

Andrzej
Iwasiewicz

Zbigniew
Paszek

statystyka

z elementami
statystycznych
metod
monitorowania
procesów



Wydawnictwo
Akademii
Ekonomicznej
w Krakowie

Andrzej
Iwasiewicz

Zbigniew
Paszek

Statystyka

z elementami
statystycznych
metod
monitorowania
procesów

Wydanie IV poprawione

Wydawnictwo
Akademii
Ekonomicznej
w Krakowie

Kraków 2004

Wydanie I: *Elementy statystyki i statystycznej kontroli jakości*,
Akademia Ekonomiczna w Krakowie, Kraków 1992

Recenzent: *Olgierd Hryniewicz*

Wydania kolejne: II zmienione: *Statystyka z elementami statystycznych metod sterowania jakością*, Kraków 1997; III poprawione i uzupełnione, Kraków 2000

Redaktor naukowy
Andrzej Sokołowski

Projekt okładki
Marcin Sokołowski

Redaktor Wydawnictwa
Agnieszka Penarska

Skrypt przeznaczony jest dla studentów wszystkich kierunków studiów

© Copyright by Akademia Ekonomiczna w Krakowie, Kraków 2004

ISBN 83-7252-238-3

Wydawnictwo Akademii Ekonomicznej w Krakowie
ul. Rakowicka 27, 31-510 Kraków

Pracownia Poligraficzna Akademii Ekonomicznej w Krakowie
ul. Rakowicka 27, 31-510 Kraków

Wydanie czwarte poprawione
Objętość 15,0 ark. wyd.
Zam. 153/2004

SPIS TREŚCI

Od autorów	7
Rozdział 1	
ELEMENTY RACHUNKU PRAWDOPODOBIENSTWA	
1.1. Wprowadzenie	9
1.2. Prawdopodobieństwo	11
1.2.1. Definicje prawdopodobieństwa	12
1.2.2. Podstawowe własności prawdopodobieństwa	15
1.2.3. Prawdopodobieństwo zupełne i wzór Bayesa	16
1.3. Jednowymiarowe zmienne losowe	18
1.3.1. Podstawowe pojęcia i określenia	18
1.3.2. Charakterystyki rozkładu jednowymiarowej zmiennej losowej	20
1.3.2.1. Funkcja prawdopodobieństwa i funkcja gęstości prawdopodobieństwa	20
1.3.2.2. Dystrybuanta	22
1.3.3. Parametry jednowymiarowej zmiennej losowej	24
1.4. Wybrane rozkłady jednowymiarowych zmiennych losowych	27
1.4.1. Rozkłady dyskretnych zmiennych losowych	27
1.4.1.1. Rozkład dwupunktowy	27
1.4.1.2. Rozkład dwumianowy	29
1.4.1.3. Rozkład Poissona	33
1.4.1.4. Rozkład hipergeometryczny	37
1.4.2. Rozkłady ciągłych zmiennych losowych	39
1.4.2.1. Rozkład jednostajny	39
1.4.2.2. Rozkład normalny	41
1.4.2.3. Rozkład wykładniczy	48
1.5. Dwuwymiarowe zmienne losowe	52
1.5.1. Funkcja prawdopodobieństwa, funkcja gęstości prawdopodobieństwa i dystrybuanta dwuwymiarowej zmiennej losowej	53
1.5.2. Rozkłady brzegowe i rozkłady warunkowe	54
1.5.3. Dwuwymiarowy rozkład normalny	58
1.5.4. Współczynnik korelacji liniowej	61
1.6. Twierdzenia graniczne	61
1.6.1. Prawo wielkich liczb Bernoulliego	62
1.6.2. Twierdzenie Moivre'a-Laplace'a	63
1.6.3. Nierówność Czebyszewa	65

Rozdział 2
METODOLOGIA BADAŃ STATYSTYCZNYCH

2.1. Wprowadzenie	67
2.2. Zbiorowość generalna	71
2.3. Próba	75
2.4. Zmienne losowe w badaniach empirycznych	81
2.5. Skale pomiarowe	85

Rozdział 3
CHARAKTERYSTYKI OPISOWE ZMIENNYCH LOSOWYCH
W BADANIACH WYCZERPUJĄCYCH

3.1. Wprowadzenie	89
3.2. Jednowymiarowe zmienne losowe	90
3.2.1. Rozkład empiryczny	90
3.2.1.1. Szereg rozdzielczy	91
3.2.1.2. Szereg kumulacyjny	98
3.2.2. Miary położenia	106
3.2.2.1. Średnia arytmetyczna	106
3.2.2.2. Średnia harmoniczna	111
3.2.2.3. Średnie potęgowe	112
3.2.2.4. Średnia geometryczna	114
3.2.2.5. Mediana	116
3.2.2.6. Modalna	119
3.2.3. Miary rozproszenia	120
3.2.3.1. Wariancja i odchylenie standardowe	121
3.2.3.2. Odchylenie przeciętne	127
3.2.3.3. Miary rozproszenia oparte na statystykach pozycyjnych	129
3.2.4. Momenty empiryczne	130
3.3. Wielowymiarowe zmienne losowe	131
3.3.1. Rozkład empiryczny dwuwymiarowej zmiennej losowej	131
3.3.2. Współzależność dwóch zmiennych losowych	135
3.3.2.1. Empiryczny współczynnik korelacji liniowej	135
3.3.2.2. Empiryczny współczynnik korelacji dwuseryjnej	139
3.3.2.3. Współczynnik skojarzenia	142
3.3.3. Regresja dwóch zmiennych	143
3.3.4. Uogólnienie współczynnika korelacji	149
3.3.5. Regresja wielowymiarowa oraz korelacja wieloraka i cząstkowa	152

Rozdział 4
ESTYMACJA PARAMETRÓW ZMIENNYCH LOSOWYCH

4.1. Wprowadzenie	163
4.2. Estymatory i ich podstawowe własności	167
4.3. Metoda największej wiarygodności	169
4.4. Punktowa estymacja parametrów jednowymiarowej zmiennej losowej	175
4.4.1. Szacowanie wartości oczekiwanej	175

