. Całą informację możemy przedstawić w postaci tablicy kontyngencji, w której np. kategorie zmiennej losowej stanowić będą kolumny, natomiast wiersze odpowiadać będą kolejnym populacjom, czyli w tym przypadku, ocenianym pracownikom; pierwszy wiersz opisuje dane wylosowane z pierwszego rozkładu, czyli odpowiadające pierwszemu pracownikowi, itd. (zatem, na przecięciu pierwszego wiersza i pierwszej kolumny jest podana liczba osób, które pierwszemu pracownikowi dały ocenę bnzd, itd.):

| | bnzd | nzd | mrn | db | bdb |
|---|---|---|---|---|---|
| Asystent nr 1 | 17 | 25 | 21 | 9 | 10 |
| Asystent nr 2 | 11 | 29 | 18 | 12 | 9 |
| Asystent nr 3 | 6 | 7 | 11 | 39 | 21 |

Do przykładu 6.2 wrócimy w podrozdz. 6.3.

Zauważmy, że znaczenie ostatniej tablicy jest inne niż tablicy w przykł. 6.1, chociaż obydwie są tej samej postaci ogólnej. Tym razem wiersze opisują zaobserwowane rozkłady kolejnych populacji (studentów uczonych przez poszczególnych asystentów) i, co więcej, liczba wszystkich obserwacji w wierszu jest z góry ustalona. Odwołując do wcześniej przyjętej notacji, w ogólności mamy $k$ populacji opisanych rozkładami o $l$ możliwych kategoriach. Z $i$-tej

populacji, $i = 1, 2, \ldots, k$, pobieramy próbkę o $n_{i\cdot}$ elementach, gdzie

$$n_{i\cdot} = \sum_{j=1}^{l} n_{ij}$$

(kropka zamiast drugiego wskaźnika w symbolu $n_{i\cdot}$ oznacza sumowowanie liczności $n_{ij}$ po wskaźniku $j$).

## 6.2. Analiza jednej zmiennej

### 6.2.1. Uwagi wstępne

Na pytanie dotyczące graficznego przedstawienia rozkładu zmiennej jakościowej odpowiedzieliśmy w p. 1.2.1. Jak zwróciliśmy już na to uwagę w poprzednim podrozdziale, również szczególnie proste przypadki danych ilościowych, gdy dane te mogą przyjmować niewiele wartości, wygodnie jest niekiedy analizować tak, jak analizujemy dane jakościowe. Co więcej, np. konstrukcja diagramów liczebności i częstości, taka jak w przykł. 1.2, nie odbiega niczym od konstrukcji wykresów słupkowych dla danych jakościowych o uporządkowanych kategoriach, gdy kolejność kategorii na wykresie powinna być zgodna z naturalnym uporządkowaniem kategorii.

Podobna uwaga dotyczy histogramu do przykł. 1.3, ponieważ histogram ten powstaje przez dyskretyzację oryginalnych danych ilościowych, czyli „skategoryzowanie" danych na kilka kategorii i tym sposobem niejako skonstruowanie zmiennej jakościowej. Podkreślmy jednak raz jeszcze ograniczenia związane z takim postępowaniem. Im więcej jest możliwych wartości zmiennej ilościowej, i im większa jest liczność próby, tym więcej informacji o próbie tracimy dokonując dyskretyzacji. W przypadku przykł. 1.4–1.6 histogramy są już tylko dalekimi od doskonałości przybliżeniami ciągłych gęstości opisujących rozkłady omawianych tam danych ilościowych. I oczywiście, o czym też już wspominaliśmy, odpowiedzenie na wiele ważnych pytań wymaga skorzystania z ilościowego charakteru danej zmiennej.

Analiza rozkładu zmiennej jakościowej zwykle nie kończy się na jego graficznym przedstawieniu. Ważnym zagadnieniem jest np. pytanie o zgodność tego rozkładu z rozkładem zadanym. Jest to pytanie z dziedziny testowania hipotez, w tym przypadku z dziedziny testów zgodności. Testowana hipoteza może być przy tym prosta lub złożona. W pierwszym przypadku hipoteza zerowa orzeka zgodność z jednym, konkretnym rozkładem, w drugim z rodziną rozkładów. Zajmiemy się najpierw pierwszym przypadkiem.

### 6.2.2. Testowanie prostej hipotezy o zgodności

Niech zmienna jakościowa $X$ ma $k$ możliwych wartości (inaczej poziomów lub kategorii), $x_1, x_2, \ldots, x_k$ i niech prawdopodobieństwo wystąpienia wartości $x_i$ wynosi $p_i$, $i = 1, 2, \ldots, k$, $P(X = x_i) = p_i$. Zakładamy, że wartości $p_i$ są nieznane. Niech ponadto będzie dany pewien ustalony rozkład prawdopodobieństwa $\{p_1^0, p_2^0, \ldots, p_k^0\}$ (innymi słowy, jest dany zbiór $k$ takich liczb nieujemnych, że $\sum_{i=1}^{p} p_i^0 = 1$). Będziemy zakładać, że wszystkie wartości $p_i^0$, $i = 1, \ldots, k$, są dodatnie (kategorie o zerowym prawdopodobieństwie wystąpienia możemy pominąć).

Rozważmy problem testowania hipotezy o zgodności rozkładu $\{p_1, p_2, \ldots, p_k\}$ z zadanym rozkładem $\{p_1^0, p_2^0, \ldots, p_k^0\}$. Hipoteza zerowa ma zatem postać:

$$H_0\colon\ p_i = p_i^0, \tag{6.1}$$

$i = 1, 2, \ldots, k$. Hipoteza alternatywna ma postać:

$$H_1\colon \text{ hipoteza } H_0 \text{ jest fałszywa}, \tag{6.2}$$

czyli rozkład $\{p_1, p_2, \ldots, p_k\}$ jest różny od zadanego rozkładu $\{p_1^0, p_2^0, \ldots, p_k^0\}$.

Trudno znaleźć podręcznik ze statystyki, w którego części poświęconej testowaniu zgodności dla danych jakościowych nie byłoby odwołania do rewolucyjnych badań zakonnika Gregora Mendla z dziedziny genetyki. Także i my zaczniemy od takiego przykładu. Ojciec Mendel przewidywał na przykład, że populacja potomków jednostek o genotypie Aa (czyli genotypie zawierającym gen A oraz gen a) będzie się charakteryzować genotypami AA, Aa oraz aa i jednostki o tych genotypach będą występować w proporcji 1:2:1 (czyli 1/4 potomków będzie miała genotyp AA, połowa Aa i 1/4 aa). Wynik taki, zdaniem o. Mendla, miał być efektem losowości zjawiska: genotyp dziecka powstaje przez losowe połączenie się w jeden genotyp po jednym genie z dwóch genotypów rodziców. Zdarza się przy tym, że jednostki o genotypie Aa mają taki sam wygląd jak np. jednostki o genotypie AA (mówimy wówczas, że gen A jest dominujący). Tak jest w przypadku grochu, na którym o. Mendel głównie przeprowadzał swoje doświadczenia, z genem A będącym genem grochu żółtego oraz genem a oznaczającym gen grochu zielonego – skutkiem dominacji genu A, groch o genoptypie Aa jest koloru żółtego.

**Przykład 6.3.** W jednym ze swoich doświadczeń o. Mendel wyhodował populację potomków grochu o genotypie Aa. W efekcie uzyskał 8023 nasiona potomków grochu Aa. Zgodnie z jego teorią, po wysianiu nasion powinien był oczekiwać otrzymania $p_1^0 = 3/4$ roślin grochu

żółtego (AA lub Aa) oraz $p_2^0 = 1/4$ roślin grochu zielonego (aa). Innymi słowy, powinien był oczekiwać, że wyrośnie mu około $(3/4)8023 = 6017{,}25$ roślin grochu żółtego (o ile mogłoby wyrosnąć coś takiego jak 0,25 rośliny!) oraz około $(1/4)8023 = 2005{,}75$ roślin grochu zielonego. W rzeczywistości jego pomocnicy naliczyli 6022 rośliny pierwszego rodzaju i 2001 roślin drugiego rodzaju. Czy otrzymany wynik może być podstawą do odrzucenia hipotezy Mendla, czy raczej ją potwierdza?

Oczywiście powyższe pytanie jest w istocie pytaniem o zgodność z założonym rozkładem prawdopodobieństwa. Według Mendla, czyli zgodnie z hipotezą zerową (6.1), prawdopodobieństwo otrzymania potomka żółtego (AA lub Aa) wynosi $p_1 = p_1^0 = 3/4$, natomiast prawdopodobieństwo otrzymania potomka zielonego (aa), $p_2 = p_2^0 = 1/4$. Naturalnym sposobem przeprowadzenia testu hipotezy (6.1) przy ogólnej alternatywie (6.2) jest oparcie się na porównaniu liczności danych każdej kategorii z licznościami oczekiwanymi przy spełnieniu hipotezy $H_0$. Przy tym porównanie nie powinno być czułe na znak odchylenia wartości zaobserwowanej od oczekiwanej, a jedynie na jego wielkość (jest nam wszystko jedno czy dany wynik jest zbyt duży czy zbyt mały w stosunku do wartości oczekiwanej). Ponadto takie porównanie powinno być jakoś znormalizowane (jeśli oczekiwana dla danej kategorii liczność wynosi 10000, zaś w wyniku doświadczenia otrzymamy 9996 wyników tej kategorii, to będzie to wynik bliski oczekiwanemu; jeśli oczekiwana liczność wynosi 5, a otrzymamy liczność 1, to będzie to wynik daleki od oczekiwań). Okazuje się, że w przypadku zachodzenia hipotezy zerowej statystyka

$$Q = \sum_{\substack{\text{wszystkie}\\ \text{kategorie}}} \frac{(\text{liczność zaobserwana} - \text{liczność oczekiwana})^2}{\text{liczność oczekiwana}} \tag{6.3}$$

ma w przybliżeniu rozkład $\chi^2$ z liczbą stopni swobody równą liczbie możliwych kategorii danych pomniejszoną o 1. Mówiąc ogólnie, duże wartości statystyki testowej $Q$ świadczą przeciw prawdziwości hipotezy zerowej.

Bardziej formalnie, niech dane mogą należeć do $k$ różnych kategorii $x_i$, $i = 1, 2, \ldots, k$. Załóżmy, że dysponujemy próbą losową $n$ danych (obserwacji) i że kategorię $x_1$ zaobserwowaliśmy $n_1$ razy, kategorię $x_2$ zaobserwowaliśmy $n_2$ razy, ..., kategorię $x_k$ zaobserwowaliśmy $n_k$ razy, skąd $n_1 + n_2 + \ldots + n_k = n$.

Jeżeli jest prawdziwa hipoteza (6.1), to statystyka

$$Q = \sum_{i=1}^{k} \frac{(n_i - np_i^0)^2}{np_i^0} \tag{6.4}$$

ma w przybliżeniu rozkład $\chi^2$ z $k-1$ stopniami swobody (rozkład prawdopodobieństwa jest określony przez $k$ parametrów związanych jednym równaniem, orzekającym, iż suma prawdopodobieństw musi być równa 1, i stąd liczba stopni swobody jest równa $k-1$).

Przybliżenie rozkładem $\chi^2$ uznajemy zwykle za dopuszczalne, gdy $np_i^0 \geq$ $\geq 5$ dla każdego $i = 1, 2, \ldots, k$ i za dobre, gdy $np_i^0 \geq 10$ dla każdego $i =$ $= 1, 2, \ldots, k$. Jeżeli liczba kategorii jest duża (np. większa niż 6), zwykle zgadzamy się stosować przybliżenie rozkładem $\chi^2$ także wtedy, gdy dla jednej lub dwóch kategorii $1 \leq np_i^0 < 5$. Test zgodności oparty na statystyce $Q$ nazywamy testem zgodności $\chi^2$ lub testem $\chi^2$ Pearsona od nazwiska jego wynalazcy, Karla Pearsona.

**Przykład 6.3 cd.** W naszym przypadku

$$Q = \frac{(6022 - 6017{,}25)^2}{6017{,}25} + \frac{(2001 - 2005{,}75)^2}{2005{,}75} = 0{,}015.$$

Wartości $np_i^0$, $i = 1, 2$, są bardzo duże i możemy oczywiście uznać, że statystyka $Q$ ma praktycznie dokładnie rozkład $\chi^2$ z 1 stopniem swobody. Przypomnijmy w tym miejscu, że zmienna losowa o rozkładzie $\chi^2$ z 1 stopniem swobody jest kwadratem standardowej zmiennej losowej normalnej. W konsekwencji, obliczenie $p$–wartości, odpowiadającej otrzymanej wartości statystyki $Q$, jest bardzo łatwe i nie wymaga nawet odwołania się do rozkładu $\chi^2$. Wystarczy zauważyć, że $\sqrt{0{,}015} = 0{,}1225$ i stąd, że szukana $p$–wartość jest równa prawdopodobieństwu

$$1 - P(-0{,}1225 < Y < 0{,}1225),$$

gdzie zmienna losowa $Y$ ma standardowy rozkład normalny. Ostatecznie otrzymujemy $p$–wartość równą w przybliżeniu 0,9.

Otrzymana $p$–wartość jest bardzo duża i oczywiście nie odrzucamy hipotezy Mendla.

Na pewno każdy Czytelnik zetknął się już z sytuacją, w której uzyskał podobnie mocne uzasadnienie testowanej przez niego hipotezy zerowej. Otrzymanie $p$–wartości rzędu 0,9 wskazuje na to, że prawdopodobieństwo otrzymania wartości statystyki $Q$ odległej od zera o więcej niż 0,015 jest bliskie 1. Uzasadnienie hipotezy

Mendla jest więc rzeczywiście bardzo przekonujące. Ale z doświadczeniami zakonnika rzecz się miała jeszcze ciekawiej niżby to wynikało z opisanego jednego doświadczenia. Sir Ronald Fisher zauważył, że doświadczenia Mendla praktycznie *zawsze* dawały wyniki doskonale potwierdzające jego hipotezy. Fisher zebrał wyniki Mendla i poddał je łącznej analizie (przypomnijmy, że suma niezależnych statystyk $\chi^2$ ma także rozkład $\chi^2$, tyle, że z odpowiednio większą liczbą stopni swobody, i przeto można skonstruować test łączny, będący niejako sumą testów pojedynczych). Fisher otrzymał dla testu łącznego $p$–wartość równą aż 0,99996! Albo więc o. Mendel miał szczęście zupełnie nieprawdopodobne, albo raczej jego pomocnicy byli świadomi celu badań i wymyślili dość niecny sposób przypodobania się pryncypałowi. Co nie przeczy temu, że Mendel był genialnym ojcem genetyki, choć najpewniej dopiero późniejsi badacze empirycznie potwierdzili słuszność tez zakonnika.

W przykładzie 6.3 jest opisana najprostsza możliwa zmienna jakościowa, bo przyjmująca zaledwie dwie wartości. Zauważmy, że przy założeniu prawdziwości hipotezy $H_0$ zaobserwowany wynik, dotyczący np. grochu zielonego, jest liczbą sukcesów w 8023 doświadczeniach Bernoulliego z prawdopodobieństwem pojedynczego sukcesu równym 1/4. Jest to zatem zaobserwowana wartość zmiennej losowej o rozkładzie dwumianowym z parametrami $n = 8023$ i $p = 0{,}25$. Liczba roślin grochu żółtego to w tym języku po prostu liczba porażek, równa różnicy między liczbą doświadczeń Bernoulliego oraz liczbą odniesionych sukcesów. Innymi słowy, w przykł. 6.1 testowaliśmy hipotezę o tym, że proporcja (sukcesów) jest równa zadanej liczbie, w naszym przypadku $p = p_0 = 0{,}25$. Okazuje się, że gdy $k = 2$, nasz test jest równoważny testowi dla proporcji z hipotezą alternatywną $p \neq p_0$ oraz statystyką (3.67) jako statystyką testową; por. zad. 6.1.

**Przykład 6.4.** W doświadczeniu z przykł. 1.2 wykonano 100 rzutów kostką do gry i otrzymano następujące liczebności wyników:

| $x_i$ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $n_i$ | 16 | 19 | 9 | 17 | 25 | 14 |

Czy istnieją podstawy do odrzucenia hipotezy, że rzuty były wykonywane uczciwą kostką, czyli do odrzucenia hipotezy o jednostajności rozkładu, $p_1^0 = p_2^0 = \cdots = p_6^0 = 1/6$?

Przy hipotezie zerowej oczekiwane liczności są takie same dla wszystkich kategorii i wynoszą 16,66. Otrzymujemy zatem

$$\begin{aligned} Q &= \frac{1}{16{,}66}[(16-16{,}66)^2 + (19-16{,}66)^2 + (9-16{,}66)^2 + \\ &\quad +(17-16{,}66)^2 + (25-16{,}66)^2 + (14-16{,}66)^2] = 8{,}48. \end{aligned}$$

Zgodnie z wcześniej podaną regułą możemy przyjąć, że statystyka $Q$ ma w przybliżeniu rozkład $\chi^2$ z 5 stopniami swobody. Ostatecznie otrzymujemy $p$–wartość równą 0,12, który to wynik nie powinien prowadzić do odrzucenia hipotezy zerowej.

Omawianie testowania zgodności przy prostej hipotezie zerowej zakończymy dwiema uwagami. Po pierwsze, jak przestrzegaliśmy już w rozdz. 3 i obecnie tylko krótko o tym wspomnimy, nie należy mylić istotności statystycznej z istotnością praktyczną. Im większa jest liczność próby, tym większa jest także czułość testu, czyli jego skłonność do odrzucania hipotezy zerowej, gdy prawdziwy rozkład jest tylko nieznacznie (nieistotnie z praktycznego punktu widzenia) różny od rozkładu postulowanego przez tę hipotezę.

Wzmianka o drugiej kwestii musi być obszerniejsza. Chodzi o stosowanie testu $\chi^2$ Pearsona do testowania zgodności z zadanym rozkładem ciągłym. Idea takiego testu jest prosta i sprowadza się do właściwej dyskretyzacji owego rozkładu.

Załóżmy, że mamy $n$ elementową próbę losową z rozkładu ciągłego zadanego ustaloną gęstością $f_0(x)$. Załóżmy dalej, że gęstość $f_0(x)$ jest dodatnia na przedziale $(a_0, a_k)$, przy czym nie wykluczamy, że $a_0 = -\infty$ albo $a_k = \infty$ (tzn. możliwe jest też, że $(a_0, a_k) = (-\infty, \infty)$). Przedział $(a_0, a_k)$ możemy podzielić na $k$ podprzedziałów, $(a_0, a_1)$, $[a_1, a_2)$, $\ldots$, $[a_{k-1}, a_k)$. Prawdopodobieństwo, że zmienna losowa $X$ o rozkładzie danym gęstością $f_0(x)$ przyjmie wartość należącą do przedziału $[a_{i-1}, a_i)$ wynosi

$$p_i^0 = P(a_{i-1} \leq X < a_i) = \int_{a_{i-1}}^{a_i} f_0(x)dx.$$

Taka procedura podzielenia przedziału $(a_0, a_k)$ umożliwia dyskretyzację zarówno zmiennej losowej $X$, jak i jej rozkładu. Mianowicie, liczbowe wartości tej zmiennej zastępujemy przedziałami – od tego momentu nie interesuje nas zaobserwowana wartość liczbowa zmiennej losowej $X$, lecz to tylko, do którego przedziału ta wartość należy; jeżeli zaobserwowaliśmy wartość $x \in [a_{i-1}, a_i)$, to mówimy, że zmienna losowa przyjęła wartość z przedziału $[a_{i-1}, a_i)$. Zarazem znamy prawdopobieństwo przyjęcia przez zmienną losową (dowolnej) wartości z tego przedziału, równe $p_i^0$. W wyniku przeprowadzonej dyskretyzacji oryginalna informacja o próbie losowej zostaje zredukowana do zapamiętania liczności obserwacji, które znalazły się w kolejnych $k$ przedziałach podziału.

W przykładzie 1.6 mamy 6 przedziałów równej długości, przy czym prawdopodobieństwo znalezienia się obserwacji w zadanym przedziale wynosi $p_i^0 = 1/6$; zebrana tam próba losowa 100 obserwacji jest opisana następującą tablicą liczności:

| | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ | $x_6$ |
|---|---|---|---|---|---|---|
| $n_i$ | 19 | 15 | 18 | 21 | 15 | 12 |

gdzie $x_i$ oznacza $i$-ty przedział, $i = 1, 2, \ldots, 6$, i $n_i$ jest liczbą obserwacji z próby, które znalazły się w przedziale $x_i$.

Rozkład, z którego wylosowano dane do przykł. 1.6 był znany – był to rozkład jednostajny na przedziale $[0, 1]$. Niemniej jednak, można by oczywiście sprawdzić, czy zastosowanie testu zgodności Pearsona (6.4) nie daje podstaw do odrzucenia takiej hiptezy $H_0$; por. zad. 6.2.

W sytuacji, gdy rozkład ciągły jest nieznany i chcemy testować jego zgodność z zadanym rozkładem, wykorzystując test $\chi^2$ Pearsona, możemy oprzeć się na opisanej dyskretyzacji. Otrzymamy w ten sposób liczności $n_i$ oraz rozkład dyskretny $p_1^0, p_2^0, \ldots, p_k^0$, odpowiadający rozkładowi ciągłemu (właściwie przybliżający rozkład ciągły), z którym zgodność chcemy testować. Aby dokonać dyskretyzacji, trzeba tylko ustalić sposób określenia brzegów przedziałów podziału $(a_0, a_1)$, $[a_1, a_2)$, $\ldots$, $[a_{k-1}, a_k)$. Obowiązuje tu następujące ogólne zalecenie: *przedziały podziału powinny być tak dobrane, aby otrzymane w konsekwencji jego dokonania prawdopodobieństwa $p_1^0, p_2^0, \ldots, p_k^0$ były sobie przynajmniej w przybliżeniu równe oraz, aby był spełniony warunek $np_i^0 \geq 5$, $i = 1, 2, \ldots, k$.*

Łatwo zauważyć, że testowanie zgodności z zadanym rozkładem ciągłym za pomocą testu $\chi^2$ jest przedsięwzięciem kontrowersyjnym, ponieważ punktem wyjścia do konstrukcji testu jest świadoma utrata informacji związana z koniecznością dokonania dyskretyzacji. Dlatego, gdy mamy do czynienia z rozkładem ciągłym, powinniśmy unikać stosowania tego testu i raczej stosować testy opisane w podrozdz. 3.4. Dopiero, gdy próba losowa jest bardzo liczna i histogram sporządzony na jej podstawie przypomina gładki rozkład ciągły, zastosowanie testu $\chi^2$ przestaje być ryzykowne. Inna sprawa, że test ten może być jedynym dającym się zastosować w danej konkretnej sytuacji. Tak jest np. wtedy, gdy dane, którymi dysponujemy, pochodzą wprawdzie z rozkładu ciągłego, ale są już zdyskretyzowane.

### 6.2.3. Testowanie złożonej hipotezy o zgodności

Niekiedy postulowany rozkład prawdopodobieństwa $p_1^0, p_2^0, \ldots, p_k^0$ zależy od nieznanych parametrów. W ogólności przyjmuje się, że rozkład ten może co najwyżej zależeć od $k-2$ nieznanych parametrów. W dalszym ciągu ograniczymy się do przypadku jednego tylko nieznanego parametru, który będziemy oznaczać symbolem $\theta$. Większa liczba nieznanych parametrów nie zmienia istoty problemu i dlatego nie będziemy się zajmować tym bardziej ogólnym przypadkiem.

Hipoteza $H_0$ przyjmuje teraz postać:

$$H_0\colon\ p_i = p_i^0(\theta), \tag{6.5}$$

$i = 1, 2, \ldots, k$, gdzie $\theta$ jest nieznanym parametrem. Hipoteza alternatywna (6.2) pozostaje niezmieniona. Obecnie nie tylko hipoteza alternatywna, ale także hipoteza zerowa jest hipotezą złożoną (dla każdej wartości parametru $\theta$ rozkład $\{p_1^0(\theta), p_2^0(\theta), \ldots, p_k^0(\theta)\}$ może być inny). W zasadzie nie interesuje nas jaka jest wartość parametru $\theta$, chodzi natomiast o weryfikację hipotezy o zgodności rozkładu z rodziną rozkładów określoną za pomocą tego parametru.

**Przykład 6.5.** Można wykazać, że jeżeli teoria Mendla losowego tworzenia się genotypów potomstwa jest słuszna i jeżeli w populacji występują genotypy AA, Aa oraz aa, zaś gen A stanowi ułamek $\theta$ wszystkich genów (czyli gen a stanowi ułamek $1-\theta$), to populacja o proporcji genotypów AA, Aa i aa równej $\theta^2{:}2\theta(1-\theta){:}(1-\theta)^2$ pozostaje w stanie równowagi. Innymi słowy, generacja potomków stanowi populację o tej samej proporcji genotypów.

Załóżmy, że eksperymentalnie stwierdzono, że populacja pewnych 500 jednostek składa się ze 110 jednostek o genotypie AA, 235 jednostek Aa oraz 155 jednostek aa. Czy istnieje podstawa do odrzucenia hipotezy Mendla?

Hipoteza Mendla pełni funkcję hipotezy zerowej, czyli ma postać

$$H_0\colon\ p_1^0(\theta) = \theta^2,\ \ p_2^0(\theta) = 2\theta(1-\theta),\ \ p_3^0(\theta) = (1-\theta)^2, \tag{6.6}$$

gdzie $p_i^0$, $i = 1, 2, 3$, jest prawdopodobieństwem wystąpienia w populacji genotypu, odpowiednio, AA, Aa i aa. Hipoteza alternatywna ma postać (6.2). Zgodnie z hipotezą (6.6), jakakolwiek by była (niestotna w tym miejscu) wartość parametru $\theta$, prawdopodobieństwa $p_1^0(\theta)$, $p_2^0(\theta)$ i $p_3^0(\theta)$ są związane podaną tam zależnością od tego parametru.

Metoda postępowania w przypadku złożonej hipotezy zerowej polega na zastąpieniu w definicji statystyki (6.4) nieznanych wartości prawdopodobieństw $p_1^0$, $p_2^0$ oraz $p_k^0$ (w naszym przykładzie $p_1^0$, $p_2^0$ i $p_3^0$) ich estymatorami. Mianowicie, w miejsce nieznanej wartości parametru $\theta$ wstawia się jego estymator największej wiarogodności, $\hat{\theta}$, i w ten sposób otrzymuje się estymatory (także największej wiarogodności) nieznanych prawdopodobieństw. Okazuje się, że w przypadku prawdziwości hipotezy (6.6) zmodyfikowana statystyka $Q$ ma w przybliżeniu rozkład $\chi^2$ z $k-2$ stopniami swobody (w przykł. 6.5

rozkład $\chi^2$ z 1 stopniem swobody). W stosunku do przypadku testowania hipotezy prostej, liczba stopni swobody jest mniejsza o 1, czyli o liczbę estymowanych parametrów. Przybliżenie rozkładem $\chi^2$ uznajemy za dopuszczalne, gdy $np_i^0(\hat{\theta}) \geq 5$ dla każdego $i = 1, 2, \ldots, k$.

**Przykład 6.5 cd.** Wylosowana próba przyniosła następujące liczności jednostek o różnych genotypach:

| Typ | AA | Aa | aa |
|---|---|---|---|
| $n_i$ | 110 | 235 | 155 |

Test hipotezy Mendla (6.6) wymaga obliczenia estymatora NW parametru $\theta$. Funkcja wiarogodności przyjmuje w naszym przykładzie postać

$$(\theta^2)^{n_1}[2\theta(1-\theta)]^{n_2}[(1-\theta)^2]^{n_3}.$$

Po zlogarytmowaniu funkcji wiarogodności łatwo jest obliczyć, że estymator NW parametru $\theta$ wynosi

$$\hat{\theta} = \frac{2n_1 + n_2}{2n},$$

gdzie, jak zwykle, $n = n_1 + n_2 + n_3$. Estymator ten jest oczywiście równy liczbie genów A zaobserwowanych w próbie $2n$ genów. W naszym przypadku

$$\hat{\theta} = 0{,}455.$$

W tej sytuacji

$$Q = \frac{(110 - 103{,}51)^2}{103{,}51} + \frac{(235 - 247{,}97)^2}{247{,}97} + \frac{(155 - 148{,}51)^2}{148{,}51} = 1{,}369.$$

Ostatecznie $p$–wartość wynosi 0,24, czyli nie znajdujemy podstaw do odrzucenia hipotezy (6.6).

Problem z $m$, $m > 1$, nieznanymi parametrami rozwiązujemy analogicznie. Najpierw obliczamy estymatory NW tych parametrów, następnie wstawiamy otrzymane oszacowania do wzorów na prawdopodobieństwa występujące w hipotezie zerowej i w końcu stosujemy statystykę (6.4). Różnica polega na tym, że przy zachodzeniu $H_0$ rozkład statystyki testowej jest w przybliżeniu rozkładem $\chi^2$ z liczbą stopni swobody równą $k - 1 - m$.

Jeżeli rozkład postulowany w hipotezie zerowej jest ciągły i zależny od $m$ nieznanych parametrów, i jeżeli jesteśmy zmuszeni zastosować test $\chi^2$, rozpoczynamy najczęściej od wyznaczenia estymatorów NW wspomnianych parametrów, następnie dokonujemy stosownej dyskretyzacji rozkładu oraz danych, i dalej postępujemy tak, jakby chodziło o test prostej hipotezy o zgodności dla zmiennych jakościowych. Jakkolwiek rozkład statystyki testowej nie jest w tym przypadku znany, to wiadomo, że przy zachodzeniu hipotezy zerowej $p$–wartość odpowiadająca otrzymanemu wynikowi zawiera się w przybliżeniu między $p$–wartością obliczoną dla rozkładu $\chi^2$ z $k-1-m$ stopniami swobody a $p$–wartością obliczoną dla rozkładu $\chi^2$ z $k-1$ stopniami swobody.

Nie trzeba dodawać, że stosowanie testu $\chi^2$, gdy hipoteza zerowa ma postać (6.5) wymaga zwiększonej ostrożności. Rzecz w tym, że już w przypadku hipotezy prostej statystyka testowa ma rozkład $\chi^2$ tylko w przybliżeniu. Gdy hipoteza ma postać (6.5), problem się komplikuje ze względu na konieczność estymacji nieznanego parametru. Zależność hipotezy zerowej od więcej niż jednego parametru przynosi kolejną komplikację zadania i opieranie się na teście $\chi^2$ wymaga jeszcze większej ostrożności. Oczywiście najbardziej krytyczni musimy być, starając się zastosować ten test, gdy rozkład postulowany w złożonej hipotezie zerowej jest ciągły.

## 6.3. Testowanie jednorodności

Załóżmy, że interesuje nas jednorodność $k$ rozkładów w sytuacji, gdy obserwacje każdej z populacji mogą należeć do $l$ kategorii. Niech ciąg $\{p_{i1}, p_{i2}, \ldots, p_{il}\}$ oznacza (nieznany) rozkład prawdopodobieństwa $i$-tej populacji, $i = 1, 2, \ldots, k$. Zauważmy, że postać tych rozkładów nas w zasadzie nie interesuje, chcemy natomiast zweryfikować prawdziwość hipotezy o równości $k$ rozkładów

$$H_0\colon\ p_{1j} = p_{2j} = \cdots = p_{kj} \text{ dla każdego } \ j = 1, 2, \ldots, l, \tag{6.7}$$

przy hipotezie alternatywnej

$$H_1\colon \text{ hipoteza } H_0 \text{ jest fałszywa.} \tag{6.8}$$

Aby móc rozwiązać postawione zadanie, musimy pobrać próbki losowe ze wszystkich populacji; z $i$-tej populacji pobieramy próbkę o liczności $n_{i\cdot}$, $i = 1, 2, \ldots, k$, przy czym $\sum_i n_{i\cdot} = n$. W wyniku dla $i$-tej próbki otrzymujemy

$n_{ij}$ obserwacji należących do $j$-tej kategorii, $j = 1, 2, \ldots, l$, gdzie oczywiście

$$\sum_{j=1}^{l} n_{ij} = n_{i\cdot}.$$

Przy spełnieniu hipotezy (6.7) wszystkie populacje mają ten sam rozkład, który możemy oznaczyć $\{p_1, p_2, \ldots, p_l\}$,

$$p_{1j} = p_{2j} = \cdots = p_{kj} = p_j$$

dla każdego $j = 1, 2, \ldots, l$. Naturalnym estymatorem (w rzeczywistości estymatorem NW) prawdopodobieństwa $p_j$ jest ułamek

$$\frac{n_{\cdot j}}{n},$$

gdzie

$$n_{\cdot j} = \sum_{i=1}^{k} n_{ij}.$$

Rzeczywiście, przy założeniu prawdziwości (6.7), wszystkie obserwacje pochodzą z tego samego rozkładu, możemy zatem mówić o próbie losowej z tego rozkładu o liczności $n$ i $n_{\cdot j}$ obserwacjach o $j$-tej kategorii[1]. Ale jeśli tak, to iloczyn

$$n_{i\cdot}\frac{n_{\cdot j}}{n}$$

jest estymatorem oczekiwanej liczby obserwacji w $i$-tym wierszu i $j$-tej kolumnie tablicy kontyngencji, czyli jest estymatorem wielkości $E(n_{ij})$.

Przedstawione rozumowanie w oczywisty sposób nawiązuje do statystyki (6.3), której powinniśmy obecnie nadać postać

$$Q = \sum_{i=1}^{k}\sum_{j=1}^{l} \frac{(n_{ij} - n_{i\cdot}n_{\cdot j}/n)^2}{n_{i\cdot}n_{\cdot j}/n}. \tag{6.9}$$

Można udowodnić, że wraz ze wzrostem wszystkich liczności $n_{i\cdot}$, $i = 1, 2, \ldots, k$, i przy zachodzeniu hipotezy zerowej (6.7), rozkład statystyki $Q$ zbliża się do rozkładu $\chi^2$ z $(k-1)(l-1)$ stopniami swobody. Podana liczba stopni swobody bierze się stąd, iż mamy $kl$ liczności $n_{ij}$, sumy w wierszach są związane $k$ równościami, sumy w kolumnach $l$ równościami, przy czym niezależnych równości jest $k + l - 1$, ponieważ $\sum n_{i\cdot} = \sum n_{\cdot j} = n$. Ostatecznie

[1] W tym miejscu warto jeszcze raz wyraźnie podkreślić, że w rozważanym zadaniu tylko sumaryczna liczba wszystkich obserwacji w kolumnie macierzy kontyngencji jest losowa – sumaryczna liczba obserwacji w wierszu jest z góry ustalona.

zatem otrzymujemy $kl - k - l + 1 = (k-1)(l-1)$ stopni swobody. Duże wartości statystyki $Q$ świadczą przeciw hipotezie $H_0$.

**Przykład 6.2 cd.** Podstawiając dane z tablicy kontyngencji do wzoru (6.9) otrzymujemy $Q = 55{,}06$ i stąd $p$–wartość mniejszą niż 0,0001. Hipotezę o jednorodności rozkładów należy odrzucić na korzyść hipotezy alternatywnej. Jest to konkluzja, która wymaga analizy przyczyn jej otrzymania. Na przykład niejednorodne mogą być populacje studentów przydzielonych prowadzącym ćwiczenia. Osoby prowadzące mogą stosować różne sposoby oceniania studentów, jedni asystenci mogą być zbyt surowi, inni zbyt łagodni. Ćwiczenia mogą być różnej jakości. W każdym razie dobro studentów wymaga wykrycia przyczyny niejednorodności i właściwej na nią reakcji.

W przykładzie 6.2 mogliśmy zastosować test $\chi^2$, ponieważ liczności $n_{i\cdot}$ były tam stosunkowo duże. Jeżeli problem jest opisywany tablicą kontyngencji o wymiarze $2 \times 2$ i liczności prób w wierszach (kolumnach) są zbyt małe, by móc polegać na teście Pearsona, można oprzeć się na tzw. dokładnym teście Fishera, którego omówienie pomijamy. Niestety konstrukcja Fishera jest zbyt skomplikowana, aby nadawała się do testowania tablic o większym wymiarze niż $2 \times 2$.

## 6.4. Analiza dwóch zmiennych losowych

### 6.4.1. Testowanie niezależności

Tym razem mamy do czynienia z jedną populacją opisaną przez parę jakościowych zmiennych losowych. Dysponujemy $n$-elementową próbą losową i każda obserwacja musi należeć do jednej z $kl$ możliwych kombinacji kategorii pierwszej i drugiej zmiennej losowej. Otrzymywaną w wyniku zebrania próby informację przedstawia się za pomocą tablicy kontyngencji wprowadzonej w podrozdz. 1.2, np. takiej jak w przykł. 6.1. Para zmiennych losowych jest scharakteryzowana pewnym, nieznanym i nas tu nie interesującym, łącznym rozkładem prawdopodobieństwa. Oznaczmy symbolem $p_{ij}$, $i = 1, 2, \ldots, k$, $j = 1, 2, \ldots, l$, prawdopodobieństwo zaobserwowania w jednym doświadczeniu $i$-tej kategorii zmiennej losowej $X$ oraz $j$-tej kategorii zmiennej losowej $Y$. Przyjmijmy też, że $p_{i\cdot}$ i $p_{\cdot j}$ oznaczają, odpowiednio, (również nieznane) rozkłady brzegowe zmiennej $X$ i zmiennej $Y$. Interesuje nas czy zmienne losowe $X$ i $Y$ są niezależne.

Zdobyta dotąd wiedza o analizie zmiennych jakościowych pozwala od razu

zaproponować test niezależności, czyli test hipotezy zerowej

$$H_0\colon\ p_{ij} = p_{i\cdot}p_{\cdot j} \tag{6.10}$$

dla wszystkich $i = 1, 2, \ldots, k$, $j = 1, 2, \ldots, l$, przy hipotezie alternatywnej

$$H_1\colon \text{ hipoteza } H_0 \text{ jest fałszywa.}$$

Nasuwa się mianowicie idea oparcia testu na statystyce typu statystyki (6.3). Przy zachodzeniu hipotezy $H_0$ oczekiwana liczba obserwacji o kombinacji kategorii $(x_i, y_j)$ wynosi

$$np_{i\cdot}p_{\cdot j}.$$

Estymatorami NW wymienionych prawdopodobieństw brzegowych są odpowiednio

$$n_{i\cdot}/n \quad \text{oraz} \quad n_{\cdot j}/n,$$

gdzie, jak poprzednio, $n_{i\cdot} = \sum_j n_{ij}$ oraz $n_{\cdot j} = \sum_j n_{ij}$. Zatem, estymatorem podanej wartości oczekiwanej jest

$$n\frac{n_{i\cdot}}{n}\frac{n_{\cdot j}}{n}.$$

Stąd, statystyka testowa powinna mieć postać:

$$Q = \sum_{i=1}^{k}\sum_{j=1}^{l}\frac{(n_{ij} - n_{i\cdot}n_{\cdot j}/n)^2}{n_{i\cdot}n_{\cdot j}/n}. \tag{6.11}$$

Formalnie jest to statystyka tej samej postaci co statystyka (6.9), ale jej znaczenie jest nieco inne. W szczególności, statystyka (6.11) została wyprowadzona z użyciem estymatora NW prawdopodobieństwa brzegowego $p_{i\cdot}$. W rozważanym wcześniej problemie jednorodności populacji pojęcie prawdopodobieństwa brzegowego $p_{i\cdot}$ nie ma sensu. Oczywiście, w obydwu przypadkach ma jednak sens mówienie o licznościach $n_{i\cdot}$, choć w teście jednorodności są to wartości ustalone, a w teście niezależności wielkości losowe. I w rezultacie używane w obydwu przypadkach statystyki oraz procedury testowe są praktycznie takie same.

Można wykazać, że jeżeli jest prawdziwa hipoteza o niezależności (6.10), to statystyka testowa (6.11) ma w przybliżeniu rozkład $\chi^2$ z $(k-1)(l-1)$ stopniami swobody. Przybliżenie jest tym dokładniejsze, im większa jest liczność próby $n$. Liczba stopni swobody jest taka sama jak w przypadku testu jednorodności, ale jej obliczenie jest inne. Tym razem, z jednej strony estymujemy $k+l$ prawdopodobieństw brzegowych, które są związane dwiema równościami, z drugiej natomiast możemy swobodnie wybrać $kl-1$ wyrazów tablicy kontyngencji (liczba $kl$ wyrazów tablicy musi być zmniejszona o 1,

ponieważ suma liczności jest ustalona, równa $n$). Estymacja $k+l$ prawdopodobieństw związanych dwiema równościami nakłada na liczności w tablicy kontyngencji $k+l-2$ niezależnych więzów. Ostatecznie zatem otrzymujemy $kl-1-(k+l-2)=(k-1)(l-1)$ stopni swobody.

**Przykład 6.6.** W późnych latach sześćdziesiątych pewną popularność wśród amerykańskiej młodzieży szkolnej i akademickiej zyskało sobie palenie marihuany. Psychologowie badali m.in., czy stosunek młodzieży do palenia tej używki jest zależny od poglądów politycznych. W jednym z eksperymentów każdy młody człowiek z losowej próbki 1349 uczniów i studentów został zapytany czy pali marihuanę (często, okazjonalnie lub nigdy) oraz jakie są jego/jej poglądy polityczne (postępowe, konserwatywne lub inne). Wyniki są podane w tablicy kontyngencji (dane cytujemy za: Devore J.L. (1982): *Probability and statistics for engineering and the sciences.* Brooks/Cole, Monterey):

| | Nigdy | Okazjonalnie | Często |
|---|---|---|---|
| Postępowe | 479 | 173 | 119 |
| Konserwatywne | 214 | 47 | 15 |
| Inne | 172 | 45 | 85 |

Hipotezę o niezależności skłonności do palenia marihuany od poglądów politycznych trzeba odrzucić – $Q = 64,65$ i dla rozkładu $\chi^2$ z 4 stopniami swobody daje to $p$–wartość mniejszą niż 0,0001.

Liczności w tablicy kontyngencji z przykł. 6.6 są dostatecznie duże, by można było zastosować przybliżenie rozkładu statystyki (6.11) rozkładem $\chi^2$. Tak jak w zadaniu testowania jednorodności, jeżeli nie można zastosować testu $\chi^2$, ale szczęśliwie tablica kontyngencji ma wymiar zaledwie $2\times 2$, pozostaje odwołanie się do dokładnego testu Fishera.

### 6.4.2. Analiza zależności

Testowanie niezależności przynosi najwięcej pożytku, gdy można z dużym przekonaniem orzec, iż nie ma podstaw do odrzucenia hipotezy zerowej, zwłaszcza gdy przynajmniej jedna ze zmiennych jest nominalna albo też, gdy hipoteza zerowa okazuje się bardzo mało prawdopodobna. Trzeba zarazem pamiętać, że w przypadku pary cech o uporządkowanych kategoriach test niezależności może okazać się zwodniczy.

Załóżmy, że dysponujemy próbą losową par obserwacji i zmienne mają po cztery kategorie uporządkowane, odpowiednio, 1, 2, 3, 4 oraz A, B, C, D.

Zebrane dane można zestawić w następującej tablicy kontyngencji:

| | A | B | C | D |
|---|---|---|---|---|
| 1 | 9 | 11 | 26 | 14 |
| 2 | 11 | 14 | 28 | 33 |
| 3 | 12 | 14 | 33 | 38 |
| 4 | 3 | 17 | 31 | 39 |

(6.12)

Test $\chi^2$ niezależności daje $p$–wartość 0,22 i, jeśli bezkrytycznie zawierzylibyśmy naszym formalnym wynikom, z przekonaniem nie odrzucilibyśmy hipotezy o niezależności obydwu cech. Wszakże konstruując test $\chi^2$ w ogóle nie korzystamy z informacji o uporządkowanym charakterze badanych cech i przeto otrzymany wynik może być nieprawidłowy, bo nie oparty na całej posiadanej wiedzy o problemie. Przyjrzyjmy się zatem bliżej tablicy (6.12). Kombinacjom kategorii $(4, \mathrm{A})$ oraz $(1, \mathrm{D})$ odpowiadają liczności wyraźnie mniejsze niż inne liczności w ich kolumnach. Ponieważ badane cechy są mierzone na skali porządkowej, może to oznaczać, że łączy je pewnego typu zależność. Żeby to jaśniej zobaczyć, umówmy się, że kategorie 1 i A odpowiadają „najmniejszym" wartościom obydwu cech, natomiast kategorie 4 i D wartościom „największym". Dostrzegamy teraz, że obserwacje o „dużych" wartościach ze względu na jedną cechę stosunkowo rzadko są związane w parę z obserwacjami o „małych" wartościach ze względu na drugą cechę. A zatem możemy mieć do czynienia z pewną zależnością monotoniczną (typu „małe z małym i duże z dużym"), której istnienia nie można wykryć posługując się testem $\chi^2$. Możliwość taką należy bezwzględnie zbadać (por. zad. 6.5).

Podany przykład wskazuje ponadto na konieczność precyzyjnego sformułowania co mamy na myśli, mówiąc o zależności cech jakościowych, jak również na potrzebę zaproponowania sposobu pomiaru tej zależności. Jest przy tym jasne, że potrzeba wprowadzenia miary zależności dotyczy zarówno przypadku pary zmiennych o cechach uporządkowanych, jak i przypadku zmiennych nominalnych.

Omówienie ograniczamy do niektórych klasycznych miar zależności dla par obserwacji mierzonych na skali nominalnej i dla par na skali porządkowej. W pełni zadowalające z metodologicznego punktu widzenia byłoby przedstawienie jednej miary albo zestawu miar, z których każda spełnia dobrze określone postulaty (czyli dobrze wiadomo, jaki typ zależności mierzy) i z których każda może być użyta do pomiaru siły zależności wiążącej zmienne dowolnych typów, np. zależność między dwiema zmiennymi o uporządkowanych kategoriach, między zmienną o uporządkowanych kategoriach oraz zmienną ilościową, itd. Próba wyjścia ku ujęciu nieklasycznemu wykracza jednak poza ramy tego podręcznika (wyczerpujące opracowanie nowoczesnej metodologii pomiaru zależności można znaleźć w monografii: T. Kowalczyk, E. Pleszczyńska, F. Ruland (2004): *Grade data analysis: models and methods.* Physica Verlag).

W dalszym ciągu tego podrozdziału będziemy zakładać, że zawsze mamy do czynienia z planem krzyżowym eksperymentu.

### Miara zależności dla danych nominalnych

Współczynnik korelacji, jak wiemy, można uznać za miarę zależności liniowej między dwiema zmiennymi ilościowymi. Pojęcie związku liniowego w przypadku pary zmiennych jakościowych nie ma sensu. Przyglądając się tablicy (6.12) zwróciliśmy uwagę, że jeżeli zmienne mają kategorie uporządkowane, to można wprowadzić pojęcie zależności monotonicznej. Gdy zmienne są nominalne, odpada także taka możliwość.

Współczynik korelacji $\rho$ zmiennych ilościowych $X$ oraz $Y$ opiera się na kowariancji między tymi zmiennymi, a więc pewnej mierze ich współzmienności. Na podstawie współczynnika determinacji $R^2$, który jest równy kwadratowi współczynnika $\rho$, możemy orzec, jaką część zmienności zmiennej losowej $Y$ może wyjaśniać liniowa funkcja zmiennej $X$ (tę właśnie część zmienności zmiennej $Y$ wyjaśnia model regresji liniowej).

W sytuacji, gdy przedmiotem naszej analizy jest para zmiennych nominalnych $X$ i $Y$, możemy się zastanowić, jak zdefiniować zmienność tych zmiennych, następnie opisać, jak np. znajomość wartości zmiennej $X$ wpływa na zmienność zmiennej $Y$ i wreszcie uśrednić ów wpływ na zmienność zmiennej $Y$ względem rozkładu zmiennej $X$. Uzasadnienie takiego postępowania jest następujące – interesuje nas średnia (ze względu na rozkład możliwych kategorii zmiennej $X$) wielkość redukcji zmienności zmiennej $Y$, wynikająca ze znajomości kategorii zmiennej $X$. Im większy jest odpowiednio wyrażony stopień takiej redukcji, tym mocniejsza jest zależność między zmiennymi. Opisane postępowanie odpowiada do pewnego stopnia wyznaczaniu współczynnika $R^2$ w analizie zmiennych ilościowych. Odnotujmy, że tak jak tam, zakładamy tu milcząco, iż zmienna $Y$ pełni funkcję zmiennej objaśnianej (odpowiedzi), zmienna $X$ zaś zmiennej objaśniającej.