4.4.1.1. Średnia arytmetyczna z próby	175
4.4.1.2. Mediana z próby	176
4.4.2. Szacowanie wariancji i odchylenia standardowego	177
4.4.2.1. Wariancja z próby	177
4.4.2.2. Odchylenie standardowe z próby	181
4.4.2.3. Szacowanie odchylenia standardowego normalnej zmiennej losowej na podstawie rozstępu z próby	182
4.5. Rozkłady podstawowych charakterystyk z próby	185
4.5.1. Rozkład średniej arytmetycznej z próby	185
4.5.2. Rozkład wariancji i odchylenia standardowego z próby	188
4.6. Przedziałowa estymacja parametrów jednowymiarowej zmiennej losowej	191
4.6.1. Ogólne zasady konstruowania przedziałów ufności	191
4.6.2. Przedział ufności dla wartości oczekiwanej	191
4.6.3. Przedział ufności dla różnicy dwóch wartości oczekiwanych	194
4.6.4. Przedziały ufności dla wariancji i odchylenia standardowego	197
4.6.5. Przedział ufności dla stosunku dwóch wariancji	199
4.6.6. Przedział ufności dla wskaźnika struktury	201
4.6.7. Przedział ufności dla różnicy między dwoma wskaźnikami struktury ..	203
4.6.8. Liczebność próby	204
4.7. Punktowa i przedziałowa estymacja współczynnika korelacji liniowej	205

Rozdział 5

WERYFIKACJA HIPOTEZ STATYSTYCZNYCH

5.1. Wprowadzenie	208
5.2. Weryfikacja hipotez dotyczących wartości oczekiwanej	214
5.2.1. Weryfikacja istotności różnicy między wartością oczekiwaną zmiennej losowej a ustaloną wartością	214
5.2.1.1. Test u	215
5.2.1.2. Test t	221
5.2.2. Weryfikacja istotności różnicy między wartościami oczekiwanymi dwóch zmiennych losowych	223
5.2.2.1. Test u	224
5.2.2.2. Test t	225
5.3. Weryfikacja hipotez dotyczących wariancji	228
5.3.1. Weryfikacja istotności różnicy między wariancją a ustaloną wartością .	228
5.3.2. Weryfikacja istotności różnicy między dwiema wariancjami	233
5.4. Weryfikacja hipotez dotyczących wskaźnika struktury	236
5.4.1. Weryfikacja istotności różnicy między wskaźnikiem struktury a ustaloną wartością	236
5.4.2. Weryfikacja istotności różnicy między dwoma wskaźnikami struktury	238
5.5. Weryfikacja istotności współczynnika korelacji	241
5.6. Weryfikacja niezależności dwóch cech za pomocą testu chi-kwadrat	244
5.7. Weryfikacja zgodności rozkładu empirycznego z rozkładem hipotetycznym za pomocą testu chi-kwadrat	249
5.8. Weryfikacja hipotezy o losowości wyników doświadczenia	254
5.9. Sekwencyjne metody weryfikacji hipotez statystycznych	257
5.9.1. Ogólne zasady funkcjonowania testów sekwencyjnych	257

5.9.2. Weryfikacja hipotez dotyczących wartości oczekiwanej normalnej zmiennej losowej	258
5.9.3. Weryfikacja hipotez dotyczących odchylenia standardowego normalnej zmiennej losowej	261
5.9.4. Weryfikacja hipotez dotyczących wskaźnika struktury	263
5.9.5. Weryfikacja hipotez dotyczących parametru rozkładu Poissona	265

Rozdział 6

ANALIZA SZEREGÓW CZASOWYCH

6.1. Wprowadzenie	268
6.2. Wskaźnikowa analiza szeregów czasowych	270
6.2.1. Wskaźniki indywidualne	270
6.2.2. Wskaźniki agregatowe	275
6.3. Procedury wyznaczania tendencji rozwojowej	279
6.3.1. Metoda średnich ruchomych	279
6.3.2. Analityczne wygładzanie szeregów czasowych	281
6.4. Wyodrębnianie wahań sezonowych (okresowych)	286
6.5. Wyodrębnianie wahań przypadkowych (losowych)	294
6.6. Monitorowanie procesów za pomocą kart kontrolnych	299
6.6.1. Karty kontrolne Shewharta	300
6.6.1.1. Karta kontrolna \bar{x}	300
6.6.1.2. Karta kontrolna $\bar{x}-s$	305
6.6.1.3. Karta kontrolna $\bar{x}-r$	308
6.6.1.4. Karta kontrolna z i karta kontrolna w	309
6.6.1.5. Karta kontrolna c i karta kontrolna u	314
6.6.2. Karty kontrolne sum skumulowanych	321
6.6.2.1. Monitorowanie wartości oczekiwanej normalnej zmiennej losowej	321
6.6.2.2. Kontrola odchylenia standardowego normalnej zmiennej losowej	333
6.6.2.3. Monitorowanie parametru p zero-jedynkowej zmiennej losowej	334
6.6.2.4. Monitorowanie parametru λ zmiennej losowej Poissona	338
Aneks. Tablice	342
I. Wartości funkcji $\Theta(u)$	346
II. Kwantyle rozkładów zmiennej losowej t_r -Studenta	348
III. Kwantyle rozkładów zmiennej losowej χ_r^2	349
IV. Kwantyle rozkładów zmiennej losowej F	351
V. Parametry rozkładu rozstępu	355
VI. Kwantyle rozkładu serii	356
VII. Fragment tablicy liczb losowych	359
VIII. Fragment tablicy liczb złotych	360
IX. Funkcja wykładnicza e^x i e^{-x}	363

OD AUTORÓW

Oddajemy do rąk Czytelników czwarte, zmodyfikowane wydanie skryptu, który – po dokonanych zmianach – nosi tytuł *Statystyka z elementami statystycznych metod monitorowania procesów*. Drugie wydanie tej pracy ukazało się w roku 1997 i było zmodyfikowaną wersją skryptu *Elementy statystyki i statystycznej kontroli jakości*, wydanego po raz pierwszy w roku 1992. Kolejne, trzecie wydanie, poprawione i uzupełnione, ukazało się w roku 2000 pt. *Statystyka z elementami statystycznych metod sterowania jakością*. Zebrane doświadczenia dydaktyczne skłoniły autorów do wprowadzenia pewnych zmian w rozdziałach 5 i 6, poświęconych weryfikacji hipotez statystycznych i analizie szeregów czasowych. W pozostałych częściach skryptu wprowadzono tylko niewielkie poprawki i usunięto zauważone błędy.

Skrypt ten – podobnie jak jego poprzednie wydania – przeznaczony jest przede wszystkim dla studentów kierunku towaroznawstwa. Po dokonanych zmianach może on być użyteczny również dla studentów innych kierunków studiów, a zwłaszcza dla słuchaczy studiów podyplomowych, zainteresowanych pogłębieniem swojej wiedzy w zakresie statystycznych metod analizy danych i monitorowania procesów.