Niech, jak zwykle, $\{p_{1\cdot}, \ldots, p_{k\cdot}\}$ oraz $\{p_{\cdot 1}, \ldots, p_{\cdot l}\}$ oznaczają rozkłady brzegowe, odpowiednio, zmiennej $X$ oraz $Y$. Jedną z miar zmienności zmiennej nominalnej $Y$ jest **współczynnik Giniego**:

$$V(Y) = \sum_{j=1}^{l} p_{\cdot j}(1 - p_{\cdot j}) = 1 - \sum_{j=1}^{l} p_{\cdot j}^2. \tag{6.13}$$

Współczynnik Giniego zmiennej $Y$ jest równy prawdopodobieństwu zdarzenia polegającego na tym, że dwie niezależne obserwacje tej zmiennej należą do dwóch różnych kategorii. Współczynnik ten osiąga wartość najmniejszą, równą zeru, gdy zmienna $Y$ może przyjmować tylko jedną wartość,

czyli gdy dla pewnego wskaźnika $j$ mamy $p_{\cdot j} = 1$. Współczynnik Giniego osiąga wartość maksymalną, gdy rozkład zmiennej $Y$ jest jednostajny, $p_{\cdot 1} = \ldots = p_{\cdot l} = 1/l$.

Chcąc określić średni wpływ znajomości kategorii zmiennej $X$ na redukcję zmienności zmiennej $Y$ zdefiniujemy najpierw warunkową zmienność zmiennej $Y$, gdy zmienna $X$ przyjęła wartość $x_i$.

Niech $\{p_{1|i}, p_{2|i}, \ldots, p_{l|i}\}$ oznacza rozkład warunkowy zmiennej $Y$ pod warunkiem, że zmienna $X$ przyjęła wartość $x_i$, gdzie $i$ może być równe $1, 2, \ldots$ lub $k$. Zmienność warunkową zmiennej $Y$, gdy zmienna $X = x_i$, definiujemy następująco:

$$V(Y|X = x_i) = 1 - \sum_{j=1}^{l} p_{j|i}^2. \tag{6.14}$$

Gdy zmienne $X$ i $Y$ są niezależne, zmienność warunkowa (6.14) jest równa zmienności bezwarunkowej (6.13) – wobec niezależności zmiennych nie ma żadnej redukcji zmienności związanej ze znajomością kategorii zmiennej $X$. Wielkość redukcji jest maksymalna (czyli zmienność warunkowa jest równa 0) wtedy i tylko wtedy, gdy $p_{j|i} = 1$ dla pewnego $j$, a zatem, gdy znajomość kategorii zmiennej $X$ jednoznacznie wyznacza kategorię zmiennej $Y$.

Podana zmienność warunkowa ma w warunku ustaloną kategorię zmiennej $X$. Ogólnie, ponieważ zmienna $X$ może przyjąć dowolną z $k$ wartości zgodnie z rozkładem $\{p_{1\cdot}, \ldots, p_{k\cdot}\}$, średnia warunkowa zmienność zmiennej $Y$ pod warunkiem zmiennej $X$ ma postać

$$E[V(Y|X)] = \sum_{i=1}^{k} p_{i\cdot} V(Y|X = x_i),$$

skąd

$$E[V(Y|X)] = 1 - \sum_{i=1}^{k} p_{i\cdot} \sum_{j=1}^{l} p_{j|i}^2 = 1 - \sum_{i=1}^{k} \sum_{j=1}^{l} p_{ij}^2 / p_{i\cdot}, \tag{6.15}$$

ponieważ $p_{j|i} = p_{ij}/p_{i\cdot}$.

Ostatecznie stopień redukcji zmienności zmiennej $Y$ przy znajomości zmiennej losowej $X$ można określić następująco:

$$\tau = \frac{V(Y) - E[V(Y|X)]}{V(Y)}. \tag{6.16}$$

Wielkość ta nosi nazwę **współczynnika $\tau$ Goodmana i Kruskala**. Ażeby współczynnik Goodmana i Kruskala był dobrze określony, wystarczy, aby

przynajmniej dwa z prawdopodobieństw brzegowych $p_{\cdot j}$ były dodatnie. Łatwo zauważyć, że $0 \leq \tau \leq 1$ oraz, że współczynnik ten jest równy zeru wtedy i tylko wtedy, gdy $X$ i $Y$ są niezależne. Jak również wynika z wcześniejszych spostrzeżeń, współczynnik $\tau$ jest równy 1 wtedy i tylko wtedy, gdy dla każdego $i$ jest spełniony dla pewnego $j$ warunek $p_{j|i} = 1$. Możemy powiedzieć, że $\tau$ Goodmana i Kruskala jest swego rodzaju odpowiednikiem współczynnika determinacji. Mianowicie, $\tau = 0$, gdy zmienne są niezależne i $\tau = 1$, gdy kategoria zmiennej $X$ określa kategorię zmiennej $Y$. Inna, ważna interpretacja współczynnika jest podana w zad. 6.3. Jak z niej wynika, współczynnik $\tau$ mierzy stopień redukcji prawdopodobieństwa błędnego przewidywania kategorii zmiennej $Y$, zawdzięczany znajomości kategorii zmiennej $X$.

O sile tak rozumianej zależności decyduje oczywiście to, jak bliska jest wartość $\tau$ jedynce. Problem jednak w tym, że jeżeli liczba kategorii zmiennej odpowiedzi $Y$ jest duża, to wartość współczynnika Giniego (6.13) będzie dla wielu rozkładów bliska jedności. Jest to prosty efekt podnoszenia do kwadratu prawdopodobieństw rozkładu brzegowego zmiennej $Y$ (por. definicję współczynnika Giniego (6.13)). Z tego samego powodu, nawet przy istotnym zróżnicowaniu wartości prawdopodobieństw warunkowych występujących w wyrażeniu (6.14) zarówno wartość tego wyrażenia, jak i w konsekwencji wyrażenia (6.15), może się także okazać bliska jedności. Ostatecznie prowadzi to wówczas do bliskiej zeru wartości współczynnika $\tau$. Niestety jest to nieunikniona wada współczynnika Goodmana i Kruskala.

Współczynnik $\tau$ Goodmana i Kruskala jest niesymetryczny w tym sensie, że różnie traktuje zmienne $X$ i $Y$ – pierwszą jak zmienną objaśniającą, drugą jak zmienną objaśnianą (odpowiedzi). Możliwa jest łatwa symetryzacja współczynnika, ale nie będziemy jej omawiać.

W praktyce, prawdopodobieństwa potrzebne do obliczenia współczynnika Goodmana i Kruskala są nieznane i posługujemy się ich dobrze nam znanymi estymatorami, otrzymanymi na podstawie liczności w tablicy kontyngencji.

**Przykład 6.7.** J. Devore i R. Peck przytaczają w podręczniku *Statistics: The exploration and analysis of data* (West Publishing Co., 1986) badanie z początku lat osiemdziesiątych, w którym badaniu poddano amerykańskie pary małżeńskie. Celem badania było orzeczenie, czy Amerykanie mają tendencję do ożenku z osobą tego samego wyznania. Każde małżeństwo było opisane parą zmiennych losowych o tych samych kategoriach, określających, odpowiednio, wyznanie męża (wiersze tablicy kontyngencji) i żony (kolumny tablicy kontyngencji). Możliwe kategorie to: katolik (K), baptysta (B), metodysta (M), luteranin (L), inne wyznanie (I) oraz bez wyznania (BW).

| | | Żona | | | | | |
|---|---|---|---|---|---|---|---|
| | | K | B | M | L | I | BW |
| Mąż | K | 722 | 29 | 24 | 35 | 51 | 16 |
| | B | 36 | 577 | 32 | 8 | 53 | 2 |
| | M | 27 | 40 | 345 | 10 | 38 | 3 |
| | L | 36 | 6 | 11 | 233 | 18 | 5 |
| | I | 79 | 44 | 45 | 20 | 772 | 17 |
| | BW | 53 | 31 | 17 | 21 | 55 | 67 |

Współczynnik $\tau$ Goodmana i Kruskala przyjmuje wartość 0,51, gdy wyznanie męża traktować jako zmienną objaśniającą $X$. Stosunkowo duża wartość współczynnika nie zaskakuje, ponieważ współzależność, a ściślej tendencja do zgodności, wyznania męża i żony jest wyraźnie zauważalna już w samej tablicy kontyngencji. Jego wartość byłaby jeszcze większa, gdyby niewierzący mężowie zwracali większą uwagę na wiarę żony. Na podstawie tablicy kontyngencji można przypuszczać, że $\tau$ Goodmana i Kruskala może mieć większą wartość, gdy za zmienną objaśniającą przyjąć wyznanie żony (por. zad. 6.4).

Zgodnie z tym co już wcześniej zauważyliśmy, współczynnik $\tau$ ma jasną interpretację. Zarazem jednak trzeba pamiętać, że tablica kontyngencji nie dostarcza nam dokładnych prawdopodobieństw występujących we wzorze (6.16) i przeto otrzymywana wartość tego współczynnika powinna być traktowana jako przybliżenie wartości prawdziwej. Jeżeli liczności zawarte w tablicy kontyngencji nie są zbyt małe, warto byłoby zatem podać przynajmniej asymptotyczny (tzn. obowiązujący dla dużych liczności) przedział ufności dla współczynnika $\tau$. Goodman i Kruskal udowodnili, że ich współczynnik ma asymptotycznie rozkład normalny o znanych parametrach i że tym samym znana jest postać przedziału ufności, ale niestety tylko wtedy, gdy plan badania jest różny od krzyżowego. Mianowicie, jeżeli dla ustalenia uwagi przyjąć, że kategorie zmiennej objaśniającej są podane w wierszach macierzy kontyngencji, to trzeba założyć, że znany jest rozkład brzegowy tej zmiennej $\{p_{i\cdot},\ i = 1, 2, \ldots, k\}$ oraz liczności w wierszach są równe $np_{i\cdot}$, $i = 1, \ldots, k$. Zagadnieniem tym nie będziemy się dokładniej zajmować.

Znane są jeszcze inne miary zależności między zmiennymi nominalnymi, takie jak miara oparta na zmienności mierzonej za pomocą tzw. entropii zamiast współczynnika Giniego oraz miara $\lambda$ Goodmana i Kruskala, których nie będziemy jednak omawiać. Wspomnimy tylko, że zaproponowane jeszcze w pierwszej połowie ubiegłego wieku miary oparte na statystyce $\chi^2$ danej wzorem (6.11), np. Craméra $V$ i Pearsona $C$, nie powinny być stosowane (wyjątkiem jest przypadek tablicy kontyngencji o wymiarach $2 \times 2$ oraz miary $V$, wówczas bowiem $V^2$ jest równe mierze $\tau$ Goodmana i Kruskala). Statystyka (6.11) została skonstruowana z myślą o testowaniu niezależności,

a nie mierzeniu siły zależności i właściwie nie wiadomo jak w tym drugim przypadku ją interpretować.

### Miara zależności dla danych o uporządkowanych kategoriach

Przykład tablicy kontyngencji (6.12) sugeruje, że w przypadku pary zmiennych o uporządkowanych kategoriach można badać, czy zmienne są niezależne, czy też monotonicznie zależne. Tę drugą możliwość określiliśmy obrazowo jako zależność typu „małe z małym i duże z dużym". Uściślimy teraz tę argumentację.

Po pierwsze, musimy zwrócić uwagę, że zamiast mówić o dużych i małych wartościach lepiej jest odwoływać się explicite do porównań i mówić o wartościach większych lub mniejszych od innych. W dalszym ciągu zamiast mówić o większych lub mniejszych wartościach, będziemy mówić o wartościach (kategoriach) o, odpowiednio, wyższej lub niższej randze.

Tablica kontyngencji zawiera informację o $n$ jednostkach, z których każda jest opisana parą kategorii – wartości jaką w przypadku danej jednostki przyjęła pierwsza zmienna oraz wartości drugiej zmiennej (jak zwykle, pierwszą zmienną oznaczać będziemy symbolem $X$, drugą symbolem $Y$).

**Definicja 6.1.** *Mówimy, że para jednostek*[2] *jest* **zgodna**, *jeżeli jest spełniony jeden z następujących warunków:*

(1) *ranga kategorii, jaką przyjmuje zmienna $X$ w przypadku pierwszej jednostki w parze, jest wyższa od rangi kategorii, jaką przyjmuje zmienna $X$ dla drugiej jednostki w parze oraz ranga kategorii, jaką przyjmuje zmienna $Y$ dla pierwszej jednostki w parze, jest wyższa od rangi kategorii, jaką przyjmuje zmienna $Y$ dla drugiej jednostki w parze jednostek;*

(2) *ranga kategorii, jaką przyjmuje zmienna $X$ w przypadku pierwszej jednostki w parze, jest niższa od rangi kategorii, jaką przyjmuje zmienna $X$ dla drugiej jednostki w parze oraz ranga kategorii, jaką przyjmuje zmienna $Y$ dla pierwszej jednostki w parze, jest niższa od rangi kategorii, jaką przyjmuje zmienna $Y$ dla drugiej jednostki w parze jednostek.*

Zatem para jest zgodna, jeśli kategoria zmiennej $X$ i kategoria zmiennej $Y$ jednej z jednostek w parze są wyższe rangą od obydwu kategorii charakteryzujących drugą z jednostek pary.

[2]Uwaga: Mówimy tu o parze **jednostek** spośród populacji $n$ jednostek; każda jednostka jest opisana parą kategorii, $x$ oraz $y$.

**Definicja 6.2.** *Para jednostek jest* **niezgodna**, *jeżeli jedna z jednostek w parze ma kategorię zmiennej $X$ wyższej rangi oraz kategorię zmiennej $Y$ niższej rangi niż druga z jednostek pary.*

**Definicja 6.3.** *Jeżeli ranga kategorii, jaką przyjmuje zmienna $X$ w przypadku pierwszej jednostki w parze, jest równa randze kategorii, jaką przyjmuje zmienna $X$ dla drugiej jednostki w parze lub ranga kategorii, jaką przyjmuje zmienna $Y$ dla pierwszej jednostki w parze, jest równa randze kategorii, jaką przyjmuje zmienna $Y$ dla drugiej jednostki w parze jednostek, to mówimy, że jednostki są* **związane**.

Miarę siły monotonicznej zależności dwóch zmiennych jakościowych o uporządkowanych kategoriach możemy oprzeć na odpowiednim porównaniu prawdopodobieństwa, że dwie losowo wybrane jednostki z rozkładu łącznego $\{p_{ij}\}$ są zgodne, z prawdopodobieństwem, że te jednostki są niezgodne. Oznaczając odpowiednio obydwa prawdopodobieństwa symbolami $\Pi_c$ i $\Pi_d$, możemy zdefiniować miarę

$$\gamma = \frac{\Pi_c - \Pi_d}{\Pi_c + \Pi_d}, \tag{6.17}$$

zwaną **miarą gamma**. Z oczywistych względów miarę gamma możemy uznać za miarę zależności monotonicznej, dodatniej gdy $\gamma > 0$ i ujemnej, gdy $\gamma < 0$.

Łatwo wykazać, że jest spełniony warunek $-1 \leq \gamma \leq 1$. Jeżeli $\Pi_c = 0$, to $\gamma = -1$, jeżeli natomiast $\Pi_d = 0$, to $\gamma = 1$.

W praktyce nieznane prawdopodobieństwa $\Pi_c$ i $\Pi_d$ zastępujemy ich estymatorami, co prowadzi do definicji estymatora $\hat{\gamma}$. Niech $C$ oznacza liczbę wszystkich par zgodnych w tablicy kontyngencji, natomiast $D$ liczbę wszystkich par niezgodnych. (Wszystkich par jest w tablicy $n(n-1)/2$, ale oczywiście w przypadku występowania jednostek związanych $C + D < n(n-1)/2$). Nietrudno zauważyć, że próbkowy odpowiednik gammy ma postać

$$\hat{\gamma} = \frac{C - D}{C + D}. \tag{6.18}$$

Przeanalizowanie ogólnej tablicy kontyngencji o wymiarach $2 \times 2$,

| | |
|---|---|
| $n_{11}$ | $n_{12}$ |
| $n_{21}$ | $n_{22}$ |

prowadzi do wniosku, że

$$C = n_{11}n_{22} \quad \text{oraz} \quad D = n_{12}n_{21}. \tag{6.19}$$

Zatem dla tablic o wymiarach $2 \times 2$ mamy

$$\hat{\gamma} = \frac{n_{11}n_{22} - n_{12}n_{21}}{n_{11}n_{22} + n_{12}n_{21}}. \tag{6.20}$$

Podany dla tablic $2\times2$ współczynnik (6.20) został zaproponowany na drodze innego rozumowania już na początku XX wieku przez Yule'a i jest dziś znany jako współczynnik $Q$ Yule'a.

**Przykład 6.1 cd.** Z równości (6.19) natychmiast wynika, że $\hat{\gamma} = -1$, ponieważ $n_{11} = 0$. Jak już podkreślaliśmy, ujemna zależność jest w tym przypadku widoczna gołym okiem. Jak możliwe było jej przeoczenie przez osoby odpowiedzialne za przygotowanie Challengera do startu? Z zeznania osoby odpowiedzialnej wynika, że najprawdopodobniej w analizie sprawowania się krytycznych pierścieni uszczelniających uwzględniono jedynie te loty wahadłowca, w czasie których doszło do ich uszkodzenia. Takie loty zdarzały się przy różnych temperaturach, niskich i wysokich, i stąd wyciągnięto wniosek, że szansa pęknięcia pierścienia nie zależy od temperatury. Uszkodzenia nie prowadziły w przeszłości do większych kłopotów i dlatego zgadzano się na kolejne starty. Gdyby wiedziano, że szansa defektu jest nieporównanie mniejsza przy wyższych temperaturach, na pewno nie zgodzono by się na start Challengera przy temperaturze 31°F.

Jak zwróciliśmy uwagę, to iż w przykł. 6.1 otrzymaliśmy $\hat{\gamma} = -1$ wynika wprost stąd, że $n_{11} = 0$. Gdybyśmy dla porównania w macierzy kontyngencji wartość $n_{11}$ zamienili na 1 i pozostałe liczności pozostawili niezmienione, otrzymalibyśmy $\hat{\gamma} = -0{,}915$.

**Przykład 6.6 cd.** Poglądy polityczne studenta można uznać za zmienną o uporządkowanych kategoriach, jeśli poglądy postępowe uznać za jeden biegun, konserwatywne za drugi i inne poglądy za wyraz pozostawania między tymi biegunami. Jaka jest wartość gamma w tym przypadku?

Obliczenia wygodnie jest prowadzić dla tablicy o wierszach i kolumnach ustawionych według kategorii o rosnących rangach. W przypadku poglądów politycznych kwestia ustalenia rosnącego uporządkowania rang albo ma także charakter polityczny, albo jest kwestią umowy – w tablicy poglądy postępowe znalazły się w pierwszym wierszu, natomiast konserwatywne w trzecim, w wyniku losowania (gdyby, wiersze odpowiadające poglądom postępowym i konserwatywnym zamienić miejscami,

otrzymalibyśmy wartość $\hat{\gamma}$ tę samą co do modułu, ale z przeciwnym znakiem):

| | Nigdy | Okazjonalnie | Często |
|---|---|---|---|
| Postępowe | 479 | 173 | 119 |
| Inne | 172 | 45 | 85 |
| Konserwatywne | 214 | 47 | 15 |

Jeżeli wiersze i kolumny są ustawione według kategorii o rosnących rangach, to dana jednostka tworzy parę zgodną z każdą jednostką położoną w tablicy niżej po prawej stronie. Stąd

$$C = 479(45 + 85 + 47 + 15) + 173(85 + 15) + 172(47 + 15) + (45)(15) =$$

$$= 120607.$$

Parę niezgodną z daną jednostką tworzy każda jednostka położona niżej po lewej stronie:

$$D = 119(172 + 45 + 214 + 47) + 173(172 + 214) + 85(214 + 47) + (45)(214) =$$

$$= 155475.$$

Ostatecznie

$$\hat{\gamma} = -0{,}126.$$

Czy otrzymany wynik wskazuje na wyraźne istnienie zależności ujemnej? Wydaje się, że zależność istnieje, ale jest słaba. Aby dokładniej odpowiedzieć na to pytanie można skorzystać z tego, że liczności w tablicy kontyngencji są duże i że wiadomo, iż $\hat{\gamma}$ ma asymptotycznie rozkład normalny o znanych parametrach. Można zatem skonstruować asymptotyczny przedział ufności dla $\hat{\gamma}$. Stosowne wzory pominiemy i zadowolimy się możliwością skorzystania z odpowiedniego pakietu statystycznego. Przedział ufności dla $\gamma$ na poziomie ufności 0,95 ma postać $[-0{,}21, -0{,}04]$. Jest to więc przypadek rzeczywiście słabej zależności ujemnej.

Trafność przypuszczenia o istnieniu zależności ujemnej (ale nie o sile zależności) można zweryfikować potraktowawszy $\hat{\gamma}$ jako statystykę testową w teście hipotezy zerowej o niezależności zmiennych $X$ i $Y$ przy hipotezie alternatywnej orzekającej ujemną zależność między zmiennymi. Problem sprowadza się wówczas do następującego: jaka jest $p$–wartość takiego testu przy zaobserwowanej wartości statystyki testowej?

Należy zatem podać sposób obliczenia wspomnianej $p$–wartości. Dla zadanej tablicy kontyngencji interesują nas takie odstępstwa tej tablicy od liczności $n_{ij}^0$, tzn. liczności oczekiwanych, gdy hipoteza zerowa jest prawdziwa, które

wskazują na istnienie zależności ujemnej. Punktem odniesienia jest hipoteza zerowa i, tym samym, oczekiwane liczności $n_{ij}^0$. Innymi słowy, punktem odniesienia są zaobserwowane liczności wierszy i kolumn, $n_{i\cdot}$ oraz $n_{\cdot j}$, $i = 1, \ldots, k$, $j = 1, \ldots, l$, ponieważ to one wyznaczają wartości $n_{ij}^0$ ($n_{ij}^0 = n_{i\cdot} n_{\cdot j}/n$; w dalszym ciągu liczności $n_{i\cdot}$ oraz $n_{\cdot j}$ będziemy nazywali brzegowymi).

Oznaczmy symbolem $T_0$ tablicę kontyngencji, będącą przedmiotem naszej analizy. Niech $T$ będzie dowolną tablicą o licznościach brzegowych równych odpowiednim licznościom brzegowym tablicy $T_0$ i niech $\hat{\gamma}(T)$ oznacza wartość statystyki $\hat{\gamma}$ dla tablicy $T$ (w naszym przykładzie $\hat{\gamma}(T_0) = -0{,}126$). Szukana $p$–wartość jest równa prawdopodobieństwu wylosowania – przy założeniu prawdziwości hipotezy zerowej – dowolnej tablicy $T$ spełniającej nierówność $\hat{\gamma}(T) \leq \hat{\gamma}(T_0)$. Jest to suma prawdopodobieństw $P_0(T)$, gdzie $P_0(T)$ jest prawdopodobieństwem otrzymania konkretnej tablicy $T$ przy zachodzeniu hipotezy zerowej, po wszystkich tablicach $T$ spełniających podaną nierówność:

$$\sum_{T:\ \hat{\gamma}(T) \leq \hat{\gamma}(T_0)} P_0(T). \tag{6.21}$$

Obliczenie takiej $p$–wartości nie jest łatwe, ale jest możliwe dzięki istniejącym pakietom komputerowym. Obliczenie $p$–wartości dla przykł. 6.6 pozostawiamy Czytelnikowi (por. zad. 6.6).

W sytuacji, gdy hipoteza alteratywna dotyczy dodatniej zależności, w (6.21) należy zmienić znak nierówności. Czytelnik zwrócił na pewno uwagę, że obliczając $p$–wartość, porównywaliśmy jedynie wartości $\hat{\gamma}$ dla tablic o tych samych licznościach brzegowych. Rzecz w tym, że porównywanie wartości, jakie przyjmuje miara $\gamma$, nie ma sensu, gdy tablice mają różne liczności $n_{i\cdot}$ oraz $n_{\cdot j}$, $i = 1, \ldots, k$, $j = 1, \ldots, l$. Zarazem oznacza to, że wszystkie obliczenia prowadziliśmy pod warunkiem, że rozkłady brzegowe są zadane. Zatem także otrzymana $p$–wartość jest warunkową $p$–wartością, prawdziwą przy zadanych rozkładach brzegowych.

Zdarza się, że liczba par związanych, równa $n(n-1)/2 - C - D$, jest duża. Można wówczas utrzymywać, że obecność tych par świadczy przeciw hipotezie o dodatniej (lub ujemnej) zależności. W takiej sytuacji rozsądne jest użycie miary $\tau$-$b$ Kendalla lub blisko z nią związanej miary $d$ Sommersa (por. zad. 6.7). Pakiety statystyczne umożliwiają obliczenie $p$–wartości, odpowiadającej zaobserwowanym wartościom tych miar (dla danej tablicy kontyngencji obydwie miary nie są równe, ale odpowiada im ta sama $p$–wartość).

## 6.5. Uwagi o poprawności wnioskowania i paradoksie Simpsona

Przedstawione w rozdz. 2 tablice kontyngencji w jednoznaczny sposób opisywały łączny rozkład prawdopodobieństwa dwóch zmiennych losowych, a stąd także rozkłady brzegowe i warunkowe. W tym rozdziale nie mamy już do czynienia z zadanymi rozkładami, a z obserwacjami, które – w przypadku krzyżowego planu eksperymentu – w łatwy sposób umożliwiają uzyskanie estymatorów NW wspomnianych prawdopodobieństw. Trzeba jednak pamiętać, że nie zawsze mamy do czynienia z planem krzyżowym, tj. klasyfikowaniem pewnej liczby niezależnie losowanych jednostek do par kategorii.

Omawiając problem jednorodności $k$ populacji zwróciliśmy uwagę, że liczby obserwacji w wierszach są ustalone. W przykładzie 6.2 asystent nr 1 miał 82 studentów, asystent nr 2 miał 79 studentów oraz asystent nr 3 miał ich 84, ponieważ tylu właśnie studentów przydzielono każdemu z nich.

Konstatacja ta ma proste, ale doniosłe konsekwencje. Przedstawimy je na przykładzie bardziej przemawiającym do wyobraźni, a mianowicie posłużymy się przykładem 2.12. Podane tam prawdopodobieństwa $P(A|B_1)$ oraz $P(A|B_2)$ mogły zostać oszacowane na podstawie eksperymentu, w którym test zaaplikowano 100 zdrowym osobom i 100 osobom chorym, otrzymując następującą tablicę kontyngencji:

| | Wynik testu − | Wynik testu + |
|---|---|---|
| Osoby zdrowe | 98 | 2 |
| Osoby chore | 1 | 99 |

Wynik testu jest zmienną losową o dwóch kategoriach, której rozkład oceniamy dla dwóch populacji, a mianowicie populacji osób zdrowych oraz populacji osób chorych. Na podstawie danych tablicy prawdopodobieństwo

$$P(A|B_1) = P(\text{wynik testu dodatni} \mid \text{osoba chora})$$

możemy ocenić jako równe 0,99 oraz prawdopodobieństwo

$$P(A|B_2) = P(\text{wynik testu dodatni} \mid \text{osoba zdrowa})$$

możemy ocenić jako równe 0,02. Ale korzystając jedynie z tablicy kontyngencji otrzymanej w podany sposób, nie ma żadnej możliwości oszacowania prawdopodobieństwa bycia osobą chorą pod warunkiem, że test dał wynik dodatni. Dla danych z przykł. 2.12 nie ma możliwości oszacowania prawdopodobieństwa $P(B_1|A)$. Mówiąc inaczej, nie wiemy nic o rozkładzie łącznym dwóch zjawisk: bycia osobą chorą lub zdrową oraz uzyskania

wyniku dodatniego lub ujemnego testu. Tym samym, tablica nie zawiera informacji o prawdopodobieństwie zachorowania, czyli o rozkładzie brzegowym zmiennej „zdrowy/chory". Prawdopodobieństwo zachorowania musi zostać oszacowane przez przeprowadzenie innego eksperymentu. Tak było w przypadku przykładu 2.12, gdzie prawdopodobieństwo to zostało podane jako znane. Dopiero wówczas możliwe było zastosowanie twierdzenia Bayesa i w ten sposób udzielenie odpowiedzi na interesujące nas pytanie.

Podsumowując: analizując tablicę kontyngencji, musimy dobrze pamiętać czy powstała w wyniku krzyżowego planu eksperymentu, czy też w wyniku innego planu, jak choćby w przykładach 2.12 i 6.2. Przy tym, z planami innymi niż krzyżowy mamy do czynienia stosunkowo często. W podrozdziale 2.5 omówiliśmy bardzo ważny eksperyment porównawczy, w którym grupę eksperymentalną poddajemy działaniu pewnego czynnika, jednocześnie dysponując grupą kontrolną i następnie porównujemy wyniki otrzymane w obydwu grupach. Czynnikiem może np. być zaaplikowanie pewnego leku, wynikiem – odpowiednio – reakcja na lek oraz na jego niezastosowanie. Taki eksperyment ma charakter **prospektywny** – reakcja jest przyszłym wynikiem naszego eksperymentu. Innym typowym planem różnym od krzyżowego jest badanie **retrospektywne**, gdy aktywny eksperyment, w którym wymuszamy działanie czynnika na grupę eksperymentalną jest np. niemożliwy do przeprowadzenia. Chcąc np. z największą dozą pewności ustalić czy palenie papierosów wpływa na powstanie choroby wieńcowej, najlepiej byłoby wylosować dwie grupy możliwie do siebie podobnych niemowląt i, gdy dorosną, jednej grupie kazać palić, powiedzmy, 20 papierosów dziennie, drugiej zaś palenia zakazać. Byłoby to wszakże postępowanie nieetyczne i na szczęście jest niemożliwe. Można natomiast wylosować duże grupy osób zdrowych oraz osób chorych na chorobę wieńcową i porównać te grupy ze względu na rozkład palaczy w obydwu grupach. (Takie badanie nosi nie tylko nazwę retrospektywnego, ale także **obserwacyjnego**, aby odróżnić je od aktywnego eksperymentu).

Wspominając o problemie wpływu palenia papierosów na zapadalność na chorobę wieńcową, raz jeszcze wracamy do zagadnienia ustalania związków przyczynowych. O sprawie chyba najważniejszej, a mianowicie o wpływie na otrzymywane wyniki zmiennych uwikłanych mówiliśmy już w podrozdz. 2.5. Za pomocą wspomnianych badań retrospektywnych ustalono ponad praktycznie wszelką wątpliwość, że wśród palaczy ryzyko zapadnięcia na chorobę wieńcową jest wielokrotnie wyższe niż wśród osób niepalących i ryzyko to rośnie wraz ze wzrostem liczby papierosów wypalanych dziennie. Ale, jak to wynika z naszych rozważań w podrozdz. 2.5, nie mówi to jeszcze o związku przyczynowym między paleniem i chorobą naczyń serca. A może to jakieś uwarunkowanie genetyczne jest prawdopodobną przyczyną jednego i drugiego? A może palacze mają tendencję do „niehigienicznego" życia

i dlatego częściej zapadają na chorobę wieńcową, natomiast fakt palenia jest tylko jednym z aspektów „złego prowadzenia się" przez te osoby? Jeśli dziś twierdzi się jednak, że samo palenie ma negatywny wpływ na zapadalność na wymienianą tu chorobę, to dlatego, że przeprowadzono bardzo wiele bardzo różnych badań ludzi i wiele eksperymentów na zwierzętach, które to podejrzenie summa summarum potwierdzają.

Mówiąc o zagadnieniu przyczynowości wróćmy jeszcze raz do błędu znacznie prostszego niż lekceważenie istnienia zmiennych uwikłanych. Powtórzmy: ani korelacja, ani prawdopodobieństwo warunkowe jednego zdarzenia pod warunkiem drugiego, nie mierzą siły związku przyczynowego. Ich definicje nie mają nic wspólnego z owym związkiem. Ostatnio pewien znakomity statystyk angielski poczuł się zmuszony przestrzec o tym w artykule naukowym analityków danych spoza kręgów statystyków. Jeżeli np. przy okazji tzw. analizy koszyka zakupów w hipermarkecie badamy związek kupna artykułu $B$ z kupnem artykułu $A$, to nie powinniśmy wiele wnosić z tego tylko, że prawdopodobieństwo $P(\text{kupienie } B \mid \text{kupienie } A)$ jest bliskie jedności. Może być to informacja interesująca, ale tylko wtedy, gdy $P(\text{kupienie } B \mid \text{niekupienie } A)$ jest bliskie zera. Tylko wtedy można mówić o istnieniu dwóch grup klientów, których zachowanie jest w podanym aspekcie różne. Zauważmy, że np. przy niezależności zdarzeń zakupu artykułu $B$ oraz zakupu artykułu $A$, prawdopodobieństwo $P(\text{kupienie } B \mid \text{kupienie } A)$ może przyjąć dowolną wartość, oczywiście zawsze równą $P(\text{kupienie } B)$.

Na koniec omówimy paradoks Simpsona. W podręcznikach statystyki najczęściej przedstawia się dziś ów paradoks albo na przykładzie niesłusznych podejrzeń władz jakiejś instytucji o dyskryminację kobiet, albo na przykładzie zbyt uproszczonej analizy jakości pracy zespołów lekarskich. My posłużymy się pierwszym przykładem (por. zad. 6.9, gdzie jest podany przykład drugiego typu).

**Paradoks Simpsona** powiada, że pominięcie w analizie zmiennej uwikłanej może zmienić – nawet diametralnie – otrzymywane związki między dwiema innymi zmiennymi jakościowymi.

Wyobraźmy sobie, że pewien wydział elektroniki i informatyki rekrutuje studentów na obydwa kierunki. Przed egzaminem kwalifikacyjnym kandydat musi podać kierunek, na który chce się dostać. Oto wyniki egzaminu:

| | Kobiety | Mężczyźni |
|---|---|---|
| Osoby odrzucone | 102 | 111 |
| Osoby przyjęte | 56 | 92 |

Zauważmy, że procent przyjętych mężczyzn jest wyższy od procentu przyjętych kobiet. W pierwszym przypadku przyjęto $100(92/203) = 45\%$ kandy-

datów, w drugim tylko $100(56/158) = 35\%$ kandydatek. Czy zatem można podejrzewać wydział o dyskryminację kobiet?

Zauważmy, że w analizie w ogóle nie pojawiła się trzecia zmienna, a mianowicie wybierany kierunek studiów. Jej uwzględnienie i jednoczesne opisanie wszystkich zmiennych w zasadzie wymaga skonstruowania tablicy kontyngencji o trzech wymiarach, czyli **tablicy trójdzielczej**, zamiast znanej nam tablicy dwudzielczej. Równie dobrze można jednak posłużyć się dwiema tablicami dwudzielczymi, pierwszą opisującą wcześniej już analizowane dwie zmienne przy zmiennej trzeciej o kategorii elektronika i drugiej opisującej dwie wcześniejsze zmienne przy trzeciej o ustalonej kategorii informatyka:

| Elektronika | | |
|---|---|---|
| | Kobiety | Mężczyźni |
| Osoby odrzucone | 11 | 71 |
| Osoby przyjęte | 12 | 73 |

| Informatyka | | |
|---|---|---|
| | Kobiety | Mężczyźni |
| Osoby odrzucone | 91 | 40 |
| Osoby przyjęte | 44 | 19 |

I co się okazuje? Na elektronikę przyjęto 52% kandydatek oraz 51% knadydatów, natomiast na informatykę 32% kandydatek i tyleż procent kandydatów. Wrażenie dyskryminacji powstało dlatego, że *związek między dwiema zmiennymi, ujawniany dla każdej kategorii trzeciej zmiennej oddzielnie, może zostać diametralnie zmieniony przez zagregowanie danych, polegające na zsumowaniu wyników dla różnych kategorii trzeciej zmiennej.*

W naszym przypadku agregacja dlatego dała fałszywe wrażenie dyskryminacji kobiet, że w jej wyniku nie mógł być dostrzeżony fakt znacznie trudniejszego wstępu na kierunek informatyczny, przy jednoczesnej popularności tego kierunku wśród kobiet (aż 85% kobiet i tylko 29% mężczyzn starało się o wstęp na informatykę).

## 6.6. Zadania

**Uwaga:** Chociaż dalej o tym nie wspominamy, niektóre zadania wymagają skorzystania z pakietu statystycznego.

**6.1.** Wykazać, że test hipotezy zerowej (6.1) przy hipotezie alternatywnej (6.2) oparty na statystyce (6.4) jest dla $k = 2$ równoważny testowi dla proporcji opartemu na statystyce (3.67).

**Wskazówka:** Zauważyć, że

$$(n_1 - np_1^0)^2 = (n - n_1 - n(1 - p_1^0))^2 = (n_2 - np_2^0)^2$$

i posługując się definicją statystyki $\chi^2$ wykazać, iż $\chi^2 = Z^2$.

**6.2.** Zastosować test Pearsona (6.4) do zweryfikowania hipotezy, że próba losowa 100 obserwacji opisanych następującą macierzą liczności

| | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ | $x_6$ |
|---|---|---|---|---|---|---|
| $n_i$ | 19 | 15 | 18 | 21 | 15 | 12 |

pochodzi z rozkładu jednostajnego, $p_1^0 = p_2^0 = \cdots = p_6^0 = 1/6$.

**6.3.** Przyjmijmy, że jednostki eksperymentalne są opisane parą zmiennych losowych $(X, Y)$. Staramy się przewidzieć kategorię zmiennej $Y$, odpowiadającą zadanej jednostce eksperymentalnej. Załóżmy najpierw, że kategorię zmiennej $Y$ przewidujemy losowo, zgodnie z rozkładem brzegowym tej zmiennej (tzn. kategorię pierwszą przewidujemy z prawdopodobieństwem $p_{\cdot 1}$ itd.).
**a)** Wykazać, że wielkość $V(Y)$ dana wzorem (6.13) jest prawdopodobieństwem błędnej predykcji.
**b)** Gdybyśmy wiedzieli, iż w przypadku interesującej nas jednostki eksperymentalnej zmienna $X$ ma $i$-tą kategorię, to nasze losowe przewidywanie kategorii zmiennej $Y$ powinno opierać się na rozkładzie warunkowym $\{p_{1|i}, \ldots, p_{l|i}\}$, dając prawdopodobieństwo błędnej predykcji wyrażone wielkością (6.14). Uzasadnić dlaczego wielkość (6.15) można uznać za prawdopodobieństwo błędnej predykcji uśrednione względem rozkładu zmiennej $X$.
**c)** Uzasadnić stwierdzenie, iż współczynnik $\tau$ Goodmana i Kruskala mierzy stopień redukcji prawdopodobieństwa błędnej predykcji kategorii zmiennej $Y$, zawdzięczany znajomości kategorii zmiennej $X$.

**Wskazówka do punktu a.** Rozpatrzeć rozbicie

$$A_j = \{\text{wybór } j\text{-tej kategorii}\},$$

$j = 1, 2, \ldots, l$, i skorzystać z twierdzenia o prawdopodobieństwie całkowitym.

**6.4.** Podać wartość estymatora współczynnika $\tau$ Goodmana i Kruskala dla danych z przykł. 6.7, gdy za zmienną objaśniającą przyjąć wyznanie żony.

**6.5.** Obliczyć $\hat{\gamma}$ dla tablicy kontyngencji (6.12). Podać asymptotyczny przedział ufności na poziomie ufności 0,95 dla $\gamma$. Opierając się na statystyce $\hat{\gamma}$ zweryfikować hipotezę o niezależności obydwu zmiennych przy hipotezie alternatywnej o ich dodatniej zależności. Skomentować otrzymane wyniki

(np. czy można stwierdzić słabą, ale „stosunkowo pewną" zależność między zmiennymi).

**6.6.** Na podstawie statystyki $\hat{\gamma}$ zweryfikować hipotezę o niezależności zmiennych losowych z przykł. 6.6 przy hipotezie alternatywnej o ich ujemnej zależności. Podać odpowiednią $p$–wartość.

**6.7.** Próbkowy odpowiednik miary $\tau$-$b$ Kendalla jest określony następująco:

$$\hat{\tau}\text{-}b = \frac{C - D}{\sqrt{[n(n-1)/2 - T_X][n(n-1)/2 - T_Y]}},$$

natomiast miary $d$ Sommersa następująco:

$$\hat{d} = \frac{C - D}{[n(n-1)/2 - T_X]},$$

gdzie

$$T_X = \sum_i n_{i\cdot}(n_{i\cdot} - 1)/2$$

oraz

$$T_Y = \sum_j n_{\cdot j}(n_{\cdot j} - 1)/2.$$

**a)** Jeżeli próba zawiera $n$ jednostek eksperymentalnych, można utworzyć $n(n-1)/2$ różnych (nieuporządkowanych) par tych jednostek. Każda jednostka jest opisywana parą zmiennych losowych. Niech $(X_a, Y_a)$, $(X_b, Y_b)$ oznacza parę jednostek $a$ i $b$. Niech

$$X_{ab} = \begin{cases} 1, & \text{gdy } X_a > X_b \\ 0, & \text{gdy } X_a = X_b \\ -1, & \text{gdy } X_a < X_b \end{cases}$$

oraz

$$Y_{ab} = \begin{cases} 1, & \text{gdy } Y_a > Y_b \\ 0, & \text{gdy } Y_a = Y_b \\ -1, & \text{gdy } Y_a < Y_b. \end{cases}$$

Sprawdzić, że $\hat{\tau}$-$b$ jest próbkowym współczynnikiem korelacji między zmiennymi $X_{ab}$ a $Y_{ab}$ (innymi słowy, miara $\tau$ Kendalla jest równa populacyjnemu współczynnikowi korelacji między $X_{ab}$ a $Y_{ab}$).
**b)** Określić za pomocą zmiennych $X_{ab}$ i $Y_{ab}$ populacyjną miarę $d$ Sommersa.
**c)** Obliczyć $\hat{\tau}$-$b$ i $d$ dla zmiennych z przykł. 6.6. Estymatory $\hat{\tau}$-$b$ i $\hat{d}$ są asymptotycznie normalne i dlatego dla ich populacyjnych pierwowzorów możliwe jest skontruowanie asymptotycznych przedziałów ufności. Podać asymptotyczne przedziały ufności na poziomie ufności 0,95 dla miar $\tau$-$b$ i $d$. Podać $p$–wartości dla testów niezależności przy alternatywie ujemnej zależności,

opartych na statystykach $\hat{\tau}$-$b$ oraz $\hat{d}$. Porównać otrzymane wyniki z wynikami opartymi na $\hat{\gamma}$.

**Wskazówka:** Zauważyć, że $E(X_{ab}) = E(Y_{ab}) = 0$ i wykorzystać fakt, iż $T_X$ jest liczbą par o własności $X_a = X_b$ oraz, że $T_Y$ jest liczbą par o własności $Y_a = Y_b$.

**6.8.** Ankiecie poddano 445 studentów. Każdy student odpowiadał na dwa pytania, jedno dotyczące palenia przezeń marihuany i drugie, dotyczące przyjmowania środków narkotyzujących i picia alkoholu przez jego rodziców. Na pierwsze pytanie można było odpowiedzieć: nigdy nie paliłem, palę okazjonalnie, palę regularnie. W przypadku drugiego pytania możliwe były odpowiedzi: nie przyjmowali żadnych używek (tzn. niczego nie zażywali i nie pili), przyjmowali używkę tylko jednego rodzaju (tzn. narkotyki lub alkohol, ale nie obydwa rodzaje używek), przyjmowali używki obydwu rodzajów. Odpowiedzi na drugie pytanie kodujemy następująco: nic, jeden, obydwa. Wyniki są zawarte w tablicy kontyngencji:

| | | Student | | |
|---|---|---|---|---|
| | | Nigdy | Okazjonalnie | Regularnie |
| Rodzice | Nic | 141 | 54 | 40 |
| | Jeden | 68 | 44 | 51 |
| | Obydwa | 17 | 11 | 19 |

Sprawdzić na podstawie miar $\hat{\gamma}$, $\tau$-$b$ i $d$, że obydwie zmienne można uznać za dodatnio zależne. (Podany przykład został zaczerpnięty z książki: R. Bartoszyński, M. Niewiadomska-Bugaj (1996): *Probability and statistical inference*. Wiley, New York.)

**6.9.** W cytowanym w przedmowie podręczniku, Moore i McCabe przytaczają przykład dwóch szpitali, A i B, w których przeprowadza się skomplikowany zabieg operacyjny. Kierowani na operację pacjenci mogą być w dobrym lub złym stanie. Operację odnotowuje się jako udaną, jeśli pacjent przeżywa 6 tygodni po zabiegu. Analiza obejmuje trzy zmienne, opisujące każdego pacjenta: szpital, w którym przeszedł zabieg (A lub B), jego stan (dobry lub zły) oraz wynik przebytej operacji (nie przeżył lub przeżył). Wyniki są zebrane w dwóch tablicach dwudzielczych, jednej dla kategorii dobry stan i drugiej dla kategorii zły stan:

| Stan dobry | | |
|---|---|---|
| | Szpital A | Szpital B |
| Nie przeżył | 6 | 8 |
| Przeżył | 594 | 592 |

| Stan zły | | |
|---|---|---|
| | Szpital A | Szpital B |
| Nie przeżył | 57 | 8 |
| Przeżył | 1443 | 192 |

Obliczyć procent pacjentów zmarłych w szpitalu A i w szpitalu B, gdy uprzedni stan pacjenta był dobry i gdy był zły. Z podanych tablic zbudować jedną tablicę dwudzielczą dla zmiennych wynik operacji oraz szpital, czyli dokonać agregacji danych przez zsumowanie wyników dla różnych kategorii zmiennej stan pacjenta. Czy ocena obydwu szpitali, mierzona procentem zmarłych pacjentów, zmieniła się? Skomentować słuszność zabiegu agregacji.

ROZDZIAŁ 7

# Metody wyboru prób z populacji skończonej

## 7.1. Metoda reprezentacyjna

### 7.1.1. Cel metody reprezentacyjnej

Rozpatrzmy pewną populację $\mathcal{O}$, np. populację studentów pewnej wyższej szkoły technicznej, składającą się ze skończonej liczby $N$ obiektów $o_1, o_2, \ldots, o_N$ i pewną interesującą nas cechę tych obiektów, którą oznaczymy przez $Y$. Wartości cechy $Y$ dla obiektów $o_1, o_2, \ldots, o_N$ wynoszą odpowiednio $y_1, y_2, \ldots, y_N$. Zajmiemy się ponownie problemem szacowania pewnego parametru $\theta$ wartości cechy $Y$. Parametr był dotychczas określony jako pewna funkcja rozkładu cechy $Y$ w populacji. Na przykład w przypadku ciągłej cechy losowej $Y$ o funkcji gęstości $f$ wartość średnia tej cechy była równa $\mu = \int x f(x)\, dx$, a więc była funkcją rozkładu określonego przez gęstość $f$. Gęstość prawdopodobieństwa określa szansę z jaką cecha $Y$ przyjmie wartości z dowolnego ustalonego przedziału. W przypadku populacji skończonej, dla której każdy element można poddać badaniu w tym sensie, że tylko od nas zależy, czy zmierzymy wartość cechy $Y$ dla konkretnego obiektu, czy też nie, nie chcemy wiązać z nią żadnego konkretnego rozkładu prawdopodobieństwa. W takim przypadku jest bardziej naturalne zdefiniowanie parametru jako pewnej funkcji wszystkich wartości cechy w populacji: $\theta = f(y_1, y_2, \ldots, y_N)$. Może nas interesować np. średnia wartość cechy zdefiniowana jako $\mu = N^{-1} \sum_{i=1}^{N} y_i$ lub jej rozproszenie zdefiniowane jako[1] $\sigma = ((N-1)^{-1/2} \sum_{i=1}^{N} (y_i - \mu)^2)^{1/2}$. Oczywiście, koncepcyjnie najprostszym sposobem oszacowania wartości parametru $\theta$ jest jego obliczenie na podstawie pomiaru wszystkich wartości cechy. Metoda taka, stosowana np. w spisach powszechnych jest bardzo kosztowna i technicznie skomplikowana,

[1] Taka definicja odchylenia standardowego jest podyktowana tradycją; zauważmy, że w tym przypadku bardziej naturalny byłby czynnik skalujący $N^{-1/2}$ zamiast czynnika $(N-1)^{-1/2}$, gdyż centrujemy wartości za pomocą wartości średniej $\mu$.

gdy liczba obiektów $N$ jest duża. Pozostaje nam metoda szacowania wartości $\theta$ na podstawie próby losowo wybranej z całej populacji.

Podkreślmy jeszcze, że rozpatrywana w badaniu populacja nie musi być żadną realnie istniejącą populacją, a jej definicja może być związana z wybraną metodą badania. Wyobraźmy sobie biologa, który chce ocenić całkowitą liczbę drzew pewnego gatunku, na przykład brzóz, znajdujących się w pewnym kompleksie leśnym. Swoje badanie wykona tak, że podzieli obszar lasu na kwadraty lub prostokąty, spośród nich wylosuje pewną ich próbę i w wybranych fragmentach zliczy dokładną liczbę brzóz. W tym przypadku rozpatrywaną populacją jest zbiór kwadratów, na jakie podzielono las, a przypisaną jednostkom populacji cechą jest liczba brzóz w kwadracie. Oczywiście, średnia liczba brzóz przypadających na jeden kwadrat nie jest celem badania, ale można na jej podstawie oszacować całkowitą liczbę brzóz w lesie (będącą już parametrem realnie istniejącej populacji, czyli rozpatrywanego kompleksu leśnego). Podobnie postępuje biolog, chcący np. ocenić całkowitą liczbę fok w Morzu Północnym, z tą jedynie różnicą, że zliczenie liczby fok przebywających w wylosowanych fragmentach obszaru odbywa się na podstawie zdjęć wykonanych podczas przelotu nad nim.

Specyfika wyboru prób z populacji skończonej polega przede wszystkim na tym, że mało użyteczna staje się koncepcja prostej próby losowej, gdyż w tym przypadku zmienne losowe zdefiniowane jako wartości cechy $Y$ dla kolejno wybranych obiektów są z reguły zależne. Jeśli z urny, w której znajduje się 9 białych i jedna czarna kula, wylosujemy kolejno 3 kule, nie wkładając ich z powrotem do urny, to losowe kolory kul w kolejnych ciągnieniach są zależne. Jeśli bowiem np. w pierwszym ciągnieniu wylosowaliśmy kulę czarną, to wiemy, że druga i trzecia wylosowana kula musi być biała. Oczywiście, zależność zmiennych odpowiadających kolejnym wybranym obiektom powoduje, że musimy na nowo przypatrzeć się własnościom estymatorów, takich jak np. średnia z próby, gdyż przy ich wyprowadzeniu korzystaliśmy w sposób istotny z założenia niezależności obserwacji. W omawianej sytuacji występuje również często specyficzna okoliczność sprzyjająca: to my decydujemy o schemacie wyboru próby i możemy tę wiedzę wykorzystać przy konstrukcji estymatorów.

Jakie ogólne postulaty powinna spełniać wybrana przez nas próba? Przede wszystkim musi być to **próba reprezentatywna** dla całej populacji w tym sensie, że dowolny obiekt ma dodatnie prawdopodobieństwo znalezienia się w niej. Jeśli rozpatrzymy wyniki ankiety internetowej dotyczącej osiągnięć polskich sportowców na olimpiadzie w Sydney, to próba składająca się z osób biorących udział w ankiecie nie będzie próbą reprezentatywną dla populacji wszystkich Polaków: osoby nie mające dostępu do Internetu nie wezmą w niej udziału, a więc mają zerową szansę znalezienia się w próbie.

W tym przypadku jest oczywiste, że na podstawie ankiety nie możemy wyciągać wniosków dotyczących opinii w populacji wszystkich Polaków. Zauważmy ponadto, że określenie populacji, dla której omawiana próba jest reprezentatywna, jest praktycznie niemożliwe: nie jest to bynajmniej populacja Polaków mających dostęp do Internetu, gdyż znaczna część tych osób z najróżniejszych powodów nigdy nie wzięłaby udziału w tej lub w ogóle żadnej ankiecie. Odrębną kwestią jest np. wiedza o takiej ankiecie i dysponowanie czasem na wzięcie w niej udziału. W rezultacie wyniki tak przeprowadzonego sondażu są praktycznie pozbawione wartości, gdyż nie możemy określić do jakiej populacji się odnoszą. *Warunek wyboru próby reprezentatywnej dla rozpatrywanej populacji jest warunkiem koniecznym dla poprawności wyciąganych wniosków o tej populacji.* Czasami rozpatruje się postulat mocniejszy od reprezentatywności, mówiący, że wszystkie próby o ustalonej liczności powinny mieć takie samo prawdopodobieństwo wyboru.[2] Zobaczymy, że postulat ten jest spełniony dla zdefiniowanego poniżej schematu prostego losowania bez zwracania, nie jest natomiast prawdziwy w przypadku innych naturalnych schematów losowania. Metodologię wyboru prób reprezentatywnych dla populacji nazywa się często **metodą reprezentacyjną**.

Na potrzeby tego podrozdziału rozpatrzmy następującą specyficzną przestrzeń prób $\mathcal{S}$.

*Przestrzeń $\mathcal{S}$ składa się ze wszystkich podzbiorów zbioru $\{1, 2, \ldots, N\}$. Zdarzenie elementarne ma postać $s = \{i_1, i_2, \ldots, i_n\}$ dla pewnego $1 \leq n \leq N$. Zbiór $s = \{i_1, i_2, \ldots, i_n\}$ zawiera indeksy elementów wybranych do próby.*

Rozpatrzenie takiej przestrzeni prób jest wygodne w sytuacji, gdy elementy w wylosowanej próbie nie powtarzają się i nie przywiązujemy wagi do kolejności, w jakiej elementy zostały wylosowane. Przestrzeń $\mathcal{S}$ składa się z $2^N$ zdarzeń elementarnych. Przypisanie prawdopodobieństwa każdemu z nich możemy oczywiście interpretować jako określenie szansy, z jaką konkretna próba zostanie wylosowana. Tak więc z samą populacją $\mathcal{O}$ nie wiążemy żadnej struktury losowej, struktura taka jest związana wyłącznie z metodą pobierania prób. Rozkład prawdopodobieństwa $P$ na przestrzeni prób $\mathcal{S}$ nazywamy tradycyjnie **planem losowania**. Jeśli plan $P$ jest taki, że przypisuje dodatnie prawdopodobieństwa tylko podzbiorom $n$-elementowym dla ustalonego $n$, to oznacza to, że tylko próby o tej liczności chcemy brać pod uwagę jako wynik losowania. Mówimy wtedy o **planie o ustalonej liczności** $n$. Stosując się do konwencji oznaczania zmiennych losowych dużymi literami, a ich konkretnych wartości małymi, przez $S$ będziemy oznaczać

[2] Próby o tej własności nazywa się czasami prostymi próbami losowymi. Podkreślmy tu, że z reguły **nie są** to proste próby losowe w sensie naszej def. 2.19.

zmienną losową o rozkładzie $P$. Zauważmy, że dokonujemy w ten sposób rozszerzenia def. 2.9, dopuszczając jako wartości zmiennej losowej nie tylko liczby rzeczywiste, ale również **zbiory** składające się ze skończonej liczby elementów. Zajście zdarzenia $\{S = s\}$ oznacza, że zmienna losowa $S$ przyjmuje wartość $s$ będącą w tej sytuacji pewnym podzbiorem przestrzeni prób, a zdarzenie $\{k \in S\}$ polega na wylosowaniu obiektu o indeksie $k$ do próby. Plan losowania określa prawdopodobieństwo wylosowania określonej próby, nie precyzuje jednak metody losowania. To ustala **schemat losowania**, podający algorytm wyboru kolejnych elementów populacji i prawdopodobieństwa wyboru w każdym ciągnieniu.

### 7.1.2. Podstawowe schematy losowania prób

**Schemat prostego losowania bez zwracania.** Przyjmijmy, że losujemy po kolei $n$ elementów populacji, przy czym wylosowany na pewnym etapie obiekt jest „odkładany na bok" i nie jest już brany pod uwagę w następnych losowaniach. Ponadto, na każdym etapie prawdopodobieństwo wylosowania dowolnego dostępnego elementu jest takie samo dla każdego z nich. Niech $\mathcal{S}_n$ będzie podzbiorem przestrzeni prób $\mathcal{S}$ składającym się z podzbiorów $n$ różnych indeksów. Przy takim schemacie losowania podzbiór $\mathcal{S}_n$ zawiera $\binom{N}{n}$ jednakowo prawdopodobnych elementów. Tak więc plan $P$ odpowiadający schematowi prostego losowania bez zwracania ma postać

$$P(s) = \begin{cases} 1/\binom{N}{n} & \text{dla } s \in \mathcal{S}_n; \\ 0 & \text{w przeciwnym przypadku.} \end{cases}$$

Zauważmy, że możemy łatwo sprawdzić, czy dla tego schematu losowania jest spełniony postulat reprezentatywności orzekający, że każdy element ma szansę znalezienia się w próbie. Oznaczmy przez $\pi_j$ prawdopodobieństwo $P(j \in S)$. Jest to tzw. **prawdopodobieństwo inkluzji** lub zawierania w próbie elementu $j$. Ponieważ spośród $\binom{N}{n}$ elementów $\mathcal{S}_n$ dokładnie $\binom{N-1}{n-1}$ zawiera element $j$, więc wynika stąd, że

$$\pi_j = \binom{N-1}{n-1} \Big/ \binom{N}{n} = n/N > 0 \qquad \text{dla dowolnego } 1 \le j \le n.$$

Podobnie, chcąc określić prawdopodobieństwo, że elementy $j$ i $k$ znajdą się w wybranej próbie, rozpatrujemy wszystkie zdarzenia elementarne z $\mathcal{S}_n$ zawierające te elementy. Ich liczba wynosi $\binom{N-2}{n-2}$. Zatem prawdopodobieństwo inkluzji dla elementów $j$ i $k$

$$\pi_{jk} = P(j, k \in S, j \neq k) = \frac{\binom{N-2}{n-2}}{\binom{N}{n}} = \frac{n(n-1)}{N(N-1)}.$$

**Schemat losowania Bernoulliego.** Przyjmijmy, że przeglądamy po kolei liczby od 1 do $N$ i każdą z nich klasyfikujemy do próby $S$ z pewnym ustalonym prawdopodobieństwem $p$ niezależnie od pozostałych. W tym przypadku liczba elementów $S$ jest losowa (co czasami jest zjawiskiem niepożądanym) i ma rozkład dwumianowy $Bin(n,p)$. Oczywiście w tym przypadku $\pi_j = p$, a $\pi_{ij} = P(i,j \in S) = P(i \in S)P(j \in S) = p^2$ dla $i \neq j$. Możemy również rozważyć **ogólny schemat Bernoulliego**, dla którego prawdopodobieństwo zaklasyfikowania $i$-tego elementu do próby zależy od indeksu tego elementu i wynosi $p_i$. Wówczas $\pi_i = p_i$, a $\pi_{ij} = p_i p_j$ dla $i \neq j$.

**Schemat losowania systematycznego.** Załóżmy, że liczebność populacji $N$ jest liczbą podzielną przez $n$ i $N/n = q > 1$. Niech $I$ będzie zmienną losową o dyskretnym rozkładzie jednostajnym na zbiorze $\{1,2,\ldots,q\}$. Losowa próba $S$ ma postać

$$S = \{I, I+q, \ldots, I+(n-1)q\}.$$

Tak więc do próby klasyfikujemy co $q$-ty element zbioru $\{1,2,\ldots,N\}$ poczynając od losowej liczby $I$. Odpowiadający plan jest planem o stałej liczności $n$. Jeśli przedstawimy dowolną liczbę $j$ w postaci $j = kq + r$, gdzie $1 \leq r \leq q$ i pamiętamy, że $P(I = r) = 1/q$, łatwo zauważymy, że $\pi_j = 1/q > 0$. Ponieważ liczby $j$ i $k$ mogą należeć jednocześnie do próby $S$ wtedy, gdy różnią się o wielokrotność liczby $q$, zatem w takiej sytuacji $\pi_{jk} = 1/q$ i $\pi_{jk} = 0$ w pozostałych przypadkach. Zauważmy, że dla $q > 1$ plan losowania systematycznego nie spełnia postulatu przypisania tego samego prawdopodobieństwa dowolnemu podzbiorowi zbioru $\{1,2,\ldots,N\}$ o liczności $n$, gdyż np. $P(\{1,2,\ldots,n\}) = 0$.

**Schemat prostego losowania ze zwracaniem.** Rozpatrzmy teraz sytuację, gdy obiekty populacji są losowane ze zwracaniem, czyli przy każdym losowaniu dysponujemy pełną pulą elementów, bez względu na to, jakie elementy zostały wylosowane poprzednio. Przyjmijmy ponadto, że w każdym losowaniu wszystkie elementy mają takie samo prawdopodobieństwo zakwalifikowania do próby równe $1/N$. Do opisu losowań ze zwracaniem wygodniej jest rozważyć inną przestrzeń prób niż $\mathcal{S}$. *Przestrzeń prób $\mathcal{S}'$ składa się z* **uporządkowanych** *ciągów indeksów $s = (i_1, i_2, \ldots, i_k)$, gdzie $1 \leq i_j \leq N$ dla $j = 1,2,\ldots,k$, a długość ciągu $k$ może być dowolną liczbą nie przekraczającą liczebności populacji $N$.* Ciąg $(i_1, i_2, \ldots, i_k)$ interpretujemy jako zapis historii losowania: w pierwszym ciągnieniu wybrano obiekt $o_{i_1}$, w drugim $o_{i_2}$, w ostatnim, $k$-tym losowaniu, obiekt $o_{i_k}$. Schemat losowania ze zwracaniem jest raczej rzadko stosowany przez praktyków i nie będziemy się nim bliżej zajmować. Zauważmy jedynie, że jeśli liczba losowanych elementów jest z góry ustalona i równa $n$ i $\mathcal{S}'_n$ jest podzbiorem przestrzeni prób $\mathcal{S}'$ składającym się z ciągów $n$ indeksów, to dla $s \in \mathcal{S}'_n$ mamy $P(s) = 1/N^n$. Ponadto, ponieważ prawdopodobieństwo, że $j$-ty element znajdzie się w próbie $r$ razy

wynosi $\binom{n}{r}N^{-r}(1-1/N)^{n-r}$, więc prawdopodobieństwo

$$\pi_j = 1 - P(j\text{-ty element wybrany 0 razy}) = 1 - \Big(1 - \frac{1}{N}\Big)^n$$

i podobnie dla $i \neq j$

$$\pi_{ij} = 1 - 2\Big(1 - \frac{1}{N}\Big)^n + \Big(1 - \frac{2}{N}\Big)^n,$$

ponieważ $P(A \cap B) = 1 - P(A') - P(B') + P(A' \cap B')$. Niech $Y_i$ oznacza wartość cechy $Y$ dla $i$-tego wylosowanego elementu w schemacie prostego losowania ze zwracaniem. Wówczas zmienna $Y_i$ ma rozkład jednostajny na zbiorze $\{y_1, y_2, \ldots, y_N\}$ i zmienne $Y_1, Y_2, \ldots, Y_n$ tworzą prostą próbę losową. Jak już wspomnieliśmy poprzednio, sytuacja, gdy próba wylosowana według pewnego planu z populacji skończonej tworzy prostą próbę losową jest raczej rzadka.

**Schemat losowania warstwowego.** Rozpatrzmy populację gospodarstw domowych w Polsce. Możemy w niej wyróżnić trzy naturalne warstwy: warstwę mieszkań w bloku, warstwę domów jednorodzinnych nie będących częścią gospodarstwa rolnego i warstwę gospodarstw rolnych. Wyboru reprezentatywnej próby z takiej populacji można dokonać, losując w każdej warstwie pewną liczbę gospodarstw według jednego ze schematów opisanych powyżej (być może innego w każdej warstwie), np. prostego losowania bez zwracania, a następnie łącząc tak wybrane próby w jedną większą próbę. Jak przekonamy się wkrótce, jeśli tylko warstwy są w miarę jednorodne, to tak wybrana próba ma z reguły mniejszą zmienność niż próby wybrane według schematów opisanych powyżej. Takie postępowanie jest również wygodne, jeśli oprócz pewnego parametru globalnego interesuje nas wartość analogicznego parametru dla poszczególnych warstw. Na przykład może nas interesować średnia ilość wody zużywanej miesięcznie przez gospodarstwo domowe, ale często bardziej interesujące jest rozpatrzenie średniej ilości wody zużywanej przez ten okres w rozbiciu na poszczególne warstwy. Nierzadkim przypadkiem jest również sytuacja występująca w omawianym przykładzie, gdy interesująca nas cecha jest silnie skorelowana ze zmienną stratyfikującą, czyli zmienną stanowiącą podstawę do tworzenia warstw. Podobnie, średnia wielkość miesięcznych wydatków na kulturę w rodzinie będzie zależała od wielkości dochodów tej rodziny. Ta ostatnia wielkość może stać się podstawą tworzenia warstw. Ponadto, stosując tak rozumiane losowanie warstwowe, zapewniamy sobie, że każda warstwa będzie reprezentowana w wylosowanej próbie.

Przyjmijmy, że interesuje nas oszacowanie wartości parametru $\theta = f(y_1, \ldots$ $\ldots, y_N)$. Intuicyjnie jest oczywiste, że powinniśmy starać się podzielić populację na takie warstwy (jeśli oczywiście określenie warstw nie jest ustalone

z góry), które są wewnętrznie możliwie jednorodne, a różnią się maksymalnie między sobą. Aby dobrze oszacować wartość parametru $\theta$ w określonej warstwie powinniśmy pobrać z niej stosunkowo dużo elementów, gdy zmienność cechy w tej warstwie jest duża i relatywnie mniej, gdy jej zmienność jest mała. Opiszmy dokładniej przedstawiony schemat. Załóżmy, że populacja $\mathcal{O}$ została podzielona na $k$ rozłącznych warstw $\mathcal{O}_1, \mathcal{O}_2, \ldots, \mathcal{O}_k$ tzn.

$$\bigcup_{i=1}^{k} \mathcal{O}_i = \mathcal{O} \qquad \text{i} \qquad \mathcal{O}_i \cap \mathcal{O}_j = \emptyset \quad \text{dla} \quad i \neq j.$$

Warstwa $\mathcal{O}_h$ zawiera $N_h$ obiektów, spośród których wylosowaliśmy $n_h$ obiektów $o_{h1}, o_{h2}, \ldots, o_{hn_h}$. Oczywiście, $N_1 + N_2 + \ldots + N_k = N$. Wartość $w_h = N_h/N$ nazywamy wagą warstwy o numerze $h$ w populacji. Populację $\mathcal{O}$ możemy traktować jako rodzinę $k$ podpopulacji $\mathcal{O}_1, \mathcal{O}_2, \ldots \mathcal{O}_k$, przy czym znaczenie podpopulacji $i$-tej jest określone przez jej wagę $w_i$. Jeśli przez $\mathcal{S}_i$ oznaczymy przestrzeń zdarzeń elementarnych dla $i$-tej warstwy i przez $P_i$ plan losowania dla niej, to dla planu $P$ losowania warstwowego zachodzi równość

$$P(s) = P_1(s_1)P_2(s_2)\cdots P_k(s_k),$$

gdzie $s \in \mathcal{S}$ jest zdarzeniem składajacym się z wszystkich obiektów będących elementami $s_1, s_2, \ldots, s_k$.

**Schemat losowania wielostopniowego.** Opisaliśmy dotychczas kilka prostych metod próbkowania. Przy ocenie parametrów dużej populacji np. mieszkańców kraju lub województwa z reguły stosuje się kombinację tych metod. Załóżmy, że interesuje nas średnie roczne zużycie energii elektrycznej w miejskim gospodarstwie domowym w Polsce. Możemy postąpić następująco. Podzielmy zbiór miast na warstwy, przyjmując np. jako pierwszą warstwę miasta do 40 tys. mieszkańców, za drugą miasta między 40 a 100 tysięcy, a za ostatnią miasta liczące ponad 100 tysięcy mieszkańców. Używając losowania warstwowego, otrzymujemy reprezentatywną próbę miast. W każdym z wybranych miast, stosując proste losowanie bez zwracania, losujemy pewną liczbę tzw. rejonów statystycznych, tj. jednostek terytorialnych ustalanych na potrzeby spisów powszechnych obejmujących kilka bloków lub ulicę domów jednorodzinnych. Następnie w wybranych rejonach statystycznych wybieramy losowo po kilka gospodarstw domowych i dla nich zbieramy informację o rocznym zużyciu energii. Podstawową zaletą losowania wielostopniowego jest to, że niepotrzebna jest kompletna lista wszystkich obiektów (w naszym przypadku gospodarstw domowych we wszystkich miastach) będących przedmiotem zainteresowania. Wykorzystujemy jedynie listy wszystkich jednostek na każdym etapie, czyli listy miast w każdej warstwie, listy rejonów statystycznych w wylosowanych miastach i listy gospodarstw domowych w wylosowanych rejonach. W efekcie takie postępowanie będzie znacznie mniej kosztowne i pracochłonne.

## 7.2. Estymatory parametrów populacji dla różnych schematów losowania

Załóżmy, że interesuje nas oszacowanie parametru $\theta$ na podstawie losowej próby $S$ otrzymanej przy użyciu planu $P$ na przestrzeni prób $\mathcal{S}$. Rozpatrzmy pewien estymator $\hat{\theta}$ parametru $\theta$. Jak przekonaliśmy się w rozdz. 3 dobroć estymatora oceniamy na podstawie jego rozkładu. W rozpatrywanym przypadku znalezienie rozkładu estymatora $\hat{\theta}$ wydaje się prostsze niż w sytuacji ogólnej. Mianowicie, ponieważ rozkład $P$ jest znany i liczba zdarzeń przestrzeni $\mathcal{S}$ jest skończona, zatem teoretycznie jesteśmy w stanie obliczyć prawą stronę równości

$$P(\hat{\theta} = c) = \sum_{s:\ \hat{\theta}(s)=c} P(s).$$

Problem w tym, że jeśli nawet plan losowania przypisuje dodatnie prawdopodobieństwa tylko części zdarzeń elementarnych, to i tak liczba takich zdarzeń jest z reguły astronomiczna i stwierdzenie, dla których z nich zachodzi równość $\hat{\theta}(s) = c$ jest nierealne. Na przykład dla prostego losowania bez zwracania pięćdziesięciu elementów z populacji 2000 obiektów, liczba zdarzeń elementarnych $s$, dla których $P(s) > 0$ wynosi $\binom{2000}{50}$ i jest równa w przybliżeniu $1{,}996 \cdot 10^{100}$. Ponadto, operację obliczenia prawej strony ostatniego równania trzeba by przeprowadzić dla każdej wartości $c$. O ile więc nie zastosujemy do oceny rozkładu $\hat{\theta}$ metod Monte Carlo opisanych w rozdz. 8, musimy uciec się do oceny parametrów tego rozkładu. Zanim przystąpimy do omówienia tego ostatniego zagadnienia, podkreślmy raz jeszcze, że znajomość rozkładu $P$ umożliwia wykorzystanie go przy konstrukcji estymatorów. W takiej sytuacji rozkład $P$ możemy traktować jako swoisty parametr rozpatrywanego oszacowania, który następnie można dobierać tak, aby uzyskać jak najbardziej korzystne własności estymatora $\hat{\theta}$.

Przeanalizujmy najbardziej standardową sytuację estymacji wartości średniej $\mu$ cechy $Y$ w populacji.

### 7.2.1. Estymator Horwitza–Thompsona wartości średniej cechy

Rozpatrzmy schemat prostego losowania $n$ elementów bez zwracania i niech $S = \{j_1, j_2, \ldots, j_n\}$. Ponieważ wszystkie podzbiory $n$-elementowe przestrzeni prób są jednakowo prawdopodobne, więc próba $\{y_{j_1}, y_{j_2}, \ldots, y_{j_n}\}$ jest z reguły próbą typowych wartości $Y$ i średnia empiryczna elementów próby $\bar{Y} = n^{-1} \sum y_{j_n} = n^{-1} \sum_{i \in S} y_i$ powinna być rozsądnym estymatorem

wartości średniej $\mu = N^{-1}\sum_{i=1}^{N} y_i$. Zauważmy, że ponieważ dla prostego losowania $n$ elementów bez zwracania prawdopodobieństwo $\pi_i = n/N$, estymator $\bar{Y}$ może być przedstawiony następująco:

$$\frac{1}{N}\sum_{i\in S}\frac{Y_i}{\pi_i} = \frac{1}{N}\sum_{1\le i\le n}\frac{y_i}{\pi_i}I_i(S), \tag{7.1}$$

gdzie $I_i(S)$ jest zmienną losową równą 1, gdy $i \in S$ i 0 w przeciwnym przypadku. W przypadku takiego ogólnego planu $P$, że[3] $\pi_i > 0$ dla dowolnego $1 \le i \le N$, estymator (7.1) nazywany jest **estymatorem Horwitza–Thompsona** i będzie oznaczany jako $\bar{Y}_{HT}$. Dla schematu prostego losowania bez zwracania pozostaniemy przy notacji $\bar{Y}$. Zauważmy, że jedynymi składnikami losowymi w definicji estymatora Horwitza–Thompsona są zmienne $I_i(S)$ przyjmujące wartości 0 i 1. Od zbadania ich własności zaczniemy badanie własności estymatora $\bar{Y}_{HT}$.

**Stwierdzenie 7.1.** *Dla dowolnego* $1 \le i \le N$, $EI_i(S) = \pi_i$. *Ponadto,* $\mathrm{Cov}(I_i(S), I_k(S)) = \pi_{ik} - \pi_i\pi_k$, *gdzie* $\pi_{ik} = P(i,k \in S)$.

Pierwsza część stwierdzenia wynika z równości $EI_i(S) = 1 \times P(i \in S) + 0 \times$ $\times P(i \in S) = \pi_i$. Podobnie

$$\mathrm{Cov}(I_i(S), I_k(S)) = EI_i(S)I_k(S) - EI_i(S)EI_k(S) =$$

$$= P(i,k \in S) - P(i \in S)P(j \in S) = \pi_{ik} - \pi_i\pi_k,$$

gdzie dla $i = k$ mamy $\pi_{kk} = \pi_k$. Możemy teraz przekonać się, że estymator $\bar{Y}_{HT}$ jest nieobciążony, znaleźć jego wariancję i nieobciążony estymator wariancji. W dalszym ciągu $\sum_k$ będzie oznaczać $\sum_{1\le k\le N}$ i podobnie $\sum_{k,l} = \sum_{1\le k,l\le N}$. Przypomnijmy, że przez błąd standardowy $SE_{\hat{\theta}}$ rozumiemy oszacowanie odchylenia standardowego $\sigma_{\hat{\theta}}$ estymatora $\hat{\theta}$.

**Twierdzenie 7.1.** *Dla dowolnego planu losowania* $P$ *takiego, że* $\pi_i > 0$ *dla* $1 \le i \le N$,

(1) $E\bar{Y}_{HT} = \mu$;

(2) $\mathrm{Var}(\bar{Y}_{HT}) = \frac{1}{N^2}\sum_{k,l} y_k y_l\left(\frac{\pi_{kl}}{\pi_k\pi_l} - 1\right) = \frac{1}{N^2}\sum_{k,l} y_k y_l \frac{\pi_{kl}}{\pi_k\pi_l} - \mu^2$;

[3] Przypomnijmy, że tylko takie plany uznajemy za prowadzące do wyboru prób reprezentatywnych.

(3) *Jeśli dla planu $P$ zachodzi $\pi_{kl} > 0$ dla wszystkich $k, l$ i*

$$SE^2_{\bar{Y}_{HT}} = \frac{1}{N^2} \sum_{k,l \in S} \frac{y_k y_l}{\pi_{kl}} \Big( \frac{\pi_{kl}}{\pi_k \pi_l} - 1 \Big),$$

*to wówczas* $E(SE^2_{\bar{Y}_{HT}}) = \mathrm{Var}(\bar{Y}_{HT})$.

Część (1) twierdzenia jest oczywista, gdyż na mocy pierwszej części stwierdzenia 7.1

$$E\bar{Y}_{HT} = \frac{1}{N} \sum_i \frac{y_i}{\pi_i} EI_i(S) = \frac{1}{N} \sum_i y_i = \mu.$$

Podobnie na mocy drugiej części stwierdzenia wariancja $\mathrm{Var}(\bar{Y}_{HT})$ jest równa

$$\frac{1}{N^2} \sum_{k,l} \frac{y_k y_l}{\pi_k \pi_l} \mathrm{Cov}(I_k(S), I_l(S)) = \frac{1}{N^2} \sum_{k,l} \frac{y_k y_l}{\pi_k \pi_l} (\pi_{kl} - \pi_k \pi_l) =$$

$$= \frac{1}{N^2} \sum_{k,l} y_k y_l \Big( \frac{\pi_{kl}}{\pi_k \pi_l} - 1 \Big).$$

Punkt (3) jest udowodniany podobnie jak punkt (1).

**Wniosek 7.1.** *Dla planu prostego losowania bez zwracania*

$$\mathrm{Var}(\bar{Y}) = \Big(1 - \frac{n}{N}\Big) \frac{\sigma^2}{n} \tag{7.2}$$

*i*

$$SE^2_{\bar{Y}} = \Big(1 - \frac{n}{N}\Big) \frac{\sigma^2_S}{n},$$

*gdzie* $\sigma^2_S = \frac{1}{n-1} \sum_{i \in S} (Y_i - \bar{Y})^2$.

Niech $\sigma_{kl} = \pi_{kl} - \pi_k \pi_l$. Zauważmy, że ze wzorów na prawdopodobieństwa inkluzji dla prostego losowania bez zwracania otrzymujemy

$$\sigma_{kk} = \pi_k (1 - \pi_k) = \frac{n}{N} - \frac{n^2}{N^2} = \frac{n(N-n)}{N^2}$$

oraz dla $k \neq l$

$$\sigma_{kl} = \pi_{kl} - \pi_k \pi_l = \frac{n(n-1)}{N(N-1)} - \frac{n^2}{N^2} = -\frac{n}{N} \Big( \frac{N-n}{N} \Big) \frac{1}{N-1}. \tag{7.3}$$

Zatem z tw. 7.1 (2) i faktu, że $\pi_i = n/N$ wynika, że

$$\mathrm{Var}(\bar{Y}) = \tfrac{1}{n^2} \{ \textstyle\sum_k \frac{n(N-n)}{N^2} y_k^2 - \sum_{k \neq l} \frac{n(N-n)}{N^2} \frac{1}{N-1} y_k y_l \} =$$

$$= \frac{1}{n}\frac{(N-n)}{N}\frac{1}{N-1}\{\frac{N-1}{N}\sum_k y_k^2 - \frac{1}{N}\sum_{k\neq l} y_k y_l\} =$$

$$= \frac{1}{n}\frac{(N-n)}{N}\frac{1}{N-1}\{\sum_k y_k^2 - N\mu^2\} = (1-\frac{n}{N})\frac{\sigma^2}{n},$$

gdzie ostatnia równość wynika z faktu, że $\sigma^2 = (N-1)^{-1}(\sum_k y_k^2 - N\mu^2)$. Podobnie uzasadnia się drugą część wniosku.

**Przykład 7.1.** W mieście liczącym 1100 mieszkańców w wieku ponad 6 lat postanowiono ocenić średni łączny czas spędzany w ciągu tygodnia na oglądaniu telewizji. W tym celu z populacji mieszkańców mających ponad 6 lat wylosowano za pomocą prostego schematu bez zwracania 110 mieszkańców i poproszono ich o zapisanie ilości godzin przeznaczonych na oglądanie telewizji każdego z 7 następnych dni. Niech $y_i$ oznacza całkowity czas poświęcony na oglądanie telewizji przez $i$-tego wylosowanego mieszkańca w ciągu 7 dni w tak uzyskanej próbie. Otrzymano następujące wyniki zbiorcze: $\sum_{i=1}^{110} y_i = 2393$ i $\sum_{i=1}^{110} y_i^2 = 53017$. Zatem wartość estymatora wartości średniej wynosi $\bar{y} = 2393/110 = 21{,}75$ godziny i $\sigma_s^2 = (109)^{-1}\sum_{i=1}^{110}(y_i - 21{,}75)^2 = = (109)^{-1}(53017 - 110\times(21{,}75)^2) = 8{,}99$. Tak więc $SE_{\bar{Y}} = ((1- -1/10)\times 8{,}99/110)^{1/2} = 0{,}27$ godziny, czyli około 16 minut.

Współczynnik $(1-n/N)$ występujący we wzorze na wariancję estymatora Horwitza–Thompsona dla schematu prostego losowania bez zwracania jest zwany **poprawką na bezzwrotność losowania**. Terminologia ta jest związana z następującym rozumowaniem. Jeśli rozpatrzymy proste losowanie $n$ elementów ze zwracaniem i zdefiniujemy estymator $\bar{Y} = n^{-1}\sum_{i=1}^n Y_i$ jako średnią cechy w próbie, to formalnie postać $\bar{Y}$ jest identyczna z postacią estymatora Horwitza–Thompsona dla prostego losowania bez zwracania. Jedyną różnicą jest to, że w pierwszym przypadku składniki $Y_i$ w sumie $\sum Y_i$ mogą się powtarzać. Ponieważ zmienne $Y_1, Y_2, \ldots, Y_n$ tworzą prostą próbę losową i $\mathrm{Var}Y_i = \sigma^2$, więc $\mathrm{Var}\bar{Y} = \tilde{\sigma}^2/n$, gdzie $\tilde{\sigma}^2 = N^{-1}\sum_{i=1}^N\sum(y_i-\mu)^2$. Zatem pomijając zaniedbywalną dla dużych $N$ różnicę między $\sigma^2$ a $\tilde{\sigma}^2$, wariancja estymatora Horwitza–Thompsona w przypadku losowania bezzwrotnego jest w przybliżeniu **mniejsza** o czynnik $(1-n/N)$ w stosunku do wariancji analogicznego estymatora dla losowania ze zwracaniem. Wynika stąd, że schemat prostego losowania ze zwracaniem prowadzi z reguły do bardziej zmiennych estymatorów niż ich odpowiedniki w schemacie prostego losowania bez zwracania.

Podamy bez dowodu jeszcze jedną postać wzoru na wariancję estymatora Horwitza–Thompsona dla planu o *ustalonej liczności* próby oraz jej alternatywny nieobciążony estymator (tzw. estymator Yates'a–Grundy'ego)

**Wniosek 7.2.** *Dla planu o ustalonej liczności próby i takiego, że* $\pi_i > 0$ *dla* $1 \leq i \leq n$

$$\mathrm{Var}(\bar{Y}_{HT}) = -1/2 \sum_{k,l} \sigma_{kl} \Big( \frac{y_k}{\pi_k} - \frac{y_l}{\pi_l} \Big)^2, \tag{7.4}$$

*gdzie* $\sigma_{kl} = \pi_{kl} - \pi_k \pi_l$. *Estymator wariancji postaci*

$$\widetilde{SE}^2_{\bar{Y}_{HT}} = -1/2 \sum_{k,l \in S} \frac{\sigma_{kl}}{\pi_{kl}} \Big( \frac{y_k}{\pi_k} - \frac{y_l}{\pi_l} \Big)^2 \tag{7.5}$$

*jest nieobciążony.*

Estymatory $SE^2_{\bar{Y}_{HT}}$ i $\widetilde{SE}^2_{\bar{Y}_{HT}}$ nie są na ogół identyczne. W przypadku planu losowania, dla którego $\sigma_{kl} \leq 0$ estymator wariancji (7.5) jest nieujemny. Na podstawie wzoru (7.3) stwierdzamy, że tak jest dla planu prostego losowania bez zwracania. Wzory na odchylenie standardowe estymatora Horwitza–Thompsona podane w tw. 7.1 (2) i wniosku 7.2 nie mają dużego praktycznego znaczenia, gdyż obliczanie występujących w nich podwójnych sum jest zbyt pracochłonne. Na przykład w przypadku, gdy próba składa się z 1000 obiektów, suma $\sum_{1 \leq k,\ l \leq 1000}$ zawiera milion składników. Jednakże w przypadku konkretnego schematu losowania, jak np. dla prostego losowania bez zwracania, ogólne postaci wzorów redukują się do prostszych wyrażeń, które można już bez trudu stosować (porównaj wniosek 7.1).

Zauważmy ponadto, że gdybyśmy byli w stanie podać taki plan pobierania prób o *stałej liczności*, dla którego prawdopodobieństwa $\pi_i$ byłyby proporcjonalne do wartości $y_i$, to ze wzoru (7.4) wynika, że wariancja estymatora Horwitza–Thompsona byłaby równa 0. Zauważmy, że podanie takiego planu wymagałoby znajomości wszystkich wartości cechy, a w konsekwencji również wartości $\mu = N^{-1} \sum_{i=1}^{N} y_i$. Wtedy jednak zadanie szacowania parametru $\mu$ byłoby pozbawione sensu. Jednakże wzór (7.4) sugeruje również, że jeśli cecha $X$ jest silnie zależna od cechy $Y$ i znamy jej wartości dla wszystkich elementów populacji, to wybór planu takiego, że $\pi_i$ jest proporcjonalne do $x_i$ przypuszczalnie zmniejszyłoby zmienność estymatora $\bar{Y}_{HT}$. Planem takim może być np. ogólny plan Bernoulliego, dla którego oczekiwana liczba elementów pobranych do próby jest równa wybranej liczności $n$. Do tego problemu wrócimy w p. 7.2.5.

### 7.2.2. Przedział ufności dla wartości średniej cechy

W sytuacji, gdy rozkład estymatora Horwitza–Thompsona jest w przybliżeniu normalny i błąd standardowy $SE_{\bar{Y}_{HT}}$ jest zadowalającym przybliżeniem jego rozproszenia $\sigma_{\bar{Y}_{HT}}$, to przedział

$$(\bar{Y}_{HT} - z_{1-\alpha/2} SE_{\bar{Y}_{HT}}, \bar{Y}_{HT} + z_{1-\alpha/2} SE_{\bar{Y}_{HT}}) \tag{7.6}$$

jest przedziałem ufności na poziomie ufności w przybliżeniu równym $1-\alpha$. Sytuacja jest o tyle bardziej skomplikowana w porównaniu z opisaną w rozdz. 3, że estymator Horwitza-Thompsona nie jest z reguły oparty na prostej próbie losowej i znacznie trudniej jest sprawdzić powyższe dwa założenia, na których bazuje konstrukcja przedziału ufności. Innymi słowy, kwestia stosowalności przedziału ufności (7.6) zależy nie tylko od liczności próby i od rozkładu wartości cechy w populacji, ale również od zastosowanego schematu losowania. Okazuje się jednak, że z reguły zależność między cechami różnych elementów próby jest słaba i przybliżenie normalne dla estymatora Horwitza–Thompsona faktycznie zachodzi przy spełnieniu analogicznego warunku do warunku zakładanego w Centralnym Twierdzeniu Granicznym dla niezależnych, ale niekoniecznie mających ten sam rozkład zmiennych losowych (tzw. warunku Lindeberga). Sformułujemy tutaj twierdzenie o przybliżeniu normalnym jedynie dla prostego losowania bez zwracania, przedstawiając wspomniany warunek w postaci uproszczonej. Formalnie, twierdzenie to dotyczy ciągu populacji skończonych o liczności $N$ rosnącej do nieskończoności. Z każdej z populacji losujemy prostą próbę losową bez zwracania o liczności $n$ również dążącej do nieskończoności i na jej podstawie obliczamy estymator $\bar{Y}_N = \bar{Y}$. Ta uwaga odnosi się do wszystkich wyników asymptotycznych przedstawionych w tym rozdziale.

**Twierdzenie 7.2.** (Centralne Twierdzenie Graniczne dla estymatora Horwitza–Thompsona w przypadku prostego losowania bez zwracania) *Załóżmy, że*

$$\max_{1\le i\le N} \frac{(y_i-\mu)^2}{\sum_{i=1}^N (y_i-\mu)^2} \to 0, \qquad gdy \quad N\to\infty \tag{7.7}$$

*oraz, że stosunek liczności próby do liczności populacji $n/N$ jest oddzielony od 0, tzn. że dla dużych $N$ stosunek $n/N$ jest zawarty w przedziale $(\epsilon, 1-\epsilon)$ dla pewnego $\epsilon < 1/2$. Wówczas przy $N\to\infty$*

$$P\left(\frac{\bar{Y}_N - E\bar{Y}_N}{\sqrt{\mathrm{Var}(\bar{Y}_N)}} \le x\right) \quad \to \quad \Phi(x)$$

*dla każdego $x$.*

Warunek (7.7) mówi o braku bardzo dużych odstępstw wartości cechy $Y$ od wartości średniej. Analogiczny do tw. 7.1 rezultat jest prawdziwy dla estymatora $N\bar{Y}_N$ wartości globalnej $N\mu = \sum_{i=1}^{N} y_i$. Oczywiście, ponieważ powyższy wynik nie dotyczy prostej próby losowej, trudno jest podać uniwersalną prostą regułę dotyczącą liczby elementów próby $n$, przy której powyższe twierdzenie można już stosować. Rozpatruje się w tym celu różnorakie reguły heurystyczne, których nie będziemy tutaj omawiali, zauważając jedynie, że podobnie jak w przypadku Centralnego Twierdzenia Granicznego dla prostych prób losowych, w sytuacji skośności rozkładu cechy $Y$ wskazane jest rozpatrzenie prób o większej liczności. Jednocześnie zauważmy, że stosowalność tw. 7.2 do estymatora średniej dla pewnej liczności próby możemy sprawdzić, stosując metodę Monte Carlo opisaną w następnym rozdziale: na podstawie wielu prób o tej liczności wylosowanych z populacji według pewnego schematu losowania obliczamy odpowiadające wartości estymatora Horwitza–Thompsona i sprawdzamy, czy jego rozkład jest rzeczywiście w przybliżeniu normalny.

Centralne Twierdzenie Graniczne umożliwia również wprowadzenie reguły wyboru minimalnej liczby elementów próby, o tej własności, że dla danej precyzji $d$ i poziomu ufności $1-\alpha$ w przybliżeniu zachodzi $P(|\hat{Y}_{HT} - \bar{Y}| \leq d) \geq$ $\geq 1-\alpha$. Twierdzenie 7.1 sugeruje, że precyzja $d$ powinna spełniać nierówność $d \geq z_{1-\alpha/2}\sqrt{n^{-1}(1-n/N)}\sigma$, co pociąga fakt, że $n$ musi być nie mniejsze od takiej liczby $n_0$, że

$$n_0 = \frac{1}{\frac{1}{N} + \frac{d^2}{z^2_{1-\alpha/2}\sigma^2}}.$$

Ponieważ nie znamy wartości $\sigma^2$, musimy ją estymować, używając w tym celu $\sigma_S^2$ lub oszacowania otrzymanego w uprzednich badaniach. Może to prowadzić do znacznej zmienności liczby $n_0$.

**Przykład 7.2.** Spośród 400 skrzyń zawierających taką samą liczbę zaworów przygotowanych do wysyłki wylosowano 20 skrzyń i znajdujące się w nich zawory poddano szczegółowej kontroli. W rezultacie wykryto następujące liczby braków:

0, 4, 3, 6, 0, 2, 1, 1, 4, 3, 5, 2, 3, 1, 2, 5, 4, 2, 1, 0.

Zauważmy, że w tym przypadku naszą populacją jest zbiór skrzyń, a interesującą nas cechą jest liczba braków w każdej skrzyni. Skonstruujmy 95% przedział ufności dla całkowitej liczby braków $N\mu$. W naszym przypadku $\bar{y} = 2{,}45$ i $\sigma_s^2 = 3{,}21$. Z faktu, że przedział ufności dla wartości $N\mu$ otrzymujemy przez pomnożenie końców przedziału w (7.6) przez $N$, wynika, że

przedział ufności ma postać

$$N\bar{y} \pm N \times z_{0,025}(3{,}21)^{1/2}(\frac{1-0{,}05}{20})^{1/2}.$$

Ponieważ $z_{0,025} = 1{,}96$, otrzymujemy stąd następujący przedział ufności na przybliżonym poziomie ufności 0,95

$$(980-400\times 1{,}96\times 1{,}79\times 0{,}218,\ \ 980+400\times 1{,}96\times 1{,}79\times 0{,}218) = (674, 1286).$$

### 7.2.3. Estymatory wartości średniej cechy oparte na cesze dodatkowej

Często jest tak, że starając się oszacować parametr cechy $Y$ dysponujemy wiedzą na temat wartości silnie skorelowanej z nią cechy $X$. Na przykład starając się oszacować średnie wydatki Polaków na zdrowie w pewnej grupie wiekowej, możemy dysponować wiedzą na temat dochodów członków tak określonej populacji. Taką wiedzę możemy wykorzystać do zwiększenia precyzji estymatorów jeszcze w inny sposób niż przez zastosowanie planu, dla którego prawdopodobieństwa $\pi_i$ są proporcjonalne do wartości $x_i$, o czym wspominaliśmy poprzednio. Przyjmiemy tutaj, że wartości cechy dodatkowej $X$ są znane przed losowaniem dla wszystkich $N$ członków populacji. Oznaczmy przez $\mu_Y$ wartość średnią cechy $Y$ i analogicznie przez $\mu_X$ znaną wartość średnią cechy dodatkowej $X$. **Estymator różnicy** (a właściwiej estymator wartości średniej oparty na estymatorze różnicy) wykorzystuje równość $\mu_Y = \mu_X + (\mu_Y - \mu_X) = \mu_X + \mu_{Y-X}$, gdzie $\mu_{Y-X}$ jest wartością średnią nowej cechy $D$ przyjmującej dla $i$-tego obiektu wartość $d_i = y_i - x_i$. Ponieważ wartość średnia $\mu_X$ jest znana, aby oszacować wartość średnią $\mu_Y$ wystarczy oszacować wartość średnią $\mu_{Y-X}$.Takie postępowanie ma oczywiście sens wtedy, gdy cechy $Y$ i $X$ są dodatnio zależne i możemy spodziewać się, że zmienność cechy $D$ będzie mniejsza od zmienności cechy $Y$. Wykorzystując do estymacji wartości średniej cechy $D$ estymator Horwitza–Thompsona otrzymujemy estymator różnicy

$$\bar{Y}_D = \mu_X + \bar{D}_{HT},$$

gdzie $\bar{D}_{HT} = N^{-1}\sum_{i\in S}(y_i - x_i)/\pi_i$. Zauważmy, że do obliczenia wartości tego estymatora wystarczająca jest znajomość wartości cechy $X$ w próbie (a nie w całej populacji) i wartości średniej $\mu_X$. Oczywiście z nieobciążoności estymatora Horwitza–Thompsona wynika nieobciążoność estymatora różnicy. Postać wariancji estymatora różnicy wynika natychmiast z tw. 7.1(2). Jest ona równa $N^{-2}\sum_{k,l} d_k d_l((\pi_{kl}/\pi_k\pi_l) - 1)$, a jej nieobciążony estymator jest równy $N^{-2}\sum_{k,l\in S}\pi_{kl}^{-1} d_k d_l((\pi_{kl}/\pi_k\pi_l) - 1)$. Dla schematu prostego

losowania bez zwracania otrzymujemy stąd następującą postać wariancji $\bar{Y}_D$

$$\mathrm{Var}(\bar{Y}_D) = \mathrm{Var}(\bar{D}) = \Big(1 - \frac{n}{N}\Big)\frac{\sigma_D^2}{n} = \Big(1 - \frac{n}{N}\Big)\frac{\{\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}\}}{n},$$

gdzie $\sigma_{XY} = (N-1)^{-1}\sum_{i=1}^{N}(x_i - \mu_X)(y_i - \mu_Y)$. Wariancja estymatora $\bar{Y}_D$ jest estymowana za pomocą wielkości

$$(1 - n/N)\sigma_{D,S}^2/n,$$

gdzie $\sigma_{D,S}^2$ oznacza wariancję próbkową dla cechy $D$. Jeśli współczynnik korelacji $r = \sigma_{XY}/\sigma_X\sigma_Y$ jest duży, to precyzja estymatora różnicy może być większa od precyzji estymatora Horwitza–Thompsona. W punkcie 7.2.5 omówimy wykorzystanie analogicznej idei do konstrukcji estymatora ilorazowego.