Skrypt składa się z sześciu rozdziałów i aneksu, w którym zamieszczono niezbędne tablice statystyczne. Rozdział 1 jest swego rodzaju repetytorium z rachunku prawdopodobieństwa i razem z rozdziałem 2, w którym zarysowano podstawowe problemy metodologii badań statystycznych, stanowi podstawę rozważań przedstawionych w dalszych częściach skryptu. W rozdziale 3 omówiono podstawowe problemy i metody statystyki opisowej, natomiast rozdziały 4 i 5 poświęcone są statystyce matematycznej. Znacznym modyfikacjom poddano rozdział 6, poświęcony analizie szeregów czasowych. Włączono do niego opis konstrukcji i funkcjonowania kart kontrolnych Shewharta oraz kart kontrolnych sum skumulowanych, które – zgodnie z ich formalną strukturą – potraktowano jako narzędzia monitorowania procesów, realizowanego w formie badań wyrwykowych (niewyczerpujących).

W każdym rozdziale zastosowano odrębną, ciągłą numerację wzorów, tablic i rysunków. W taki sam sposób ponumerowano również przykłady. Wzory przytoczone w przykładach mają osobną numerację. Do numerowania tych wzorów zastosowano cyfry rzymskie, zapisywane za pomocą małych liter alfabetu łacińskiego, na przykład: (i), (ii), (iii), (iv). Koniec każdego przykładu oznaczono

symbolem ■. Symbol □ oznacza natomiast, że przykład będzie kontynuowany w dalszym wykładzie.

Czytelników, którzy pragną poszerzyć swoją wiedzę z zakresu statystyki, a także statystycznych metod monitorowania procesów, odsyłamy do następujących podręczników:

– A.D. Aczel, *Statystyka w zarządzaniu*, Wydawnictwo Naukowe PWN, Warszawa 2000,

– J.R. Benjamin, C.A. Cornell, *Rachunek prawdopodobieństwa, statystyka matematyczna i teoria decyzji dla inżynierów*, WNT, Warszawa 1977,

– H. Bierman, C.P. Bonini, W.H. Hausman, *Quantitative Analysis for Business Decisions*, Irwin, Boston 1991,

– D. Bobrowski, *Probabilistyka w zastosowaniach technicznych*, wyd. II poprawione i zmienione, WNT, Warszawa 1986,

– S. Brandt, *Analiza danych*, Wydawnictwo Naukowe PWN, Warszawa 1998,

– J. Czermiński, A. Iwasiewicz, Z. Paszek, A. Sikorski, *Metody statystyczne dla chemików*, wyd. II zmienione, PWN, Warszawa 1992,

– Z. Hellwig, *Elementy rachunku prawdopodobieństwa i statystyki matematycznej*, Wydawnictwo Naukowe PWN, Warszawa 1995,

– G. Keller, B. Warrack, H. Bartel, *Statistics for Management and Economics. A System Approach*, Wadsworth Publ. Co., Wadsworth, Belmont 1988,

– J. Koronacki, J. Mielniczuk, *Statystyka dla studentów kierunków technicznych i przyrodniczych*, WNT, Warszawa 2001,

– W. Mendenhall, J.E. Reinmuth, R. Beaver, D. Duhan, *Statistics for Management and Economics*, Duxbury Press, Boston 1986,

– J. Steczkowski, *Metoda reprezentacyjna w badaniu zjawisk ekonomiczno-społecznych*, PWN, Warszawa 1996.

Tych Czytelników, którzy zechcą pogłębić swoje wiadomości z zakresu statystycznych metod sterowania procesami, odsyłamy do następujących książek:

– H. Gitlow, S. Gitlow, A. Oppenheim, R. Oppenheim, *Tools and Methods for the Improvement of Quality*, Irwin, Homewood, Boston 1989,

– O. Hryniewicz, *Nowoczesne metody statystycznego sterowania jakością*, Omnitech Press, PAN, Instytut Badań Systemowych, Warszawa 1996,

– A. Iwasiewicz, *Statystyczna kontrola jakości w toku produkcji; systemy i procedury*, PWN, Warszawa 1985,

– A. Iwasiewicz, *Zarządzanie jakością*, Wydawnictwo Naukowe PWN, Warszawa–Kraków 1999 (rozdz. 5–9),

– D.C. Montgomery, *Introduction to Statistical Quality Control*, John Wiley & Sons, New York 1991,

– J.R. Thompson, J. Koronacki, *Statystyczne sterowanie procesem. Metoda Deminga etapowej optymalizacji jakości*, Akademicka Oficyna Wydawnicza PLJ, Warszawa 1994.

Rozdział 1

ELEMENTY RACHUNKU PRAWDOPODOBIENSTWA

1.1. Wprowadzenie

Znakomita większość przedsięwzięć empirycznych realizowanych, albo tylko obserwowanych, w ekonomii i zarządzaniu, a także w technice i w innych obszarach działalności człowieka, to doświadczenia losowe. Jest to jedno z głównych pojęć w dziedzinie praktycznych zastosowań rachunku prawdopodobieństwa i statystyki. Doświadczeniem losowym jest każde przedsięwzięcie empiryczne, rzeczywiste albo symulacyjne, którego wyniku nie można przewidzieć mimo sprecyzowania warunków, w których jest ono realizowane. Doświadczeniem losowym jest więc, na przykład, każde z wymienionych poniżej działań:

- egzamin ze statystyki,
- rzut kostką do gry,
- inwestowanie w akcje spółek notowanych na giełdzie,
- działalność przedsiębiorstwa w ustalonym okresie rozliczeniowym,
- oczekiwanie na przystanku tramwajowym.

W żadnym z wymienionych przypadków, ani też w żadnej innej podobnej sytuacji, nie jesteśmy w stanie przewidzieć dokładnego wyniku doświadczenia. Nie jesteśmy w stanie tego dokonać, albowiem o wyniku doświadczenia decydują nie tylko czynniki systematyczne, w formie warunków początkowych albo parametryzujących przebieg doświadczenia, ale również spłot różnego rodzaju czynników przypadkowych (losowych). Na podstawie ogólnej wiedzy o badanym zjawisku można natomiast podać zbiór wszystkich możliwych wyników realizowanego, albo obserwowanego, doświadczenia. Zwróćmy uwagę na pierwszy z wymienionych powyżej przykładów. Przystępując do egzaminu ze statystyki można uzyskać jedną spośród sześciu następujących ocen:

- bardzo dobrze,
- dobrze plus,
- dobrze,
- dostatecznie plus,
- dostatecznie,
- niedostatecznie.