**Estymator regresyjny.** W celu dalszego zmniejszenia zmienności estymatora różnicy można go zmodyfikować następująco. Zapiszmy estymator różnicy w postaci

$$\bar{Y}_D = \bar{Y}_{HT} + (\mu_X - \bar{X}_{HT})$$

i zastąpmy cechę $X$ cechą $X' = \beta_1 X$, gdzie $\beta_1$ jest dowolną stałą. Oczywiście, estymator różnicy na podstawie cechy dodatkowej $X'$ ma postać

$$\bar{Y}_D = \bar{Y}_{HT} + \beta_1(\mu_X - \bar{X}_{HT}).$$

Minimalizując wariancję estymatora $\bar{Y}_D$ względem $\beta_1$, otrzymamy

$$\beta_1 = \frac{\sigma_{XY}}{\sigma_X^2} = \frac{\sum_{i=1}^{N}(x_i - \mu_X)(y_i - \mu_Y)}{\sum_{i=1}^{N}(x_i - \mu_X)^2}. \tag{7.8}$$

Ponieważ parametr $\beta_1$ jest nieznany, więc zastępujemy go jego estymatorem

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{X}_{HT})(y_i - \bar{Y}_{HT})}{\sum_{i=1}^{N}(x_i - \bar{X}_{HT})^2}. \tag{7.9}$$

Estymator $\bar{Y}_D$ z $\beta_1$ zastąpionym estymatorem $b_1$ nazywamy **estymatorem regresyjnym** i będziemy oznaczali $\bar{Y}_{lr}$. Zauważmy, że nie jest on niczym innym jak wartością prostej regresji dla danych $\{(x_i, y_i)\}_{i\in S}$ przesuniętej tak, aby przechodziła przez punkt $(\bar{X}_{HT}, \bar{Y}_{HT})$, obliczoną dla wartości zmiennej objaśniającej równej $\mu_X$. W przypadku zastosowania schematu losowania bez zwracania przesunięcie jest równe 0, gdyż w tej sytuacji prosta regresji przechodzi przez punkt $(\bar{X}_{HT}, \bar{Y}_{HT}) = (\bar{X}, \bar{Y})$. Mamy następujące stwierdzenie.

**Stwierdzenie 7.2.** *Dla prostego losowania bez zwracania i $n \to \infty$ mamy*

$$E\bar{Y}_{lr} \approx \mu_Y$$

*i*

$$\mathrm{Var}(\bar{Y}_{rl}) \approx (1 - n/N)\sigma_\epsilon^2/n,$$

*gdzie* $\sigma_\epsilon^2 = (N-1)^{-1}\sum_{i=1}^N (y_i - \beta_0 - \beta_1 x_i)^2$ *i* $\beta_0 = \mu_Y - \beta_1\mu_X$. $x_n \approx y_n$ *oznacza, że* $x_n/y_n \to 1$ *przy* $n \to \infty$. *Estymator wariancji otrzymamy, zastępując* $\sigma_\epsilon^2$ *w ostatnim wzorze estymatorem* $\hat{\sigma}_\epsilon^2 = (n-1)^{-1}\sum_{i=1}^n (y_i - -b_0 - b_1 x_i)^2$, *gdzie* $b_1$ *jest zdefiniowane we wzorze (7.9) i* $b_0 = \bar{Y} - b_1\bar{X}$.

Stwierdzenia tego nie będziemy tutaj dowodzić. Zauważmy, że wprowadzenie estymatora regresyjnego sugeruje, że jego wariancja powinna być nie większa od wariancji estymatora opartego na estymatorze różnicy. W punkcie 7.2.5 porównamy zachowanie się estymatora ilorazowego z estymatorem Horwitza–Thompsona i wprowadzonym tam estymatorem ilorazowym średniej, będącym również przykładem estymatora opartego na cesze dodatkowej. Ponieważ zarówno estymator regresyjny, jak i estymator ilorazowy są obciążone, właściwą miarą do ich porównania jest błąd średniokwadratowy określony w def. 2.25. Jednakże okazuje się, że w obydwu przypadkach wielkość kwadratu obciążenia tych estymatorów można pominąć w porównaniu z ich wariancją. Zatem na mocy stwierdzenia 2.11 porównanie błędów średniokwadratowych sprowadza się do porównania wariancji.

**Przykład 7.3.** Obszar eksperymentalny został podzielony na $N = 50$ kwadratów i w każdym z nich, przy zastosowaniu pewnej ilości nawozu sztucznego ($x$) otrzymano pewną wielkość zbioru ($y$); obie wartości są podane w kilogramach. Spośród wszystkich kwadratów, stosując schemat prostego losowania bez zwracania, wybrano 5, dla których zapisano zarówno wielkość plonu, jak i ilość zastosowanego nawozu. Wiadomo, że średnia wartość zastosowanego nawozu wynosiła $\mu_X = 50$ kg. Wyniki są przedstawione w tab. 7.1.

**Tabela 7.1. Wielkość plonu i ilość zastosowanego nawozu dla 5 kwadratów wylosowanych spośród 50 kwadratów eksperymentalnych**

| | | | | | |
|---|---|---|---|---|---|
| $x$: ilość nawozu sztucznego | 20 | 40 | 60 | 80 | 100 |
| $y$: wielkość plonu | 621 | 712 | 793 | 937 | 1007 |

Ponieważ wartość estymatora $b_1$ wielkości $\beta_1$ otrzymana jako wartość nachylenia prostej MNK dla zmiennej $y$ względem zmiennej $x$ wynosi 4,98, więc

wartość estymatora regresyjnego jest równa $\bar{y} + b_1(\mu_X - \bar{x}) = 814 + 4{,}98\times$ $\times(50 - 60) = 764{,}20$. Podobnie wartość $MSE = (n-2)^{-1}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ dla dopasowanego modelu wynosi 357,2, a zatem wartość błędu standardowego estymatora $\bar{Y}_{lr}$ jest równa $(0{,}9 \times (3/4)/5)^{1/2} \times (MSE)^{1/2} = 6{,}94$. W podobny sposób otrzymujemy, że wartość estymatora różnicy wynosi $\bar{y}_D = 50 + \bar{d} = 50 + 754 = 804$, a błąd standardowy tego estymatora wynosi $(0{,}9/5)^{1/2}\sigma_{D,s} = 53{,}9$, gdzie $\sigma^2_{D,s}$ jest wariancją próbkową dla różnic $y_i - x_i$, $i = 1, ..., 5$. Dla porównania obliczmy wartość $\bar{y}$. Wynosi ona $\bar{y} = 814$, a błąd standardowy średniej na podstawie wniosku 7.1 jest równy 67,23.

### 7.2.4. Estymator proporcji

Rozpatrzmy ponownie problem szacowania proporcji elementów populacji mających interesującą nas własność np. proporcji rodzin zamieszkujących w małym mieście i posiadających samochód, w populacji wszystkich rodzin zamieszkujących w tym mieście. W przeciwieństwie do p. 2.4.3 nie chcemy zakładać teraz, że rozpatrywana populacja jest tak duża, że praktycznie możemy uznać ją za nieskończoną. Niech

$$p = (\text{liczba obiektów mających rozpatrywaną własność})/N$$

będzie wartością proporcji i niech $\hat{p}$ będzie odpowiednią częstością elementów z rozpatrywaną własnością w próbie $n$-elementowej otrzymanej w schemacie prostego losowania bez zwracania. Zauważmy, że jeśli przez $y_i$ oznaczymy cechę przyjmującą wartość 1, gdy element $o_i$ ma rozpatrywaną własność i 0 w przeciwnym przypadku, to $p = N^{-1}\sum_{i=1}^{N} y_i = \mu_Y$. Ponadto, częstość $\hat{p}$ jest estymatorem Horwitza–Thompsona średniej $\mu_Y = p$. Zatem estymacja proporcji jest szczególnym przypadkiem estymacji wartości średniej. Wynika z tego natychmiast, że częstość $\hat{p}$ jest nieobciążonym estymatorem proporcji. Obliczmy wariancję częstości $\hat{p}$. Oznaczmy przez $A$ zbiór indeksów obiektów populacji mających rozpatrywaną własność. Wtedy proporcja $p$ jest równa liczbie elementów zbioru $A$ podzielonej przez $N$. Zatem

$$\frac{(N-1)}{N}\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \mu)^2 = \frac{1}{N}\Big(\sum_{i\in A}(1-p)^2 + \sum_{i\in A'}(0-p)^2\Big) =$$

$$= p(1-p)^2 + (1-p)p^2 = p(1-p).$$

Zatem na podstawie wniosku 7.1 otrzymujemy następujące stwierdzenie.

**Stwierdzenie 7.3.** *Dla schematu prostego losowania bez zwracania*

$$\mathrm{Var}(\hat{p}) = \frac{N-n}{N-1}\frac{p(1-p)}{n} \tag{7.10}$$

*i*

$$SE^2_{\hat{p}} = \Big(1 - \frac{n}{N}\Big)\frac{\hat{p}(1-\hat{p})}{n-1}. \tag{7.11}$$

Zastępując w mianowniku (7.10) $N-1$ przez $N$ stwierdzamy, że wariancja $\hat{p}$ jest w przybliżeniu równa $\big(1 - n/N\big)p(1-p)/n$. Zatem różni się ona w przybliżeniu o poprawkę na bezzwrotność losowania od wariancji częstości cechy obliczonej na postawie prostej próby losowej z populacji podanej w stwierdzeniu 2.10. Ponadto, z postaci (7.6) przedziału ufności dla wartości średniej wynika, że przedział

$$(\hat{p} - z_{1-\alpha/2}SE_{\hat{p}}, \hat{p} + z_{1-\alpha/2}SE_{\hat{p}})$$

jest przedziałem ufności dla proporcji $p$ na poziomie w przybliżeniu równym $1-\alpha$.

**Przykład 7.4.** W pewnej szkole technicznej, w której studiuje 2000 studentów, przeprowadzono ankietę, której jedno z pytań dotyczyło przeczytania co najmniej jednej książki nietechnicznej w ciągu ostatnich dwóch miesięcy. Spośród 100 studentów szkoły wybranych według schematu prostego losowania bez zwracania 23 odpowiedziało na to pytanie twierdząco, a 77 przecząco.

Znajdźmy przedział ufności na poziomie 95% dla proporcji $p$ studentów tej szkoły, którzy przeczytali przynajmniej jedną książkę nietechniczną w ciągu ostatnich dwóch miesięcy. Oczywiście, $\hat{p} = 23/100 = 0{,}23$, a odpowiedni przedział ufności wynosi

$$\Big(0{,}23 - 1{,}96 \times \Big((1-\tfrac{1}{20})(\tfrac{0{,}23\times 0{,}77}{99})\Big)^{1/2},$$

$$0{,}23 + 1{,}96 \times \Big((1-\tfrac{1}{20})(\tfrac{0{,}23\times 0{,}77}{99})\Big)^{1/2}\Big) =$$

$$= (0{,}149\,,\ 0{,}311).$$

Przedział ufności obliczony przy założeniu, że populację studentów możemy traktować jako nieskończoną, wynosi na podstawie wzoru (3.34) (0,148, 0,312), zatem niewiele różni się od przedziału otrzymanego wyżej. Różnica między tymi przedziałami staje się jednak większa, gdy rośnie stosunek $n/N$.

Załóżmy przykładowo, że takie same wyniki otrzymano ankietując stuelementową próbę spośród całkowitej liczby $N = 200$ studentów pewnej innej szkoły. W tej sytuacji $n/N = 0{,}5$ (w porównaniu z pierwotnym stosunkiem równym 0,05) i przedział ufności ma postać $(0{,}23 - 0{,}058\,,\ 0{,}23 + 0{,}058) = (0{,}172\,,\ 0{,}281)$.

### 7.2.5. Estymacja ilorazu wartości średnich

Rozpatrzmy jeszcze problem estymacji ilorazu dwóch wartości średnich cech $X$ i $Y$ określonych dla wszystkich elementów populacji. Niech

$$R = \frac{\mu_Y}{\mu_X} = \frac{\frac{1}{N}\sum_{i=1}^{N} Y_i}{\frac{1}{N}\sum_{i=1}^{N} X_i}, \tag{7.12}$$

gdzie $X_i$ i $Y_i$ jest odpowiednio wartością cechy $X$ i $Y$ dla $i$-tego elementu populacji. Przykładem sytuacji, w której może interesować nas estymacja ilorazu $R$ jest ocena średnich wydatków na kulturę w kwietniu 2001 r. na jednego członka gospodarstwa domowego. Oznaczając przez $Y_i$ wydatki na kulturę w tym miesiącu w $i$-tym gospodarstwie domowym, a przez $X_i$ liczbę członków tego gospodarstwa, widzimy, że interesująca nas wielkość ma postać ilorazu $R$ zdefiniowanego w (7.12). Rozpatrzymy jedynie schemat prostego losowania $n$ elementów bez zwracania. Na podstawie próby konstruujemy estymatory Horwitza–Thompsona $\bar{Y}$ i $\bar{X}$ dla, odpowiednio, wartości średniej $\mu_Y$ i $\mu_X$. Naturalnym estymatorem ilorazu $R$ jest iloraz odpowiednich estymatorów Horwitza–Thompsona $\hat{R} = \frac{\bar{Y}}{\bar{X}}$. W następnym stwierdzeniu przedstawiamy przybliżone wzory na wartość oczekiwaną i wariancję estymatora $\hat{R}$. Niech $f = n/N$ będzie stosunkiem liczności próby do liczności populacji.

**Stwierdzenie 7.4.** *Dla prostego losowania bez zwracania i $n \to \infty$ mamy*

$$E\hat{R} \approx R$$

*i*

$$\mathrm{Var}(\hat{R}) \approx \frac{1-f}{(N-1)n(\mu_X)^2}\sum_{i=1}^{N}(y_i - Rx_i)^2,$$

*gdzie $w_n \approx z_n$ oznacza, że iloraz $w_n/z_n \to 1$, gdy $n \to \infty$.*

Dowód tego stwierdzenia przedstawimy w zarysie. Opiera się on na analizie wyrazów liniowych rozwinięcia Taylora ilorazu definiującego $\hat{R}$. Mianowicie oznaczając przez $h(y, x) = y/x$, mamy

$$\hat{R} = h(\bar{Y}, \bar{X}) =$$

$$= h(\mu_Y, \mu_X) + \frac{\partial h(\mu_Y, \mu_X)}{\partial x}(\bar{X} - \mu_X) + \frac{\partial h(\mu_Y, \mu_X)}{\partial y}(\bar{Y} - \mu_Y) + w_n =$$

$$= R - \frac{R}{\mu_X}(\bar{X} - \mu_X) + \frac{1}{\mu_X}(\bar{Y} - \mu_Y) + w_n, \tag{7.13}$$

gdzie $w_n$ jest resztą we wzorze Taylora. Oznaczając przez $\tilde{R}$ wielkość $\hat{R} - w_n$ udowodnimy, że wartość oczekiwana i wariancja wielkości $\tilde{R}$ są opisane wyrażeniami podanymi w stwierdzeniu. Nie będziemy tutaj dowodzić, że reszta $w_n$ ma pomijalny wpływ na wartość oczekiwaną i wariancję estymatora $\hat{R}$. Na podstawie wzoru (7.13) otrzymujemy

$$\tilde{R} - R = \frac{\bar{Y} - \mu_Y - R(\bar{X} - \mu_X)}{\mu_X} = \frac{\bar{Y} - R\bar{X}}{\mu_X}.$$

Wartość oczekiwana wyrażenia po prawej stronie jest oczywiście równa 0, gdyż estymator Horwitza–Thompsona jest nieobciążonym estymatorem średniej. Zauważmy dalej, że $\bar{Y} - R\bar{X} = \bar{D}$, gdzie $\bar{D}$ jest średnią próbkową dla cechy $D = Y - RX$, której wartość dla $i$-tego elementu populacji wynosi $D_i = Y_i - RX_i$. Na podstawie wniosku 7.1 otrzymujemy

$$\mathrm{Var}(\tilde{R}) = \frac{1}{\mu_X^2}\frac{\sigma_D^2(1-f)}{n},$$

gdzie $\tilde{\sigma}_D^2 = 1/(N-1)\sum_{i=1}^{N}(D_i)^2$, gdyż średnia cechy $D$ w populacji wynosi 0.

Na mocy stwierdzenia naturalnym kandydatem na błąd standardowy estymatora $\hat{R}$ jest pierwiastek z wyrażenia

$$SE_{\hat{R}}^2 = \frac{1-f}{n(\bar{X})^2} \quad \frac{\sum_{i\in S}(Y_i - \hat{R}X_i)^2}{(n-1)} =$$

$$= \frac{1-f}{n(\bar{X})^2} \quad \frac{(\sum_{i\in S} Y_i^2 - 2\hat{R}\sum_{i\in S} X_iY_i + \hat{R}^2\sum_{i\in S} X_i^2)}{(n-1)}. \tag{7.14}$$

Zauważmy, że ponieważ na podstawie tw. 7.1 rozkłady $\bar{X}$ i $\bar{Y}$ są w przybliżeniu normalne, na podstawie równości (7.13) możemy się spodziewać, że rozkład $\hat{R}$ będzie też w przybliżeniu normalny, o ile tylko wpływ reszty $w_n$ można pominąć. Wtedy przedział ufności $(\hat{R} - z_{1-\alpha/2}SE_{\hat{R}}, \hat{R} + z_{1-\alpha/2}SE_{\hat{R}})$ ma poziom ufności w przybliżeniu równy $1 - \alpha$.

**Przykład 7.5.** Przyjmijmy, że chcemy ocenić średnią liczbę dzieci przypadających na jednego wychowawcę w przedszkolu w dużym mieście wojewódzkim. Ze 150 działających tam przedszkoli wybrano w prostym losowaniu bez zwracania 15 i otrzymano następujące wartości zbiorcze dla obydwu cech (przez $y_i$ oznaczmy liczbę dzieci w $i$-tym przedszkolu, a przez $x_i$ odpowiednią liczbę wychowawców):

$\sum_{i\in s} y_i = 309, \quad \sum_{i\in s} x_i = 39,$

$\sum_{i\in s} y_i^2 = 4236, \quad \sum_{i\in s} x_i y_i = 794, \quad \sum_{i\in s} x_i^2 = 151.$

Tak więc $\bar{x} = 2{,}6$ , $\bar{y} = 20{,}6$ i $\hat{R} = 7{,}92$. Na podstawie (7.14) otrzymujemy, że wartość błędu standardowego jest równa pierwiastkowi z wyrażenia

$$SE^2_{\hat{R}} = \frac{0{,}9}{15 \times (2{,}6)^2} \frac{4236 - 2 \times 7{,}92 \times 794 + (7{,}92)^2 \times 151}{14} = 0{,}716.$$

Zatem $SE_{\hat{R}} = 0{,}84$. Przyjmując, że rozkład $\hat{R}$ jest w przybliżeniu normalny, przedział ufności na poziomie ufności w przybliżeniu równym 0,95 wynosi (6,27, 9,57).

Zauważmy, że w przypadku, gdy znamy wartość średnią $\mu_X$ cechy $X$, estymator $\hat{R}$ ilorazu $R$ prowadzi do alternatywnego estymatora średniej $\mu_Y$ zwanego **estymatorem ilorazowym**

$$\bar{Y}_R = \hat{R} \times \mu_X,$$

dla którego wariancja w przypadku schematu losowania bez zwracania wynosi w przybliżeniu $\mathrm{Var}(\bar{Y}_R) = n^{-1}(1-n/N)(\sigma_Y^2 + R^2\sigma_X^2 - 2R\sigma_X\sigma_Y)$. Prawą stronę tego wyrażenia możemy w naturalny sposób estymować za pomocą wyrażenia $SE^2_{\hat{R}}(\mu_X^2)$. Oczywiście, ze stwierdzenia 7.4 wynika, że użycie estymatora ilorazowego ma sens głównie wtedy, gdy związek między cechami $X$ i $Y$ jest w przybliżeniu liniowy i odpowiednia prosta regresji przechodzi w przybliżeniu przez 0.

Rozpatrzyliśmy zatem cztery estymatory średniej populacji: estymator Horwitza–Thompsona oraz 3 estymatory złożone oparte na cesze dodatkowej: estymator różnicy, regresyjny i estymator ilorazowy. Z dokładniejszej analizy wynika, że najmniejszy błąd średniokwadratowy spośród nich ma z reguły estymator regresyjny. W praktyce spośród estymatorów używa się oprócz estymatora regresyjnego również estymatora ilorazowego.

**Przykład 7.6.** Załóżmy, że z populacji 200 elementowej została pobrana 15 elementowa próba zgodnie ze schematem losowania bez zwracania, dla której oprócz wartości interesującej nas cechy $Y$ zarejestrowano również wartości związanej z nią cechy $X$. Wartości obydwu cech przedstawiono w tab. 7.2. Wiadomo ponadto, że wartość średnia cechy $X$ w populacji wynosi $\mu_X = 54$. Odpowiednie wartości estymatorów średniej wynoszą $\bar{y}_{HT} = 110{,}3$, $\bar{y}_R = 107{,}5$ i $\bar{y}_{lr} = 109{,}5$. Odpowiednie wartości błędów standardowych wynoszą 4,98 dla estymatora Horwitza–Thompsona, 4,74 dla estymatora ilorazowego i 4,09 dla estymatora regresyjnego.

**Tabela 7.2. Wartości cech $Y$ i $X$ dla 15-elementowej próby pobranej z 200-elementowej populacji**

| Numer | $Y$ | $X$ |
|---|---|---|
| 1 | 122,712 | 82,047 |
| 2 | 69,167 | 16,874 |
| 3 | 123,926 | 51,865 |
| 4 | 93,418 | 47,365 |
| 5 | 157,105 | 79,732 |
| 6 | 92,633 | 53,807 |
| 7 | 115,970 | 37,709 |
| 8 | 114,749 | 82,493 |
| 9 | 112,566 | 70,884 |
| 10 | 108,265 | 23,454 |
| 11 | 93,909 | 54,999 |
| 12 | 92,628 | 58,821 |
| 13 | 118,758 | 43,577 |
| 14 | 113,658 | 70,783 |
| 15 | 125,567 | 57,065 |

### 7.2.6. Estymatory średniej dla schematu losowania warstwowego

Rozpatrzmy teraz estymator Horwitza–Thompsona dla strategii losowania warstwowego, gdy podpróby w warstwach są losowane przy użyciu prostego losowania bez zwracania. Przypomnijmy, że jednym z głównych powodów rozpatrzenia tego typu losowania jest chęć zmniejszenia zmienności estymatorów interesujących nas parametrów. Niech $\pi_{hi}$ oznacza prawdopodobieństwo wylosowania obiektu $o_{hi}$ z $h$-tej warstwy i niech $y_{hi}$ będzie wartością cechy $Y$ dla tego obiektu. Załóżmy, że liczności poszczególnych warstw są znane. Oczywiście, ponieważ z $h$-tej warstwy losujemy $n_h$ elementów, $\pi_{hi} = n_h/N_h$ i estymator Horwitza–Thompsona ma postać

$$\bar{Y}_{HT} = \frac{1}{N}\sum_{h=1}^{k}\sum_{i=1}^{n_h}\frac{N_h}{n_h}y_{hi} = \sum_{h=1}^{k}\frac{N_h}{N}\frac{1}{n_h}\sum_{i=1}^{n_h}y_{hi} = \sum_{h=1}^{k} w_h\bar{Y}_h, \tag{7.15}$$

gdzie $\bar{Y}_h$ jest średnią próbkową dla elementów wylosowanych z $h$-tej warstwy. Zauważmy, że jeśli proporcja $n_h/n$ elementów wylosowanych z $h$-tej warstwy jest równa proporcji $w_h = N_h/N$ elementów $h$-tej warstwy w całej populacji (takie losowanie nazywamy **proporcjonalnym**), to estymator $\bar{Y}_{HT}$ redukuje się do średniej próbkowej z wylosowanej próby. Ponieważ losowania w różnych warstwach są niezależne, z równości (7.15) i wniosku 7.1 otrzymujemy część (1) następującego wniosku.

**Wniosek 7.3.** (1) *Dla schematu losowania warstwowego z niezależnym losowaniem bez zwracania w różnych warstwach*

$$\mathrm{Var}(\bar{Y}_{HT}) = \sum_{h=1}^{k} w_h^2\Big(1 - \frac{n_h}{N_h}\Big)\frac{\sigma_h^2}{n_h},$$

*gdzie parametr $\sigma_h^2$ jest zdefiniowany analogicznie do wariancji $\sigma^2$ dla elementów $h$-tej warstwy. Ponadto, nieobciążony estymator* $\mathrm{Var}(\bar{Y}_{HT})$ *ma postać*

$$SE^2_{\bar{Y}_{HT}} = \sum_{h=1}^{k} w_h^2\Big(1 - \frac{n_h}{N_h}\Big)\frac{\sigma_{Sh}^2}{n_h},$$

*gdzie estymator $\sigma_{Sh}^2$ jest zdefiniowany analogicznie do wariancji próbkowej $\sigma_S^2$ dla próby z $h$-tej warstwy.*

(2) *Jeśli założymy dodatkowo, że losowanie jest proporcjonalne (tj. $n_h/n = w_h$), to otrzymamy*

$$\mathrm{Var}(\bar{Y}_{HT}) = \frac{1}{n}\Big(1 - \frac{n}{N}\Big)\sum_{h=1}^{k} w_h\sigma_h^2$$

*i nieobciążony estmator wariancji $\bar{Y}_{HT}$ ma w tym przypadku postać*

$$SE^2_{\bar{Y}_{HT}} = \frac{1}{n}\Big(1 - \frac{n}{N}\Big)\sum_{h=1}^{k} w_h\tilde{\sigma}_{Sh}^2.$$

Część (2) wniosku wynika z następującego prostego przeliczenia:

$$\mathrm{Var}(\bar{Y}_{HT}) = \sum_{h=1}^{k} w_h\frac{N_h}{N}\Big(1 - \frac{n_h}{N_h}\Big)\frac{\sigma_h^2}{n_h} = \Big(\frac{1}{n} - \frac{1}{N}\Big)\sum_{h=1}^{k} w_h\sigma_h^2 =$$

$$= \frac{1}{n}\Big(1 - \frac{n}{N}\Big)\sum_{h=1}^{k} w_h\sigma_h^2.$$

**Przykład 7.7.** W fabryce samochodów jest zatrudnionych 830 pracowników produkcyjnych, którzy pracują tam nie dłużej niż rok, 520 pracowników pracujących dłużej niż rok, ale nie dłużej niż dwa lata i 340 pracowników pracujących dłużej niż dwa lata. Podzbiory te potraktowano jako naturalne warstwy populacji pracowników produkcyjnych. Efektywność pracownika jest określana liczbą z przedziału $[2, 5]$. W celu określenia średniej efektywności pracowników produkcyjnych, pobrano 20-elementową próbę pracowników z pierwszej warstwy i 10-elementowe próby z kolejnych dwóch warstw. Otrzymano następujące wyniki dla poszczególnych warstw:

$$\bar{y}_1 = 3{,}64 \quad \bar{y}_2 = 4{,}23 \quad \bar{y}_3 = 3{,}94$$

i

$$\sigma^2_{s_1} = 1{,}26 \quad \sigma^2_{s_2} = 0{,}89 \quad \sigma^2_{s_3} = 1{,}21.$$

Liczność populacji $N = 830 + 520 + 340 = 1690$. Estymator Horwitza–Thompsona wartości średniej dla planu losowania warstwowego zgodnie ze wzorem (7.15) przyjmuje wartość

$$\bar{y}_{HT} = (830/1690) \times 3{,}64 + (520/1690) \times 4{,}23 + (340/1690) \times 3{,}94 = 3{,}88$$

a jego błąd standardowy $SE_{\bar{Y}_{HT}}$ wynosi

$$\left( \left(\frac{830}{1690}\right)^2 \left(1 - \frac{20}{830}\right) \frac{1{,}26}{20} + \left(\frac{520}{1690)}\right)^2 \left(1 - \frac{10}{520}\right) \frac{0{,}89}{10} + \left(\frac{340}{1690}\right)^2 \left(1 - \frac{10}{340}\right) \frac{1{,}21}{10} \right)^{1/2} =$$

$$= 0{,}167.$$

**Problem optymalnej alokacji.** Rozpatrzmy jeszcze problem wyboru, dla ustalonej liczności próby $n$, liczności $n_1, n_2, \ldots, n_k$ elementów wybranych w schemacie losowania bez zwracania z poszczególnych warstw, tak aby wariancja $\mathrm{Var}(\bar{Y}_{HT})$ była minimalna. Rozwiązanie tego problemu jest zwane **metodą alokacji Neymana**[4]. Odpowiednie liczności elementów wybranych z poszczególnych warstw wynoszą

$$n_h = n \frac{N_h \sigma_h}{\sum_{j=1}^{k} N_j \sigma_j} \qquad \text{dla } h = 1, \ldots, k.$$

[4]Rozwiązanie to zostało wcześniej znalezione przez Czuprowa.

Zauważmy, że w przypadku, gdy wariancje rozpatrywanej cechy w poszczególnych warstwach są równe, metoda alokacji Neymana redukuje się do losowania proporcjonalnego (z niezależnym losowaniem bez zwracania w warstwach). Ponieważ wartości $n_h$ w powyższym wzorze nie muszą być liczbami naturalnymi, w praktyce rozważa się odpowiednie najbliższe im liczby naturalne. Oczywiście z definicji alokacji optymalnej wynika, że odpowiadająca jej wariancja estymatora Horwitza–Thompsona jest nie większa od wartości wariancji tegoż estymatora w przypadku losowania proporcjonalnego, jak również w przypadku prostego losowania bez zwracania próby $n$-elementowej. Zauważmy jednak, że odchylenia standardowe $\sigma_h$ nie są znane i w praktyce musimy je estymować używając wartości $\sigma_{Sh}$ lub wielkości estymatorów otrzymanych w uprzednich badaniach ankietowych. Dlatego tak zmodyfikowana metoda alokacji nie musi mieć już najmniejszej wariancji spośród wszystkich możliwych metod. W tym kontekście interesująca może być informacja, że po pominięciu wyrazów typu $N_h^{-1}$ (co możemy zrobić dla dostatecznie dużych wartości $N_h$) wariancja estymatora Horwitza–Thompsona dla losowania proporcjonalnego jest nie większa od wariancji tegoż estymatora dla schematu losowania bez zwracania. Zatem z punktu widzenia zmienności estymatorów średniej schemat losowania proporcjonalnego ma przewagę nad schematem prostego losowania bez zwracania.

## 7.3. Zadania

**7.1.** W małym miasteczku liczącym 450 gospodarstw domowych wybrano za pomocą schematu prostego losowania bez zwracania 20 spośród nich i zanotowano liczbę dzieci poniżej 18 lat w każdym gospodarstwie. Otrzymane wyniki wynoszą

1, 2, 1, 0, 0, 3, 1, 1, 2, 4, 2, 2, 0, 3, 1, 2, 3, 1, 2, 1.

**a)** Znaleźć estymator średniej liczby dzieci w jednym gospodarstwie domowym i przedział ufności dla tego parametru na poziomie ufności 0,95.

**b)** Znaleźć estymator całkowitej liczby dzieci w mieście i przedział ufności dla tej liczby na tym samym co w punkcie a poziomie ufności.

**7.2.** Z 320 restauracji i barów szybkiej obsługi znajdujących się w dużym mieście wylosowano w schemacie losowania bez zwracania 40 i wylosowane obiekty poddano inspekcji. W 9 z nich stwierdzono nieprzestrzeganie zasad higieny przy przygotowywaniu posiłków.

**a)** Skonstruować przedział ufności na poziomie ufności 95% dla proporcji $p$ wszystkich restauracji nie przestrzegających zasad higieny.

**b)** Ile wynosi minimalna liczność próby, przy której z prawdopodobieństwem

0,95 otrzymany estymator nie będzie odstawał od prawdziwej wartości $p$ o więcej niż 0,1?

**Wskazówka:** wykorzystać uwagę z zad. 3.11.

**7.3.** Rozpatrzmy sytuację opisaną w przykł. 7.2 i załóżmy dodatkowo, że każda skrzynia zawiera 100 zaworów. Przyjmijmy, że badaną populacją jest zbiór wszystkich zaworów w skrzyniach i załóżmy, że przypisanie zaworów do skrzyń jest losowe. Z jakim planem losowania (opisanym w przykładzie) mamy do czynienia? Oszacować proporcję elementów wadliwych w partii oraz podać błąd standardowy tego oszacowania.

**7.4.** Załóżmy, że obszar doświadczalny w kształcie kwadratu $100 \times 100$ metrów został podzielony na 100 kwadratów o wielkości $10 \times 10$ metrów. Zaproponować realizację schematu prostego losowania bez zwracania 10 kwadratów przy użyciu generatora liczb losowych. Przeprowadzić wybór próby 25 razy i w każdym przypadku przedstawić graficznie, które kwadraty zostały wybrane. W ilu przypadkach wygenerowane próby zawierają kwadraty stykające się bokami lub wierzchołkami? Czy jest to zgodne z Twoją ideą *losowego* wyboru kwadratów?

**7.5.** Z populacji $N$ myszy znajdujących się na pewnym obszarze złapano $n$ spośród nich, zaobrączkowano i wypuszczono. Dwa dni poźniej powrócono na to samo miejsce i odłowiono $y$ myszy. Stwierdzono, że $x$ spośród nich jest zaobrączkowanych.

**a)** Znaleźć rozkład zmiennej $X$ liczby zaobrączkowanych myszy odłowionych za drugim razem, przy ustalonych licznościach prób $n$ i $y$.

**b)** Interpretując bycie zaobrączkowanym jako cechę populacji myszy, wyznaczyć wartość oczekiwaną i wariancję zmiennej $X$.

**c)** Przyrównując wartość $x$ do wartości oczekiwanej $EX$ otrzymanej w punkcie b zaproponować estymator $\hat{N}$ całkowitej liczby myszy w obszarze.

**d)** Wyznaczyć estymator $\hat{N}$ dla $n = 50$, $y = 100$ i $x = 15$. Znaleźć przedział $[a, b]$, dla którego $P(a \leq \tilde{X} \leq b)$ jest w przybliżeniu równe 0,9, gdzie $\tilde{X}$ jest liczbą zaobrączkowanych myszy odłowionych w niezależnym eksperymencie dla tych samych liczności $n$ i $y$.

**7.6.** W szkole średniej liczącej 250 uczniów przeprowadzono następujący eksperyment. 15 uczniów, wybranych w schemacie prostego losowania bez zwracania, poproszono o ocenę czasu $x$, który przeznaczą w ciągu najbliższego tygodnia na oglądanie telewizji, a następnie o zanotowanie rzeczywistego czasu $y$, przeznaczonego na tę czynność. Otrzymano następujące wyniki:

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $x$ | 3,0 | 7,0 | 10,5 | 4,0 | 10,0 | 12,0 | 14,0 | 0 | 5,0 | 8,0 |
| $y$ | 2,9 | 7,0 | 16,2 | 4,8 | 12,0 | 14,4 | 16,3 | 0 | 5,1 | 8,6 |
| $x$ | 11,0 | 12,0 | 10,5 | 7,0 | 14,0 | | | | | |
| $y$ | 13,2 | 14,5 | 13,3 | 8,5 | 22,5 | | | | | |

Obliczyć wartości estymatora Horwitza–Thompsona oraz estymatora regresyjnego dla średniej wartości czasu w tygodniu przeznaczonego na oglądanie telewizji w tej szkole. Porównać błędy standardowe obydwu estymatorów. Skomentować wyniki.

**7.7.** Spośród klientów dużego supermarketu wybrano losową próbę 10 osób, które poproszono o podanie wielkości miesięcznych zarobków lub emerytury ($x$) i kwoty wydawanej w ciągu miesiąca na żywność ($y$). Otrzymano następujące wyniki:

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $x$ | 700 | 2500 | 1800 | 1700 | 1200 | 500 | 600 | 3500 | 1800 | 900 |
| $y$ | 400 | 1100 | 800 | 900 | 600 | 300 | 250 | 1700 | 600 | 600 |

Zakładając, że wielkość populacji klientów supermarketu jest efektywnie nieskończona, oszacować średnią część $R$ dochodów przeznaczonych miesięcznie na zakup żywności oraz podać wielkość błędu standardowego dla tego oszacowania. Skonstruować przedział ufności dla wielkości $R$ na poziomie ufności 95%.

**7.8.** Montownia samochodów sprowadza pewien typ uszczelek od dwóch dostawców A i B, przy czym od pierwszego z nich pochodzi 75% dostaw, a od drugiego pozostałe 25% dostaw. Dział kontroli jakości w montowni zbadał jakość 100 uszczelek pobierając próbę według schematu proporcjonalnego losowania warstwowego z prostym losowaniem bez zwracania w warstwach i otrzymał następujące wyniki: wśród uszczelek pochodzących od $A$ wykryto 8 braków, a wśród uszczelek pochodzących od B 1 brak.

**a)** Oszacować proporcję braków i błąd standardowy oszacowania, zakładając, że wielkość dostaw uszczelek jest efektywnie nieskończona.

**b)** Założyć, że oszacowanie $\hat{p}$ w punkcie a otrzymano w schemacie prostego losowania bez zwracania próby 100-elementowej i ponownie oszacować błąd standardowy tego oszacowania. Porównać wyniki.

**7.9.** Chcemy się dowiedzieć, jakie są średnie roczne nakłady na cele socjalne na jednego pracownika w przedsiębiorstwach. Rozpatrywana populacja składa się z 40 dużych, 100 średnich i 250 małych przedsiębiorstw. Na podstawie poprzednio zebranych danych wiadomo, że odchylenia standardowe w poszczególnych warstwach wynoszą odpowiednio 200, 50 i 100 złotych. Wyznaczyć wielkości prób dla poszczególnych warstw dla schematu alokacji Neymana, gdy przyjmuje się, że liczność całej próby wyniesie $n = 100$. Obliczyć wariancję estymatora średniej dla tego schematu.

ROZDZIAŁ 8

# Metoda Monte Carlo

## 8.1. Wprowadzenie

W poprzednich rozdziałach zastanawialiśmy się głównie nad tym, jakie własności ma pewna funkcja danych $T(\mathbf{X})$, gdzie $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ jest prostą próbą losową z pewnego znanego (model probabilistyczny) lub znacznie częściej nieznanego (model statystyczny) rozkładu. Badania ograniczyliśmy do prostych funkcji $T$, gdyż z prób ich wykonania dla bardziej skomplikowanych funkcji lub w sytuacji złożonego charakteru zjawiska losowego wynikałoby, że zbadanie tych własności jest często niemożliwe lub możliwe tylko przy dużych licznościach prób. Konieczność znalezienia rozwiązań właśnie w takich przypadkach i rozwój technologii komputerowej doprowadziły do powstania koncepcji inscenizacji czy też inaczej symulacji komputerowej badanego zjawiska. Własności funkcji $T(\mathbf{X})$ badamy, obserwując jej zachowanie dla wielu powtórzeń eksperymentu losowego. Metoda nosi nazwę **metody Monte Carlo**. Gdy mechanizm zjawiska jest znany lub założony w rozpatrywanym modelu, wykonanie tego zadania jest prostsze; zajmiemy się tym w podrozdz. 8.2 i 8.3. W sytuacji modelu statystycznego sytuacja staje się znacznie bardziej skomplikowana. W podrozdziale 8.4 omówimy wykorzystanie metody permutacyjnej do porównania dwóch rozkładów prawdopodobieństwa. W podrozdziale 8.5 przedstawimy pewne podejście do rozwiązania ogólnego problemu będące adaptacją metody Monte Carlo do modelu statystycznego.

## 8.2. Generatory liczb pseudolosowych

### 8.2.1. Generatory liczb pseudolosowych z rozkładu jednostajnego

Omówienie metod Monte Carlo zaczniemy od omówienia kwestii podstawowej do ich implementacji: jak generować proste próby losowe z interesującego nas rozkładu prawdopodobieństwa zadanego przez dystrybuantę $F$? Rozpatrzmy najpierw sytuację, gdy $F$ jest dystrybuantą rozkładu jednostajnego $U[0,1]$ na odcinku $[0,1]$. Wkrótce dowiemy się, że dysponując prostą próbą losową z rozkładu jednostajnego i umiejąc obliczyć kwantyle dowolnego rzędu z interesującego nas rozkładu $F$, jesteśmy w stanie wygenerować prostą próbę losową z tego ostatniego rozkładu. Tak naprawdę nie generuje się prostych prób losowych z rozkładu jednostajnego $U[0,1]$, ale ciągi *deterministyczne*, których zachowanie dobrze imituje zachowanie prostej próby losowej z wymienionego rozkładu. Jest kilka przyczyn, dla których nie korzysta się tutaj z rzeczywistych wyników eksperymentu losowego, takich jak odpowiednio zinterpretowane rezultaty losowań ze zwracaniem kolejnych liczb ze zbioru $\{0,1,2,\ldots,9\}$. Jednym z powodów jest konieczność spełnienia postulatu możliwości odtworzenia konkretnej próby: często chcemy, np. w celu weryfikacji przeprowadzonych obliczeń, wygenerować powtórnie analizowaną próbę (lub, częściej, wiele prób). Oczywiście nie jest to możliwe, gdy analizowana próba jest realizacją prostej próby losowej, możemy wówczas co najwyżej wygenerować kolejną realizację tej próby, której konkretne wartości będą z reguły inne od wartości w próbie wylosowanej poprzednio. Ważna jest również dla nas możliwość łatwego generowania bardzo wielu prób z danego rozkładu; jest to konieczny warunek stosowalności większości metod Monte Carlo.

W celu generacji ciągu podobnego do realizacji prostej próby losowej z rozkładu jednostajnego z reguły rozpatruje się następującą metodę: dla pewnej wybranej przez nas funkcji $G$ określonej na odcinku $[0,1]$ i o wartościach z tego odcinka i dla początkowej wartości $u_0 \in [0,1]$, zdefiniujmy

$$u_1 = G(u_0),\ u_2 = G(u_1),\ \ldots, u_i = G(u_{i-1}),\ i = 1,2,\ldots \tag{8.1}$$

Zauważmy, że tak zdefiniowany ciąg $u_i$ spełnia postulat powtarzalności: znając początkową (startową) wartość $u_0$ oraz funkcję $G$ jesteśmy w stanie wygenerować ponownie wszystkie elementy tego ciągu. Prowadzi to nas do następującej nieformalnej definicji dużej klasy generatorów.

***Generatorem liczb pseudolosowych z rozkładu jednostajnego*** $U[0,1]$ *nazywamy algorytm, który dla pewnej funkcji* $G$ *i wartości początkowej* $u_0 \in$ $\in [0,1]$ *oblicza wartości* $u_n$ *zdefiniowane w (8.1). Generator ma przy tym*

*taką własność: dla każdej wartości $n$, wartości $u_1, u_2, \ldots, u_n$ imitują zachowanie realizacji prostej próby losowej $U_1, U_2, \ldots, U_n$ z rozkładu jednostajnego $U[0,1]$, w tym sensie, że standardowy zestaw testów nie odrzuca hipotezy $H_0$, iż $u_1, u_2, \ldots, u_n$ jest realizacją próby $U_1, U_2, \ldots, U_n$.*

Nie precyzujemy tu dokładnie o jakie standardowe testy nam chodzi; hipoteza $H_0$ nie powinna być odrzucana przez żaden ogólnie stosowany test zgodności jak np. test Craméra–von Misesa czy adaptacyjny test Neymana, czy też testy niezależności elementów próby, jak test Ljunga–Boxa i jego modyfikacje stosowane w analizie szeregów czasowych. Oczywiście, postulat imitowania zachowania prostej próby losowej pociąga za sobą warunek, że nie powinna istnieć żadna wykrywalna na podstawie danych zależność między wartością $u_i$ a pozostałymi wartościami $u_1, u_2, \ldots, u_{i-1}, u_{i+1}, \ldots, u_n$ oraz, że dla dużych $n$ częstość wpadania obserwacji do dowolnego przedziału $[a,b] \subset [0,1]$ powinna być w przybliżeniu równa jego długości $|b-a|$. Ta druga własność jest związana z faktem, że wszelkie własności probabilistyczne prostych prób losowych, w szczególności prawo wielkich liczb, powinny być w przybliżeniu spełnione dla prób pseudolosowych. Nie będziemy omawiali szczegółowo konstrukcji zadowalających przekształceń $G$, odsyłając do literatury przedmiotu (por. np. Wieczorkowski R., Zieliński R. (1997): *Komputerowe generatory liczb losowych.* Warszawa, WNT; Deak I. (1990): *Random number generators and simulation.* Budapest, Akademiai Kiado). Podamy jedynie przykład typowego generatora $G$ o „dobrych" własnościach (S. Park, K. Miller (1988): *Comm. ACM 31*, s. 1191–1201). Niech $A = 2147483647$ i $u_n = G(u_{n-1}) = G_0(A \times u_{n-1})/A$, gdzie

$$G_0(u) \text{ jest resztą z dzielenia liczby } (16807 \times u) \text{ przez } A,$$

$u_0 = C/A$ i $C$ jest dowolną liczbą naturalną z przedziału $(1, A-1)$. Tak więc w rzeczywistości generujemy pseudolosowe liczby naturalne z przedziału $(1, A-1)$, a następnie przekształcamy je tak, aby były liczbami z odcinka $(0,1)$.

Dysponując próbą pseudolosową $U_1, U_2, \ldots, U_n$ z rozkładu jednostajnego możemy rozwiązać wiele problemów wyboru pozornie niezwiązanych z generacją liczb losowych. Rozważmy np. problem wyboru losowego $k$-elementowego podzbioru spośród liczb $\{1, 2, \ldots, n\}$. Wystarczy w tym celu wybrać indeksy $l_1, l_2, \ldots, l_k$ odpowiadające $k$ największym liczbom spośród $U_1, U_2, \ldots, U_n$. Ponieważ indeksy zmiennych w próbie losowej są niezwiązane z ich wielkością, otrzymujemy w ten sposób $k$-elementowy losowy podzbiór zbioru $\{1, 2, \ldots, n\}$.

Omówimy teraz metody generacji prób pseudolosowych z innych rozkładów przy użyciu próby pseudolosowej z rozkładu jednostajnego. Dokładniej pokażemy, jak otrzymać prostą próbę losową z rozkładu $F$, dysponując prostą

próbą losową $U_1, U_2, \ldots, U_n$ z rozkładu $U[0,1]$, a następnie zastosujemy tę metodę do próby *pseudolosowej* z rozkładu $U[0,1]$. Dla uproszczenia załóżmy, że dystrybunta $F$ jest funkcją ściśle rosnącą, tzn. $F(x) < F(y)$ dla $x < y$, w tym przypadku kwantyl jest określony jednoznacznie. Przypadek ogólny wymaga modyfikacji definicji kwantyla.

### 8.2.2. Metoda przekształcenia kwantylowego

Metoda ta jest oparta na następującym prostym, ale użytecznym fakcie.

**Stwierdzenie 8.1.** *Niech $U_1, U_2, \ldots, U_n$ będzie prostą próbą losową z rozkładu $U[0,1]$. Wówczas ciąg kwantyli rzędu $U_1, U_2, \ldots, U_n$ dla rozkładu $F$, to jest ciąg $q_F(U_1), q_F(U_2), \ldots, q_F(U_n)$, jest prostą próbą losową z rozkładu $F$.*

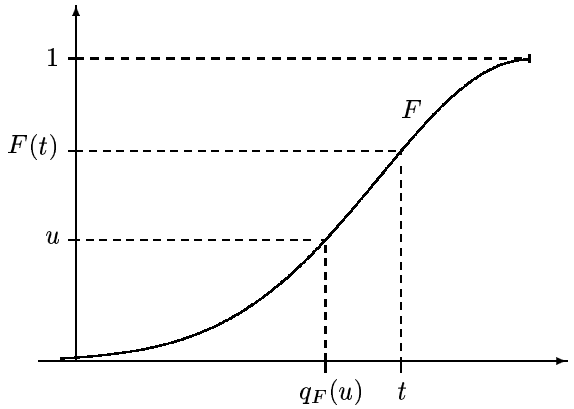
Oczywiście, przekształcając zmienne niezależne $U_1, U_2, \ldots, U_n$ za pomocą przekształcenia $q_F(\cdot)$, otrzymamy zmienne niezależne o tym samym rozkładzie, zatem wystarczy sprawdzić, czy rozkład pierwszego elementu nowej próby $q_F(U_1)$ jest rzeczywiście zadany przez dystrybuantę $F$. Ponieważ dystrybuanta $F$ jest ściśle rosnąca, mamy (por. rys. 8.1) $A = \{u\colon\ q_F(u) \leq \leq t\} = \{u\colon\ u \leq F(t)\}$, a zatem

$$P(q_F(U_1) \leq t) = P(U_1 \leq F(t)) = F(t),$$

gdzie ostatnia równość wynika z faktu, że zmienna $U_1$ ma rozkład jednostajny $U[0,1]$. Rzeczywiście zatem dystrybuanta zmiennej $q_F(U_1)$ dla wartości $t$ jest równa $F(t)$.