Jest to zbiór wszystkich możliwych i wzajemnie wykluczających się wyników egzaminu, a oceny te nie mogą być rozłożone na prostsze. W teorii rachunku prawdopodobieństwa zdarzenia takie nazywane są zdarzeniami elementarnymi, a ich zbiór nazywany jest przestrzenią zdarzeń elementarnych. Zbiór ten oznaczany jest symbolem Ω , a jego elementy (zdarzenia elementarne) symbolem ω . W rozważanym przypadku mamy więc:

$$\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5, \omega_6\},$$

gdzie:

$$\begin{array}{ll} \omega_1 - \text{bardzo dobrze,} & \omega_4 - \text{dostatecznie plus,} \\ \omega_2 - \text{dobrze plus,} & \omega_5 - \text{dostatecznie,} \\ \omega_3 - \text{dobrze,} & \omega_6 - \text{niedostatecznie.} \end{array}$$

Każde zdarzenie elementarne jest zdarzeniem losowym, ale nie zachodzi relacja odwrotna. Nie każde zdarzenie losowe jest zdarzeniem elementarnym. W ogólnym ujęciu problemu zdarzenie losowe definiuje się jako podzbiór przestrzeni zdarzeń elementarnych Ω . Dodajmy, że może to być również podzbiór pusty (zdarzenie niemożliwe) oraz podzbiór równy całej przestrzeni zdarzeń elementarnych (zdarzenie pewne). Zdarzenie losowe oznacza się zwykle dużymi literami z początku alfabetu łacińskiego (A, B, C, \dots). W rozważanym przykładzie szczególnie często definiowane są dwa zdarzenia losowe, a mianowicie:

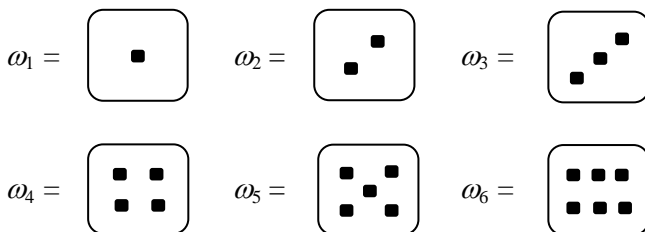
$$A_1 = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5\}, \quad A_2 = \{\omega_6\}.$$

Zdarzenie losowe A_1 oznacza pozytywny wynik egzaminu, natomiast zdarzenie losowe A_2 oznacza wynik negatywny.

Również drugi z wyróżnionych powyżej przykładów, a mianowicie rzut kostką do gry, jest doświadczeniem losowym o sześćelementowej przestrzeni zdarzeń elementarnych. W tym przypadku zdarzeniem elementarnym jest liczba „oczek” na górnej płaszczyźnie kostki. Mamy więc, tak samo jak poprzednio:

$$\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5, \omega_6\},$$

ale obecnie



Na podstawie tak określonej przestrzeni zdarzeń elementarnych można zdefiniować takie – na przykład – zdarzenie losowe:

A – parzysta liczba „oczek”, B – nieparzysta liczba „oczek”.

Mamy więc:

$$A = \{\omega_2, \omega_4, \omega_6\}, \quad B = \{\omega_1, \omega_3, \omega_5\}.$$

W praktycznych zastosowaniach rachunku prawdopodobieństwa i statystyki przestrzeń zdarzeń elementarnych Ω jest zawsze zbiorem co najwyżej przeliczalnym. W takiej sytuacji każdy podzbiór zbioru Ω jest zdarzeniem losowym. W teorii rachunku prawdopodobieństwa rozważa się również takie przypadki, gdy Ω jest zbiorem nieprzeliczalnym. W tej sytuacji nie każdy podzbiór przestrzeni zdarzeń elementarnych Ω musi być zdarzeniem losowym. Dlatego też w rozważaniach teoretycznych dodatkowo definiuje się taką rodzinę podzbiorów zbioru Ω , której elementami są zdarzenia losowe. Bierze się przy tym pod uwagę najmniejsze, przeliczalnie addytywne ciało zdarzeń. Takie ciało nazywane jest borelowskim ciałem zdarzeń i jest ono zwykle oznaczane symbolem \mathcal{B} . Przestrzeń zdarzeń elementarnych nie wyczerpuje wszystkich problemów probabilistycznych związanych z doświadczeniem losowym. Pełniejszą charakterystyką doświadczenia losowego jest przestrzeń probabilistyczna postaci:

$$(\Omega, \mathcal{B}, P), \tag{1.1}$$

albo

$$(\Omega, P), \tag{1.2}$$

gdzie P jest miarą probabilistyczną, pozwalającą mierzyć prawdopodobieństwa pojawiania się poszczególnych zdarzeń losowych, określonych jako podzbiory przestrzeni zdarzeń elementarnych Ω .

1.2. Prawdopodobieństwo

Prawdopodobieństwo zdarzenia losowego jest unormowaną miarą szansy zrealizowania się tego zdarzenia jako rezultatu doświadczenia losowego. Wyznaczanie wartości prawdopodobieństw wymaga operowania definicjami możliwymi do empirycznego zastosowania. Definicje te omówimy poniżej. Obecnie zauważmy, że każda operacyjna definicja prawdopodobieństwa powinna spełniać

warunki wynikające z trzech aksjomatów, sformułowanych przez A.N. Kołmogorowa¹. Przypomnijmy te aksjomaty:

– Każdemu zdarzeniu losowemu $A \in \mathbf{B}$ przyporządkowana jest jednoznacznie nieujemna liczba $P(A)$, nazywana prawdopodobieństwem zdarzenia A . Tak więc:

$$P(A) \geq 0. \quad (1.3)$$

– Prawdopodobieństwo zdarzenia pewnego, czyli zdarzenia obejmującego cały zbiór zdarzeń elementarnych Ω , równe jest jedności. Tak więc:

$$P(\Omega) = 1. \quad (1.4)$$

– Dla każdego ciągu A_1, A_2, A_3, \dots , parami rozłącznych zdarzeń, czyli takich zdarzeń, że:

$$\bigwedge_{\substack{i,j \\ i \neq j}} A_i \cap A_j = \emptyset \quad (1.5)$$

prawdziwa jest zależność:

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i). \quad (1.6)$$

Funkcja P spełniająca te aksjomaty jest miarą probabilistyczną określoną na zbiorze \mathbf{B} .

1.2.1. Definicje prawdopodobieństwa

Definicja klasyczna

Nazwą tą określaną jest najwcześniejsza historycznie definicja, podana w roku 1812 przez Laplace'a². Definicję tę można stosować wówczas, gdy Ω jest skończonym zbiorem, jednakowo możliwych i wzajemnie wykluczających się zdarzeń. Jeśli zbiór Ω zawiera n zdarzeń ω i jeśli zdarzenie losowe A jest m -elementowym podzbiorem zbioru Ω , to:

$$P(A) = \frac{m}{n}. \quad (1.7)$$

¹ A.N. Kołmogorow, *Osnovnyje ponjatja teorji vierojatnostiej*, Izd. Nauka, Moskwa 1974.