Przypomnijmy, że dla funkcji ściśle rosnącej znalezienie kwantyla $q_F(s)$ sprowadza się do znalezienia pierwiastka $x_0$ równania $F(x) = s$. Zadanie to daje się czasami prosto rozwiązać.

**Przykład 8.1. Generacja rozkładu wykładniczego.** Dla rozkładu wykładniczego z parametrem $\lambda$ dystrybuanta $F(x)$ wynosi $1 - e^{-\lambda x}$ dla $x \geq 0$. Zatem pierwiastkiem równania $F(x) = s$ dla $0 < s < 1$ jest $q_F(s) = -\log(1-s)/\lambda$. Tak więc na mocy stwierdzenia 8.1 próba $X_i = -\log(1-U_i)/\lambda$, $i = 1, \ldots, n$, jest prostą próbą losową z rozkładu wykładniczego o parametrze $\lambda$. Ponieważ zmienna $1 - U_i$ ma taki sam rozkład jak zmienna $U_i$, próba $X_i' = -\log(U_i)/\lambda$ jest również prostą próbą losową z rozkładu wykładniczego.




Rys. 8.1. Metoda przekształcenia kwantylowego. Losujemy $u$ z rozkładu jednostajnego na $(0,1)$. Wówczas $q_F(u)$ jest liczbą losową z rozkładu o dystrybuancie $F$

### 8.2.3. Metoda oparta na reprezentacji zmiennych losowych

Często nawet w przypadku, gdy dystrybuanta $F$ jest zadana w formie analitycznej, obliczenie odpowiedniej funkcji kwantylowej bywa trudne i wtedy zastosowanie metody transformacji kwantylowej jest niepraktyczne. Dotyczy to oczywiście również sytuacji, gdy dystrybuanta $F$ nie ma jawnej postaci. Tak jest na przykład w przypadku generacji prób z rozkładu normalnego; aby obliczyć wartość $\Phi(s)$ trzeba uciec się do metod aproksymacji. W takiej sytuacji do generacji próby pseudolosowej pomocne bywa wyrażenie zmiennej losowej o szukanym rozkładzie jako funkcji zmiennych losowych, których generacja jest stosunkowo prosta. Rozpatrzmy następujący przykład.

**Przykład 8.2. Generacja rozkładu $\chi^2$.** Zauważmy, że z podanego w p. 3.3.2 wzoru na gęstość rozkładu $\chi^2$ z $n$ stopniami swobody, który oznaczymy przez $\chi^2_n$, łatwo wynika, iż rozkład $\chi^2$ z dwoma stopniami swobody jest niczym innym jak rozkładem wykładniczym z parametrem $1/2$. Zatem zgodnie z dyskusją w przykł. 8.1, zmienna $X = -2\log(U)$ ma rozkład $\chi^2_2$, jeśli tylko rozkładem $U$ jest rozkład jednostajny $U[0,1]$. Jednocześnie ze wzoru (3.32) wynika, że zmienna losowa o rozkładzie $\chi^2_{2k}$ jest sumą $k$ niezależnych zmiennych losowych o rozkładzie $\chi^2_2$ każda. Zatem jeśli $U_1, U_2, \ldots, U_k$ są niezależnymi zmiennymi losowymi o rozkładzie $U[0,1]$, to $X_{2k} = -2\sum_{i=1}^{k} \log U_i$ ma rozkład $\chi^2_{2k}$. Fakt ten umożliwia prostą generację wartości z rozkładu $\chi^2$ o parzystej

liczbie stopni swobody. Generację rozkładu $\chi^2$ o $2k+1$ stopniach swobody można oprzeć na wzorze (3.32), z której wynika, że taki rozkład ma zmienna $X_{2k+1} = X_{2k} + Z^2$, gdzie $Z$ jest niezależną od $X_{2k}$ zmienną o rozkładzie $N(0,1)$. Generację rozkładu normalnego omawiamy poniżej.

Sformułujmy najpierw następujący fakt.

**STWIERDZENIE 8.2.** *Niech $(X_1, X_2)$ będzie parą niezależnych zmiennych losowych o rozkładzie $N(0,1)$ i niech $(R, \theta)$ będzie jej przedstawieniem we współrzędnych biegunowych*

$$R = \sqrt{X_1^2 + X_2^2}, \qquad \theta = \arccos \frac{X_1}{\sqrt{X_1^2 + X_2^2}}.$$

*Wówczas zmienna $R^2$ ma rozkład wykładniczy z parametrem $1/2$, $\theta$ ma rozkład $U[0, 2\pi]$ i zmienne $R$ i $\theta$ są niezależne.*

Stwierdzenia tego nie udowadniamy, zauważmy jednak, że jest ono intuicyjnie oczywiste. Wszystkie kierunki wektora łączącego punkt $(0,0)$ z losowym punktem $(X_1, X_2)$ są jednakowo prawdopodobne, co tłumaczy, dlaczego $\theta \sim U[0, 2\pi]$. Jednocześnie $\theta$ nie powinno zależeć od długości wektora $R$. Fakt, że zmienna $R^2$ ma rozkład wykładniczy został uzasadniony powyżej. Ze stwierdzenia na podstawie przykł. 8.1 otrzymujemy prostą metodę generacji dwóch niezależnych zmiennych o rozkładzie $N(0,1)$.

**Przykład 8.3. Generacja rozkładu normalnego.**

**1.** *Wygeneruj dwie niezależne zmienne losowe $U_1, U_2$ o rozkładzie $U[0,1]$.*

**2.** $X_1 = (-2\log U_1)^{1/2} \cos(2\pi U_2)\,, X_2 = (-2\log U_1)^{1/2} \sin(2\pi U_2)$.

*$X_1, X_2$ są zmiennymi niezależnymi o rozkładzie $N(0,1)$.*

Powyższy algorytm wynika ze stwierdzenia 8.2 po przedstawieniu pary $(X_1, X_2)$ w reprezentacji biegunowej $(R\cos\theta,\ R\sin\theta)$. Zmienna $R$ ma taki sam rozkład jak zmienna $-2\log U_1$ i jest niezależna od zmiennej $\theta$ mającej taki sam rozkład jak zmienna $2\pi U_2$.

Inna metoda przybliżonej generacji rozkładu normalnego polega na zauważeniu, że w myśl Centralnego Twierdzenia Granicznego suma dwunastu zmiennych losowych $U_i$ o rozkładzie jednostajnym $U[0,1]$ ma w przybliżeniu rozkład $N(12\times(1/2),\{12\times(1/12)\}^{1/2})$, zatem zmienna $\sum_{i=1}^{12} U_i - 6$ ma w przybliżeniu rozkład $N(0,1)$.

**Przykład 8.4. Generacja rozkładu Poissona.** Pamiętając z punktu 2.2.3, że liczba zdarzeń w odcinku $[0,\lambda]$ dla procesu Poissona z parametrem 1 ma rozkład Poissona $P(\lambda)$ i korzystając z własności, że czasy między kolejnymi zdarzeniami w tym procesie są niezależnymi zmiennymi losowymi o rozkładzie wykładniczym z parametrem 1, możemy łatwo wygenerować $X$ o rozkładzie $P(\lambda)$. Mianowicie, jeśli $W_1, W_2, \ldots$ są niezależnymi zmiennymi losowymi o rozkładzie wykładniczym z parametrem 1, to definiujemy $X = k$, gdzie $k$ jest taką największą liczbą, że $W_1 + W_2 + \ldots + W_k \leq \lambda$ (przyjmujemy $X = 0$, gdy $W_1 > \lambda$). Oczywiście, ponieważ $EX = \lambda$, średnia liczba użytych zmiennych $X_i$ wynosi $\lambda$ i metoda ta jest niepraktyczna dla dużych wartości $\lambda$.

**Przykład 8.5. Generacja rozkładu jednostajnego w $k$ punktach.** Omówienie generacji tego rozkładu ograniczymy do uwagi, że jeśli $U$ jest zmienną losową o rozkładzie $U[0,1]$, to zmienna $W = i$, gdy $U \in ((i-1)/k, i/k]$ dla $i = 1, \ldots, k$ jest zmienną o rozkładzie jednostajnym na zbiorze $\{1, \ldots, k\}$. Generacja tego rozkładu ma duże znaczenie przy zastosowaniu metod bootstrap omówionych dalej. Rozpatrując odcinki o długościach $p_1, p_2, \ldots, p_k$ możemy powyższą metodę łatwo zaadaptować do generacji takiego dowolnego dyskretnego rozkładu $X$, że $P(X = i) = p_i$, jeśli tylko $0 < p_i < 1, i = 1, \ldots, k$ i suma $p_i$ wynosi 1.

**Przykład 8.6. Generacja mieszaniny rozkładów.** Jeśli umiemy generować liczby pseudolosowe z rozkładów $F_1, F_2, \ldots, F_k$, to możemy łatwo generować liczby pseudolosowe z mieszaniny $F = \sum_{i=1}^{k} p_i F_i$, gdzie $p_i$ są dodatnimi wartościami sumującymi się do 1. Mianowicie, korzystając z poprzedniego przykładu, generujemy taki losowy indeks $I$, że $P(I = i) = p_i, i = 1, 2, \ldots, k$. Jeśli $I = j$, to następnie generujemy liczbę z rozkładu $F_j$. Uzasadnienie tej metody na podstawie wzoru o prawdopodobieństwie całkowitym pozostawiamy Czytelnikowi.

**Przykład 8.7. Generacja dwuwymiarowego rozkładu normalnego.** Generacja tego rozkładu opiera się na stwierdzeniu, że jeśli $Z_1, Z_2$ są niezależnymi zmiennymi losowymi o rozkładzie $N(0,1)$, to zmienne losowe

$$X_1 = \mu_1 + \sigma_1 Z_1, \quad X_2 = \mu_2 + \sigma_2(\rho Z_1 + \sqrt{1-\rho^2} Z_2)$$

mają dwuwymiarowy rozkład normalny $N(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$. Uzasadnienie tego faktu uzyskujemy, jeśli zauważymy, że wektor $(X_1, X_2)$ otrzymuje się za pomocą przekształcenia liniowego wektora $(Z_1, Z_2)$ o standardowym dwuwymiarowym rozkładzie normalnym. Zatem wektor $(X_1, X_2)$ ma rozkład dwuwymiarowy normalny. Ponadto, rozkład wektora $(X_1, X_2)$ ma takie same parametry jak rozkład $N(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$. Przykładowo wariancja zmiennej $X_2$ jest równa

$$\begin{aligned}\mathrm{Var}(\sigma_2(\rho Z_1 + \sqrt{1-\rho^2} Z_2)) &= \sigma_2^2(\mathrm{Var}(\rho Z_1) + \mathrm{Var}(\sqrt{1-\rho^2} Z_2)) = \\ &= \sigma_2^2(\rho^2 + (1-\rho^2)) = \sigma_2^2.\end{aligned}$$

### 8.2.4. Metoda eliminacji

Na zakończenie podrozdziału o metodach generacji omówimy jeszcze jedną popularną metodę generacji rozkładu $F$ zadanego przez gęstość $f$. Warunkiem jej zastosowania jest znajomość gęstości $g$ i takiej stałej $M$, że $f(s) \leq Mg(s)$ dla wszystkich $s$ oraz umiejętność generacji prostej próby losowej z rozkładu o gęstości $g$. Algorytm postępowania jest następujący.

**1.** *Generuj niezależne zmienne* $X \sim g$ *i* $U \sim U[0,1]$.

**2.** *Jeśli* $U \leq f(X)/Mg(X)$, *przyjmij* $Y = X$.

**3.** *W przeciwnym przypadku wróć do punktu 1.*

*Zmienna* $Y$ *ma gęstość* $f$.

Metodę tę można łatwo uzasadnić, jeśli zauważymy, że

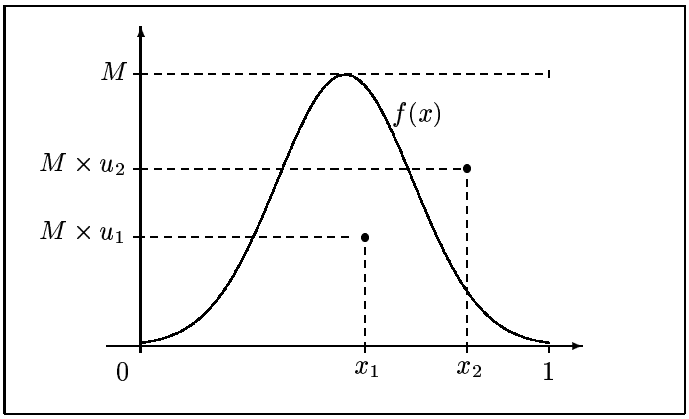
$$P(Y \leq y) = P\left(X \leq y | U \leq \frac{f(X)}{Mg(X)}\right) = \frac{P(X \leq y, U \leq \frac{f(X)}{Mg(X)})}{P(U \leq \frac{f(X)}{Mg(X)})}.$$

Po zapisaniu ostatniego wyrażenia w postaci całkowej i scałkowaniu wyrażenia w liczniku najpierw względem gęstości zmiennej $U$ otrzymujemy

$$P(Y \leq y) = \frac{\int_{-\infty}^{y} \int_{0}^{f(x)/Mg(x)} du\ g(x)dx}{\int_{-\infty}^{\infty} \int_{0}^{f(x)/Mg(x)} du\ g(x)dx} = \frac{\frac{1}{M}\int_{-\infty}^{y} f(x)dx}{\frac{1}{M}\int_{-\infty}^{\infty} f(x)dx} =$$

$$= \int_{-\infty}^{y} f(x)dx = F(y).$$

Działanie metody można prosto uzmysłowić sobie w sytuacji, gdy gęstość $f$ jest dodatnia tylko na odcinku $[0,1]$ i ograniczona na tym odcinku przez $M$. Wówczas generujemy parę niezależnych zmiennych losowych $X, U$ o rozkładzie $U[0,1]$ i analizujemy położenie punktu $(X, MU)$ o rozkładzie jednostajnym na kwadracie $[0,1] \times [0,M]$. Jeśli punkt ten leży pod wykresem gęstości $f$, to przyjmujemy $Y = X$, jeśli nie, to punkt odrzucamy i losujemy nowy aż do uzyskania punktu o tej własności (por. rys. 8.2).




Rys. 8.2. Metoda eliminacji. Punkt $(x_1, M \times u_1)$ leży pod wykresem gęstości $f(x)$: przyjmujemy $y = x_1$ jako liczbę losową z rozkładu $F$ o gęstości $f(x)$. Punkt $(x_2, M \times u_2)$ leży ponad wykresem gęstości $f(x)$: pomijamy $x_2$

Zauważmy, że w algorytmie eliminacji liczba generacji $X$ i $U$ jest związana z faktem, jak dobrze $Mg(x)$ szacuje od góry $f(x)$ – im lepsze przybliżenie tym średnio liczba generacji potrzebna do otrzymania próby o liczności $n$ będzie mniejsza. Metoda eliminacji daje m.in. inny sposób generacji standardowego rozkładu normalnego. Niech $\tilde{\phi}(z)$ równe $2\sqrt{2\pi}^{-1}\exp(-z^2/2)$ dla $z \geq 0$ i równe 0 w przeciwnym przypadku, będzie gęstością zmiennej losowej $W = |Z|$, gdzie $Z$ jest zmienną o rozkładzie standardowym normalnym. Rozpatrując jako $g$ np. gęstość rozkładu wykładniczego z parametrem $\lambda = 1$ i $M = \sup_s \tilde{\phi}(s)/g(s) = \sqrt{2e/\pi} \approx 1{,}32$ możemy za pomocą metody eliminacji wygenerować rozkład zmiennej $W$. Opatrując zmienną $W$ losowym znakiem w zależności od wyniku niezależnego rzutu monetą, otrzymamy zmienną o standardowym rozkładzie normalnym.

## 8.3. Szacowanie parametrów rozkładu metodą Monte Carlo

### 8.3.1. Estymatory parametrów rozkładu otrzymane metodą Monte Carlo

Omówimy teraz koncepcyjnie prostą, ale bardzo użyteczną metodę szacowania parametrów rozkładu metodą Monte Carlo. Metoda ta pochodzi od Metropolisa i Ulama. Załóżmy, że interesuje nas pewien parametr $\theta$ rozkładu zmiennej losowej $X$ i że $\theta$ można przedstawić jako $\theta = Eh(X)$, gdzie $h$ jest pewną znaną funkcją. Przyjmiemy ponadto, że możemy łatwo wygenerować próbę z rozkładu $X$. Będziemy dalej często mówić o prostej próbie losowej, pamiętając, że w jej zastępstwie możemy użyć (i z reguły tak robimy) próby pseudolosowej z tego rozkładu. W najprostszym przypadku, gdy $h(x) = x$, parametr $\theta$ jest po prostu wartością oczekiwaną $EX$. Możemy łatwo wyobrazić sobie inne przykłady:

a) Parametr $\theta$ jest odchyleniem standardowym zmiennej losowej $X$ o wartości oczekiwanej $\mu$: $\theta = (EX^2 - \mu^2)^{1/2}$. W tym przypadku oczywiście $h(x) = (x^2 - \mu^2)^{1/2}$.

b) Parametr $\theta$ jest wartością dystrybuanty zmiennej losowej $X$ dla pewnego $t$: $\theta = P(X \leq t) = F(t)$. Zauważmy, że $\theta$ jest wartością oczekiwaną zmiennej $h(X)$, gdzie $h(x) = 1$, gdy $x \leq t$ i 0 w przeciwnym przypadku. Rzeczywiście,

$$Eh(X) = 1 \times P(X \leq t) + 0 \times P(X > t) = F(t).$$

c) Niech $X$ będzie zmienną dyskretną o wartościach $x_1, x_2, \ldots$, $p_k = P(X = x_k)$ i $\theta = p_{k_0}$ dla pewnego $k_0$. W tym przypadku $\theta = Eh(X)$, gdzie $h(x) = 1$ gdy $x = x_{k_0}$ i $h(x) = 0$ w przeciwnym przypadku.

**Definicja 8.1.** *Niech dla pewnej wartości $m$ $X_1, X_2, \ldots, X_m$ będzie próbą pseudolosową z rozkładu zmiennej losowej $X$. Średnią $\bar{h} = m^{-1}(h(X_1) + h(X_2) + \ldots + h(X_m))$ nazywamy estymatorem wielkości $Eh(X) = \theta$ otrzymanym metodą Monte Carlo.*

Oczywiście wielkość $\bar{h}$ jest niczym innym niż średnią próbkową obliczoną na podstawie próby pseudolosowej $h(X_1), h(X_2), \ldots, h(X_m)$ i sensowność jej użycia do oszacowania parametru $\theta$ wynika z prawa wielkich liczb. Dodanie w definicji określenia metodą Monte Carlo jest związane z faktem, że mamy do czynienia z próbą pseudolosową oraz ma podkreślać odmienną niż w zwykłym problemie statystycznym specyfikę zagadnienia. Naszym problemem jest tutaj estymacja parametru $\theta$, któremu niekoniecznie musimy przypisywać interpretację probabilistyczną. Na przykład może interesować nas

oszacowanie całki $\int_0^1 h(s)ds$. Zauważenie, że całka ta jest wartością oczekiwaną zmiennej losowej $h(X)$, gdzie zmienna $X$ ma rozkład jednostajny na odcinku $[0,1]$ daje nam możliwość innego jej szacowania niż za pomocą klasycznej metody trapezów czy jej ulepszeń. Możliwość ta opiera się na generacji próby losowej lub pseudolosowej z rozkładu jednostajnego, przekształceniu elementów próby za pomocą funkcji $h$ i obliczeniu estymatora $\bar{h}$ parametru $\theta$.

### 8.3.2. Błędy standardowe estymatorów i przedziały ufności

Oczywiście, błąd przybliżenia $\bar{h}-\theta$ jest losowy w odróżnieniu od błędu w metodzie trapezów, co stanowi pewną trudność, gdyż nie można stwierdzić z całkowitą pewnością jak duża jest odchyłka wartości $\bar{h}$ od $\theta$. Mamy jednak wypracowane narzędzia oceny zmienności rozkładu $\bar{h}-\theta$. W szczególności, możemy łatwo oszacować jego wariancję, pamiętając, że $\mathrm{Var}(\bar{h}-\theta) = \mathrm{Var}\bar{h} = m^{-1}\mathrm{Var}(h(X))$. Prowadzi to do następującej definicji błędu standardowego estymatora $\bar{h}$

$$S_{\bar{h}} = \sqrt{\frac{1}{m(m-1)}\sum_{j=1}^{m}(h(X_j)-\bar{h})^2}.$$

Ponieważ wybór liczności próby pseudolosowej zależy od nas, możemy wybrać go tak, żeby błąd standardowy $S_{\bar{h}}$ odchylenia $\bar{h}-\theta$ był z dużym prawdopodobieństwem mały.

Rozpatrzmy jeszcze problem konstrukcji przedziału ufności dla parametru $\theta$. Wiemy, że dla dużych $m$ studentyzowana średnia $(\bar{h}-\theta)/S_{\bar{h}}$ ma w przybliżeniu rozkład normalny. Można wykorzystać to do konstrukcji przedziału ufności dla parametru $\theta$ na poziomie ufności w przybliżeniu równym $1-\alpha$. Mianowicie, analogicznie jak w p. 3.3.1 stwierdzamy, że taki przedział ufności ma postać

$$(\bar{h} - z_{1-\alpha/2}S_{\bar{h}}, \bar{h} + z_{1-\alpha/2}S_{\bar{h}}).$$

W przypadku szacowania wielkości proporcji $\theta = P(X \le t)$ (przypadek (b)) mamy sytuację szczególną: możemy oszacować z góry długość przedziału ufności niezależnie od wylosowanej próby. Rozważmy ten problem dokładniej. Mamy

$$\bar{h} = \hat{F}(t) = \frac{1}{m}\sum_{i=1}^{m} h(X_i) = \frac{\text{liczba obserwacji} \le t}{m}.$$

Wielkość $\hat{F}(t)$ jest dystrybuantą empiryczną w punkcie $t$ zdefiniowaną w p. 3.4.2 (def. 3.5). Zauważmy, że w tej sytuacji $m\hat{F}(t)$ jest zmienną

losową o rozkładzie dwumianowym $Bin(m, F(t))$. Korzystając z nierówności $s(1-s) \leq 1/4$ dla $s \in (0,1)$ mamy

$$\mathrm{Var}\bar{h} = \frac{F(t)(1-F(t))}{m} \leq \frac{1}{4m}.$$

Zatem, aby długość przedziału ufności na poziomie ufności 0,95 nie przekraczała jednej tysięcznej musimy mieć $(2 \times 1{,}96)/(2 \times \sqrt{m}) \leq 10^{-3}$, a stąd $m \geq (1{,}96 \times 10^3)^2 = 3{,}84 \times 10^6$. Pamiętajmy, że nie jest to specjalnie duża liczba, jeśli generacja wartości z rozkładu $F$ jest łatwa, oraz że zastosowane oszacowanie $F(s)(1-F(s)) \leq 1/4$ jest dokładne tylko dla takich $s$, że $F(s) \approx 1/2$, a więc w okolicy mediany. Zatem dla $s$ znajdujących się w ogonie rozkładu $F$ potrzebujemy faktycznie znacznie mniej obserwacji dla równie precyzyjnego oszacowania wartości dystrybuanty. Metodę tę można np. zastosować do aproksymacji dystrybuanty rozkładu normalnego $N(0,1)$.

Oczywiście, obliczenie wartości dystrybuanty lub prawdopodobieństw przyjęcia pewnych wartości metodą Monte Carlo jest sensowne tylko wtedy, gdy bezpośrednie obliczenie tych wartości jest żmudne. Taka sytuacja występuje jednak często, np. w schematach wielostopniowego losowania.

Zauważmy, że za pomocą metody Monte Carlo możemy rozwiązać zadanie ogólniejsze od powyżej rozpatrywanych, mianowicie wyznaczyć dla dowolnej ciągłej zmiennej losowej $W$ taki przedział $[q_{\alpha/2}, q_{1-\alpha/2}]$, że $P(q_{\alpha/2} \leq W \leq q_{1-\alpha/2}) = 1-\alpha$. Oczywiście, wystarczy w tym celu wyznaczyć kwantyle $q_{\alpha/2}$ i $q_{1-\alpha/2}$ rzędu $\alpha/2$ i $1-\alpha/2$ tego rozkładu. W tym celu generujemy próbę z rozkładu zmiennej losowej $W$ o dużej liczności $k$, a następnie obliczamy statystyki porządkowe rzędu $[k\alpha/2]$ i $[k(1-\alpha/2)]$ w tej próbie. Oznaczmy odpowiednie statystyki przez $\tilde{w}_{k,\alpha/2}$ i $\tilde{w}_{k,1-\alpha/2}$. Wówczas $P(\tilde{w}_{k,\alpha/2} \leq W \leq \tilde{w}_{k,1-\alpha/2}) \approx 1-\alpha$ i różnica wielkości po obu stronach dąży do 0, gdy liczność $k$ dąży do nieskończoności. Zmienna losowa $W$ nie musi być zadana w sposób analityczny, wystarczy, że umiemy wygenerować próbę o dowolnej liczności z jej rozkładu. Przypuśćmy na przykład, że interesuje nas znalezienie kwantyli rozkładu dziesiątej statystyki porządkowej z prostej próby losowej o liczności 30 z rozkładu $N(0,1)$. Generując $k$ prób o liczności 30 z tego rozkładu i rozpatrując dziesiątą statystykę porządkową w każdej z nich otrzymamy $k$-elementową próbę z rozkładu dziesiątej statystyki porządkowej w próbie 30-elementowej z rozkładu $N(0,1)$. Metodę tę wykorzystamy w zad. 8.8.

**Przykład 8.8.** Rozważmy eksperyment polegający na losowaniu 100 elementów z partii 2000 elementów, spośród których 200 jest wadliwych. Jeśli interesuje nas prawdopodobieństwo, że liczba elementów wadliwych $X$ w wylosowanej próbie jest nie mniejsza niż na przykład

14, to wiemy, że

$$p = P(X \geq 14) = \sum_{k=14}^{100} \frac{\binom{200}{k}\binom{1800}{100-k}}{\binom{2000}{100}} = 0{,}118. \tag{8.2}$$

Obliczenie tego prawdopodobieństwa jest dosyć żmudne. Znacznie prostszą metodą jest oszacowanie tego prawdopodobieństwa przez wylosowanie dużej liczby, na przykład dziesięciu tysięcy, stuelementowych prób z rozpatrywanej partii 2000 elementów i obliczenie częstości tych spośród nich, które zawierają nie mniej niż 14 elementów wadliwych. W wyniku takiej operacji otrzymano oszacowanie prawdopodobieństwa $p$ równe $\hat{p} = 0{,}117$ z błędem standardowym $\sqrt{\hat{p}(1-\hat{p})/100} = 0{,}0321$.

Zauważmy, że bardzo ważną sytuacją, do której można stosować opisaną w przykładzie metodę jest obliczanie $p$–wartości lub mocy rozpatrywanego testu. Obie z tych wielkości mają bowiem interpretację prawdopodobieństwa wystąpienia pewnych wartości statystyki testowej w sytuacji, gdy hipoteza zerowa, bądź odpowiednio alternatywna, jest prawdziwa. Na przykład sytuację opisaną w przykł. 8.8 można przedstawić następująco. Producent twierdzi, że partia 2000 elementów zawiera nie więcej niż 10% braków, to jest nie więcej niż 200 elementów wadliwych. Testowanie elementów w próbie 100 elementowej wykazało 14 braków. Jeśli liczba braków $X$ jest użyta jako statystyka testowa do testowania $H_0$: $p = 0{,}1$ przeciwko alternatywie $H_1$: $p > 0{,}1$, gdzie $p$ oznacza prawdopodobieństwo, że losowo wybrany element jest wadliwy, to $p$–wartość tej statystyki jest określona w (8.2).

### 8.3.3. Modelowanie eksperymentów losowych metodą Monte Carlo

Modelowanie metodą Monte Carlo ma duże znaczenie przy analizie skomplikowanych eksperymentów losowych. Można się tu zetknąć z dwoma typami problemów. Pierwszy dotyczy skomplikowanych systemów losowych, w których wprawdzie proces losowy jest ściśle sprecyzowany lub dobrze poznany, ale ze względu na stopień złożoności trudno jest obliczyć wartości interesujących nas parametrów zmiennych wynikowych. Taka sytuacja jest omówiona w przykł. 8.9. Drugi typ problemów pojawia się wtedy, gdy zasadniczą trudnością jest stworzenie adekwatnego matematycznego modelu będącego przedmiotem naszego zainteresowania. Po zaproponowaniu pewnego hipotetycznego modelu możemy starać się go sfalsyfikować generując za pomocą metody Monte Carlo wyniki modelowe i porównując je z danymi eksperymentalnymi. To podejście ma ogromne znaczenie m.in. w medycynie,

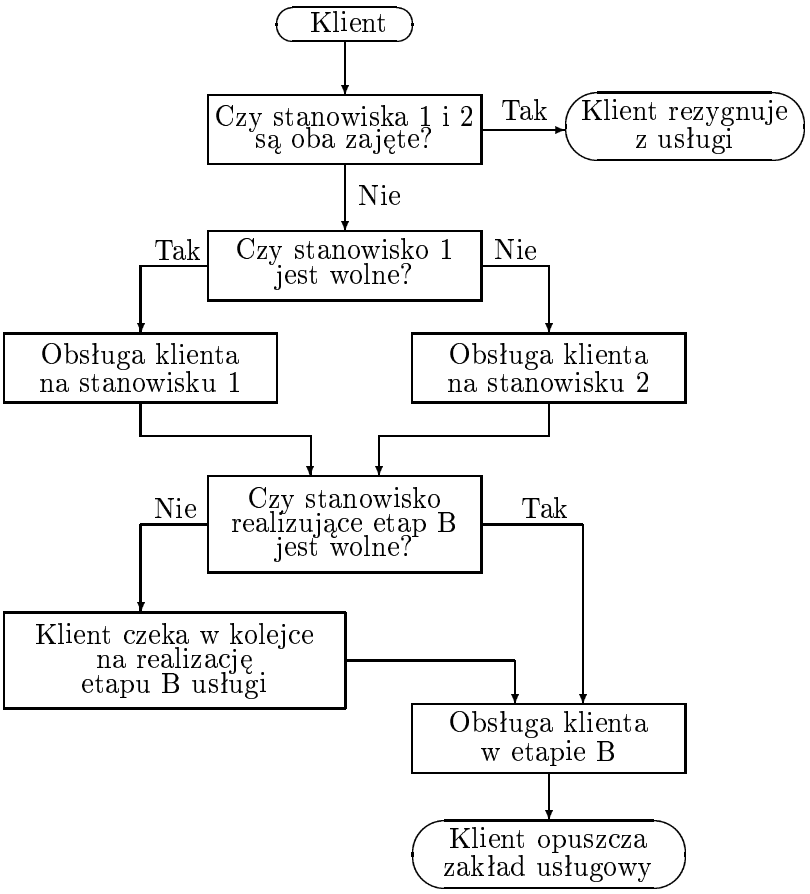
biologii i ekonometrii, gdzie z reguły proponowane modele są zbyt skomplikowane, aby pokusić się o rozwiązania analityczne. Jako przykład można podać modele rozrostu tkanki rakowej czy rozprzestrzenianie się AIDS.

**Przykład 8.9.** Rozpatrzmy usługę składającą się z dwóch następujących po sobie etapów, które nazwiemy A i B. Czas realizacji etapu A (w godzinach) jest losowy i ma rozkład wykładniczy z parametrem 2, czas realizacji etapu B jest również losowy i ma rozkład jednostajny na odcinku $(1/6, 1/3)$. Zakład świadczący usługę przewidział dwa niezależnie działające stanowiska realizujące etap A oznaczone 1 i 2 i jedno stanowisko realizujące etap B. Zapotrzebowanie na usługę można opisać za pomocą czasów zgłaszania się klientów do zakładu; przyjmijmy, że momenty zgłoszeń możemy adekwatnie modelować procesem Poissona z parametrem $\lambda = 1$. Działanie systemu przedstawiono na rys. 8.3. Klienci przybywają w losowych momentach czasu, gdy wolne jest jedno z dwóch stanowisk A, są natychmiast przyjmowani (jeśli oba stanowiska są wolne, to obsługuje ich stanowisko o niższym numerze), a jeśli oba stanowiska są zajęte, rezygnują z realizacji usługi. Fakt ich rezygnacji sprawia, że rozkład faktycznych chwil pojawienia się klientów na etapie A nie jest już opisany procesem Poissona z parametrem $\lambda$. Po zakończeniu etapu A klienci czekają w kolejności na realizację etapu B. Zdefiniujmy czas $W$ obsługi klienta jako sumę czasów realizacji etapu A i etapu B oraz czasu oczekiwania między etapem A a etapem B.

Załóżmy dla uproszczenia, że zakład realizujący usługę działa 24 godziny na dobę bez żadnych przerw. W takim przypadku okazuje się, że po dłuższym okresie działania sytuacja stabilizuje się w tym sensie, że losowe czasy obsługi kolejnych klientów mają w przybliżeniu takie same rozkłady. Interesuje nas średni czas obsługi $EW$ w sytuacji ustabilizowanej oraz przedział $[a, b]$ taki, dla którego $P(a \leq W \leq b) = 0{,}90$.

Rozwiązanie tego zagadnienia metodą analityczną opartą na obliczeniu rozkładu czasu usługi jest skomplikowane ze względu na istnienie wielu możliwych historii realizacji usługi. Przybywający klient może być natychmiast obsłużony przez jedno z dwóch stanowisk etapu A, oba stanowiska mogą być jednak zajęte i wtedy rezygnuje z usługi. Po realizacji etapu A klient może oczekiwać w kolejce na etap B bądź być natychmiast przyjęty. Z tego względu najlepiej rozwiązać to zagadnienie metodą Monte Carlo. Prześledźmy historię obsługi 1000 kolejnych klientów poczynając od tysiąc pierwszego klienta, pozostawiając czas obsługi pierwszych tysiąca klientów na ustabilizowanie się systemu. Niech $T_{1001}, T_{1002}, \ldots, T_{2000}$ będą niezależnymi zmiennymi losowymi o rozkładzie wykładniczym z parametrem 1. Zmienna

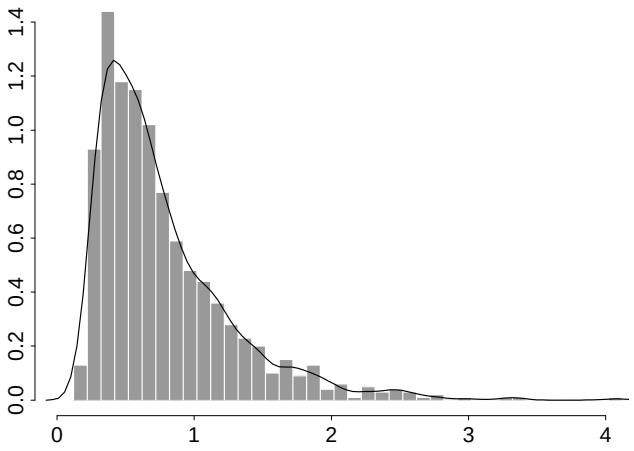
$T_i$, $i = 1001, 1002, \ldots, 2000$ będzie oznaczać przedział czasowy między przybyciem klienta o numerze $(i-1)$ i klienta o numerze $i$. Niech ponadto $X_i^{(1)}$ i $X_i^{(2)}$ oznaczają czasy obsługi $i$-tej obsługi na stanowisku 1 i 2, a $Y_i$ czas $i$-tej obsługi na stanowisku $B$.




Rys. 8.3. Schemat działania zakładu usługowego z przykł. 8.9

Histogram czasów obsługi 1000 klientów przedstawiono na rys. 8.4; średni czas obsługi wynosi 0,775 godziny, a odchylenie standardowe wynosi 0,495 godziny. Ponieważ rozkład czasu obsługi odbiega znacznie od rozkładu normalnego, przedział ufności dla czasu obsługi skonstruujemy na podstawie odpowiednich kwantyli z rzeczywistego rozkładu $W$. Z rysunku 8.4 otrzymujemy, że kwantyl rzędu 0,05 czasu obsługi $W$ jest równy 0,281, a kwantyl rzędu 0,95 jest równy 1,766, zatem przedział ufności dla $W$ na poziomie ufności 0,90 jest równy

$$[0{,}281,\ 1{,}766].$$




Rys. 8.4. Histogram czasów obsługi 1000 klientów z naniesionym estymatorem jądrowym gęstości z parametrem wygładzającym $h = 0{,}35$

Zauważmy, że stosując przybliżenie normalne, otrzymalibyśmy przedział $[-0{,}040,\ 1{,}589]$.

Na zakończenie tego przykładu zauważmy, że fakt, czy sytuacja w systemie się stabilizuje, można również sprawdzić przy użyciu metody Monte Carlo. Wystarczy w tym celu wygenerować dużą liczbę potoków 2000 klientów i porównać histogramy czasów obsługi klientów począwszy od tysiąc pierwszego klienta w każdym potoku. Bliskość tych histogramów będzie świadczyć o stabilizacji sytuacji w systemie po upływie czasu obsługi 1000 pierwszych klientów.

## 8.4. Testy permutacyjne

### 8.4.1. Testowanie jednorodności

**Przykład 8.10.** Dane w tabeli 8.1 dotyczą czasu wykonania (w sekundach) prostej czynności manualnej przez siedem dziewczynek i pięciu chłopców w tym samym wieku. Przypuśćmy, że rozkład czasu wykonania tej czynności jest losowy i w populacji dziewczynek jest opisany przez dystrybuantę $F$, a w populacji chłopców przez dystrybuantę $G$. Naszym celem jest stwierdzenie, czy rozkład czasu wykonania czynności zależy od płci, a zatem testujemy hipotezę $H_0$: $F = G$ przeciwko hipotezie alternatywnej $H_1$: $F \neq G$.

**Tabela 8.1.**

| | Czasy wykonania prostej czynności manualnej | | | | | | |
|---|---|---|---|---|---|---|---|
| Dziewczynki | **x**: 36,2 | 31,2 | 28,2 | 39,4 | 30,3 | 38,6 | 42,2 |
| Chłopcy | **y**: 28,9 | 32,6 | 26,0 | 28,8 | 34,7 | | |

Test hipotezy chcielibyśmy oprzeć na porównaniu pewnych wskaźników próbkowych obu rozkładów. Rozpatrzmy sytuację najprostszą, gdy porównujemy ich wartości średnie. Zajmiemy się tym przypadkiem, gdyż jest on nam dobrze znany z rozdz. 4. Podkreślmy jednak mocno, że omawiane podejście można zastosować do porównania dowolnych innych parametrów jak mediany czy wariancje, a fakt występowania różnic między rozkładami w praktyce sprowadza się do różnicy pewnych ich wskaźników. Inne podejście polegające na porównaniu rozkładów $F$ i $G$ opiszemy w rozdz. 9. Przyjmijmy, że nasza statystyka testowa jest różnicą średnich czasów wykonania czynności, a zatem dla rozpatrywanych prób $\hat{\theta}_0 = \bar{x} - \bar{y} = 35{,}15 - 30{,}2 = 4{,}95$. W rozdziale 3 omówiliśmy test dotyczący porównania dwóch wartości średnich w różnych populacjach. W przypadku, gdy wiemy, że rozkłady wyników w obydwu populacjach mają rozkład normalny i różnią się co najwyżej wartością oczekiwaną, test taki byłby równoważny testowi równości rozkładów. W przypadku naszych danych nie możemy jednak racjonalnie wnioskować o normalności rozkładów ani o równości ich wariancji, wielkości prób, $n = 7$ i $m = 5$ są na to zbyt małe. Małe liczności prób uniemożliwiają również stosowanie testu opartego na przybliżeniu normalnym. Opiszemy tutaj pewne podejście do tego zagadnienia zwane **testem permutacyjnym**. Inne bliskie mu alternatywne podejście, zwane testem rangowym, zostanie opisane w p. 9.2.1. Zauważmy, ze rozpatrywaną hipotezę zerową można sformułować jako hipotezę niezależności czasów wykonania czynności od płci. Dlatego naturalne wydaje się następujące postępowanie: losowo przypiszmy $n$ spośród wszystkich $n + m$ otrzymanych wyników do populacji dziewczynek, a pozostałe do populacji chłopców i przy takim przyporządkowaniu obliczmy odpowiednią wartość $\hat{\theta}$. W tym przypadku analizujemy na pewno sytuację opisaną hipotezą $H_0$, gdyż losowo przypisane do populacji wyniki są niezależne od płci. Powtarzając tę operację, otrzymamy ciąg wartości statystyk przy spełnionej hipotezie $H_0$ i będziemy w stanie stwierdzić, na ile wartość 4,95 jest wśród nich typowa. Opiszmy teraz dokładniej przedstawione postępowanie.

Połączmy próby **x** i **y** w jedną i uporządkujmy wartości od najmniejszej do największej. Otrzymany wektor oznaczmy przez **v**. Przyporządkujmy mu wektor **r** składający się z zer i jedynek, przy czym jedynka na pewnej współrzędnej wektora **r** oznacza, że na tym samym miejscu wektora **v** znajduje się jedna z wartości $x$, a zero, że jedna z wartości $y$. Tak więc uporządkowana i połączona próba **v** i wektor **r** przynależności do odpowiedniej próby mają

postać

$$\mathbf{v} = (26{,}0,\ 28{,}2,\ 28{,}8,\ 28{,}9,\ 30{,}3,\ 31{,}2,\ 32{,}6,\ 34{,}7,\ 36{,}2,\ 38{,}6,\ 39{,}4,\ 42{,}2)$$

i

$$\mathbf{r} = (0,\ 1,\ 0,\ 0,\ 1,\ 1,\ 0,\ 0,\ 1,\ 1,\ 1,\ 1).$$

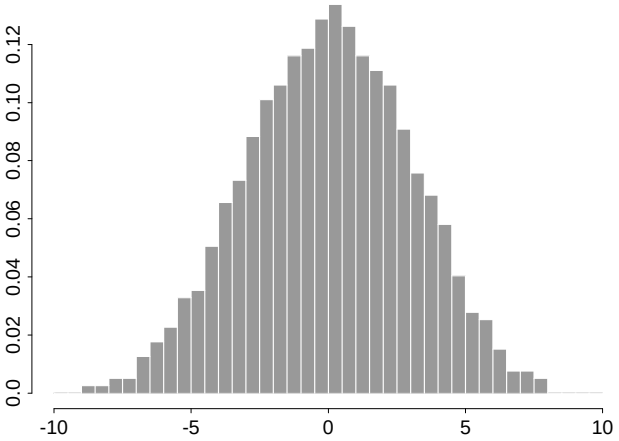
Wprowadźmy jeszcze następujące oznaczenia: niech $\bar{X}$ i $\bar{Y}$ będą średnimi próbkowymi czasu wykonania dla dwóch prostych prób losowych z populacji dziewczynek i chłopców o liczności $n = 7$ i $m = 5$ odpowiednio i niech $\mathbf{V}$ i $\mathbf{R}$ będą zmiennymi losowymi zdefiniowanymi analogicznie do $\mathbf{v}$ i $\mathbf{r}$. Ponadto, niech $\hat{\theta} = \bar{X} - \bar{Y}$ i $N = \binom{n+m}{n}$ oznacza liczbę różnych ciągów zerojedynkowych o długości $n+m$ zawierających $n$ jedynek. Rzeczywiście, liczba takich ciągów jest równa liczbie wyborów $n$ miejsc, na których będą stać jedynki spośród $n+m$ miejsc. Test permutacyjny jest oparty na następującym prostym stwierdzeniu.

**Stwierdzenie 8.3.** *Przy spełnieniu hipotezy $H_0$ i przy ustalonej wartości $\mathbf{V} = \mathbf{v}$ zmienna losowa $\mathbf{R}$ ma rozkład jednostajny na zbiorze wszystkich ciagów zerojedynkowych o długości $n+m$ zawierających $n$ jedynek.*

Dowodu stwierdzenia nie podajemy; zauważmy jednak, że jego teza dla $m = n$ jest intuicyjna: jeśli rozkłady obydwu populacji są takie same, to prawdopodobieństwo, że najmniejsza wartość w połączonej próbie pochodzi z populacji dziewczynek jest równe prawdopodobieństwu, że pochodzi z populacji chłopców. Podobnie można uzasadnić, że prawdopodobieństwo, że zmienna $\mathbf{R}$ jest równa dowolnemu ustalonemu ciągowi zerojedynkowemu jest równe $1/N$. Zdefiniujmy teraz

$$p_{\mathbf{v}} = P_{H_0}(|\hat{\theta}| \geq |\hat{\theta}_0| \text{ dla ustalonej wartości } \mathbf{V} = \mathbf{v}) = P(|\hat{\theta}| \geq 4{,}95 | \mathbf{V} = \mathbf{v}).$$

Zauważmy, że powyższa definicja jest analogiczna do definicji $p$–wartości jako prawdopodobieństwa otrzymania wartości nie mniej ekstremalnej od faktycznie otrzymanej wartości statystyki przy spełnionej hipotezie $H_0$. Jedyna różnica polega na ustaleniu dopuszczalnego zbioru wartości połączonych prób $\mathbf{x}$ i $\mathbf{y}$. Oznacza to, że dopuszczamy tylko sytuację, gdy każda z możliwych wartości musi być jedną ze współrzędnych wektora $\mathbf{v}$, a zmianie może ulegać jedynie wektor przynależności $\mathbf{r}$. Ten zabieg powoduje, że na podstawie stwierdzenia można obliczyć taką warunkową $p$–wartość, w odróżnieniu od zwykłej $p$–wartości $P_{H_0}(|\hat{\theta}| \geq 4{,}95)$, która nie jest określona, gdyż to prawdopodobieństwo zależy od rozkładu $F = G$, którego hipoteza zerowa nie specyfikuje. Mianowicie, niech $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_N$ będą wartościami zmiennej




Rys. 8.5. Histogram różnicy średnich otrzymany przy zastosowaniu dokładnego testu permutacyjnego (przykł. 8.10)

losowej $\hat{\theta}$ dla ustalonego wektora $\mathbf{V} = \mathbf{v}$. Ciąg ten zawiera (z ewentualnymi powtórzeniami) wszystkie możliwe wartości statystyki $\hat{\theta}$ dla $\mathbf{V} = \mathbf{v}$. Wówczas na podstawie stwierdzenia

$$p_{\mathbf{v}} = \frac{\#\{|\hat{\theta}_i| \geq 4{,}95,\ i = 1, 2, \ldots, N\}}{N}.$$

W rozpatrywanym przykładzie $N = \binom{12}{5} = 792$ i wartość $p_{\mathbf{v}}$ można obliczyć, rozpatrując wszystkie możliwe ciągi składające się z zer i jedynek i zliczając frakcję odpowiadających wartości $\hat{\theta}_i$ nie mniejszych co do wartości bezwzględnej od wartości $|\hat{\theta}| = 4{,}95$. W ten sposób otrzymujemy $p_{\mathbf{v}} = 0{,}101$. Nie mamy zatem podstaw do odrzucenia hipotezy o niezależności czasu wykonania czynności od płci. Rozkład różnic średnich dla testu permutacyjnego przy spełnionej hipotezie zerowej przedstawiono na rys. 8.5. Taką procedurę nazywamy dokładnym testem permutacyjnym (randomizacyjnym). W sytuacji, gdy wartość $N$ jest duża, wartość $p_{\mathbf{v}}$ można przybliżyć przez analogiczną frakcję dla losowo wybranej dużej liczby $B$ ciągów tzn.

$$\hat{p}_{\mathbf{v}} = \frac{\#\{|\hat{\theta}_i| \geq 4{,}95,\ i = 1, 2, \ldots, B\}}{B},$$

gdzie $\hat{\theta}_i$ są wartościami zmiennej $\hat{\theta}$ dla $B$ losowo wybranych ciągów zerojedynkowych zawierających $n$ jedynek spośród $N$ możliwych takich ciągów. W tym przypadku mówimy o przybliżonym teście permutacyjnym (randomizacyjnym). Zauważmy, że omawiana metodologia nie zależy od postaci

statystyki $\hat{\theta}$, która może być dowolnym wskaźnikiem oceny nierówności rozkładów. Przykładowo, statystyka $\hat{\theta}$ może być zdefiniowana jako różnica median lub iloraz wariancji próbkowych w obydwu próbach.