² Pierre Simon de Laplace (1749–1827), francuski astronom, matematyk i fizyk.

Definicja ta ma pewien defekt logiczny, polegający na tym, że do definiowania prawdopodobieństwa wykorzystuje się pojęcie prawdopodobieństwa. Wszak zdarzenia „jednakowo możliwe” to zdarzenia jednakowo prawdopodobne. Nie zmienia to jednak faktu, że za pomocą tej definicji można poprawnie – a więc w sposób zgodny z podanymi powyżej aksjomatami – wyznaczyć prawdopodobieństwa w schematach urnowych, a także w zagadnieniach praktycznych, które można modelować za pomocą schematów urnowych³.

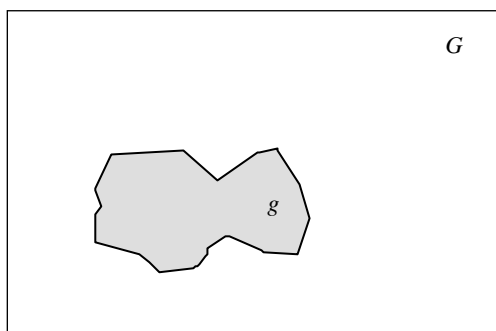
Definicja geometryczna

Za pomocą definicji klasycznej nie można wyznaczyć prawdopodobieństwa zdarzenia losowego w sytuacji pokazanej na rys. 1.1. Ponieważ zarówno obszar G , jak i obszar g zawierają nieskończenie wiele punktów, zatem – posługując się klasyczną definicją – nie można wyznaczyć prawdopodobieństwa zdarzenia losowego A polegającego na tym, że wybierając losowo punkt z obszaru G , wybierzemy punkt z obszaru g . Trudność tę usuwa definicja geometryczna, według której:

$$P(A) = \frac{\text{mes } g}{\text{mes } G}, \quad (1.8)$$

gdzie $\text{mes } g$ oznacza miarę obszaru g , natomiast $\text{mes } G$ jest miarą obszaru G ⁴.

Zauważmy, że definicja geometryczna nie usuwa wspomnianego powyżej defektu logicznego definicji klasycznej. Również tu wszystkie zdarzenia elementarne, czyli wszystkie punkty w obszarze G , muszą być „równouprawnione”.



Rys. 1.1. Geometryczna interpretacja prawdopodobieństwa

³ Pod pojęciem schematu urnowego rozumiemy tu wszelkie modele fizyczne o odpowiednio zorganizowanej strukturze stochastycznej.

⁴ „mes” jest skrótem angielskiego słowa *measure* – miara.

Definicje częstościowe

U podstaw częstościowych definicji prawdopodobieństwa leży pojęcie częstości względnej. Charakterystykę tę definiujemy następująco:

$$V_n(A) = \frac{n(A)}{n}, \quad (1.9)$$

gdzie:

$V_n(A)$ – częstość względna zdarzenia A w hipotetycznie jednorodnym, n -elementowym ciągu doświadczeń losowych,

n – liczba wykonanych doświadczeń,

$n(A)$ – liczba doświadczeń, w których zrealizowało się zdarzenie A .

Ponieważ $n(A) \leq n$, zatem:

$$0 \leq V_n(A) \leq 1. \quad (1.10)$$

Częstość względna jest charakterystyką, która odnosi się do przeszłości, podczas gdy prawdopodobieństwo – jak to już powiedzieliśmy powyżej – ma charakteryzować szanse zrealizowania się danego zdarzenia A . Jeśli jednak warunki, w których realizowane jest doświadczenie, pozostają nie zmienione, to częstość względną $V_n(A)$ można wykorzystać jako przybliżenie prawdopodobieństwa $P(A)$, a przybliżenie to jest tym lepsze, im większa jest liczba obserwacji n . Stwierdzenie to – znajdujące swe teoretyczne uzasadnienie w prawie wielkich liczb Bernoulliego (zob. punkt 1.6.1) – legło u podstaw dwóch częstościowych definicji prawdopodobieństwa. Według tzw. definicji statystycznej prawdopodobieństwo zdarzenia losowego A jest to pewna nieznaną liczbą $P(A)$ z przedziału $[0; 1]$, wokół której oscylują częstości względne $V_n(A)$ i do której przybliżają się one ze wzrostem liczby obserwacji n . Na tej samej konstrukcji logicznej oparta jest definicja podana przez Misesa⁵. Definicja ta głosi, że:

$$P(A) = \lim_{n \rightarrow \infty} V_n(A). \quad (1.11)$$

Definicji tej zarzuca się brak logicznej spójności, przejawiający się w pomieszaniu elementów empirycznych i teoretycznych.

⁵ Richard von Mises (1883–1953), austriacki statystyk, urodzony we Lwowie. Podana przez niego definicja prawdopodobieństwa ma postać: $p_1 = \lim_{N \rightarrow \infty} N_1/N$, gdzie: p_1 – prawdopodobieństwo uzyskania liczby 1 w ciągu zer i jedynek, N_1 – liczebność zbioru jedynek, $N = N_1 + N_0$, przy czym N_0 jest liczebnością zbioru zer.

Dodajmy, że ocena prawdopodobieństwa za pomocą odpowiedniej częstości względnej jest podstawową metodą postępowania we wszystkich naukach empirycznych.

1.2.2. Podstawowe własności prawdopodobieństwa

1. Prawdopodobieństwo dowolnego zdarzenia losowego A jest liczbą spełniającą następującą, podwójną nierówność:

$$0 \leq P(A) \leq 1. \quad (1.12)$$

2. Prawdopodobieństwo zdarzenia niemożliwego, tj. takiego, które jest pustym podzbiorem zbioru Ω , równa się zeru:

$$P(\emptyset) = 0. \quad (1.13)$$

3. Suma prawdopodobieństwa zdarzenia A i zdarzenia przeciwnego \bar{A} równa się jedności:

$$P(A) + P(\bar{A}) = 1. \quad (1.14)$$

4. Jeżeli zdarzenie losowe A zawiera się w zdarzeniu losowym B , czyli $A \subset B$, to:

$$P(A) \leq P(B). \quad (1.15)$$

5. Prawdopodobieństwo sumy dwóch zdarzeń A i B równa się sumie prawdopodobieństw tych zdarzeń, pomniejszonej o prawdopodobieństwo ich iloczynu:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B). \quad (1.16)$$

W przypadku większej liczby zdarzeń, na przykład trzech, wzór (1.16) przyjmie następującą postać:

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C). \quad (1.17)$$

6. Prawdopodobieństwo warunkowe. Prawdopodobieństwem warunkowym lub inaczej prawdopodobieństwem względnym $P(A|B)$ zdarzenia A pod warunkiem, że zaszło zdarzenie B , nazywamy iloraz prawdopodobieństwa iloczynu zdarzeń A i B oraz prawdopodobieństwa zdarzenia B :

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \quad P(B) > 0 \quad (1.18)$$

stąd

$$P(A \cap B) = P(B) \cdot P(A|B), \quad (1.19)$$

co oznacza, że prawdopodobieństwo łącznego zajścia zdarzeń A i B równa się iloczynowi prawdopodobieństwa zdarzenia B i prawdopodobieństwa warunkowego zdarzenia A , pod warunkiem, że zaszło zdarzenie B .