### 8.4.2. Testowanie niezależności cech

Test permutacyjny może być użyteczny również w innych sytuacjach. Przypuśćmy, że interesuje nas zależność wyników egzaminów końcowych z modelowania matematycznego ($X$) i inżynierii oprogramowania ($Y$) w pewnej wyższej szkole technicznej. Dysponując próbą wyników $(x_i, y_i)$, $i = 1, 2, \ldots, n$ dla $n$ studentów możemy rozumować następująco. Obliczmy współczynnik korelacji $r$ lub inny wskaźnik zależności dla naszych danych. Jeśli zmienne $X$ i $Y$ nie są zależne (hipoteza $H_0$), to fakt, że współrzędne w parze $(x_i, y_i)$ odpowiadają wynikom tego samego studenta nie ma znaczenia i permutując wartości pierwszej współrzędnej, a drugą pozostawiając bez zmian powinniśmy otrzymać „podobną" wartość współczynnika korelacji. Rozpatrując $B$ permutacji i obliczając odpowiadające wartości współczynnika korelacji $r_1, r_2, \ldots, r_B$, możemy stwierdzić na ile typowa jest wśród nich oryginalna wartość $r$.

## 8.5. Estymacja rozkładu statystyki metodą bootstrap

Zauważmy, że metody przedstawione w podrozdz. 8.3 można łatwo wykorzystać do oszacowania rozkładu statystyki $\hat{\theta} = T(X_1, X_2, \ldots, X_n)$, gdzie $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ jest prostą próbą losową ze znanego rozkładu $F$. Wystarczy w tym celu uzyskać odpowiednio dużo wartości zmiennej $\hat{\theta}$, co odpowiada obliczeniu wartości statystyki $T$ dla wielu niezależnych prób. Mianowicie, niech dla $i = 1, 2, \ldots, k$, $\mathbf{X}_i = (X_{i1}, X_{i2}, \ldots, X_{in})$ będzie $i$-tą prostą próbą losową z rozkładu $F$ i $\hat{\theta}_i = T(\mathbf{X}_i)$. Wówczas histogram wartości $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k$ jest dla dużej liczby $k$ generowanych prób dobrym przybliżeniem rozkładu statystyki $\hat{\theta}$. Będziemy go nazywali **estymatorem rozkładu $\hat{\theta}$ otrzymanym metodą Monte Carlo**. Jednak z reguły nie znamy rozkładu $F$ i dysponujemy jedynie pojedynczą realizacją prostej próby losowej z tego rozkładu. Wtedy nie możemy wykorzystać powyższej konstrukcji. Co możemy zrobić? Jak wiemy, w sytuacji, gdy statystyka $T(x_1, x_2, \ldots, x_n)$ jest sumą $x_1 + x_2 + \ldots + x_n$, a rozkład $F$ jest rozkładem normalnym z nieznanymi parametrami, to przez standaryzację lub studentyzację możemy sprowadzić rozkład $\hat{\theta}$ do znanego rozkładu. Jeśli rozkład $F$ nie jest normalny, ale liczność próby $n$ jest duża, to możemy obliczyć przybliżony rozkład $\hat{\theta}$,

korzystając z przybliżenia normalnego, prawdziwego dla dowolnego rozkładu $F$ mającego skończoną wariancję. Dla małej liczności prób i dla statystyk $T(x_1, x_2, \ldots, x_n)$ o bardziej skomplikowanej postaci niż suma współrzędnych tych metod nie możemy stosować. Jeśli rozkład, z którego jest generowana próba jest znany z dokładnością do pewnych parametrów, np. wiemy, że jest to rozkład wykładniczy z parametrem $\lambda$, to możemy uzyskać estymator $\hat{\lambda}$ na podstawie próby, a następnie przybliżyć rozkład estymatora $\hat{\theta}$, generując wiele prób z rozkładu wykładniczego $\mathcal{E}(\hat{\lambda})$ i na ich podstawie obliczając wartości estymatora. Metoda ta zawodzi, gdy nic nie wiemy o postaci parametrycznej rozkładu.

### 8.5.1. Zasada bootstrap

Efron zaproponował następującą adaptację metody Monte Carlo umożliwiającą oszacowanie rozkładu $\hat{\theta}$ również w tej sytuacji. Załóżmy, że dysponujemy realizacją $x_1, x_2, \ldots, x_n$ prostej próby losowej i niech $\hat{F}$ oznacza dystrybuantę empiryczną dla tej próby zdefiniowaną w p. 3.4.2 (def. 3.5). Ponieważ $\hat{F}$ jest znanym przybliżeniem nieznanego rozkładu $F$, więc zastosujmy wyżej opisaną metodę estymacji rozkładu $\hat{\theta}$ do $\hat{F}$ zamiast do $F$, tzn. oceńmy rozkład estymatora $\hat{\theta}$ na podstawie wielu prób wygenerowanych z rozkładu $\hat{F}$. Prowadzi to do następującej definicji

**Definicja 8.2.** *Prostą próbę losową* $\mathbf{X}^* = (X_1^*, X_2^*, \ldots, X_n^*)$ *z rozkładu* $\hat{F}$ *dla ustalonej realizacji* $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ *nazywamy* ***próbą typu bootstrap*** *lub prościej* ***próbą bootstrap***.

W celu otrzymania realizacji próby bootstrap[1] dokonuje się $n$-krotnego losowania ze zwracaniem spośród elementów oryginalnej próby. Cała operacja jest przeprowadzana dla ustalonych wartości $\{x_1, x_2, \ldots, x_n\}$, a losowość w próbie $\mathbf{X}^*$ jest związana jedynie z losowymi wyborami pewnego elementu spośród $x_1, x_2, \ldots, x_n$ w każdym z $n$ ciągnięć. Tak więc zaobserwowaną realizację $x_1, x_2, \ldots, x_n$ traktujemy jako populację, z której czerpiemy proste próby losowe. Ten dziwny na pierwszy rzut oka pomysł okazał się bardzo pożyteczny i umożliwił częściowe ominięcie problemu dysponowania jedynie pojedynczą realizacją z rozkładu $F$. Zauważmy, że ponieważ dystrybuanta $\hat{F}$ przypisuje każdemu z punktów $x_i, i = 1, 2, \ldots, n$ prawdopodobieństwo wyboru równe $1/n$, oznacza to, że każda ze zmiennych $X_i^*$ niezależnie od pozostałych przyjmuje dowolną wartość próby z jednakowym prawdopodobieństwem. Próba bootstrap składa się z elementów oryginalnej próby, z których pewne mogą być w niej uwzględnione zero razy, pewne mogą być

[1] Określenie metoda bootstrap pochodzi od wyrażenia angielskiego: *pull oneself up by one's bootstraps* - wydobyć się z opresji używając własnych sił.

uwzględnione jeden raz itd. Zauważmy, że w próbie $\mathbf{X}^*$ elementy z reguły się powtarzają; prawdopodobieństwo, że każdy z elementów $x_i$ pojawi się tam dokładnie jeden raz, tzn. że $\mathbf{X}^* = \mathbf{x}$ wynosi tylko $n!/n^n$. Powtarzanie wartości nie powinno nas specjalnie dziwić, podobnie ma się sytuacja przy obserwacji liczby pojawiających się oczek przy kolejnych rzutach kostką. W obu przypadkach liczba potencjalnych wartości jest skończona. Oczywiście, moglibyśmy losować inną liczbę obserwacji z rozkładu $\hat{F}$ niż $n$, dla ustalenia uwagi przyjmijmy jednak, że liczności pierwotnej próby i próby bootstrap są równe. Indeks $*$ w oznaczeniu zmiennnej $X_i^*$ ma służyć odróżnieniu tej zmiennej od zmiennej $X_i$; pierwsza z nich ma rozkład jednostajny w punktach $x_1, x_2, \ldots, x_n$, druga ma rozkład $F$.

Uzasadnieniem propozycji Efrona jest następująca reguła, którą udowadnia się dla takich statystyk $T(\mathbf{X})$, że interesujący nas parametr $\theta$ spełnia równość $\theta = \tilde{T}(F)$ oraz jego estymator $\hat{\theta}$ ma postać $\hat{\theta} = T(\mathbf{X}) = \tilde{T}(\hat{F})$ dla pewnej funkcji $\tilde{T}$. Oznacza to, że nasz parametr $\theta$ jest pewną funkcją dystrybuanty, a jego estymator otrzymujemy, podstawiając jako wartość tej funkcji dystrybuantę empiryczną $\hat{F}$ w miejsce dystrybuanty $F$. Najprostszym przykładem takiej sytuacji jest wartość dystrybuanty w pewnym punkcie $\theta = F(t_0)$, której estymatorem jest $\hat{\theta} = \hat{F}(t_0)$. Również wartość średnia wraz ze swoim naturalnym estymatorem ma taką reprezentację: $\mu = T(F) = \int x dF(x)$ i $\bar{X} = T(F_n) = \int x d\hat{F}(x)$. Rzeczywiście, ponieważ dystrybuanta empiryczna jest funkcją przedziałami stałą i mającą skoki równe $1/n$ tylko w poszczególnych obserwacjach, mamy $\int x d\hat{F}(x) = n^{-1}(X_1 + X_2 + \ldots + X_n) = \bar{X}$. Podobnie ma się rzecz dla odchylenienia standardowego i kwantyli. Praktycznie wszystkie interesujące nas parametry i ich estymatory mają wymaganą reprezentację.

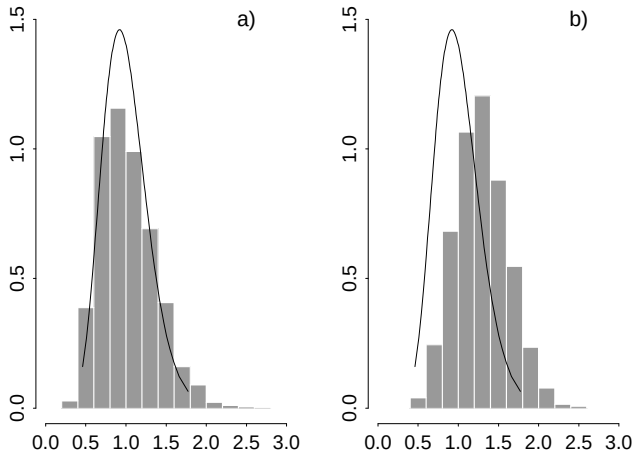
**Zasada bootstrap.** *Rozkład statystyki* $(T(\mathbf{X}^*) - \hat{\theta})$ *dla próby bootstrap przy ustalonych wartościach realizacji* $x_1, x_2, \ldots, x_n$ *jest dla regularnych statystyk* $T$ *bliski rozkładowi statystyki* $(T(\mathbf{X}) - \theta)$.

Zasada bootstrap mówi, że kształt rozkładu $T(\mathbf{X}^*)$ przy ustalonych wartościach realizacji $x_1, x_2, \ldots, x_n$ jest bliski kształtowi rozkładu $T(\mathbf{X})$. Fakt, że położenie rozkładu statystyki $T(X^*)$ jest przesunięte względem położenia rozkładu statystyki $T(\mathbf{X})$ o wielkość $\hat{\theta} - \theta$ jest dla porównania ich kształtu nieistotny. Tak więc ocena rozkładu $\theta = T(\mathbf{X})$ wygląda następująco:

– *wylosuj niezależne próby losowe bootstrap* $\mathbf{X}_1^*, \ldots, \mathbf{X}_k^*$ *na podstawie realizacji* $x_1, x_2, \ldots, x_n$,

– *oblicz wartości* $\hat{\theta}_1^* = T(\mathbf{X}_1^*) - \hat{\theta}$, $\hat{\theta}_2^* = T(\mathbf{X}_2^*) - \hat{\theta}$, $\ldots$, $\hat{\theta}_k^* = T(\mathbf{X}_k^*) - \hat{\theta}$,

– *skonstruuj histogram wartości* $\hat{\theta}_1^*, \hat{\theta}_2^*, \ldots, \hat{\theta}_k^*$. *Jest on przybliżeniem rozkładu statystyki* $\hat{\theta} - \theta$.



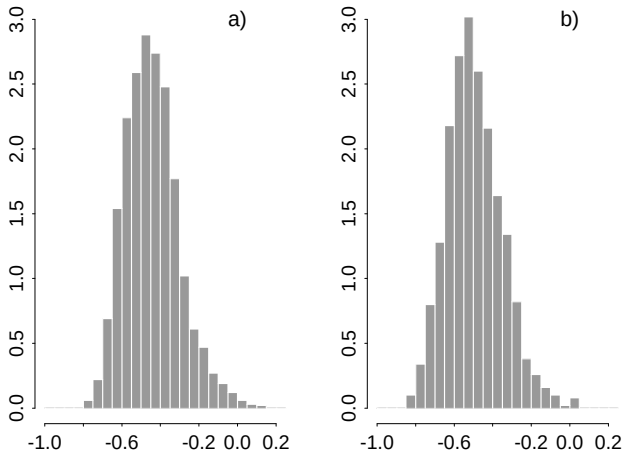

Rys. 8.6. Estymatory rozkładu średniej próbkowej otrzymane metodą bootstrap (przykł. 8.11)

Nazywamy go **estymatorem rozkładu $\hat{\theta}$ otrzymanym metodą bootstrap**. Z reguły potrzebujemy co najmniej $k = 1000$ prób bootstrap do otrzymania zadowalającego przybliżenia rozkładu $T(\mathbf{X}^*) - \hat{\theta}$. Zauważmy, że łatwo można sobie wyobrazić modyfikacje metody bootstrap, np. zamiast losować próby z rozkładu $\hat{F}$, możemy je losować z pewnej wygładzonej wersji dystrybuanty empirycznej. Takich użytecznych w pewnych sytuacjach modyfikacji nie będziemy tu omawiać.

**Przykład 8.11.** Na rysunku 8.6a przedstawiono estymator rozkładu średniej próbkowej otrzymany metodą bootstrap na podstawie próby o liczności $n = 25$ z rozkładu $\chi^2_1$. Do estymacji użyto $k = 2000$ prób typu bootstrap. Na rysunku przedstawiono też prawdziwą gęstość rozkładu średniej.[2] Na rysunku 8.6b przedstawiono analogiczny estymator rozkładu otrzymany na podstawie innej dwudziestopięcioelementowej próby z rozkładu $\chi^2_1$. Zauważmy, że kształty rozkładów na rysunkach a i b są bardzo podobne.

**Przykład 8.12.** Na rysunku 8.7a przedstawiono estymator typu bootstrap rozkładu współczynnika korelacji dla trzydziestoelementowej próby pochodzącej z dwuwymiarowego rozkładu normalnego $N(0, 0, 1, 1, -0{,}5)$. Użyto $k = 2000$ prób bootstrap. Na rysunku 8.7b przedsta-

[2] Zauważmy, że średnia ma taki sam rozkład jak zmienna $X/25$, gdzie $X \sim \chi^2_{25}$.

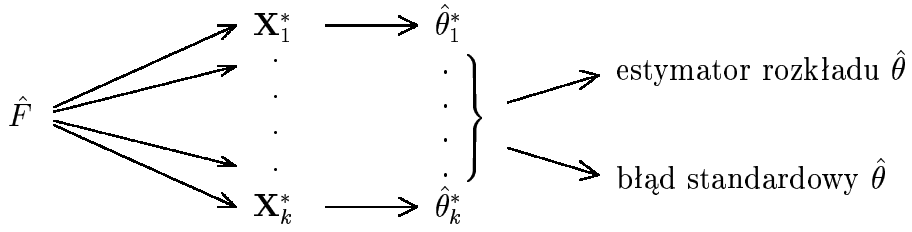



Rys. 8.7. a) Estymator rozkładu współczynnika korelacji otrzymany metodą bootstrap, b) rozkład współczynnika korelacji (przykł. 8.12)

wiono przybliżony rozkład współczynnika korelacji obliczony na podstawie $l = 1000$ prostych prób losowych o liczności $n = 30$ z rozkładu $N(0, 0, 1, 1, -0,5)$.

### 8.5.2. Błąd standardowy typu bootstrap

Powyższe postępowanie można przedstawić za pomocą następującego diagramu:




Rys. 8.8. Diagram przedstawiający realizację metody bootstrap

Omówmy prawą dolną część diagramu. Jest on również oparty na zasadzie bootstrap. Jeśli rozkład $T(\mathbf{X})$ możemy przybliżać przez odpowiednio przesunięty rozkład $T(\mathbf{X}^*)$, to w szczególności dowolny parametr kształtu pierwszego rozkładu może być przybliżany przez analogiczny parametr drugiego rozkładu. Wiemy bowiem, że parametr kształtu nie zależy od usytuowania rozkładu. Dla nas ważna jest przede wszystkim ocena zmienności

estymatora $\hat{\theta} = T(\mathbf{X})$, gdyż dysponując jedynie wartościami $x_1, x_2, \ldots, x_n$, nie jesteśmy z reguły w stanie zrobić tego bezpośrednio. Rozpatrzmy odchylenie standardowe $\sigma_{\hat{\theta}}$ jako miarę zmienności estymatora $\hat{\theta}$. Dysponując zmiennymi $\hat{\theta}_1^*, \hat{\theta}_2^*, \ldots, \hat{\theta}_k^*$, możemy wprowadzić następujące pojęcie:

**Definicja 8.3.** ***Błędem standardowym typu bootstrap*** *estymatora $\hat{\theta}$ nazywamy*

$$SE_{\hat{\theta}^*} = \sqrt{\frac{1}{k-1}\sum_{i=1}^{k}(\hat{\theta}_i^* - \bar{\theta}^*)^2},$$

*gdzie* $\bar{\theta}^* = k^{-1}\sum_{i=1}^{k}\hat{\theta}_i^*$.

Tak więc błąd standardowy typu bootstrap estymatora $\hat{\theta}$ jest niczym innym niż odchyleniem standardowym obliczonym na podstawie próby $\hat{\theta}_1^* = T(\mathbf{X}_1^*), \hat{\theta}_2^* = T(\mathbf{X}_2^*), \ldots, \hat{\theta}_k^* = T(\mathbf{X}_k^*)$. Stosuje się go jako ocenę odchylenia standardowego $\sigma_{\hat{\theta}}$. Z reguły do oceny odchylenia standardowego potrzebujemy mniej prób bootstrap niż do oceny całego rozkładu; zwykle $k = 50 - 200$. Zauważmy, że nasze postępowanie jest zupełnie analogiczne do szacowania parametrów rozkładu metodą Monte Carlo. Jedyna, ale zasadnicza różnica polega na tym, że ponieważ nie znamy rozkładu $F$ i nie możemy wygenerować z niego prób, zastępujemy go rozkładem $\hat{F}$ i z niego generujemy próby.

Metoda bootstrap, choć z pozoru uniwersalna, nie zawsze musi dawać dobre rezultaty.

**Przykład 8.13.** Niech $x_1, x_2, \ldots, x_{100}$ będzie realizacją prostej próby losowej $X_1, X_2, \ldots, X_n$ z rozkładu jednostajnego na przedziale $[0, \theta]$, gdzie prawy koniec przedziału $\theta$ jest nieznanym parametrem. Za estymator $\theta$ przyjmiemy największą obserwację w próbie: $\hat{\theta} = X_{(n)}$. Zauważmy, że dystrybuanta estymatora $\hat{\theta}$ ma postać $G(t) = P(X_{(n)} \leq t) = \prod_{i=1}^{n} P(X_i \leq t) = t^n$ dla $0 \leq t \leq 1$ i jest funkcją ciągłą na tym przedziale. Niech $X_1^*, X_2^*, \ldots, X_n^*$ będzie próbą bootstrap z rozkładu $\hat{F}$. Wtedy odpowiednik $\hat{\theta}^*$ estymatora $\hat{\theta}$ określony na podstawie próby bootstrap ma postać $\hat{\theta}^* = X_{(n)}^*$, gdzie $X_{(n)}^*$ jest największą obserwacją w próbie bootstrap. Obliczmy prawdopodobieństwo $P(\hat{\theta}^* = x_{(n)})$. Jest ono oczywiście równe prawdopodobieństwu, że obserwacja $x_{(n)}$ będzie zawarta w próbie bootstrap, a zatem

$$P(\hat{\theta}^* = x_{(n)}) = 1 - \left(\frac{n-1}{n}\right)^n \to 1 - e^{-1}.$$

Ponieważ z wyprowadzonej powyżej postaci dystrybuanty $G$ wynika, że rozkład $\hat{\theta} = X_{(n)}$ jest ciągły, a rozkład $\hat{\theta}^* = X^*_{(n)}$ ma nawet dla dużych $n$ skok w punkcie $x_{(n)}$, rozkłady te nie mogą być sobie bliskie. Zasada bootstrap w tym przypadku nie może być spełniona. Można to wyjaśnić zauważywszy, że interesujący nas parametr $\theta$ jest związany z zachowaniem dystrybuanty $F$ dla takich punktów $t$, że $F(t)$ jest bliskie 1. Dla takich $t$ dystrybuanta $\hat{F}$ nie jest dobrym oszacowaniem dystrybuanty $F$ i nie można się spodziewać, że próba wygenerowana z rozkładu $\hat{F}$ będzie pomocna przy ocenie parametru $\theta$.

### 8.5.3. Przedziały ufności typu bootstrap

Omówimy trzy rodzaje przedziałów ufności typu bootstrap.

**Przedział ufności typu bootstrap oparty na przybliżeniu normalnym.** Ten typ przedziału jest stosowany w sytuacji, gdy rozkład $\hat{\theta}^*$ jest w przybliżeniu normalny (co możemy bez trudu ocenić konstruując histogram wartości $\hat{\theta}^*_1, \hat{\theta}^*_2, \ldots, \hat{\theta}^*_k$ dla odpowiednio dużej liczby $k$). Wtedy na mocy zasady bootstrap w przybliżeniu normalny jest również rozkład $\hat{\theta}$ i przedział ufności dla parametru $\theta$ na poziomie ufności $1-\alpha$ ma postać

$$(\hat{\theta} - z_{1-\frac{\alpha}{2}}\sigma_{\hat{\theta}}, \hat{\theta} + z_{1-\frac{\alpha}{2}}\sigma_{\hat{\theta}}). \tag{8.3}$$

Zastępując w nim nieznane odchylenie estymatora $\hat{\sigma}$ bliskim mu błędem standardowym $SE_{\hat{\theta}^*}$ typu bootstrap, otrzymujemy **przedział ufności typu bootstrap oparty na przybliżeniu normalnym**

$$(\hat{\theta} - z_{1-\frac{\alpha}{2}}SE_{\hat{\theta}^*}, \hat{\theta} + z_{1-\frac{\alpha}{2}}SE_{\hat{\theta}^*}) \tag{8.4}$$

z przybliżonym poziomiem ufności $1-\alpha$.

**Przykład 8.11 cd.** Stosując wzór (8.4) do danych z przykł. 8.11, otrzymamy, że przedział ufności typu bootstrap oparty na przybliżeniu normalnym z przybliżonym poziomem ufności 0,95 dla parametru $\theta = 1$ wynosi (0,352, 1,691) dla próby o histogramie na rys. 8.6a oraz (0,656, 1,928) dla próby o histogramie na rys. 8.6b.

**Percentylowy przedział ufności typu bootstrap.** Oznaczmy przez $q^*_\alpha$ kwantyl rzędu $\alpha$ z rozkładu $\hat{\theta}^* - \hat{\theta}$. Wówczas

$$P_{\hat{F}}(q^*_{\frac{\alpha}{2}} \leq \hat{\theta}^* - \hat{\theta} \leq q^*_{1-\frac{\alpha}{2}}) = 1-\alpha$$

i na podstawie zasady bootstrap prawdopodobieństwo występujące po lewej stronie powyższego wzoru jest w przybliżeniu równe $P_F(q^*_{\frac{\alpha}{2}} \leq \hat{\theta} - \theta \leq q^*_{1-\frac{\alpha}{2}})$. Zatem

$$P_F(q^*_{\frac{\alpha}{2}} \leq \hat{\theta} - \theta \leq q^*_{1-\frac{\alpha}{2}}) = 1 - \alpha$$

i

$$P_F(\hat{\theta} - q^*_{1-\frac{\alpha}{2}} \leq \theta \leq \hat{\theta} - q^*_{\frac{\alpha}{2}}) = 1 - \alpha.$$

Przedział

$$(\hat{\theta} - q^*_{1-\frac{\alpha}{2}}, \hat{\theta} - q^*_{\frac{\alpha}{2}}), \tag{8.5}$$

nazywany **przedziałem percentylowym**, może być używany przy dowolnym kształcie rozkładu $\hat{\theta}$.

**Przykład 8.12 cd.** Stosując wzór (8.5) do danych z przykł. 8.12, otrzymamy, że percentylowy przedział ufności typu bootstrap wynosi $(-0{,}762, -0{,}212)$.

**Przedział ufności typu bootstrap oparty na studentyzacji**. Przedział ten jest oparty na podobnej idei co przedział percentylowy, ale zastosowanej do statystyk studentyzowanych. Rozpatrzmy $i$-tą próbę bootstrap $\mathbf{X}^*_i$ i obliczony na jej podstawie estymator $\hat{\theta}^*_i$. Jak oszacować odchylenie standardowe $\sigma_{\hat{\theta}^*_i}$? Można to zrobić podobnie, jak robiliśmy to w przypadku oceny odchylenia standardowego $\sigma_{\hat{\theta}}$: wylosujmy pewną liczbę, powiedzmy $l$, prób bootstrap z próby $\mathbf{X}^*_i$ i obliczmy, korzystając z def. 8.3, błąd standardowy $SE_{\hat{\theta}^*_i}$. Zauważmy, że rozpatrujemy tu dodatkowy poziom metody Efrona, generując próby bootstap „drugiego rzędu" na podstawie próby bootstrap „pierwszego rzędu". Następnie korzysta się z odpowiednika **zasady bootstrap dla statystyk studentyzowanych**, mówiącej, że rozkłady zmiennych

$$\frac{\hat{\theta} - \theta}{SE_{\hat{\theta}}} \quad i \quad \frac{\hat{\theta}^* - \hat{\theta}}{SE_{\hat{\theta}^*}}$$

są bliskie. Ponieważ nie znamy błędu standardowego $SE_{\hat{\theta}}$, analogicznie jak w przypadku pierwszego przedziału zastępujemy go przez bliski mu błąd standardowy $SE_{\hat{\theta}^*}$. Możemy wtedy powiedzieć, że rozkłady

$$\frac{\hat{\theta} - \theta}{SE_{\hat{\theta}^*}} \quad i \quad \frac{\hat{\theta}^* - \hat{\theta}}{SE_{\hat{\theta}^*}}$$

są bliskie. Drugi z rozkładów $(\hat{\theta}^* - \hat{\theta})/SE_{\hat{\theta}^*}$ możemy w naturalny sposób ocenić na podstawie próby $(\hat{\theta}^*_i - \hat{\theta})/SE_{\hat{\theta}^*_i}$, $i = 1, 2, \ldots, k$. Oznaczmy przez $\tilde{q}^*_\alpha$ kwantyl próbkowy obliczony na podstawie próby $(\hat{\theta}^*_i - \hat{\theta})/SE_{\hat{\theta}^*_i}$, $i = 1, 2, \ldots, k$. Korzystając z powyższej obserwacji i rozumując jak przy konstrukcji przedziału percentylowego, otrzymamy następujący przedział ufności dla parametru $\theta$ postaci

$$(\hat{\theta} - \tilde{q}^*_{1-\frac{\alpha}{2}} SE_{\hat{\theta}^*}, \hat{\theta} - \tilde{q}^*_{\frac{\alpha}{2}} SE_{\hat{\theta}^*}) \tag{8.6}$$

o przybliżonym poziomie ufności $1-\alpha$. Zauważmy, że przedział (8.6) ma taką samą postać jak przedział (8.4) z tą różnicą, że kwantyle rozkładu normalnego zostały zastąpione kwantylami z próby $(\hat{\theta}_i^* - \hat{\theta})/SE_{\hat{\theta}_i^*}$, $i = 1, 2, \ldots, k$. Tak więc rolę tabeli rozkładu normalnego odgrywa tu tabela sporządzona na podstawie danych. Wyznaczenie tego przedziału jest najbardziej kosztowne obliczeniowo. Załóżmy, że dokonaliśmy $k = 1000$ losowań prób bootstrap i $l = 25$ dla obliczenia błędu standardowego $SE_{\hat{\theta}_i^*}, i = 1, 2, \ldots, 1000$. Łącznie potrzebujemy więc generacji 25000 prób bootstrap. Z badań symulacyjnych wynika jednak, że ten właśnie przedział ufności ma poziom ufności bliższy wartości $1-\alpha$ niż poziom ufności dla przedzialu percentylowego. Przedział (8.6) jest szczególnie użyteczny, gdy estymator $\hat{\theta}$ jest estymatorem środka rozkładu.

### 8.5.4. Testowanie hipotez przy użyciu metody bootstrap

Podstawowa różnica między tradycyjnym podejściem do testowania hipotez a podejściem opartym na metodzie bootstrap polega na wykorzystaniu w tym drugim rozkładu typu bootstrap statystyki testowej przy spełnionej hipotezie $H_0$.

Powróćmy do analizy czasów wykonania czynności manualnej w próbie dziewczynek w przykł. 8.10 i rozpatrzmy problem testowania hipotezy $H_0$, że wartość oczekiwana wykonania tej czynności w populacji dziewczynek wynosi $\mu = 40$ sekund przeciwko hipotezie alternatywnej, że $\mu \neq 40$ sekund. Gdy jest uzasadnione przyjęcie założenia, że czasy wykonania mają rozkład normalny, do testowania hipotezy $H_0$ można użyć statystyki $T$

$$T = \frac{\bar{X} - 40}{S_{\bar{X}}},$$

która przy spełnieniu hipotezy $H_0$ ma rozkład $t$ Studenta z sześcioma stopniami swobody. W przypadku niemożności przyjęcia tego założenia możemy uciec się do metody bootstrap. Niech $t = (\bar{x} - 40)/s_{\bar{x}}$ będzie wartością próbkową statystyki $T$. Na mocy zasady bootstrap dla statystyk studentyzowanych przy spełnieniu hipotezy zerowej rozkład statystyki $T$ jest bliski rozkładowi statystyki $T^* = (\bar{X}^* - \bar{x})/SE_{\bar{X}^*}$, gdzie $\bar{X}^*$ oznacza średnią z próby bootstrap i $\bar{x} = 35{,}15$. Podobnie jak poprzednio, używając $k$ prób bootstrap możemy oszacować rozkład zmiennej $T^*$ i obliczyć przybliżone prawdopodobieństwo $p = P(|T^*| \geq |t|) \approx P(|T| \geq |t|)$ przy ustalonej próbie $x_1, x_2, \ldots, x_7$. Dla $k = 5000$ otrzymujemy $p = 0{,}367$; przy założeniu, że $T \sim t_6$, $P(|T| \geq |t|) = 0{,}396$. A zatem w tym przypadku wynik testu metodą bootstrap różni się bardzo nieznacznie od wyniku zwykłego testu $t$ Studenta.

Zauważmy, że metody bootstrap można użyć również do testowania hipotezy o równości rozkładów w problemie dwóch prób rozpatrzonym w przykł. 8.10. Różnica z testem permutacyjnym polega na losowaniu prób bootstrap (a więc *ze zwracaniem*) o liczności $m+n$ spośród współrzędnych wektora $\mathbf{v}$ i przypisaniu pierwszych $n$ spośród nich do populacji dziewczynek, a pozostałych $m$ do populacji chłopców. Ta różnica w losowaniu okazuje się w tym przypadku na tyle nieznaczna, że wynik obu testów jest podobny: estymator prawdopodobieństwa $p_{\mathbf{v}}$ otrzymany metodą bootstrap dla $k = 5000$ wynosi 0,097. Zauważmy jednak, że uzasadnienie stosowania testu permutacyjnego opierało się na stwierdzeniu 8.3, w którym zakładaliśmy, że jest spełniona hipoteza o równości rozkładów w obu populacjach. Stwierdzenie 8.3 przestaje być prawdziwe, gdy hipoteza zerowa mówi nie o równości rozkładów, ale tylko o równości pewnych ich parametrów, np. wartości średnich. W tej sytuacji nadal jednak możliwe jest użycie metody bootstrap, która nie opiera się na tym stwierdzeniu.

Analogicznie, zamiast testu permutacyjnego hipotezy o niezależności, o którym była mowa w p. 8.4.1, można zaproponować test niezależności typu bootstrap. Różnica miedzy testami polega na tym, że w teście typu bootstrap nie permutujemy losowo wyników pierwszego egzaminu, ale wielokrotnie losujemy dwie niezależne próby ze zwracaniem spośród wartości $(x_i)$ i $(y_i)$, $i = 1, 2, \ldots, n$ i na ich podstawie obliczamy wartości współczynnika korelacji $r^*$. Używając uzyskanych w ten sposób wielu wartości $r^*$ przybliżamy rozkład współczynnika korelacji przy założonej niezależności wyników obu egzaminów. Przyjmijmy na przykład, że wartość współczynnika korelacji między wynikami egzaminów wynosi 0,45. Jeśli wylosowaliśmy 1000 prób typu bootstrap i na podstawie histogramu obliczonych współczynników korelacji $r^*$ stwierdzamy, że współczynnik $r = 0{,}45$ odpowiada kwantylowi rzędu 0,982 tzn. tylko w przypadku osiemnastu prób bootstrap współczynnik korelacji był nie mniejszy od 0,45, to $p$–wartość testu bootstrap hipotezy $H_0$: $r = 0$ przeciwko $H_1$: $r > 0$ jest w przybliżeniu równa 0,018. Hipoteza $H_0$ zostanie zatem odrzucona na poziomie istotności $\alpha = 0{,}05$.

## 8.6. Zadania

**8.1.** Na podstawie definicji rozkładu $F$ Snedecora podanego w p. 3.3.2 wygenerować 500 elementową próbę z rozkładu $F_{2,4}$. Oszacować na podstawie otrzymanej próby kwantyle tego rozkładu rzędu 0,1, 0,2, ..., 0,9 i porównać z dokładnymi wartościami podanymi w tablicach. Skomentować dokładność przybliżenia w funkcji rzędu kwantyla.

**8.2.** Rozkładem logarytmiczno-normalnym z parametrami $(\mu, \sigma)$ nazywamy rozkład zmiennej losowej $X$ o tej własności, że przekształcona zmienna lo-

sowa $\ln(X)$ ma rozkład normalny o średniej $\mu$ i odchyleniu standarowym $\sigma$. Rozkłady logarytmiczno-normalne są często używane do modelowania czasów wystąpienia w produktach pęknięć spowodowanych zmęczeniem materiałowym.

**a)** Na podstawie powyższej definicji i wybranej metody generacji rozkładu normalnego, wygenerować próbę o liczności 1000 z rozkładu logarytmiczno-normalnego z parametrami $(2, 1)$. Skonstruować histogram rozkładu, oszacować jego medianę, średnią, wariancję oraz kwantyle rzędu 0,05 i 0,95. Skomentować otrzymane wyniki.

**b)** Wygenerować próby 200-elementowe z rozkładów logarytmiczno-normalnych z parametrami $(0, 1), (0{,}5, 1), \ldots, (4{,}5, 1), (5, 1)$ i obliczyć odpowiednie średnie próbkowe. Sporządzić wykres średnich próbkowych w funkcji parametru $\mu$. Zaproponować przekształcenie zmiennej $y$, którego można użyć do otrzymania przybliżonej zależności liniowej wykresu. Dokonać zaproponowanego przekształcenia danych i wyznaczyć estymator wyrazu wolnego i współczynnika nachylenia zależności liniowej metodą najmniejszych kwadratów. Na podstawie otrzymanych wyników oszacować średnią rozkładu logarytmiczno-normalnego z parametrami (0,8, 1). Wygenerować 200-elementową próbę z tego ostatniego rozkładu, obliczyć jego średnią próbkową i porównać z uprzednio otrzymaną wartością oszacowania. Porównać z wartością oczekiwaną rozkładu logarytmiczno-normalnego (zad. 3.10).

**8.3.** Rozpatrzyć rozkład zmiennej losowej

$$Y = I_1 X_1 + I_2 X_2,$$

gdzie $X_1, X_2, I_1, I_2$ są niezależnymi zmiennymi losowymi, $X_1$ ma rozkład normalny o średniej 0 i odchyleniu standardowym $\sigma = \sqrt{2}$, $X_2$ ma rozkład jednostajny na odcinku $(1, 2)$, $I_1$ ma rozkład Bernoulliego z $P(I_1 = 1) = 0{,}9$, a $I_2$ ma rozkład Bernoulliego z $P(I_2 = 1) = 0{,}1$.

**a)** Obliczyć średnią i wariancję rozkładu $Y$.

**b)** Wygenerować 100 prób o liczności 100 każda, obliczyć średnią i wariancję próbkową każdej próby i średnią wszystkich średnich i wariancji. Porównać z wynikami otrzymanymi w punkcie a.

**8.4.** Rozpatrzyć test hipotezy, że średnia $\mu$ rozkładu normalnego $N(\mu, 1)$ jest równa 0 przeciwko hipotezie alternatywnej, że $\mu \neq 0$ na podstawie prostej próby losowej o liczności 15 z tego rozkładu.

**a)** Sformułować postać statystyki testowej i obszaru krytycznego dla tego testu na poziomie istotności $0, 05$. Na podstawie 1000 prób 15 elementowych z rozkładu $N(0, 1)$ obliczyć frakcję odrzuceń hipotezy zerowej i porównać z zadanym poziomem istotności.

**b)** Rozpatrzyć obszar krytyczny postaci $\{\sqrt{15}|\bar{x}/s| \geq z_{0,975}\}$. Odpowiada to sytuacji, gdy nie znamy wartości rozproszenia rozkładu i traktujemy (niesłusznie) próbkową wartość rozproszenia jako prawdziwą wartość $\sigma$. Obliczyć frakcję odrzuceń w punkcie a w przypadku użycia tego obszaru. Porównać z wynikiem otrzymanym w punkcie a i skomentować.

**c)** Rozpatrzyć sytuację, gdy prawdziwy rozkład jest rozkładem normalnym $N(\mu, 1)$, gdzie $\mu = 1, 2, 3$. Postępując analogicznie jak w punkcie a obliczyć frakcję odrzuceń dla testu hipotezy $H_0$: $\mu = 0$. Skomentować wyniki. Jaka charakterystyka testu odpowiada otrzymanym frakcjom?

**8.5.** Wygenerować 200 15-elementowych prób z dwuwymiarowego rozkładu normalnego $N(0;0;1;1;0{,}5)$. Na ich podstawie sporządzić histogram współczynnika korelacji. Porównać z histogramem otrzymanym dla prób z rozkładu $N(0;0;1;1;0)$. W obu przypadkach dokonać przekształcenia współczynnika korelacji $f(r) = r\sqrt{13}/(1 - r^2)$ (por. zad. 4.7) i skonstruować histogramy dla otrzymanych wyników. Porównać otrzymane histogramy z gęstością rozkładu $t$ Studenta z 13 stopniami swobody.

**8.6.** Rozpatrzmy układ elektroniczny składający się z takich dziesięciu obwodów scalonych, że dla każdego z nich czas bezawaryjnego działania jest rozkładem wykładniczym z parametrem $\lambda = 1$. Przyjmijmy, że czasy bezawaryjnej pracy poszczególnych obwodów są niezależne i układ działa do momentu awarii któregokolwiek z obwodów. Interesuje nas rozkład czasu $T$ działania układu.

**a)** Skonstruować histogram czasu $T$ na podstawie 500 dziesięcioelementowych prób z rozkładu wykładniczego.

**b)** Wyprowadzić postać dystrybuanty czasu $T$. Na podstawie postaci tego rozkładu i korzystając z metody przekształcenia kwantylowego, skonstruować histogram rozkładu $T$ odpowiednio przekształcając 500-elementową próbę wygenerowaną z rozkładu jednostajnego.

**8.7. a)** Zaproponować metodę oszacowania całki $\int_0^1 e^{-x^2}\,dx$ metodą Monte Carlo i użyć jej dla próby o liczności $n = 500$. Porównać z wartością dokładną całki.

**b)** Wykonać punkt a w celu oszacowania całki $\sqrt{2\pi}^{-1} \int_{-\infty}^{\infty} x^6 e^{-x^2/2}\,dx$.

**8.8.** W celu uniknięcia trudności interpretacyjnych dla wykresu kwantylowego często wykorzystuje się następującą metodę. Dla ustalonej liczności $n$ próby rozpatrzmy 100 prób o tej liczności z rozkładu normalnego $N(0,1)$ i na ich podstawie oceńmy kwantyle rozkładu $k$-tej statystyki pozycyjnej $X_{k:n}$ dla próby z rozkładu normalnego i dla $k = 1, 2, \ldots, n$. Przyjmijmy, że kolejne próby są zapisane w kolumnach macierzy $n \times 100$. Najpierw uporządkujmy elementy od najmniejszego do największego *w kolumnach*; $k$-ty wiersz upo-

rządkowanej macierzy będzie odpowiadał próbie z rozkładu zmiennej losowej $X_{k:n}$. Rozpatrując piąty i dziewięćdziesiąty piąty element co do wielkości element $k$-tego *wiersza* otrzymamy kwantyle $q_{i,\,0{,}05}$ i $q_{i,\,0{,}95}$ rzędu 0,05 i 0,95 z tego rozkładu. Przeprowadzić cały eksperyment dla $n = 35$. Sporządzić wykresy $(q_{k,\,0{,}05}, z_{k/35})$ oraz $(q_{k,\,0{,}95}, z_{k/35})$ dla $k = 1, 2, \ldots, 34$, gdzie $z_{k/35}$ jest kwantylem rzędu $k/35$ standardowego rozkładu normalnego. Zinterpretować wykresy. Dokonać dodatkowego losowania 35-elementowej próby ze standardowego rozkładu normalnego, sporządzić dla niej wykres kwantylowy i sprawdzić, czy znajduje się w obszarze wyznaczonym przez wykresy sporządzone poprzednio. Przeprowadzić analogiczny eksperyment z 35 elementową próbą z rozkładu wykładniczego z parametrem 1, przesuniętym tak, aby jego średnia była równa 0. Uzasadnić, że przedział $(q_{k,\,0{,}05}, q_{k,\,0{,}95})$ zawiera z prawdopodobieństwem 0,9 $k$-ty punkt wykresu kwantylowego dla prostej próby losowej z rozkładu normalnego $N(0,1)$ i $k = 1, 2, \ldots, 34$.

**8.9.** Rozpatrzyć przykł. 4.1 dotyczący związku między wynikami kolokwium a wynikami egzaminu i skonstruować percentylowy przedział ufności metodą bootstrap dla współczynnika nachylenia $\beta_1$. Metoda bootstrap polega w tym przypadku na generacji prób typu bootstrap $\mathbf{e}_i^*$, $i = 1, 2, \ldots, k$, z rozkładu rezyduów $\mathbf{e} = (e_1, e_2, \ldots, e_n)'$ i utworzeniu próby bootstrap obserwacji $\mathbf{Y}_i^* = \hat{\mathbf{Y}} + \mathbf{e}_i^*$, na podstawie której estymujemy parametry $\beta_0$ i $\beta_1$. Dokonując wyboru $k = 500$ prób bootstrap skonstruować histogram dla wartości estymatorów współczynnika nachylenia i 95% percentylowy przedział ufności na podstawie tego histogramu. Porównać go z przedziałem ufności (1,5, 1,8) otrzymanym w p. 4.2.4.

**8.10.** W przykładzie 9.3 punktu 9.4.2 następnego rozdziału rozpatruje się zależność między łączną ilością opadów a ilością trawy zebranej z pewnego pastwiska podgórskiego w ciągu 6 lat. Używając współczynnika korelacji jako statystyki testowej przeprowadzić test permutacyjny i test typu bootstrap zależności między rozpatrywanymi cechami.

**8.11.** Na podstawie danych z przykł. 1.4 dotyczących latencji L3-N33 dla kończyny lewej i stosując metodę bootstrap, przetestować hipotezę, że wartość średnia rozkładu latencji wynosi 30 milisekund. Porównać wynik z wynikiem testu $t$ zastosowanego do tych danych.

ROZDZIAŁ 9

# Metody rangowe

## 9.1. Wprowadzenie

W rozdziale 3 omówiliśmy metody testowania hipotez dotyczących wielkości wartości średniej cechy w populacji i równości wartości średnich dwóch cech w sytuacji, gdy rozkład rozpatrywanej cechy, lub w drugim przypadku, rozkłady obydwu cech, są normalne. Podkreśliliśmy, że test $t$ stosowany w takich sytuacjach ma ważną własność odporności na umiarkowane odstępstwa rozkładów od rozkładu normalnego, zwłaszcza gdy wielkości rozpatrywanych prób nie są zbyt małe. Jednakże, oczywiście nie wszystkie rozkłady można uznać za umiarkowanie odległe od rozkładu normalnego, a przyjęcie takiego założenia staje się szczególnie karkołomne dla małych prób. Ponadto test $t$ dla dwóch prób wymaga założenia o równości wariancji w obydwu populacjach. Drastyczne niespełnienie jednego z powyższych założeń prowadzi do znacznego odstępstwa rzeczywistego poziomu istotności od zadanego poziomu istotności testu, a w przypadku, gdy posługujemy się $p$–wartością do niemożności poprawnego jej zinterpretowania. Zachodzi zatem pytanie, czy i w jaki sposób w takiej sytuacji możemy porównać pewne parametry rozkładów lub same rozkłady. Istnieje kilka metod postępowania w tym przypadku, jedna z nich polega na przekształceniu cechy tak, aby jej rozkład stał się w przybliżeniu normalny i wnioskowaniu na podstawie rozkładu przekształconego. Metody tej nie będziemy tutaj bliżej omawiali. Omówimy natomiast metody testowania, których własności przy spełnieniu hipotezy zerowej nie zależą w ogóle od rozkładu rozpatrywanych cech. Ich ideę przedstawimy najpierw dla sytuacji porównania rozkładów dwóch cech.