Podobnie można napisać:

$$P(B|A) = \frac{P(A \cap B)}{P(A)}, \quad P(A) > 0 \quad (1.20)$$

stąd

$$P(A \cap B) = P(A) \cdot P(B|A). \quad (1.21)$$

7. Prawdopodobieństwo iloczynu dwóch zdarzeń niezależnych. Zdarzenia A i B są niezależne, jeżeli zajście zdarzenia A nie wpływa na prawdopodobieństwo zajścia zdarzenia B , i odwrotnie. W konsekwencji, wzór (1.19) przybierze postać:

$$P(A \cap B) = P(A) \cdot P(B), \quad (1.22)$$

bowiem

$$P(A) = P(A|B)$$

oraz

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

W przypadku k niezależnych zdarzeń losowych mamy:

$$\begin{aligned} P(A_1 \cap A_2 \cap \dots \cap A_k) &= P\left(\bigcap_{i=1}^k A_i\right) = \prod_{i=1}^k P(A_i) = \\ &= P(A_1) \cdot P(A_2) \cdot \dots \cdot P(A_k). \end{aligned} \quad (1.23)$$

1.2.3. Prawdopodobieństwo zupełne i wzór Bayesa

Układ zupełny zdarzeń

Ciąg zdarzeń losowych $A_1, A_2, \dots, A_i, \dots, A_n$ nazywamy zupełnym układem zdarzeń, jeśli:

$$\bigwedge_{\substack{i,j \\ i \neq j}} A_i \cap A_j = \emptyset \quad (1.24)$$

oraz jeśli:

$$A_1 \cup A_2 \cup \dots \cup A_i \cup \dots \cup A_n = \bigcup_{i=1}^n A_i = \Omega. \quad (1.25)$$

Warunek (1.24) oznacza, że zdarzenia wykluczają się parami. Warunek (1.25) oznacza natomiast, że rozważany ciąg zdarzeń wyczerpuje przestrzeń zdarzeń elementarnych, związaną z danym doświadczeniem losowym. Zauważmy, że w szczególności zdarzenia przeciwne (A, \bar{A}) tworzą zupełny układ zdarzeń.

Prawdopodobieństwo zupełne

Jeżeli zdarzenia $A_1, \dots, A_i, \dots, A_n$ tworzą układ zupełny zdarzeń, to prawdopodobieństwo zdarzenia B , mogącego się zrealizować pod warunkiem zajścia poszczególnych zdarzeń A_i , można obliczyć za pomocą następującego wzoru:

$$P(B) = \sum_{i=1}^n P(A_i) \cdot P(B|A_i). \quad (1.26)$$

Wzór Bayesa

Jeżeli zdarzenia $A_1, \dots, A_i, \dots, A_n$ tworzą układ zupełny zdarzeń oraz $P(B) > 0$, to wówczas prawdopodobieństwo zajścia dowolnego zdarzenia A_i pod warunkiem, że zaszło zdarzenie B , można obliczyć ze wzoru:

$$P(A_i|B) = \frac{P(A_i) \cdot P(B|A_i)}{\sum_{i=1}^n P(A_i) \cdot P(B|A_i)}. \quad (1.27)$$

We wzorze (1.27) wyrażenie $P(A_i|B)$ nazywamy niekiedy prawdopodobieństwem *a posteriori*, gdyż podaje prawdopodobieństwo zrealizowania się zdarzenia A_i dopiero po wystąpieniu zdarzenia B , natomiast wyrażenie $P(A_i)$ – prawdopodobieństwem *a priori*.

Przykład 1.1

W pewnej fabryce dwa automaty produkują ten sam gatunek śrub, które wrzucane są do wspólnego pojemnika. Wiadomo przy tym, że 60% produkcji pochodzi z automatu I, a 40% z automatu II. Stwierdzono ponadto, że 95% śrub wytworzonych przez automat I odpowiada wymaganiom jakościowym, a 5% stanowią śruby z wadami. Analogiczne liczby dla automatu II wynoszą 97% i 3%. Obliczyć prawdopodobieństwo tego, że śruba losowo wybrana z pojemnika pochodzi z automatu I, jeżeli stwierdzono, że odpowiada ona wymaganiom jakościowym.

Wprowadzamy następujące oznaczenia:

A_1 – zdarzenie polegające na tym, że losowo wybrana śruba pochodzi z automatu I,

A_2 – zdarzenie polegające na tym, że losowo wybrana śruba pochodzi z automatu II,

B – zdarzenie polegające na tym, że losowo wybrana śruba spełnia wymagania jakościowe.

Stąd

$$P(A_1) = 0,6; \quad P(A_2) = 0,4; \quad P(B|A_1) = 0,95; \quad P(B|A_2) = 0,97;$$

$$\begin{aligned} P(A_1|B) &= \frac{P(A_1) \cdot P(B|A_1)}{P(A_1) \cdot P(B|A_1) + P(A_2) \cdot P(B|A_2)} = \\ &= \frac{0,6 \cdot 0,95}{0,6 \cdot 0,95 + 0,4 \cdot 0,97} = \frac{0,57}{0,57 + 0,388} = \frac{0,57}{0,958} \approx 0,59. \blacksquare \end{aligned}$$

1.3. Jednowymiarowe zmienne losowe

1.3.1. Podstawowe pojęcia i określenia

Z każdym doświadczeniem losowym związana jest określona przestrzeń probabilistyczna (Ω, \mathcal{B}, P) . Mając tę przestrzeń można definiować zmienne losowe, opisujące liczbowo badany obiekt lub zjawisko.

Jednowymiarowa zmienna losowa X jest to funkcja rzeczywista postaci:

$$X: \Omega \rightarrow \mathbf{R}, \tag{1.28}$$

mająca tę własność, że dla każdej liczby rzeczywistej $x \in \mathbf{R}$ zbiór zdarzeń elementarnych $\omega \in \Omega$, dla których:

$$X(\omega) < x \quad (1.29)$$

jest zdarzeniem losowym, czyli elementem borelowskiego ciała zdarzeń \mathcal{B} .