## 9.2. Porównanie rozkładu cech w dwóch populacjach

Załóżmy, że mamy do czynienia z dwiema cechami $X$ i $Y$ o rozkładach określonych przez dystrybuanty $F$ i $G$ odpowiednio. Rozkłady te chcemy porównać na podstawie dwóch niezależnych prób losowych $X_1, X_2, \ldots, X_{n_1}$ i $Y_1, Y_2, \ldots, Y_{n_2}$, przy czym pierwsza próba o liczności $n_1$ pochodzi z rozkładu $F$, a druga próba o liczności $n_2$ pochodzi z rozkładu $G$. Liczności $n_1$ i $n_2$ mogą, ale nie muszą być równe. Załóżmy na początku, że dystrybuanty $F$ i $G$ są ciągłe. Wtedy prawdopobieństwo, że w połączonej próbie $X_1, X_2, \ldots, X_{n_1}, Y_1, Y_2, \ldots, Y_{n_2}$ powtórzą się te same wartości jest zerowe. Ponieważ nie chcemy nic zakładać o kształcie ani o położeniu rozkładów $F$ i $G$, a więc nie chcemy przyjąć ich konkretnej postaci parametrycznej, taką sytuację nazywamy **sytuacją nieparametryczną**, a metody wnioskowania statystycznego używane w takim przypadku nazywamy **metodami nieparametrycznymi**. W rozdziale tym omówimy własności pewnej ważnej klasy metod nieparametrycznych zwanych metodami rangowymi. Zastanówmy się przez chwilę, jakie postaci hipotezy zerowej i alternatywnej są tutaj naturalne. W przypadku dwóch rozkładów normalnych $N(\theta_0, \sigma)$ i $N(\theta_1, \sigma)$, testując hipotezę $H_0$: $\theta_0 = \theta_1$, testowaliśmy w istocie hipotezę o równości obydwu rozkładów, gdyż odchylenie standardowe, które wraz z wartością średnią jednoznacznie wyznacza rozkład normalny, jest w tym przypadku takie samo dla obu rozkładów. W sytuacji nieparametrycznej właśnie ta równoważna postać hipotezy jest rozpatrywana jako hipoteza zerowa $H_0$: $F = G$. Zauważmy następnie, że w przypadku rozkładów normalnych jednostronna hipoteza alternatywna $H_1$: $\theta_1 > \theta_0$ jest równoważna faktowi, że dla każdej ustalonej wartości $x$ prawdopodobieństwo $P(Y > x)$ przyjęcia wartości większej od $x$ przez zmienną $Y$ jest nie mniejsze niż prawdopodobieństwo $P(X > x)$ przyjęcia wartości większej od $x$ przez zmienną $X$, i jednocześnie, że prawdopodobieństwa te są różne dla pewnej wartości $x_0$ (a więc hipoteza zerowa jest różna od hipotezy alternatywnej). W sytuacji nieparametrycznej tak właśnie formułujemy interesującą nas alternatywną hipotezę jednostronną. Za pomocą dystrybuant $F$ i $G$ można zapisać ją następująco:

$$G(t) \leq F(t) \quad \text{dla wszystkich } t \text{ i } G \neq F. \tag{9.1}$$

W przypadku, gdy zmienna $Y$ oznacza np. czas przeżycia pacjenta po zastosowaniu nowej terapii, a $X$ jest analogicznym czasem przeżycia pacjenta w grupie kontrolnej, hipoteza alternatywna (9.1) mówi, że pacjent po zastosowaniu terapii ma nie mniejsze szanse przeżycia roku, dwóch lat i ogólnie czasu $t$ w porównaniu z szansami przeżycia analogicznego czasu przez pa-

cjenta z grupy kontrolnej i szanse te są większe dla pacjenta z grupy eksperymentalnej dla pewnej wartości $t$.

W przypadku, gdy mamy przesłanki do przyjęcia, że zastosowanie terapii nie zmieniło istotnie kształtu rozkładu czasu przeżycia, a jedynie spowodowało jego przesunięcie w prawo, czyli zwiększenie jego parametru położenia, oznacza to, że nasza hipoteza alternatywna może być sformułowana w sposób bardziej szczegółowy niż w przypadku (9.1), mianowicie

$$G(t) = F(t - \Delta) \quad \text{dla wszystkich } t, \text{ gdzie } \Delta > 0. \tag{9.2}$$

### 9.2.1. Test Wilcoxona

Ideę konstrukcji statystyki testowej Wilcoxona do testowania hipotezy $H_0$: $F = G$ przeciwko alternatywie (9.1) lub (9.2) przedstawimy, omawiając następujący przykład.

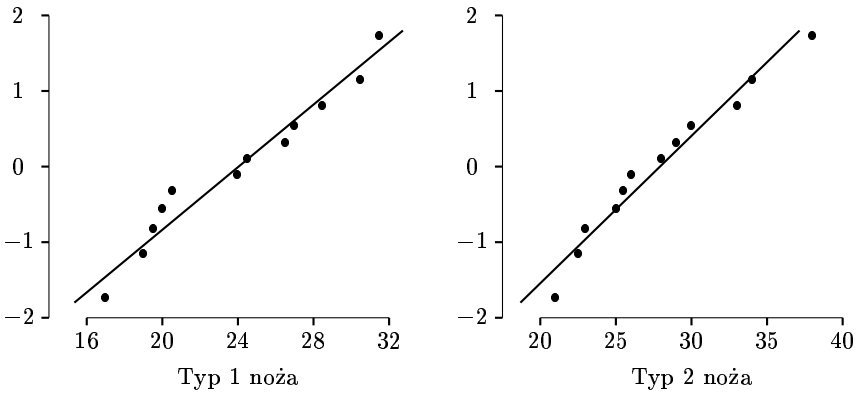
**Przykład 9.1.** Rozpatrzmy dwa rodzaje noża automatu tokarskiego, różniące się rodzajem stali, z której są wykonane. Nazwijmy je nożami typu 1 i typu 2. W celu porównania ich jakości rozpatrujemy zużycie noży podczas obróbki prętów metalowych. Jako interesującą nas cechę rozpatrujemy trwałość noża rozumianą jako czas jego prawidłowej pracy po zamontowaniu w automacie. Oznaczmy przez $F$ dystrybuantę rozkładu trwałości noża typu 1, a przez $G$ noża typu 2. Istnieje przypuszczenie, że użycie stopu wyprodukowanego według nowej technologii w nożu typu 2 zwiększa trwałość noża. Wydaje się więc uzasadnione testownie hipotezy o równości dystrybuant $F$ i $G$ przeciwko alternatywie (9.1).

Zauważmy, że przeprowadzenie eksperymentu w celu zebrania danych służących do testownia naszej hipotezy $H_0$ wymaga pewnej ostrożności. Powinniśmy wziąć pod uwagę fakt, że trwałość noży jest silnie zależna od twardości obrabianych prętów. Ponieważ praktycznie niemożliwe jest rozpatrzenie partii prętów o ustalonej twardości, konieczne jest przeprowadzenie zrandomizowanego eksperymentu, w którym pręty zostaną losowo przypisane typowi noża. Jest to zgodne z naszą dyskusją w podrozdz. 2.5 i prowadzi do jednakowych rozkładów twardości prętów przypisanych obu typom noża. W sytuacji jednak, gdy zmienność rozkładu twardości jest duża, może to utrudnić testowanie hipotezy $H_0$ nawet przy poprawnym eksperymencie zrandomizowanym, w szczególności może spowodować zwiększenie szans nieodrzucenia hipotezy

zerowej, mimo że jest ona fałszywa, czyli szans popełnienia błędu drugiego rodzaju. W celu uwzględnienia tego problemu można następująco zmodyfikować plan eksperymentu. Z partii prętów wybieramy losowo pewną liczbę prętów, i każdy z nich dzielimy na połowę. Następnie jedną z części przypisujemy losowo nożowi typu 1, a drugą nożowi typu 2. W ten sposób eliminujemy wpływ różnicy twardości prętów obrabianych przy użyciu różnych typów noży na ich trwałość.

W eksperymencie tym dla losowo wybranych prętów, z których każdy podzielono na dwie części, otrzymano następujące wyniki (w godzinach pracy noża):

| | Liczba godzin pracy noża |
|---|---|
| Typ 1 noża | 28,5 20 26,5 24 30,5 27 19 19,5 31,5 20,5 17 24,5 |
| Typ 2 noża | 34 22,5 25 29 38 25,5 28 23 33 26 21 30 |




Rys. 9.1. Wykresy kwantylowe dla prób trwałości dwu typów noży tokarskich

Wykresy kwantylowe dla prób trwałości obu typów noży przedstawiono na rys. 9.1. Widzimy, że rozkład trwałości noża typu 1 przypuszczalnie znacznie odbiega od rozkładu normalnego. Naszym celem jest testowanie równości obu rozkładów bez użycia tego założenia. Przypomnijmy, że gdy jest spełniona hipoteza alternatywna, możemy spodziewać się, że trwałości noża typu 2 mają tendencję do systematycznego przewyższania trwałości noży typu 1. Inaczej mówiąc, gdybyśmy uszeregowali *wszystkie* otrzymane trwałości noży typu 1 i noży typu 2 w jeden ciąg od wartości najmniejszej do największej, to wartości odpowiadające trwałościom noży typu 2 będą miały tendencję do zajmowania dalszych miejsc w ciągu w porównaniu z miejscami odpowiadającymi trwałościom noży typu 1. Uporządkowany ciąg wartości jest przedstawiony poniżej, przy czym trwałości noży typu 2 zostały wyróżnione pogrubioną czcionką.

17 19 19,5 20 20,5 **21 22,5 23** 24,0 24,5 **25 25,5 26** 26,5 27 **28** 28,5 **29 30** 30,5 31,5 **33 34 38**.

Fakt, że wartości odpowiadające drugiemu typowi noża mają tendencję do zajmowania dalszych miejsc w porównaniu z miejscami odpowiadającymi trwałościom noży typu 1 można wyrazić inaczej. Mianowicie, przypiszmy każdej obserwacji jej rangę w *połączonym* zbiorze obserwacji, to jest numer miejsca, które zajmuje ona w tym ciągu. Tak więc wartość 17 ma rangę równą 1, wartość 20 rangę równą 4, a największa wartość w połaczonej próbie równa 38 ma rangę równą 24. Obserwacje odpowiadające typowi 2 noża mają rangi

6 7 8 11 12 13 16 18 19 22 23 24.

Zastąpienie konkretnych wartości ich rangami umożliwia odstąpienie od przyjmowania założeń o parametrycznej postaci rozkładów cech, w szczególności założenia o ich normalności. Zauważmy ponadto, że małe zmiany wartości w próbach (w tym przypadku mniejsze od 0,5 godziny) nie spowodują zmiany uszeregowania obserwacji w połączonej próbie, a zatem również wartości ich rang nie ulegną zmianie. W tym sensie rangi są odporne na małe zmiany wartości obserwacji, które mogą wynikać na przykład z błędów pomiarowych. Zauważmy jednocześnie, że hipoteza alternatywna może być adekwatnie sformułowana za pomocą rang. W przypadku spełnienia hipotezy alternatywnej możemy spodziewać się, że suma rang wartości odpowiadających typowi 2 noża będzie znacznie przewyższać oczekiwaną wartość sumy rang przy spełnieniu hipotezy zerowej. W naszym przykładzie wartość ta wynosi $w = 6 + 7 + \ldots + 24 = 179$ i jest większa od wartości 150 równej wartości oczekiwanej rangi pojedynczego elementu przy spełnieniu hipotezy $H_0$ (to jest wartości 12,5) pomnożonej przez liczbę elementów drugiej próby równą 12. Ta intuicja prowadzi nas do rozpatrzenia własności statystyki zdefiniowanej analogicznie w przypadku ogólnym.

**Definicja 9.1.** ***Statystyką Wilcoxona*** *$W$ dla testowania hipotezy $H_0$: $F = G$ przeciwko alternatywie (9.1) lub (9.2) nazywamy sumę rang obserwacji $Y_1, Y_2, \ldots, Y_{n_2}$ w połączonej próbie $X_1, X_2, \ldots, X_{n_1}, Y_1, Y_2, \ldots, Y_{n_2}$.*

Czasami statystykę Wilcoxona wprowadza się w trochę inny sposób, udowadniając, że jest ona równa przekształconej liniowo wartości statystyki $t$ dla dwóch prób składających się odpowiednio z rang pierwszej i drugiej próby pierwotnej w próbie połączonej. Wydawać by się mogło, że w związku z tym zachowanie testu opartego na statystyce Wilcoxona będzie znacznie gorsze niż testu $t$, gdyż zastąpienie faktycznych wartości rangami oznacza dużą stratę informacji. Okazuje się, że tak nie jest, nawet dla sytuacji, gdy oba

rozkłady są rozkładami normalnymi z taką samą wariancją użycie testu rangowego powoduje tylko niewielkie zmniejszenie mocy testu. Rzecz w tym, że dla weryfikacji spełnienia alternatyw typu (9.2) istotne znaczenie mają tylko rangi, a nie konkretne wartości obserwacji w obydwu próbach. Zaznaczmy jednak, że ponieważ rozpatrywane w tym rozdziale statystyki bazują na rangach, a nie na wartościach w próbie, możemy spodziewać się stosunkowo dużej ich zmienności.

### 9.2.2. Własności statystyki Wilcoxona

Zauważmy, że bez względu na to czy jest spełniona czy nie hipoteza o równości rozkładów, możemy zawsze ograniczyć wartość statystyki $W$ z dołu i z góry

$$n_2(n_2+1)/2 \le W \le n_1 n_2 + n_2(n_2+1)/2.$$

Własność tę łatwo uzasadnić, zauważając, że minimalna wartość $W$ jest przyjmowana wtedy, gdy wszystkie elementy drugiej próby $Y_1, Y_2, \ldots, Y_{n_2}$ poprzedzają najmniejszy element z pierwszej próby. Analogicznie największą wartość $W$ otrzymujemy wtedy, gdy wszystkie wartości drugiej próby przewyższają największy element pierwszej próby. Minimalna i maksymalna wartość $W$ służą jako użyteczne punkty odniesienia określające zakres zmienności $W$. Jak udowodnimy za chwilę, wartość oczekiwana $W$ przy spełnieniu hipotezy $H_0$ leży w środku między tymi wartościami. Dla przykładu 9.1 na podstawie powyższych nierówności otrzymujemy, że minimalna wartość $W$ jest równa 78, a maksymalna 222, zatem wartość $W$ leży w trzeciej ćwiartce przedziału wyznaczonego przez wartość minimalną i maksymalną. Przedstawmy teraz własności statystyki Wilcoxona, gdy jest spełniona hipoteza $H_0$.

**Stwierdzenie 9.1.** *Jeśli $F = G$ i dystrybuanta $F$ jest ciągła,*

(1) *rozkład statystyki Wilcoxona $W$ nie zależy od dystrybuanty $F$ (i oczywiście od równej jej dystrybuanty $G$); $EW = n_2(n_2+n_1+1)/2$ i $\mathrm{Var}(W) = n_1 n_2(n_1+n_2+1)/12$.*

(2) *Ponadto dla dowolnej liczby $t$ $P((W-EW)/\sqrt{\mathrm{Var}(W)} \le t) \to \Phi(t)$, gdy $\min(n_1, n_2) \to \infty$ i $\Phi$ oznacza dystrybuantę standardowego rozkładu normalnego.*

Uzasadnimy tylko własność (1). Przyjmijmy $N = n_1 + n_2$. Zauważmy, że w przypadku spełnienia hipotezy $H_0$ rozkład rang elementów drugiej populacji jest równy rozkładowi $n_2$ liczb wylosowanych bez zwracania ze skończonej populacji $\{1, 2, \ldots, N\}$. Uzasadnia to niezależność rozkładu statystyki Wilcoxona $W$ od rozkładu $F$. Ponadto, zauważmy, że zmienna losowa $W/n_2$

jest estymatorem Horwitza–Thompsona wartości średniej $\mu$ populacji dla tego schematu losowania. Na mocy wniosku 7.1 jest to estymator nieobciążony średniej $\mu$, a zatem

$$E\Big(\frac{W}{n_2}\Big) = \mu = \frac{(1+2+\ldots+N)}{N} = \frac{N+1}{2}.$$

Ponadto na mocy tego wniosku, wariancja $\mathrm{Var}(W/n_2)$ jest równa

$$\frac{(1-\frac{n_2}{N})}{n_2} \times \frac{1}{N-1}\sum_{i=1}^{N}\Big(i-\frac{N+1}{2}\Big)^2 = \frac{n_1}{Nn_2}\frac{1}{(N-1)}\frac{N(N^2-1)}{12} = \frac{n_1}{n_2}\frac{N+1}{12},$$

z czego wynika część (1) stwierdzenia. Części (2) nie będziemy tutaj dowodzili, zaznaczając na marginesie, że nie wynika ona bezpośrednio z Centralnego Twierdzenia Granicznego, gdyż rangi odpowiadające obserwacjom drugiej próby mają wprawdzie ten sam rozkład, ale *nie* są niezależnymi zmiennymi losowymi.

Ponieważ na mocy stwierdzenia 9.1 rozkład statystyki Wilcoxona przy spełnionej hipotezie $H_0$ nie zależy od rozkładu $F$ populacji, możliwe jest zatem wyznaczenie tego rozkładu dla dowolnych liczności prób $n_1, n_2$. Jednocześnie na podstawie punktu (2) stwierdzenia, gdy obie z liczb $n_1$ i $n_2$ są duże, rozkład statystyki $W$ jest bliski rozkładowi normalnemu. Z reguły przyjmuje się, że gdy liczności $n_1$ i $n_2$ są obie większe od 8, przybliżenie normalne jest dostatecznie dokładne. Większość pakietów umożliwia jednak wyznaczenie dokładnej $p$–wartości dla otrzymanej wartości statystyki $W$ przy alternatywie (9.1) lub (9.2).

**Przykład 9.1 cd.** Dla danych z przykładu wartość statystyki Wilcoxona wynosi $w = 179$ i przy spełnionej hipotezie zerowej $P(W \geq \geq 179) = 0{,}052$. Tak więc istnieją dosyć przekonywające przesłanki przemawiające za odrzuceniem hipotezy o równości rozkładów obydwu populacji na rzecz hipotezy (9.1). Zobaczmy, ile wynosi przybliżona $p$–wartość obliczona na podstawie przybliżenia normalnego. Ponieważ dla $F = G$ na mocy stwierdzenia 9.1, $EW = (12 \times 25)/2 = 150$ i $\mathrm{Var}(W) = 12 \times 25 = 300$, więc stosując poprawkę w przybliżeniu normalnym (rozkład $W$ jest dyskretny!), otrzymujemy

$$P(W \geq 179) = P\Big(\frac{W-150}{\sqrt{300}} \geq \frac{178{,}5-150}{\sqrt{300}}\Big) = 0{,}0505.$$

Zauważmy ma marginesie, że dla testu $t$, którego nie powinniśmy w tym przypadku stosować, otrzymujemy $p$–wartość równą 0,034.

Statystykę Wilcoxona możemy również stosować do testowania hipotezy $H_0$ przeciwko alternatywie

$$G(t) \geq F(t) \quad \text{dla wszystkich } t \text{ i } F \neq G. \tag{9.3}$$

Alternatywa (9.3) odpowiada sytuacji, gdy wartości zmiennej $Y$ są systematycznie mniejsze od wartości zmiennej $X$. Oczywiście w tym przypadku zbiór krytyczny będzie się składał z odpowiednio małych wartości statystyki Wilcoxona. W celu znalezienia w tablicach $p$–wartości dla takiego przypadku, często stosuje się fakt, iż rozkład statystyki jest symetryczny względem swojej wartości oczekiwanej $n_2(n_1 + n_2 + 1)/2$, z czego wynika, że $P(W \geq w) = P(W \leq n_2(n_1 + n_2 + 1) - w)$. Analogicznie, statystyki Wilcoxona możemy użyć do testowania hipotezy dwustronnej mówiącej, że zachodzi albo alternatywa (9.2), albo alternatywa (9.3). Zbiór krytyczny jest wtedy symetryczny względem wartości $n_2(n_1+n_2+1)/2$ i składa się z odpowiednio dużych i małych wartości statystyki Wilcoxona. Skonstatujmy jeszcze, że w niektórych pakietach zamiast statystyki Wilcoxona istnieje możliwość obliczenia tzw. statystyki Manna–Whitneya $U$ równej liczbie takich par $1 \leq i \leq n_1, 1 \leq j \leq n_2$, że $Y_j - X_i > 0$. Bez szczegółowego omawiania tej ostatniej, stwierdźmy tutaj jedynie, że jest ona równa statystyce Wilcoxona pomniejszonej o wartość $n_2(n_2 + 1)/2$, tak więc są one sobie równoważne[1]. Oznacza to, że wyniki testowania przy zastosowaniu obu statystyk są takie same.

Można stwierdzić, że gdy oba rozpatrywane rozkłady są rozkładami normalnymi o tej samej wariancji i różnią się tylko wartością średnią, to moc testu $t$, który ma optymalne własności dla tej sytuacji, tylko nieznacznie przewyższa moc testu Wilcoxona. Natomiast, gdy odstępstwa od normalności są znaczne, moc testu Wilcoxona może być znacząco większa od mocy testu $t$. W szczególności dzieje się tak w sytuacji modelu (9.1), gdy dystrybuanta $F$ odpowiada rozkładowi wyraźnie skośnemu.

Dotąd przyjmowaliśmy, że w połączonej próbie nie powtarzają się takie same obserwacje. Mówimy w takiej sytuacji, że w próbie nie ma **obserwacji związanych**[2]. W przypadku ciągłych rozkładów $F$ i $G$ zdarzenie polegające na wystąpieniu co najmniej jednej pary obserwacji związanych jest równe 0. Możliwość użycia statystyki Wilcoxona nie ogranicza się jednak tylko do rozkładów ciągłych. Dla rozkładów dyskretnych musimy jednak uwzględnić możliwość wystąpienia obserwacji związanych i zmodyfikować definicję rangi obserwacji. W sytuacji wystąpienia kilku równych sobie obserwacji jako rangę każdej z nich przyjmujemy średnią rang im przypisanych,

[1] Wynika to z faktu, że ranga w próbie połączonej $j$-tej statystyki pozycyjnej $Y_{(j)}$ jest równa $j$ plus liczba takich $X_i$, że $X_i \leq Y_{(j)}$.

[2] Zwanych również więzami lub powiązaniami.

gdy wszystkie obserwacje ustawimy w porządku niemalejącym, przy czym kolejność obserwacji równych sobie jest dowolna. Tak więc dla uporządkowanego ciągu obserwacji 1, 5, 5, 5, 11, 13, odpowiednie rangi wynoszą 1, 3, 3, 3, 5, 6. Dla tak zmodyfikowanej statystyki $W$ i dyskretnego rozkładu $F$ jest konieczna adaptacja stwierdzenia 9.1. Nie będziemy tego tutaj robić, zaznaczmy jedynie, że część (2) stwierdzenia pozostaje prawdziwa, jeśli do standaryzacji użyjemy zmodyfikowanej wartości średniej i zmodyfikowanego odchylenia standardowego. Część pakietów dopuszcza obliczenie przybliżonej wartości $p$–wartości dla statystyki Wilcoxona w przypadku występowania więzów na podstawie przybliżenia normalnego.

### 9.2.3. Estymacja parametru przesunięcia $\Delta$

Rozpatrzmy sytuację, gdy wiemy, że dla pewnego $\Delta > 0$ jest spełniony warunek (9.2), tzn. rozkład $G$ powstaje przez przesunięcie w prawo rozkładu $F$. Zauważmy po pierwsze, że w takiej sytuacji możemy prosto zaadaptować test Wilcoxona do testowania hipotezy $H_0$: $\Delta = \Delta_0$. W tym celu wystarczy zauważyć, że założenie (9.2) odpowiada stwierdzeniu, że zmienna $X_i + \Delta$ ma rozkład $G$. Inaczej mówiąc, jeśli $\Delta = \Delta_0$, to zmienna $Y_i - \Delta_0$ ma rozkład $F$. Tak więc, w celu testowania hipotezy $H_0$: $\Delta = \Delta_0$ przeciwko hipotezie $\Delta > \Delta_0$, wystarczy zastosować test Wilcoxona do prób $X_1, X_2, \ldots, X_{n_1}$ i $Y_1 - \Delta_0, Y_2 - \Delta_0, \ldots, Y_{n_2} - \Delta_0$ z odpowiednim zbiorem krytycznym postaci $\{W \geq w\}$. Podobnego rozumowania można użyć do estymacji parametru przesunięcia $\Delta$. Mianowicie zauważmy, że gdy zachodzi warunek (9.2) i $\Delta = 0$, to mediana rozkładu zmiennej $Y_i - X_j$ dla dowolnych indeksów $i, j$ jest równa 0. Ogólnie, dla dowolnej wartości parametru $\Delta$ mediana rozkładu zmiennej $(Y_i - \Delta) - X_j$ będzie równa 0. Ponieważ dla dowolnej zmiennej $Z$ mediana rozkładu zmiennej $Z + a$ jest równa medianie zmiennej $Z$ plus $a$, otrzymujemy, że naturalnym estymatorem $\Delta$ jest mediana z próby

$$D_{ij} = Y_j - X_i \qquad i = 1, 2, \ldots, n_1, j = 1, 2, \ldots, n_2$$

o liczności $n_1 n_2$. Nosi on nazwę **estymatora Hodgesa–Lehmanna wielkości przesunięcia w problemie dwóch prób**. Używając podobnego rozumowania, możemy uzasadnić następującą konstrukcję przedziału ufności dla wielkości $\Delta$. Mianowicie, niech $\alpha$ będzie taką wartością, że dla pewnej wartości $w_\alpha$

$$P(w_\alpha \leq W - n_2(n_2+1)/2 \leq n_1 n_2 - w_\alpha) = P(w_\alpha \leq U \leq n_1 n_2 - w_\alpha) = 1 - \alpha,$$

gdzie $U$ jest statystyką Manna–Whitneya zdefiniowaną poprzednio. Oczywiście, ponieważ statystyka Wilcoxona przyjmuje wartości naturalne, $w_\alpha$ jest też liczbą naturalną. Przedział ufności dla $\Delta$ na poziomie ufności $1 - \alpha$ ma

postać $[D_{(w_\alpha)}, D_{(n_1 n_2 - w_\alpha + 1)})$, gdzie $D_{(k)}$ oznacza $k$-tą statystykę pozycyjną w ciągu $\{D_{ij}\}$. Przykład konstrukcji przedziału ufności dla parametru $\Delta$ będzie rozpatrzony w zad. 9.3.

### 9.2.4. Test Kołmogorowa–Smirnowa

Zauważmy na zakończenie tego podrozdziału, że dla problemu porównania rozkładów w dwóch populacjach istnieje naturalny odpowiednik testu zgodności Kołmogorowa omówionego w p. 3.4.2. Nosi on nazwę **testu Kołmogorowa–Smirnowa** i stosuje się go czasami do testowania $H_0$: $F = G$ przeciwko alternatywie $H_1$: $F \neq G$. Statystyka testowa tego testu jest równa

$$T_{KS} = \sup_{x \in R} |F_{n_1}(x) - G_{n_2}(x)|,$$

gdzie $F_{n_1}$ i $G_{n_2}$ są odpowiednio dystrybuantami empirycznymi prób $X_1, X_2, \ldots, X_{n_1}$ i $Y_1, Y_2, \ldots, Y_{n_2}$. Okazuje się, że wartość statystyki $T_{KS}$ jest równa maksymalnej wartości różnicy dystrybuant empirycznych obliczonych w punktach połączonej próby. Przy spełnieniu hipotezy zerowej i gdy dystrybuanta $F$ jest ciągła, rozkład statystyki Kołmogorowa–Smirnowa nie zależy od rozkładu $F$. Nie jest to dziwne, jeśli przekonamy się, że jest to w istocie statystyka oparta na rangach. Oznaczmy przez $R_i$ rangę obserwacji $X_i$ w próbie $X_1, X_2, \ldots, X_{n_1}$, a przez $\tilde{R}_i$ rangę tej obserwacji w próbie połączonej. Wtedy łatwo przekonać się, że

$$|F_{n_1}(X_i) - G_{n_2}(X_i)| = \left| \frac{R_i}{n_1} - \frac{(\tilde{R}_i - R_i)}{n_2} \right|.$$

Stąd prosto już wynika, że statystyka $T_{KS}$ jest oparta na rangach. Rozkład asymptotyczny tej statystyki przy spełnionej hipotezie $H_0$ został wyprowadzony przez Smirnowa. Testu tego nie będziemy tutaj dokładnie omawiać, gdyż podobnie jak w przypadku testu zgodności Kołmogorowa, własności mocy tego testu są często niezadowalające.

## 9.3. Testy porównania rozkładów dla par obserwacji

### 9.3.1. Test Wilcoxona dla par obserwacji

Przedstawmy teraz test Wilcoxona dla par obserwacji $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$, gdzie pary są wzajemnie niezależne, ale zmienne w parze mogą być zależne. Załóżmy ponadto, że pary mają ten sam rozkład dwuwymiarowy. Przypomnijmy, że sytuacja ta odpowiada np. pomiarowi pewnej

zmiennej dla tych samych jednostek przed i po zastosowaniu terapii. Test, który omówimy jest również zwany testem znakowanych rang. W odróżnieniu od ujęcia przedstawionego w podrozdz. 3.4 nie chcemy zakładać, że rozkład par $D_i = X_i - Y_i$ jest normalny. Podobnie jak w poprzednim punkcie chcielibyśmy bez przyjmowania założeń parametrycznych zbadać równość rozkładów w obydwu populacjach przeciwko hipotezie alternatywnej (9.2) lub (9.3). Interesuje nas zatem testowanie tej samej hipotezy co poprzednio, jedyna różnica polega na zmianie schematu obserwacji: nie dysponujemy teraz dwiema niezależnymi próbami z obu populacji. Ponieważ będziemy chcieli oprzeć nasz test na zachowaniu się różnic $D_i$, musimy nasze uprzednio sformułowane hipotezy tak zmodyfikować, aby odnosiły się one do wspólnego rozkładu $F$ tych różnic. Hipotezę $H_0$ o równości rozkładów zastępuje się hipotezą ogólniejszą, że zmienne $D_i = X_i - Y_i$ i $-D_i = Y_i - X_i$ będą miały ten sam rozkład, albo inaczej, że rozkład zadany przez dystrybuantę $F$ jest symetryczny względem 0. Zauważmy, że dystrybuanta w punkcie $t$ zmiennej $-D_i$ ma postać $1-F(-t)$, tak więc naszą hipotezę zerową możemy sformułować następująco:

$$H_0\colon\ F(t) = 1 - F(-t) \quad \text{dla każdego } t. \tag{9.4}$$

Odpowiednikiem hipotezy (9.1) jest hipoteza, że wartości zmiennej $-D_i$ mają tendencję do systematycznego przewyższania wartości zmiennej $D_i$, czyli, że zachodzi hipoteza o lewostronnej skośności rozkładu $F$

$$H_1\colon\ 1 - F(-t) \leq F(t) \text{ dla każdego } t, \tag{9.5}$$

a odpowiednikiem hipotezy (9.3) jest hipoteza o prawostronnej skośności rozkładu $F$

$$H_1\colon\ 1 - F(-t) \geq F(t) \text{ dla każdego } t. \tag{9.6}$$

Rozpatrzmy następujący przykład.

**Przykład 9.2.** W celu zbadania szybkości procesu uczenia się myszy wykonano eksperyment polegający na pomiarze czasu potrzebnego myszom laboratoryjnym na znalezienie wyjścia z labiryntu, do którego były wpuszczone ustaloną liczbę razy. Codziennie wykonano po jednym eksperymencie. Uczenie się przez mysz topografii jedynej drogi prowadzącej do wyjścia w labiryncie odpowiadałoby temu, że czasy potrzebne na pokonanie labiryntu w kolejnych dniach byłyby systematycznie mniejsze od analogicznych czasów pokonania labiryntu w dniach poprzedzających. W poniższej tabeli przedstawiono czasy pokonania labiryntu (w sekundach) dla sześciu myszy odpowiednio pierwszego i dziesiątego dnia eksperymentu.

| | Czas pokonania labiryntu [s] | | | | | | |
|---|---|---|---|---|---|---|---|
| Pierwszy dzień eksperymentu | $x_i$ : | 85 | 122 | 162 | 206 | 121 | 250 |
| Dziesiąty dzień eksperymentu | $y_i$ : | 87 | 82 | 158 | 96 | 131 | 121 |

Ciąg różnic $d_i$ między czasami pokonania labiryntu pierwszego i dziesiątego dnia (czyli czas dnia pierwszego minus czas dnia dziesiątego) ma postać

$d_i$: $-2$ 40 4 110 $-10$ 129.

Jeśli czasy przebycia labiryntu uległy systematycznemu skróceniu, to dodatnie różnice powinny bardziej odstawać od zera niż różnice ujemne. W celu stwierdzenia, czy tak rzeczywiście jest uporządkujmy wartości bezwzględne $|d_i|$:

$|d_i|$: 2 **4** 10 **40** **110** **129**,

gdzie zostały wyróżnione wartości bezwzględne odpowiadające różnicom dodatnim. Statystyka testowa w tym przypadku opiera się na wartościach wyróżnionych.

**Definicja 9.2.** ***Statystyka testowa Wilcoxona*** $W^+$ *dla par obserwacji jest zdefiniowana jako suma rang wartości bezwzględnych różnic odpowiadających różnicom dodatnim.*

W naszym przykładzie statystyka $W^+$ przyjmuje wartość $w^+ = 2+4+5+6 = 17$. Oczywiście w tym przypadku spodziewamy się, że czasy pobytu w labiryncie po dziesięciu dniach uległy systematycznemu skróceniu, a zatem interesuje nas czy dane wskazują na spełnianie alternatywy (9.6). Można udowodnić, że z dokładnością do stałej statystyka $W^+$ równa się sumie wszystkich rang pomnożonych przez znaki odpowiadających im różnic. Dlatego czasami nazywa się statystykę Wilcoxona **statystyką znakowanych rang**.

### 9.3.2. Własności statystyki Wilcoxona dla par obserwacji

Możemy teraz sformułować stwierdzenie analogiczne do stwierdzenia 9.1.

**Stwierdzenie 9.2.** *Załóżmy, że dystrybuanta $F$ jest ciągła i zachodzi hipoteza (9.4).*

(1) *Wówczas rozkład statystyki $W^+$ nie zależy od dystrybuanty $F$; $EW^+ = n(n+1)/4$ i $\mathrm{Var}(W^+) = n(n+1)(2n+1)/24$.*

(2) *Ponadto dla dowolnej liczby $t$ $P((W^+ - EW^+)/\sqrt{\mathrm{Var}(W)} \le t) \to \Phi(t)$, gdy $n \to \infty$.*

Stwierdzenia tego nie będziemy dowodzić. Skonstatujmy jedynie, że z reguły zaleca się stosowanie przybliżenia normalnego dla $n \geq 25$. Zauważmy ponadto, że możliwe wartości statystyki $W^+$ są liczbami naturalnymi z przedziału $[0, n(n+1)/2]$ i rozkład tej statystyki jest symetryczny względem środka tego przedziału. Dla dyskretnego rozkładu $F$ problem przypisania rang takim samym wartościom różnic rozwiązujemy jak poprzednio, przypisując każdej z takich samych wartości średnią z rang. Zerowe wartości różnic się odrzuca. Oczywiście, w takim przypadku postać wartości oczekiwanej i wariancji w stwierdzeniu 9.2 (1) ulega zmianie, ale część (2) tego stwierdzenia dla zmodyfikowanych wartości wariancji i wartości średniej pozostaje prawdziwa.

**Przykład 9.2 cd.** W rozpatrywanym przykładzie $EW^+ = (6 \times 7)/$ $/4 = 10{,}5$ i $\sigma_{W^+} = (6 \times 7 \times 13/24)^{1/2} = 4{,}77$. Poniżej zamieszczono wydruk pakietu S-Plus dla tego przykładu. Dokładna $p$–wartość dla statystyki Wilcoxona $W^+$ dla testowania hipotezy $H_0$ przeciwko alternatywie (9.6) równa $P(W^+ \geq 17)$ wynosi w tym przypadku 0,109 i nie może stanowić przesłanki do odrzucenia hipotezy $H_0$. Zauważmy, że przy uwzględnieniu poprawki standaryzowana wartość $W^+$ wynosi $(16{,}5 - 10{,}5)/4{,}77 = 1{,}05$ i odpowiadająca $p$–wartość przy założeniu rozkładu normalnego dla statystyki $W^+$ wynosi 0,147, a więc różni się znacznie od prawdziwej $p$–wartości.

**Tabela 9.1. Wydruk obliczeń za pomocą pakietu S-Plus dotyczących testu Wilcoxona dla przykł. 9.2**

```
x <- c(85, 122, 162, 206, 121, 250)
y <- c(87, 82, 158, 96, 131, 121)
wilcox.test(x, y, alternative = ''greater'', exact = T, paired = T)

        Exact Wilcoxon signed-rank test

data:  x and y
signed-rank statistic V = 17, n = 6, p-value = 0.1094
alternative hypothesis: true mu is greater than 0
```

Komentarza wymaga postępowanie przy testowaniu hipotezy $H_0$ przeciwko alternatywie (9.5). Oczywiście w tym przypadku zbiór krytyczny ma postać $\{W^+ \leq w\}$ dla pewnego $w$. To postępowanie jest równoważne modyfikacji statystyki Wilcoxona dla par jako sumy rang wartości bezwzględnych różnic odpowiadających różnicom *ujemnym* i rozpatrzenie takiego samego zbioru krytycznego, jak przy hipotezie alternatywnej (9.6) lub zamianie miejscami zmienych $X_i$ i $Y_i$ bez jakiejkolwiek zmiany definicji statystyki Wilcoxona

i metody konstrukcji zbioru krytycznego. W różnych podręcznikach statystyki przyjmuje się różne konwencje definiowania statystyki Wilcoxona dla par obserwacji i w związku z tym warto zwrócić uwagę, jak jest sformułowana hipoteza alternatywna i czy różnice $d_i$ oznaczają różnice między wartościami pierwszej i drugiej współrzędnej próby, czy też odwrotnie.

### 9.3.3. Estymacja parametru przesunięcia $\Delta$

Rozpatrzmy teraz sytuację, gdy mamy przesłanki do przyjęcia, że rozkład obserwacji w grupie po zastosowaniu terapii jest przesuniętym o stałą $\Delta$ rozkładem obserwacji w tej grupie przed zastosowaniem terapii. Z założenia tego wynika, że rozkład różnic spełnia warunek

$$F(t) = F_0(t - \Delta) \qquad \text{dla wszystkich } t, \tag{9.7}$$

gdzie $F_0$ jest dystrybuantą rozkładu symetrycznego względem zera i $\Delta > 0$ odpowiada sytuacji podobnej do tej w przykładzie, gdy wartości w pierwszej grupie systematycznie przewyższają wartości w drugiej. Przy założeniu, że jest spełniony warunek (9.7) hipotezę $H_0$: $\Delta = 0$, testujemy używając statystyki Wilcoxona $W^+$, identycznie jak to uczyniliśmy poprzednio. W przypadku, gdy hipoteza zerowa zostanie odrzucona, interesujące staje się oszacowanie wartości $\Delta$, która może odpowiadać szybkości uczenia się lub stopniowi efektywności terapii. Powtarzając rozumowanie z p. 9.2.3, można stwierdzić, że dobre intuicyjne uzasadnienie ma estymator zdefiniowany jako mediana wartości

$$B_{ij} = \frac{1}{2}(D_i + D_j) \qquad \text{dla } i \le j \le n.$$

Estymator ten jest zwany **estymatorem Hodgesa–Lehmanna wielkości przesunięcia $\Delta$ dla par obserwacji**. Uśrednianie występujące w jego definicji ma na celu poprawę precyzji estymacji $\Delta$. W przypadku spełnienia założenia (9.7) możemy analogicznie jak robiliśmy to w przypadku hipotezy (9.2) zbudować przedział ufności dla parametru $\Delta$, posługując się przy tym zamiast statystyką $W$ statystyką Wilcoxona $W^+$ dla par obserwacji. Wystarczy w tym celu dla ustalonego $\alpha$, używając rozkładu $W^+$ lub jego przybliżenia normalnego, znaleźć taką liczbę $w_\alpha^+$, że prawdopodobieństwo $P(w_\alpha^+ \le W^+ \le n(n+1)/2 - w_\alpha^+)$ jest w przybliżeniu równe $1-\alpha$. Wówczas przedział ufności na poziomie w przybliżeniu równym $1-\alpha$ ma postać

$$[B_{(w_\alpha^+)}, B_{(n(n+1)/2 - w_\alpha^+ - 1)})$$

gdzie $B_{(k)}$ jest $k$-tą statystykę pozycyjną w ciągu $\{B_{ij}\}$ dla $1 \le i < j \le n$.

### 9.3.4. Test znaków

Przy okazji omawiania testu Wilcoxona dla par, warto zauważyć, że przy takim schemacie obserwacyjnym można skonstruować prosty test nieparametryczny dla hipotezy, że **mediana rozkładu różnicy współrzędnych pary jest równa 0**. Test ten nie jest oparty na rangach. Zauważmy mianowicie, że przy spełnieniu tej hipotezy $p = P(D_i < 0) = 1/2$, jeśli tylko rozkład różnic $D_i$ jest ciągły. W takiej sytuacji zmienna $S$ równa liczbie dodatnich $D_i$, ma rozkład dwumianowy z parametrami $n$ i $1/2$. Nie interesują nas zatem, tak jak w teście Wilcoxona, wielkości różnic ujemnych, a jedynie ich liczba. Tak więc statystyki $S$ możemy użyć do testowania równoważnej hipotezy $H_0$: $p = 1/2$ (odpowiadającej medianie rozkładu różnicy równej 0) przeciwko alternatywie $H_1$: $p > 1/2$ (równoważnej temu, że mediana rozkładu różnicy jest mniejsza od 0). Oczywiście test ten, zwany testem znaków ma szanse być porównywalnie dobrym testem do testu Wilcoxona, gdy oba rozkłady są tylko przesunięte względem siebie, ale nie różnią się kształtem, tak jak się to postuluje w hipotezie (9.2). Jednocześnie jest intuicyjne, że jeśli mediana rozkładu różnicy jest równa zeru, ale rozkład ten nie jest symetryczny, to test Wilcoxona będzie miał znacznie większą moc niż test znaków. Analogiczne rozumowanie można zastosować do testowania hipotezy, że mediana rozpatrywanego rozkładu ciągłego jest równa pewnej stałej $\mu_0$. W tym celu rozpatruje się statystykę równą liczbie elementów próby większych od $\mu_0$ (por. zad. 9.5).

## 9.4. Rangowe testy niezależności

Rozpatrzmy teraz sytuację omówioną we wstępie do rozdz. 4, gdy dysponujemy prostą próbą losową $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ z rozkładu pary zmiennych losowych $(X, Y)$. Jak tam stwierdziliśmy, istotnym problemem jest stwierdzenie, czy zmienne losowe $X$ i $Y$ są zależne. Możemy to uczynić, analizując strukturę wykresu rozproszenia oraz obliczając próbkowy współczynnik korelacji $r$. Jak wiemy, ten ostatni daje nam ilościową ocenę siły zależności, w przypadku, gdy zależność jest w przybliżeniu liniowa. Jednak formalne testy dotyczące wartości współczynnika korelacji są możliwe do przeprowadzenia wtedy, gdy rozkład pary $(X, Y)$ jest normalny lub liczność próby $n$ jest duża (por. zad. 4.7(b)). W pozostałych przypadkach możemy posłużyć się metodą Monte Carlo opisaną w rozdz. 7.

### 9.4.1. Współczynnik korelacji Spearmana

Alternatywnie możemy rozpatrzeć przekształcenie rangowe obu prób $X$ i $Y$. Postępujemy analogicznie do metody wspomnianej w p. 9.2.1, polegającej na obliczeniu statystyki $t$ dla prób rang, co prowadziło tam do statystyki Wilcoxona. Zastąpmy mianowicie parę $(X_i, Y_i)$ parą $(Q_i, R_i)$, gdzie $Q_i$ jest rangą obserwacji $X_i$ w próbie $X_1, X_2, \ldots, X_n$ i $R_i$ jest rangą obserwacji $Y_i$ w próbie $Y_1, Y_2, \ldots, Y_n$ i obliczmy współczynnik korelacji dla ciągu par rang $(Q_i, R_i)$, $i = 1, 2, \ldots, n$. Zakładamy przy tym jedynie, że rozkład pary $(X, Y)$ jest ciągły. Prowadzi to do następującej definicji

**Definicja 9.3.** ***Rangowym współczynnikiem korelacji Spearmana*** *dla próby* $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ *nazywamy wartość*

$$r_S = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(Q_i - \bar{Q})(R_i - \bar{R})}{\left[\frac{1}{n-1}\sum_{i=1}^{n}(Q_i - \bar{Q})^2 \frac{1}{n-1}\sum_{i=1}^{n}(R_i - \bar{R})^2\right]^{1/2}} \tag{9.8}$$

*będącą próbkowym współczynnikiem korelacji zdefiniowanym w def. 4.1 dla próby* $(Q_1, R_1), (Q_2, R_2), \ldots, (Q_n, R_n)$. $\bar{Q}$ *i* $\bar{R}$ *oznaczają średnie próbkowe w odpowiednich próbach rang.*

Proste przekształcenia algebraiczne definicji współczynnika korelacji $r_S$ oparte na fakcie, że $\bar{Q} = \bar{R} = (n+1)/2$ i $S_Q^2 = S_R^2 = n(n+1)/12$ prowadzą do następującej postaci współczynnika korelacji Spearmana

$$r_S = \frac{12}{n(n^2-1)}\sum_{i=1}^{n} Q_i R_i - \frac{3(n+1)}{n-1}. \tag{9.9}$$

Statystyki $r_S$ chcemy użyć do testowania hipotezy $H_0$: $X$ i $Y$ są niezależne, przeciwko alternatywie $H_1$: $X$ i $Y$ są dodatnio zależne.

Jako hipotezę alternatywną możemy również przyjąć hipotezę jednostronną mówiącą o ujemnej zależności $X$ i $Y$ lub hipotezę dwustronną o dodatniej bądź ujemnej zależności tych zmiennych. Podstawową rolę odgrywa tutaj stwierdzenie analogiczne do stwierdzeń 9.1(1) i 9.2(1).

**Stwierdzenie 9.3.** *Dla niezależnych zmiennych losowych* $X$ *i* $Y$ *rozkład współczynnika korelacji Spearmana* $r_S$ *nie zależy od rozkładu zmiennych* $(X, Y)$.

Stwierdzenie to łatwo uzasadnić, jeśli zauważymy, że wartość sumy $\sum_{i=1}^{n} Q_i R_i$ się nie zmieni, gdy rangi obliczymy dla następujacej permutacji obserwacji: rozpatrzmy jako pierwszą obserwację $(X_{i_1}, Y_{i_1})$, dla której

wartość $X_{i_1}$ w próbie $X_1, X_2, \ldots, X_n$ jest najmniejsza (tj. ranga $R_{i_1}$ jest równa 1), jako drugą obserwację $(X_{i_2}, Y_{i_2})$, taką że ranga $R_{i_2}$ jest równa 2 i tak dalej. Oznaczając przez $S_1, S_2, \ldots, S_n$ rangi dla kolejnych wartości drugiej współrzędnej równych $Y_{i_1}, Y_{i_2}, \ldots, Y_{i_n}$, otrzymujemy, że odpowiednikiem rangi $(Q_i, R_i)$ dla przepermutowanej próby jest $(i, S_i)$ i $\sum_{i=1}^n R_i Q_i = \sum_{i=1}^n i S_i$. Jednocześnie, gdy zmienne $X$ i $Y$ są niezależne, uporządkowanie pierwszej współrzędnej nie ma wpływu na uporządkowanie drugiej, a zatem rozkład zmiennych $S_1, S_2, \ldots, S_n$ jest taki sam jak rozkład $n$ kolejno wylosowanych liczb bez zwracania ze zbioru $\{1, 2, \ldots, n\}$. Tak więc $P(S_1 = s_1, S_2 = s_2, \ldots, S_n = s_n) = 1/n!$, można zatem obliczyć dokładny rozkład statystyki Spearmana przy spełnieniu hipotezy zerowej. Wartość oczekiwana tej statystyki jest w takim przypadku oczywiście równa 0, a wariancja wynosi $1/(n-1)$. Kwantyle tego rozkładu są podane w tablicach, można je również wyznaczyć za pomocą większości pakietów statystycznych.

### 9.4.2. Współczynnik Kendalla

Rangowy współczynnik korelacji Spearmana nie jest jedyną rangową miarą zależności dla próby dwuwymiarowej. Jedną z innych miar rangowych jest próbkowy współczynnik Kendalla $\hat{\tau}$ równy

$$\hat{\tau} = \frac{2}{n(n-1)} \sum_{1 \le i < j \le n} J((X_i, Y_i), (X_j, Y_j)),$$

gdzie $J((X_i, Y_i), (X_j, Y_j)) = 1$, gdy para $(X_i, Y_i)$ jest zgodna z parą $(X_j, Y_j)$, tzn. gdy $(X_i - X_j)(Y_i - Y_j) > 0$ i $J((X_i, Y_i), (X_j, Y_j)) = -1$, gdy $(X_i - X_j)(Y_i - Y_j) < 0$. Statystyka Kendalla $\hat{\tau}$ jest zatem różnicą miedzy liczbą zgodnych i niezgodnych par w próbie podzieloną przez liczbę wszystkich nieuporządkowanych par równą $n(n-1)/2$.[3] Oczywiście $\hat{\tau}$ jest statystyką rangową, gdyż $J((X_i, Y_i), (X_j, Y_j)) = J((R_i, Q_i), (R_j, Q_j))$. Łatwo stwierdzić, że wartością oczekiwaną współczynnika $\hat{\tau}$ jest

$$\tau = P((X_1 - X_2)(Y_1 - Y_2) > 0) - P((X_1 - X_2)(Y_1 - Y_2) < 0) =$$

$$= 2P((X_1 - X_2)(Y_1 - Y_2) > 0) - 1.$$

W przypadku, kiedy zmienne $X$ i $Y$ są niezależne, $\tau = 0$. Podobnie jak dla statystyki Spearmana, w przypadku niezależności rozkład statystyki Kendalla jest niezależny od rozkładu ciągłych zmiennych losowych $X$ oraz $Y$ i ponadto, $\mathrm{Var}(\hat{\tau}) = 2(2n+5)/9n(n-1)$. Dla dużych liczności próby $(n > 40)$

[3] W przypadku cech dyskretnych odejmuje się od liczby $n(n-1)$ liczbę więzów (por. definicję $\tau$-$b$ Kendalla w zad. 6.7).

standaryzowane statystyki $r_s$ i $\hat{\tau}$ mają w przybliżeniu standardowy rozkład normalny.

**Przykład 9.3.** W wyniku przeprowadzonych badań dotyczących związku między łączną ilością opadów (w cm) w miesiącach maj–sierpień a ilością trawy (w kg/ar) zebranej z pewnego pastwiska podgórskiego w tym okresie otrzymano następujące wyniki dla kolejnych sześciu lat

| | | | | | | |
|---|---|---|---|---|---|---|
| Ilość opadów | 22,21 | 17,8 | 9,63 | 27,81 | 11,80 | 35,74 |
| Ilość trawy | 301 | 366 | 201 | 421 | 252 | 408 |

Próbkowy współczynnik korelacji wynosi 0,87. Z wykresu rozproszenia obu zmiennych można wnosić w przybliżeniu liniowy charakter dodatniej zależności. Opierając się na współczynniku korelacji Spearmana przetestujmy hipotezę o niezależności obu zmiennych przeciwko hipotezie o dodatniej zależności ilości zebranej trawy od ilości opadów. Rangi próby uporządkowanej według wielkości pierwszej współrzędnej wyglądają następująco:

$i$ : 1 2 3 4 5 6

$S_i$ : 1 2 4 3 6 5

Odpowiednia wartość $\sum_{i=1}^{6} iS_i = \sum_{i=1}^{6} P_iQ_i = 89$ i na mocy (9.9) wartość $r_S$ wynosi $(12\times 89)/6\times 35-(3\times 7)/5 = 0{,}8857$, skąd odpowiednia $p$–wartość wynosi 0,0167. Analogicznie, wartość próbkowego współczynnika Kendalla wynosi $\hat{\tau} = 0{,}733$ i odpowiednia $p$–wartość jest równa 0,028. Obie $p$–wartości rozpatrywanych testów niezależności, choć istotnie różne, przemawiają za odrzuceniem hipotezy o niezależności na rzecz dodatniej zależności między zmiennymi.

## 9.5. Porównanie rozkładów cech w wielu populacjach

Rozpatrzmy teraz uogólnienie problemu porównania rozkładów dwóch populacji rozpatrzonego w podrozdz. 9.2 na sytuację, gdy rozważamy $k$ populacji, gdzie $k > 2$, o rozkładach zadanych odpowiednio przez dystrybuanty $F_1, F_2, \ldots, F_k$. Chcemy testować hipotezę o ich równości $H_0$: $F_1 = = F_2 = \ldots = F_k$ bez zakładania, że rozkłady $F_1, F_2, \ldots, F_k$ są rozkładami normalnymi o tej samej wariancji. W przypadku przyjęcia tego założenia hipoteza zerowa sprowadza się do testowania równości wartości średnich $k$ rozkładów normalnych. Problem ten został dokładnie omówiony w rozdz. 5.

Rozpatrywana tam hipoteza alternatywna była zaprzeczeniem hipotezy zerowej i mówiła, że pewne dwie spośród $k$ rozpatrywanych wartości średnich nie są sobie równe. W przypadku nieparametrycznym załóżmy jedynie, że rozkłady zadane przez dystrybuanty $F_1, F_2, \ldots, F_k$ są ciągłe. W takiej sytuacji hipoteza alternatywna jest uogólnieniem hipotezy alternatywnej dla przypadku rozkładów normalnych i mówi, że dla każdej pary populacji $i$ i $j$ zachodzi

$$F_i(t) \leq F_j(t) \quad \text{dla wszystkich } t \quad \text{albo} \quad F_i(t) \geq F_j(t) \quad \text{dla wszystkich } t \tag{9.10}$$

i jednocześnie dla pewnych dwóch populacji ich dystrybuanty są różne. Z relacji (9.10) wynika, że wartości losowane z $i$-tej populacji systematycznie przewyższają wartości losowane z $j$ tej populacji lub odwrotnie, wartości z $i$-tej populacji są systematycznie mniejsze od wartości z $j$-tej populacji.

### 9.5.1. Test Kruskala–Wallisa

Podobnie jak w przypadku statystyki Wilcoxona dla problemu dwóch prób chcielibyśmy oprzeć konstrukcję statystyki na rangach elementów poszczególnych prób w próbie połączonej. Niech $R_{i1}, R_{i2}, \ldots, R_{in_i}$ oznaczają rangi elementów $Y_{i1}, Y_{i2}, \ldots, Y_{in_i}$ $i$-tej próby w próbie połączonej składającej się ze wszystkich elementów $k$ prób i niech $\bar{R}_i$ oznacza odpowiednią średnią próbkową $\bar{R}_i = n_i^{-1} \sum_{j=1}^{n_i} R_{ij}$. Ponadto, niech $n$ będzie sumą liczności wszystkich prób $n = \sum_{i=1}^{k} n_i$. Wprowadźmy następującą definicję.

**Definicja 9.4.** ***Statystyką testową Kruskala–Wallisa*** *do testowania hipotezy o równości rozkładów $F_1, F_2, \ldots, F_k$ przeciwko hipotezie alternatywnej (9.10) nazywamy statystykę*

$$T = \frac{12}{n(n+1)} \sum_{i=1}^{k} n_i [\bar{R}_i - \frac{1}{2}(n+1)]^2. \tag{9.11}$$

Statystyka $T$ jest więc pewną miarą odstępstwa średnich próbkowych $\bar{R}_i$ od swojej wartości średniej równej $(n+1)/2$ przy spełnieniu hipotezy $H_0$. Tak więc duża wartość $T$ może wskazywać na niespełnienie hipotezy $H_0$. Aby uzyskać inną interpretację, skonstatujmy, że statystyka $T$ może być przedstawiona jako monotoniczne odwzorowanie statystyki $\tilde{F}$ obliczonej na podstawie rang i odpowiadającej zdefiniowanej w (5.7) statystyce $F$ (licznik statystyki $\tilde{F}$ odpowiada zmienności międzygrupowej $SSA$, natomiast mianownik odpowiada zmienności wewnątrzgrupowej $SSE$)

$$\tilde{F} = \frac{\sum_{i=1}^{k} n_i(\bar{R}_i - \bar{R})^2}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(R_{ij} - \bar{R}_i)^2},$$

gdzie $\bar{R} = 1/n \sum_{i=1}^{n} R_{ij} = (n+1)/2$. Aby stwierdzić, że tak faktycznie jest, należy zauważyć, że wyrażenie $\tilde{F}$ można zapisać jako $SSA/(SST - SSA)$, gdzie $SSA = \sum_{i=1}^{k} n_i(\bar{R}_i - \bar{R})^2$, $SST = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(R_{ij} - \bar{R})^2$ i że $SST$ jest stałą zależną tylko od liczności $n$. Oczywiście, podobnie jak poprzednio, rozkład statystyki $T$ przy spełnieniu hipotezy $H_0$ nie zależy od wspólnego rozkładu cechy w populacjach. Wynika to z faktu, że w tej sytuacji każdy układ rang dla $k$ prób ma takie samo prawdopodobieństwo równe $1/n!$. Znalezienie rozkładu $T$ polega zatem na obliczeniu liczby układów rang dla których otrzymuje się określoną wartość $T = t$. Oczywiście, takie obliczenia są żmudne i w praktyce dla większych liczności prób stosuje się następujące przybliżenie

**Stwierdzenie 9.4.** *Przy spełnionej hipotezie $H_0$ i ciągłym rozkładzie cechy w populacjach mamy*

$$P(T \leq t) \to P(\mathcal{X}^2_{k-1} \leq t) \quad gdy \;\; t \to \infty,$$

*gdzie $\mathcal{X}^2_{k-1}$ oznacza zmienną o rozkładzie $\chi^2$ z $k-1$ stopniami swobody.*

Przybliżenie to stosuje się dla $k = 3$, gdy wszystkie liczności prób są większe od 5 i dla $k > 3$, gdy wszystkie liczności prób są większe od 4.

**Przykład 9.4.** Rozpatrzmy 4 próby o liczności 10 każda wygenerowane z rozkładu jednostajnego, przy czym w przypadku trzech pierwszych prób jest to rozkład jednostajny na przedziale $(0,1)$, a w przypadku czwartej próby rozkład jednostajny na przedziale (0,2, 1,2). Dane są przedstawione w tab. 9.2.

W przypadku rozkładów jednostajnych określonych na odcinku o tej samej długości hipoteza $H_0$ jest równoważna hipotezie o równości wartości średnich. Oczywiście, żeby móc to stwierdzić, musielibyśmy znać dokładną postać rozkładów, czego nie chcemy w tym miejscu zakładać. Rozpatrzmy rangi $R_{ij}, i = 1, \ldots, 4, j = 1, \ldots, 10$ elementów poszczególnych prób w próbie połączonej. Rangi te są przedstawione w tab. 9.3. Widać z niej np., że elementem najmniejszym spośród 40 elementów jest przedostatni element pierwszej próby równy 0,062, a największym elementem jest przedostatni element ostatniej próby równy 1,198. Wektor średnich próbkowych

$(\bar{R}_1, \bar{R}_2, \bar{R}_3, \bar{R}_4)$ jest równy (15,1, 21,8, 15,2, 29,9). Próbkowa średnia rang $\bar{R}$ wszystkich obserwacji jest równa

$$(15{,}1 + 21{,}8 + 15{,}2 + 29{,}9)/4 = (1 + 2 + \ldots + 40)/40 = 20{,}5.$$

W naszym przykładzie, zgodnie z def. 9.4, wartość statystyki Kruskala–Wallisa $T$ wynosi

$$\frac{12}{40 \times 41}\Big(4(15{,}1-20{,}5)^2+4(21{,}8-20{,}5)^2+4(15{,}2-20{,}5)^2+(29{,}9-20{,}5)^2\Big) =$$

$$= 10{,}778.$$

Ponieważ wszystkie liczności prób przekraczają 4, więc na mocy stwierdzenia statystyka $T$ ma w przybliżeniu rozkład $\chi^2$ z $k-1 = 4-1 = 3$ stopniami swobody. Odpowiednia $p$–wartość $P(F \geq 10{,}778)$ jest zatem w przybliżeniu równa 0,013, a więc na podstawie analizowanych danych jesteśmy w stanie odrzucić hipotezę, że wszystkie rozkłady są równe.

**Tabela 9.2. Cztery próby z rozkładu jednostajnego**

| Nr | Próba 1 z (0, 1) | Próba 2 z (0, 1) | Próba 3 z (0, 1) | Próba 4 z (0,2, 1,2) |
|---|---|---|---|---|
| 1 | 0,715 | 0,494 | 0,688 | 0,502 |
| 2 | 0,117 | 0,631 | 0,630 | 0,505 |
| 3 | 0,294 | 0,911 | 0,446 | 1,114 |
| 4 | 0,198 | 0,717 | 0,148 | 1,000 |
| 5 | 0,938 | 0,848 | 0,995 | 0,886 |
| 6 | 0,666 | 0,765 | 0,709 | 0,798 |
| 7 | 0,824 | 0,112 | 0,187 | 1,105 |
| 8 | 0,248 | 0,706 | 0,236 | 0,854 |
| 9 | 0,062 | 0,922 | 0,384 | 1,198 |
| 10 | 0,553 | 0,541 | 0,577 | 0,800 |

**Tabela 9.3. Rangi elementów prób z tab. 9.2 w próbie połączonej**

| Nr | Próba 1 | Próba 2 | Próba 3 | Próba 4 |
|---|---|---|---|---|
| 1 | 24 | 12 | 21 | 13 |
| 2 | 3 | 19 | 18 | 14 |
| 3 | 9 | 33 | 11 | 39 |
| 4 | 6 | 25 | 4 | 37 |
| 5 | 35 | 30 | 36 | 32 |
| 6 | 20 | 26 | 23 | 27 |
| 7 | 29 | 2 | 5 | 38 |
| 8 | 8 | 22 | 7 | 31 |
| 9 | 1 | 34 | 10 | 40 |
| 10 | 16 | 15 | 17 | 28 |

### 9.5.2. Porównania wielokrotne

W przypadku takim, jak w przykładzie, gdy test Kruskala–Wallisa odrzuca hipotezę o równości rozkładów $F_1, F_2, \ldots, F_k$ może nas interesować uporządkowanie tych rozkładów w sensie definicji relacji (9.10). Można w tym celu posłużyć się obserwacją, że jeśli $F_i(t) \leq F_j(t)$ dla wszystkich $t$, to oczekiwana wartość $r_i$ rangi dowolnego elementu $i$-tej próby jest nie mniejsza niż oczekiwana wartość $r_j$ rangi dowolnego elementu $j$-tej próby w próbie połączonej i starać się skonstruować przedział ufności dla różnicy $r_i - r_j$ przy spełnieniu hipotezy $H_0$. Oczywiście w tym przypadku $r_i - r_j = 0$, a więc jeśli skonstruowany przedział ufności na poziomie ufności $1-\alpha$ leży, powiedzmy, na prawo od zera i nie zawiera go, to na poziomie istotności $\alpha$ możemy odrzucić hipotezę o równości rozkładów $F_i$ i $F_j$ na rzecz hipotezy $F_i(t) \leq F_j(t)$ dla wszystkich $t$ i jednocześnie $F_i \neq F_j$. Aby skonstruować przedział ufności, wykorzystamy fakt, że dla każdej pary $1 \leq i < j \leq k$, $\bar{R}_j - \bar{R}_i$ przy spełnieniu hipotezy $H_0$ ma w przybliżeniu rozkład normalny $N(0, n(n+1)(n_i^{-1} + n_j^{-1})/12)$. Nie będziemy tego faktu tutaj dowodzić, zauważmy jedynie, że dla $k = 2$ wynika on łatwo ze stwierdzenia 9.1, gdyż w tym przypadku różnica $\bar{R}_1 - \bar{R}_2$ różni się o stałą od wielkości $(n_1^{-1} + n_2^{-1})W$. Otrzymujemy stąd łatwo, że przedział

$$\left(\bar{R}_i - \bar{R}_j - z_{1-\alpha/2}\frac{n(n+1)}{12}\Big(\frac{1}{n_i}+\frac{1}{n_j}\Big), \bar{R}_i - \bar{R}_j + z_{1-\alpha/2}\frac{n(n+1)}{12}\Big(\frac{1}{n_i}+\frac{1}{n_j}\Big)\right)$$

powinien z prawdopodobieństwem równym w przybliżeniu $1-\alpha$ zawierać 0. Ponieważ w celu ustalenia porządku wartości $r_1, r_2, \ldots, r_k$ konstruujemy $p = k(k-1)/2$ takich przedziałów, czyli dokonujemy $p$ porównań, zgodnie z procedurą Bonferroniego opisaną w p. 4.3.3, powinniśmy przy konstrukcji pojedynczego przedziału ufności zastąpić $\alpha$ przez $\alpha/p$ w celu otrzymania wyniku, o którego prawdziwości możemy orzekać z prawdopodobieństwem nie mniejszym niż $1-\alpha$. Oczywiście dla dużej liczby $p$ taka procedura może być mało efektywna, gdyż w takiej sytuacji poziom istotności indywidualnego testu równy $\alpha/p$ będzie bardzo mały, co spowoduje, że indywidualny test będzie rzadko odrzucał hipotezę zerową, gdy jest ona nieprawdziwa.

**Przykład 9.4 cd.** Zastosujmy powyższą metodę do uporządkowania rozkładów $F_1, F_2, F_3$ i $F_4$ z prawdopodobieństwem $1-\alpha = 0{,}90$. Odchylenie standardowe każdej różnicy $\bar{R}_i - \bar{R}_j$ wynosi $((40 \times 41)/12 \times \times (2/10))^{1/2} = 5{,}23$. Dla $\alpha/(2 \times 6) = 0{,}0083$ odpowiedni kwantyl rozkładu normalnego $z_{0{,}0083}$ wynosi $-2{,}39$, a zatem przedział ufności dla wartości $r_i - r_j$ wynosi $\bar{R}_i - \bar{R}_j \pm 2{,}39 \times 5{,}23$. Otrzymujemy więc, że jedyne przedziały, które nie zawierają 0, to przedziały dla wartości

$r_4 - r_1$ i $r_4 - r_3$. Ponieważ przedziały te leżą na prawo od zera, wynika stąd, że z prawdopodobieństwem w przybliżeniu równym 0,90 możemy wnioskować, że $F_4 \leq F_1$ i $F_4 \leq F_3$, natomiast nie możemy nic orzekać na temat uporządkowania pozostałych par rozkładów.

Zauważmy na zakończenie tego punktu, że w przypadku analizy zrandomizowanego układu blokowego, do testowania braku wpływu czynnika można rozpatrzeć rangowy odpowiednik testu $F$ z rozdz. 5, zastępując wartości obserwacji w każdym bloku przez ich rangi. Test taki nosi nazwę **testu Friedmana.**

## 9.6. Metody rangowe dla modelu regresji liniowej

Omówimy krótko typowe zastosowanie metod rangowych w testowaniu hipotez o współczynnikach regresji liniowej. Rozpatrzmy sytuację, gdy jest spełniony model regresji jednokrotnej opisany równaniem (4.12), przy czym założymy, że dystrybuanta błędów jest dystrybuantą ciągłą. Jest to jedyne dodatkowe założenie, które czynimy o rozkładzie błędów, poza zawsze przyjmowanym założeniem, że wartość średnia i wariancja błędów są równe odpowiednio 0 i $\sigma^2$. Przyjmijmy teraz, że podobnie jak w p. 4.2.4 chcemy testować hipotezę $H_0$: $\beta_1 = 0$ przeciwko jednostronnej hipotezie alternatywnej $H_1$: $\beta_1 > 0$. Załóżmy dla uproszczenia, że obserwacje zostały ponumerowane w ten sposób, że $x_1 < x_2 < \cdots < x_n$. Zauważmy, że w przypadku spełnienia hipotezy alternatywnej rozkład zmiennej $Y_j$ jest dla $j > i$ rozkładem zmiennej $Y_i$ przesuniętym w prawo o wielkość $\beta_1(x_j - x_i)$. Ta obserwacja podobnie jak poprzednio nasuwa pomysł wykorzystania rang zmiennych $Y_i$ do testowania hipotezy $H_0$ przeciwko alternatywie $H_1$. Oznaczmy przez $R_i$ rangę obserwacji $Y_i$ w próbie $Y_1, Y_2, \ldots, Y_n$. Ponieważ dla hipotezy alternatywnej zmienna objaśniana jest dodatnio zależna od zmiennej objaśniającej, możemy przypuszczać, że w takiej sytuacji, współczynnik nachylenia prostej MNK obliczony dla par $(x_i, R_i), i = 1, 2, \ldots, n$ będzie dodatni. Podkreślmy, że w rzeczywistości nie ma żadnego powodu, dla którego pary $(x_i, R_i), i = 1, 2, \ldots, n$ miałyby spełniać równanie regresji liniowej (4.12). To co robimy, jest jedynie przedstawieniem intuicyjnego rozumowania prowadzącego do definicji statystyki testowej. Współczynnik nachylenia prostej MNK dla takich danych ma postać $b_1 = \sum_{i=1}^n R_i(x_i - \bar{x}) / \sum_{i=1}^n (x_i - \bar{x})^2$. Ponieważ $\sum_{i=1}^n R_i \bar{x} = n(n+1)\bar{x}/2$ i $x_i$ są stałymi, zmienna $b_1$ jest

przekształceniem afinicznym (por. przypis w p. 4.2.3) zmiennej

$$U = \sum_{i=1}^{n} R_i x_i. \tag{9.12}$$

Hipotezę $H_0$ chcielibyśmy odrzucić na rzecz hipotezy $H_1$, gdy wartość $U$ jest większa niż typowa wartość $U$ przy spełnionej hipotezie $H_0$. Trudnością przy wyprowadzeniu dokładnego rozkładu $U$ jest fakt, że rozkład ten zależy od wartości $x_i, i = 1, 2, \ldots, n$. Rozkład ten jest z reguły stablicowany dla specyficznych wartości zmiennych objaśniających, np. dla często występującej w sytuacji eksperymentalnej sytuacji, gdy $x_i$ są kolejnymi punktami kratowymi tj. takimi, że dla pewnego $\delta > 0$ różnice $x_i - x_{i-1} = \delta$ dla $i = 1, 2, \ldots, n$. W ogólnym przypadku musimy uciec się albo do symulacyjnego przybliżenia rozkładu $U$, albo, dla dostatecznie dużych $n$, do zastosowania przybliżenia normalnego. Ponieważ $ER_i = (1+2+\ldots+n)/n = (n+1)/2$, łatwo stąd wynika, że $EU = n(n+1)\bar{x}/2$. Nieco bardziej skomplikowanych rachunków wymaga udowodnienie, że

$$\mathrm{Var}(U) = \frac{1}{12} n(n+1) \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

Okazuje się, że przy $n \to \infty$ i gdy jest spełniony warunek (7.7) o braku dużych odstępstw od wartości średniej dla próby zmiennych objaśniających $\{x_i\}$, to

$$P\Big(\frac{U - EU}{\sqrt{\mathrm{Var}U}} \le t\Big) \to \Phi(t) \quad \text{dla wszystkich } t.$$

Powyższa własność w rutynowy sposób prowadzi nas do konstrukcji testu dla hipotezy $H_0$ przeciwko alternatywie $H_1$, mającego w przybliżeniu zadany poziom istotności $\alpha$. Przykład zastosowania tej metody jest rozpatrzony w zad. 9.5.

## 9.7. Zadania

**9.1.** Na wstępnym etapie przygotowania produkcji tłoków nowego typu otrzymano następujące 10 pomiarów jego średnicy, pochodzących z dwóch różnych obrabiarek

| | Średnica tłoka | | | | |
|---|---|---|---|---|---|
| Obrabiarka 1 | 8,613 | 9,769 | 8,844 | 9,322 | 9,501 |
| | 9,138 | 9,585 | 8,556 | 8,722 | 9,291 |
| Obrabiarka 2 | 8,710 | 10,579 | 9,794 | 10,935 | 8,798 |
| | 9,420 | 10,010 | 9,415 | 9,114 | 10,002 |

Sporządzić wykresy normalne dla obydwu prób i stwierdzić, czy można uznać je za pochodzące z rozkładu normalnego o tej samej wariancji. Za pomocą odpowiedniego testu, przetestować hipotezę o równości rozkładów przeciwko hipotezie alternatywnej mówiącej, że odpowiednie rozkłady są różne.

**9.2.** W celu porównania skażenia wody w dwóch jeziorach pobrano po 6 próbek z każdego z nich i zmierzono w nich zawartość ołowiu (w $\mu$g/l):

| | Zawartość ołowiu (w $\mu$g/l) | | | | | |
|---|---|---|---|---|---|---|
| Jezioro 1 | 21,0 | 17,5 | 23,1 | 26,2 | 20,1 | 19,8 |
| Jezioro 2 | 17,3 | 21,0 | 18,2 | 18,8 | 19,0 | 18,1 |

Za pomocą testu Wilcoxona, stwierdzić, czy rozkłady zawartości ołowiu w obu jeziorach można uznać za różne.

**9.3.** Rozpatrzmy sytuację opisaną w przykł. 2.29, gdy staramy się zbadać, czy spożycie płynu XXL wpływa na wyniki egzaminów. W poniższej tabeli podano sumę punktów uzyskanych w sesji egzaminacyjnej przez 5 losowo wybranych studentów, ktorzy nie stosowali płynu XXL (pierwszy wiersz) i 5 losowo wybranych, którzy go stosowali (drugi wiersz). Przy założeniu, że spożycie płynu XXL wpływa jedynie na przesunięcie rozkładu sumy punktów uzyskanych w sesji egzaminacyjnej, czyli, że prawdziwy jest model (9.2), obliczyć wartość estymatora parametru $\Delta$ i skonstruować dla niego przedział ufności na poziomie ufności 0,944. Czy na podstawie skonstruowanego przedziału ufności możemy wnioskować o dodatnim wpływie spożycia płynu XXL na wyniki egzaminów?

| | Suma punktów | | | | |
|---|---|---|---|---|---|
| Bez stosowania płynu XXL | 27 | 23 | 37,5 | 32 | 36 |
| Przy stosowaniu płynu XXL | 47 | 35 | 41 | 30 | 45 |

**9.4.** Rozpatrzmy dane dotyczące poziomu cukru we krwi na czczo i w godzinę po podaniu dawki glukozy dla grupy kobiet w trzecim trzymestrze ciąży z zad. 1.4. Przeprowadzić test hipotezy, że rozkład różnic poziomów glukozy $F$ jest rozkładem symetrycznym przeciwko alternatywie (9.7). Skonstruować przedział ufności dla parametru przesunięcia $\Delta$.

**9.5.** Dla danych z zad. 4.4 przeprowadzić test hipotezy $\beta_1 = 0$, używając metody rangowej opisanej w podrozdz. 9.6 dla całego zbioru danych oraz po usunięciu dwóch obserwacji odstających (wykorzystać przybliżenie normalne). Porównać wyniki otrzymane w obu przypadkach.

**9.6.** Osiem próbek gleby z pewnego rejonu poddano badaniom chemicznym w celu ustalenia ich kwasowości mierzonej współczynnikiem pH. Otrzymano następujące wyniki:

$$6{,}8 \quad 7{,}8 \quad 6{,}9 \quad 6{,}4 \quad 7{,}5 \quad 8{,}4 \quad 7{,}4 \quad 7{,}1.$$

Na podstawie poprzednich badań przyjęto, że mediana rozkładu pH dla gleby w tej okolicy wynosi $\mu = 7$. Po zauważeniu, że przy przyjętej hipotezie $H_0$ liczba elementów próby większych od 7 ma rozkład dwumianowy z parametrami $n = 8$ i $p = 1/2$, skonstruować obszar krytyczny odpowiedniego testu dla hipotezy alternatywnej $\mu \neq 5$ i użyć go do zweryfikowania hipotezy na poziomie $\alpha = 0{,}05$.

**9.7.** W badaniu rozpatruje się czas funkcjonowania żarówek pochodzących od 3 różnych producentów. Próby składają się z 10 żarówek każdego typu. Czasy funkcjonowania (w godzinach) są podane w tabeli

| | Czas działania żarówki [h] | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Producent I | 437 | 244 | 490 | 178 | 2062 | 70 | 359 | 3001 | 91 | 75 |
| Producent II | 1075 | 526 | 344 | 76 | 481 | 2495 | 922 | 75 | 372 | 35 |
| Producent III | 1791 | 263 | 461 | 141 | 554 | 265 | 134 | 856 | 489 | 648 |

Użyć testu Kruskala–Wallisa do testowania hipotezy o równości wszystkich trzech rozkładów przeciwko hipotezie alternatywnej (9.10).

**9.8.** Dane przedstawione poniżej przedstawiają zależność między średnią liczbą kroków na sekundę $x$ a średnią odległością pokonywaną w ciągu sekundy $y$ dla siedmiu biegaczy długodystansowych. Użyć statystyk testowych Spearmana i Kendalla do testowania hipotezy o niezależności tych wielkości przeciwko alternatywie mówiącej o ich dodatniej zależności.

| $x$ | 3,5 | 3,2 | 3,0 | 4,1 | 3,1 | 3,8 | 3,3 |
|---|---|---|---|---|---|---|---|
| $y$ | 4,7 | 3,9 | 4,8 | 5,1 | 4,1 | 4,9 | 4,6 |

# Przedmowa

Tablica I. Dystrybuanta $\Phi(u)$ rozkładu normalnego dla $u \geq 0$. Dla $u < 0$ $\Phi(u)$ obliczamy ze wzoru $\Phi(u) = 1 - \Phi(-u)$

| $u$ | 0,00 | 0,01 | 0,02 | 0,03 | 0,04 | 0,05 | 0,06 | 0,07 | 0,08 | 0,09 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0,0 | ,5000 | ,5040 | ,5080 | ,5120 | ,5160 | ,5199 | ,5239 | ,5279 | ,5319 | ,5359 |
| 0,1 | ,5398 | ,5438 | ,5478 | ,5517 | ,5557 | ,5596 | ,5636 | ,5675 | ,5714 | ,5753 |
| 0,2 | ,5793 | ,5832 | ,5871 | ,5910 | ,5948 | ,5987 | ,6026 | ,6064 | ,6103 | ,6141 |
| 0,3 | ,6179 | ,6217 | ,6255 | ,6293 | ,6331 | ,6368 | ,6406 | ,6443 | ,6480 | ,6517 |
| 0,4 | ,6554 | ,6591 | ,6628 | ,6664 | ,6700 | ,6736 | ,6772 | ,6808 | ,6844 | ,6879 |
| 0,5 | ,6915 | ,6950 | ,6985 | ,7019 | ,7054 | ,7088 | ,7123 | ,7157 | ,7190 | ,7224 |
| 0,6 | ,7257 | ,7291 | ,7324 | ,7357 | ,7389 | ,7422 | ,7454 | ,7486 | ,7517 | ,7549 |
| 0,7 | ,7580 | ,7611 | ,7642 | ,7673 | ,7704 | ,7734 | ,7764 | ,7794 | ,7823 | ,7852 |
| 0,8 | ,7881 | ,7910 | ,7939 | ,7967 | ,7995 | ,8023 | ,8051 | ,8078 | ,8106 | ,8133 |
| 0,9 | ,8159 | ,8186 | ,8212 | ,8238 | ,8364 | ,8289 | ,8315 | ,8340 | ,8365 | ,8389 |
| 1,0 | ,8413 | ,8438 | ,8461 | ,8485 | ,8508 | ,8531 | ,8554 | ,8577 | ,8599 | ,8621 |
| 1,1 | ,8643 | ,8665 | ,8686 | ,8708 | ,8729 | ,8749 | ,8770 | ,8790 | ,8810 | ,8830 |
| 1,2 | ,8849 | ,8869 | ,8888 | ,8907 | ,8925 | ,8944 | ,8962 | ,8980 | ,8997 | ,9015 |
| 1,3 | ,9032 | ,9049 | ,9066 | ,9082 | ,9099 | ,9115 | ,9131 | ,9147 | ,9162 | ,9177 |
| 1,4 | ,9192 | ,9207 | ,9222 | ,9236 | ,9251 | ,9265 | ,9279 | ,9292 | ,9306 | ,9319 |
| 1,5 | ,9332 | ,9345 | ,9357 | ,9370 | ,9382 | ,9394 | ,9406 | ,9418 | ,9429 | ,9441 |
| 1,6 | ,9452 | ,9463 | ,9474 | ,9484 | ,9495 | ,9505 | ,9515 | ,9525 | ,9535 | ,9545 |
| 1,7 | ,9554 | ,9564 | ,9573 | ,9582 | ,9591 | ,9599 | ,9608 | ,9616 | ,9625 | ,9633 |
| 1,8 | ,9641 | ,9649 | ,9656 | ,9664 | ,9671 | ,9678 | ,9686 | ,9693 | ,9699 | ,9706 |
| 1,9 | ,9713 | ,9719 | ,9726 | ,9732 | ,9738 | ,9744 | ,9750 | ,9756 | ,9761 | ,9767 |
| 2,0 | ,9772 | ,9778 | ,9783 | ,9788 | ,9793 | ,9798 | ,9803 | ,9808 | ,9812 | ,9817 |
| 2,1 | ,9821 | ,9826 | ,9830 | ,9834 | ,9838 | ,9842 | ,9846 | ,9850 | ,9854 | ,9857 |
| 2,2 | ,9861 | ,9864 | ,9868 | ,9871 | ,9875 | ,9878 | ,9881 | ,9884 | ,9887 | ,9890 |
| 2,3 | ,9893 | ,9896 | ,9898 | ,9901 | ,9904 | ,9906 | ,9909 | ,9911 | ,9913 | ,9916 |
| 2,4 | ,9918 | ,9920 | ,9922 | ,9925 | ,9927 | ,9929 | ,9931 | ,9932 | ,9934 | ,9936 |
| 2,5 | ,9938 | ,9940 | ,9941 | ,9943 | ,9945 | ,9946 | ,9948 | ,9949 | ,9951 | ,9952 |
| 2,6 | ,9953 | ,9955 | ,9956 | ,9957 | ,9959 | ,9960 | ,9961 | ,9962 | ,9963 | ,9964 |
| 2,7 | ,9965 | ,9966 | ,9967 | ,9968 | ,9969 | ,9970 | ,9971 | ,9972 | ,9973 | ,9974 |
| 2,8 | ,9974 | ,9975 | ,9976 | ,9977 | ,9977 | ,9978 | ,9979 | ,9979 | ,9980 | ,9981 |
| 2,9 | ,9981 | ,9982 | ,9982 | ,9983 | ,9984 | ,9984 | ,9985 | ,9985 | ,9986 | ,9986 |

Tablica II. Kwantyle $t_{\alpha,n}$ rozkładu $t$ Studenta z $n$ stopniami swobody.
Dla $0 < \alpha < 0{,}5 \quad t_{\alpha,n} = -t_{1-\alpha,n}$

| $n$ | $\alpha = 0,9$ | $\alpha = 0,95$ | $\alpha = 0,975$ | $\alpha = 0,99$ | $\alpha = 0,995$ |
|---|---|---|---|---|---|
| 1 | 3,0777 | 6,3138 | 12,7062 | 31,8205 | 63,6569 |
| 2 | 1,8856 | 2,9200 | 4,3027 | 6,9646 | 9,9248 |
| 3 | 1,6378 | 2,3534 | 3,1825 | 4,5407 | 5,8410 |
| 4 | 1,5332 | 2,1318 | 2,7764 | 3,7470 | 4,6041 |
| 5 | 1,4759 | 2,0151 | 2,5706 | 3,3649 | 4,0321 |
| 6 | 1,4398 | 1,9432 | 2,4489 | 3,1247 | 3,7075 |
| 7 | 1,4149 | 1,8946 | 2,3646 | 2,9980 | 3,4995 |
| 8 | 1,3969 | 1,8595 | 2,3060 | 2,8965 | 3,3554 |
| 9 | 1,3830 | 1,8331 | 2,2622 | 2,8215 | 3,2498 |
| 10 | 1,3722 | 1,8125 | 2,2281 | 2,7638 | 3,1693 |
| 11 | 1,3634 | 1,7959 | 2,2010 | 2,7181 | 3,1058 |
| 12 | 1,3562 | 1,7823 | 2,1788 | 2,6810 | 3,0545 |
| 13 | 1,3502 | 1,7709 | 2,1604 | 2,6503 | 3,0123 |
| 14 | 1,3450 | 1,7613 | 2,1448 | 2,6245 | 2,9768 |
| 15 | 1,3406 | 1,7530 | 2,1315 | 2,6025 | 2,9467 |
| 16 | 1,3368 | 1,7459 | 2,1199 | 2,5835 | 2,9208 |
| 17 | 1,3334 | 1,7396 | 2,1098 | 2,5669 | 2,8982 |
| 18 | 1,3304 | 1,7341 | 2,1009 | 2,5524 | 2,8784 |
| 19 | 1,3278 | 1,7291 | 2,0930 | 2,5395 | 2,8610 |
| 20 | 1,3253 | 1,7247 | 2,0860 | 2,5280 | 2,8453 |
| 21 | 1,3232 | 1,7207 | 2,0796 | 2,5176 | 2,8314 |
| 22 | 1,3213 | 1,7172 | 2,0739 | 2,5083 | 2,8187 |
| 23 | 1,3194 | 1,7139 | 2,0687 | 2,4999 | 2,8074 |
| 24 | 1,3178 | 1,7109 | 2,0639 | 2,4921 | 2,7969 |
| 25 | 1,3164 | 1,7081 | 2,0595 | 2,4851 | 2,7874 |
| 26 | 1,3150 | 1,7056 | 2,0556 | 2,4786 | 2,7787 |
| 27 | 1,3137 | 1,7033 | 2,0519 | 2,4727 | 2,7707 |
| 28 | 1,3125 | 1,7011 | 2,0484 | 2,4671 | 2,7633 |
| 29 | 1,3114 | 1,6991 | 2,0452 | 2,4620 | 2,7564 |
| 30 | 1,3104 | 1,6973 | 2,0423 | 2,4573 | 2,7500 |
| 31 | 1,3095 | 1,6955 | 2,0395 | 2,4528 | 2,7440 |
| 32 | 1,3086 | 1,6939 | 2,0369 | 2,4487 | 2,7385 |
| 33 | 1,3077 | 1,6923 | 2,0345 | 2,4448 | 2,7333 |
| 34 | 1,3070 | 1,6909 | 2,0322 | 2,4411 | 2,7284 |
| 35 | 1,3062 | 1,6896 | 2,0301 | 2,4377 | 2,7238 |
| 36 | 1,3055 | 1,6883 | 2,0281 | 2,4345 | 2,7195 |
| 37 | 1,3048 | 1,6871 | 2,0262 | 2,4314 | 2,7154 |
| 38 | 1,3043 | 1,6860 | 2,0244 | 2,4285 | 2,7116 |
| 39 | 1,3036 | 1,6849 | 2,0227 | 2,4258 | 2,7079 |
| 40 | 1,3031 | 1,6839 | 2,0211 | 2,4232 | 2,7045 |

Tablica III. Kwantyle rzędu $\alpha$ rozkładu $\chi^2$ z $n$ stopniami swobody

| | $\alpha$ | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $n$ | 0,005 | 0,01 | 0,025 | 0,05 | 0,95 | 0,975 | 0,99 | 0,995 |
| 1 | 0,000 | 0,000 | 0,001 | 0,004 | 3,841 | 5,024 | 6,635 | 7,879 |
| 2 | 0,010 | 0,020 | 0,051 | 0,103 | 5,991 | 7,378 | 9,210 | 10,597 |
| 3 | 0,072 | 0,115 | 0,216 | 0,352 | 7,815 | 9,348 | 11,345 | 12,838 |
| 4 | 0,207 | 0,297 | 0,484 | 0,711 | 9,488 | 11,143 | 13,277 | 14,860 |
| 5 | 0,412 | 0,554 | 0,831 | 1,145 | 11,070 | 12,832 | 15,086 | 16,750 |
| 6 | 0,676 | 0,872 | 1,237 | 1,635 | 12,592 | 14,449 | 16,812 | 18,548 |
| 7 | 0,989 | 1,239 | 1,690 | 2,167 | 14,067 | 16,013 | 18,475 | 20,278 |
| 8 | 1,344 | 1,646 | 2,180 | 2,733 | 15,507 | 17,535 | 20,090 | 21,955 |
| 9 | 1,735 | 2,088 | 2,700 | 3,325 | 16,919 | 19,023 | 21,666 | 23,589 |
| 10 | 2,156 | 2,558 | 3,247 | 3,940 | 18,307 | 20,483 | 23,209 | 25,188 |
| 11 | 2,603 | 3,053 | 3,816 | 4,575 | 19,675 | 21,920 | 24,725 | 26,757 |
| 12 | 3,074 | 3,571 | 4,404 | 5,226 | 21,026 | 23,336 | 26,217 | 28,300 |
| 13 | 3,565 | 4,107 | 5,009 | 5,892 | 22,362 | 24,735 | 27,688 | 29,819 |
| 14 | 4,075 | 4,660 | 5,629 | 6,571 | 23,685 | 26,119 | 29,141 | 31,319 |
| 15 | 4,601 | 5,229 | 6,262 | 7,261 | 24,996 | 27,488 | 30,578 | 32,801 |
| 16 | 5,142 | 5,812 | 6,908 | 7,962 | 26,296 | 28,845 | 32,000 | 34,267 |
| 17 | 5,697 | 6,408 | 7,564 | 8,672 | 27,587 | 30,191 | 33,409 | 35,718 |
| 18 | 6,265 | 7,015 | 8,231 | 9,390 | 28,869 | 31,536 | 34,805 | 37,156 |
| 19 | 6,844 | 7,633 | 8,907 | 10,117 | 30,144 | 32,852 | 36,191 | 38,582 |
| 20 | 7,434 | 8,260 | 9,591 | 10,851 | 31,410 | 34,170 | 37,566 | 39,997 |
| 21 | 8,034 | 8,897 | 10,283 | 11,591 | 32,671 | 35,479 | 38,932 | 41,401 |
| 22 | 8,643 | 9,542 | 10,982 | 12,338 | 33,924 | 36,781 | 40,289 | 42,796 |
| 23 | 9,260 | 10,196 | 11,688 | 13,091 | 35,172 | 38,076 | 41,638 | 44,181 |
| 24 | 9,886 | 10,856 | 12,401 | 13,848 | 36,415 | 39,364 | 42,980 | 45,558 |
| 25 | 10,520 | 11,524 | 13,120 | 14,611 | 37,652 | 40,646 | 44,314 | 46,928 |
| 26 | 11,160 | 12,198 | 13,884 | 15,379 | 38,885 | 41,923 | 45,642 | 48,290 |
| 27 | 11,808 | 12,879 | 14,573 | 16,151 | 40,113 | 43,194 | 46,963 | 49,645 |
| 28 | 12,461 | 13,365 | 15,308 | 16,928 | 41,337 | 44,461 | 48,278 | 50,993 |
| 29 | 13,121 | 14,256 | 16,047 | 17,708 | 42,557 | 45,722 | 49,588 | 52,336 |
| 30 | 13,787 | 14,953 | 16,791 | 18,493 | 43,773 | 46,979 | 50,892 | 53,672 |
| 31 | 14,458 | 15,655 | 17,539 | 19,281 | 44,985 | 48,232 | 52,191 | 55,003 |
| 32 | 15,134 | 16,362 | 18,291 | 20,072 | 46,194 | 49,480 | 53,486 | 56,328 |
| 33 | 15,815 | 17,073 | 19,047 | 20,867 | 47,400 | 50,725 | 54,776 | 57,648 |
| 34 | 16,501 | 17,789 | 19,806 | 21,664 | 48,602 | 51,966 | 56,061 | 58,964 |
| 35 | 17,192 | 18,509 | 20,569 | 22,465 | 49,802 | 53,203 | 57,342 | 60,275 |
| 36 | 17,887 | 19,233 | 21,336 | 23,269 | 50,998 | 54,437 | 58,619 | 61,581 |
| 37 | 18,586 | 19,960 | 22,106 | 24,075 | 52,192 | 55,668 | 59,892 | 62,882 |
| 38 | 19,289 | 20,691 | 22,878 | 24,884 | 53,384 | 56,895 | 61,162 | 64,181 |
| 39 | 19,996 | 21,426 | 23,654 | 25,695 | 54,572 | 58,120 | 62,428 | 65,476 |
| 40 | 20,707 | 22,164 | 24,433 | 26,509 | 55,758 | 59,342 | 63,691 | 66,766 |
| 50 | 27,991 | 29,707 | 32,357 | 34,764 | 67,505 | 71,420 | 76,154 | 79,490 |
| 60 | 35,535 | 37,485 | 40,482 | 43,188 | 79,082 | 83,298 | 88,379 | 91,952 |
| 70 | 43,275 | 45,442 | 48,758 | 51,739 | 90,531 | 95,023 | 100,425 | 104,215 |
| 80 | 51,172 | 53,540 | 57,153 | 60,391 | 101,879 | 106,629 | 112,329 | 116,321 |
| 90 | 59,196 | 61,754 | 65,647 | 69,126 | 113,145 | 118,136 | 124,116 | 128,299 |
| 100 | 67,328 | 70,075 | 74,222 | 77,929 | 124,342 | 129,561 | 135,807 | 140,169 |

Tablica IV. Kwantyle $f_{0,95,n_1,n_2}$ rzędu 0,95 rozkładu $F$ Snedecora z $(n_1, n_2)$ stopniami swobody

| | | $n_1$ | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 4 | 6 | 8 | 10 | 12 | 24 | $\infty$ |
| | 1 | 161,4 | 199,5 | 224,6 | 234,0 | 238,9 | 241,9 | 243,9 | 249,1 | 254,3 |
| | 2 | 18,51 | 19,00 | 19,25 | 19,33 | 19,37 | 19,40 | 19,41 | 19,45 | 19,50 |
| | 3 | 10,13 | 9,55 | 9,12 | 8,94 | 8,85 | 8,79 | 8,74 | 8,64 | 8,53 |
| | 4 | 7,71 | 6,94 | 6,39 | 6,16 | 6,04 | 5,96 | 5,91 | 5,77 | 5,63 |
| | 5 | 6,61 | 5,79 | 5,19 | 4,95 | 4,82 | 4,74 | 4,68 | 4,53 | 4,36 |
| | 6 | 5,99 | 5,14 | 4,53 | 4,28 | 4,15 | 4,06 | 4,00 | 3,84 | 3,67 |
| | 7 | 5,59 | 4,74 | 4,12 | 3,87 | 3,73 | 3,64 | 3,57 | 3,41 | 3,23 |
| | 8 | 5,32 | 4,46 | 3,84 | 3,58 | 3,44 | 3,25 | 3,28 | 3,12 | 2,93 |
| | 9 | 5,12 | 4,26 | 3,63 | 3,37 | 3,23 | 3,14 | 3,07 | 2,90 | 2,71 |
| | 10 | 4,96 | 4,10 | 3,48 | 3,22 | 3,07 | 2,98 | 2,91 | 2,74 | 2,54 |
| | 11 | 4,84 | 3,98 | 3,36 | 3,09 | 2,95 | 2,85 | 2,79 | 2,61 | 2,40 |
| $n_2$ | 12 | 4,75 | 3,89 | 3,26 | 3,00 | 2,85 | 2,75 | 2,69 | 2,51 | 2,30 |
| | 13 | 4,67 | 3,81 | 3,18 | 2,92 | 2,77 | 2,67 | 2,60 | 2,42 | 2,21 |
| | 14 | 4,60 | 3,74 | 3,11 | 2,85 | 2,70 | 2,60 | 2,53 | 2,35 | 2,13 |
| | 15 | 4,54 | 3,68 | 3,06 | 2,79 | 2,64 | 2,54 | 2,48 | 2,29 | 2,07 |
| | 16 | 4,49 | 3,63 | 3,01 | 2,74 | 2,59 | 2,49 | 2,42 | 2,24 | 2,01 |
| | 17 | 4,45 | 3,59 | 2,96 | 2,70 | 2,55 | 2,45 | 2,38 | 2,19 | 1,96 |
| | 18 | 4,41 | 3,55 | 2,93 | 2,66 | 2,51 | 2,41 | 2,34 | 2,15 | 1,92 |
| | 19 | 4,38 | 3,52 | 2,90 | 2,63 | 2,48 | 2,38 | 2,31 | 2,11 | 1,88 |
| | 20 | 4,35 | 3,49 | 2,87 | 2,60 | 2,45 | 2,35 | 2,28 | 2,08 | 1,84 |
| | 25 | 4,24 | 3,39 | 2,76 | 2,49 | 2,34 | 2,24 | 2,16 | 1,96 | 1,71 |
| | 30 | 4,17 | 3,32 | 2,69 | 2,42 | 2,27 | 2,16 | 2,09 | 1,89 | 1,62 |
| | 40 | 4,08 | 3,23 | 2,61 | 2,34 | 2,18 | 2,08 | 2,00 | 1,79 | 1,51 |
| | 60 | 4,00 | 3,15 | 2,53 | 2,25 | 2,10 | 1,99 | 1,92 | 1,70 | 1,39 |
| | 120 | 3,92 | 3,07 | 2,45 | 2,17 | 2,02 | 1,91 | 1,83 | 1,61 | 1,25 |
| | $\infty$ | 3,84 | 3,00 | 2,37 | 2,10 | 1,94 | 1,83 | 1,75 | 1,52 | 1,00 |

# Skorowidz