Zmienne losowe oznaczamy najczęściej dużymi, końcowymi literami alfabetu łacińskiego, na przykład X , T , Y , Z . Wartości liczbowe, jakie te zmienne przyjmują, oznaczamy odpowiednio małymi literami, używając w miarę potrzeby indeksów. Na przykład, kolejne realizacje zmiennej losowej Y mogą być oznaczone następująco: y_1, \dots, y_n .

Rozróżniamy dwa typy zmiennych losowych, a mianowicie:

- zmienne losowe skokowe (dyskretne),
- zmienne losowe ciągłe.

Zmienne losowe skokowe mogą przyjmować wartości wyrażające się tylko niektórymi liczbami rzeczywistymi z określonych przedziałów; najczęściej liczbami całkowitymi nieujemnymi. Zbiór możliwych wartości zmiennej losowej dyskretnej jest w konsekwencji zawsze przeliczalny, a niekiedy także skończony. Przykładem zmiennej losowej dyskretnej o przeliczalnym zbiorze wartości może być liczba wad technologicznych w jednej sztuce wyrobu, liczba wypadków komunikacyjnych w ustalonym przedziale czasu, na określonym obszarze kraju, liczba cykli pracy wyłącznika elektrycznego itp. Każda z tych zmiennych losowych może przyjmować wartości wyrażające się liczbami całkowitymi nieujemnymi: $0, 1, \dots$ W żadnym z tych przypadków nie można jednak w sposób ścisły, wynikający z istoty badanego zjawiska, wskazać wartości najwyższej.

Z inną sytuacją mamy do czynienia w przypadku takiej zmiennej losowej dyskretnej, jak liczba sztuk wadliwych w partii towaru o liczności N . Tak zdefiniowana zmienna losowa może przyjmować wartości $0, 1, \dots, N$. Nie może bowiem być więcej sztuk wadliwych niż liczność całej partii towaru.

Zmienne losowe ciągłe mogą przyjmować wartości wyrażające się dowolnymi liczbami rzeczywistymi z określonych przedziałów. Przykładem zmiennej losowej ciągłej może być temperatura jakiegoś ciała, stężenie procentowe roztworu, czas bezawaryjnej pracy określonego elementu jakiegoś urządzenia itp. Jeżeli temperaturę jakiegoś materiału będziemy mierzyć w kelwinach (K), to wówczas może ona wyrazić się dowolną liczbą rzeczywistą z przedziału $[-273; \tau_r]$, gdzie τ_r oznacza temperaturę termicznego rozkładu materiału. Jeżeli natomiast pominiemy pewne ograniczenia wynikające z fizycznych właściwości roztworu, to wtedy jego procentowe stężenie może się wyrazić każdą liczbą rzeczywistą z przedziału $[0; 100]$. Czas pracy określonego urządzenia może się wyrazić dowolną liczbą rzeczywistą z przedziału $[0; +\infty)$.

Należy tutaj podkreślić, że ciągłość zmiennej losowej jest pojęciem, które w pełni funkcjonuje tylko w modelu matematycznym. Doświadczenie udyskretynia bowiem te modele i w praktyce wszystkie zmienne losowe mają charakter skokowy, z tym, że skokowość ta jest wyraźniej lub słabiej zaznaczona. Jeżeli – na przykład – wspomnianą wyżej temperaturę będziemy mierzyć za pomocą termometru o działce elementarnej 0,01 K, to wówczas uzyskany zbiór wyników będzie „praktycznie ciągły”. Jeśli natomiast zastosowany termometr będzie miał działkę elementarną 1 K, to wtedy uzyskiwane wyniki będą wyraźnie dyskretne. Obydwa zbiory wyników będą jednak nieciągłe w sensie matematycznym.

1.3.2. Charakterystyki rozkładu jednowymiarowej zmiennej losowej

Każda zmienna losowa postaci (1.28) jest funkcją, która przyporządkowuje wartości liczbowe zdarzeniom generowanym przez doświadczenie losowe. Zdarzenia te realizują się z prawdopodobieństwami określonymi przez probabilistyczny mechanizm doświadczenia. W konsekwencji, poszczególne wartości zmiennej losowej realizują się z takimi samymi prawdopodobieństwami jak odpowiadające im zdarzenia losowe. Regułę, według której jednostkowa masa prawdopodobieństwa rozłożona jest na poszczególne wartości zmiennej losowej, albo na pewne agregaty tych wartości, nazywamy rozkładem prawdopodobieństwa (lub krócej: rozkładem) zmiennej losowej. Charakterystyki wykorzystywane do opisu rozkładu omówione są poniżej.

1.3.2.1. Funkcja prawdopodobieństwa i funkcja gęstości prawdopodobieństwa

Rozkład prawdopodobieństwa zmiennej losowej dyskretnej może być dany za pomocą funkcji prawdopodobieństwa:

$$P(Y = y_i) = p(y_i) = p_i; \quad i = 1, \dots, k \quad (1.30)$$

takiej, że:

$$\sum_{i=1}^k p_i = 1. \quad (1.31)$$

Jest to więc funkcja przyporządkowująca wartości prawdopodobieństwa (p_i) poszczególnym wartościom y_i zmiennej losowej Y . Funkcja ta może być określona analitycznie za pomocą wzoru, przez wymienienie wszystkich par (y_i, p_i) , lub też graficznie.

Przykład 1.2

Na podstawie długotrwałych obserwacji procesu produkcji pewnego wyrobu sztukowego stwierdzono, że:

5% produkowanego wyrobu nie odpowiada wymaganiom jakościowym (zdarzenie losowe A_1);

40% odpowiada wymaganiom jakościowym i należy do gatunku pierwszego (zdarzenie losowe A_2);

55% odpowiada wymaganiom jakościowym i należy do gatunku drugiego (zdarzenie losowe A_3).

Określimy następującą zmienną losową:

$$Y = \begin{cases} 1 & \text{– gdy zachodzi zdarzenie } A_1, \\ 2 & \text{– gdy zachodzi zdarzenie } A_2, \\ 3 & \text{– gdy zachodzi zdarzenie } A_3. \end{cases}$$

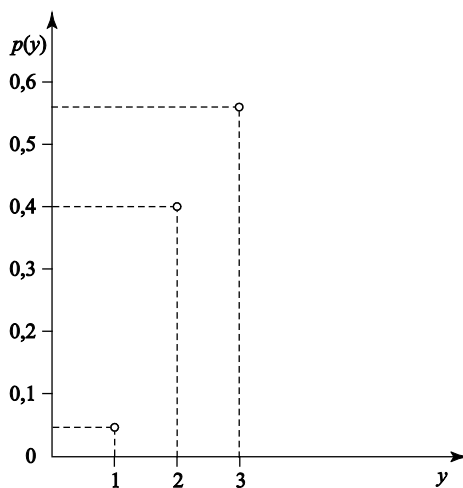
Funkcja prawdopodobieństwa tej zmiennej losowej przedstawia się następująco:

$$P(Y = 1) = 0,05,$$

$$P(Y = 2) = 0,40,$$

$$P(Y = 3) = 0,55.$$

Wykres tej funkcji pokazano na rys. 1.2. ■



Rys. 1.2. Przykład funkcji prawdopodobieństwa dyskretnej zmiennej losowej

Jeżeli zbiór wartości zmiennej losowej jest nieskończony i nie można podać postaci analitycznej funkcji prawdopodobieństwa, to wtedy określenie tej funkcji dwoma podanymi sposobami napotyka pewne trudności. Stosujemy wówczas postępowanie polegające na określeniu kolejnych par: $(y_1, p_1), \dots, (y_i, p_i), \dots$, aż do momentu, gdy spełniona jest relacja:

$$1 - \sum_i p_i < \varepsilon, \quad (1.32)$$

w której ε oznacza dowolnie małą, dodatnią liczbę rzeczywistą.

W przypadku zmiennych losowych ciągłych nie można określać funkcji prawdopodobieństwa ze względu na nieprzeliczalność zbiorów wartości tych zmiennych. Dowodzi się, że prawdopodobieństwo zdarzenia losowego polegającego na tym, że ciągła zmienna losowa Y przyjmie wartość równą konkretnej liczbie rzeczywistej y_0 , jest zawsze równe zeru:

$$P(Y = y_0) = 0. \quad (1.33)$$

Twierdzenie to ma oczywistą interpretację intuicyjną. W każdym, dowolnie wąskim przedziale na osi liczb rzeczywistych, znajduje się nieskończenie wiele tych liczb. Pytając zatem o prawdopodobieństwo pojawienia się jednej, konkretnej liczby rzeczywistej, pytamy o prawdopodobieństwo wystąpienia jednej spośród nieskończenie wielu wartości. Wobec powyższego, w przypadku zmiennych losowych ciągłych, zamiast funkcji prawdopodobieństwa określamy funkcję gęstości prawdopodobieństwa $f(y)$, którą definiujemy następująco:

$$f(y) = \lim_{\Delta y \rightarrow 0} \frac{P(y \leq Y < y + \Delta y)}{\Delta y}, \quad (1.34)$$

przy czym

$$\int_a^b f(y) dy = 1, \quad (1.35)$$

gdzie:

a i b – kres dolny i kres górny przedziału zmienności zmiennej losowej Y ,
 $f(y) \geq 0$.

1.3.2.2. Dystrybuanta

Dystrybuanta zmiennej losowej definiowana jest w sposób następujący:

$$F_Y(x) = P(Y < x), \quad x \in \mathbf{R}. \quad (1.36)$$

Wartościami dystrybuanty są więc prawdopodobieństwa zdarzeń losowych polegających na tym, że zmienna losowa Y przybierze wartość mniejszą niż liczba rzeczywista x .

W przypadku zmiennych losowych dyskretnych stosujemy niekiedy zmodyfikowaną definicję dystrybuanty, a mianowicie:

$$F_Y(x) = P(Y \leq x), \quad x \in \mathbf{R}. \quad (1.37)$$

Jeżeli nie ma wątpliwości, jakiej zmiennej losowej dotyczy dystrybuanta, to wówczas w jej symbolu pomijamy indeks oznaczający tę zmienną, pisząc po prostu $F(x)$, zamiast $F_Y(x)$ czy $F_Z(x)$.

Między dystrybuantą a funkcją prawdopodobieństwa i funkcją gęstości prawdopodobieństwa zachodzą następujące związki:

$$F(x) = \sum_{y_i \leq x} p_i, \quad (1.38)$$

$$F(x) = \int_{-\infty}^x f(y) dy. \quad (1.39)$$

Ze wzoru (1.39) wynika bezpośrednio, że:

$$f_Y(x) = F'_Y(x), \quad (1.40)$$

$$f_Y(x) = \lim_{\Delta x \rightarrow 0} \frac{P(x < Y < x + \Delta x)}{\Delta x} = \lim_{\Delta x \rightarrow 0} \frac{F(x + \Delta x) - F(x)}{\Delta x}. \quad (1.41)$$

Dystrybuanta dowolnej zmiennej losowej ma następujące własności:

$$1. \quad 0 \leq F(x) \leq 1, \quad \text{dla wszystkich } x; \quad (1.42)$$

$$2. \quad \text{Jeśli } x_2 > x_1, \quad \text{to } F(x_2) \geq F(x_1); \quad (1.43)$$

$$3. \quad \lim_{x \rightarrow -\infty} F(x) = F(-\infty) = 0; \quad (1.44)$$

$$4. \quad \lim_{x \rightarrow +\infty} F(x) = F(+\infty) = 1. \quad (1.45)$$

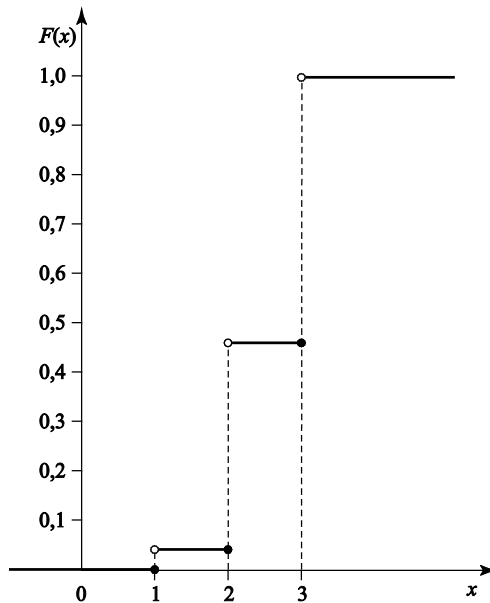
Przykład 1.3

Na podstawie funkcji prawdopodobieństwa zmiennej losowej X , określonej w przykładzie 1.2, można wyznaczyć dystrybuantę tej zmiennej losowej.

Stosując definicję wyrażoną wzorem (1.36), otrzymujemy:

$$F(x) = \begin{cases} 0 & \text{dla } x \leq 1, \\ 0,05 & \text{dla } 1 < x \leq 2, \\ 0,45 & \text{dla } 2 < x \leq 3, \\ 1 & \text{dla } x > 3. \end{cases}$$

Wykres tej dystrybuanty przedstawiono na rys. 1.3. ■



Rys. 1.3. Przykład dystrybuanty zmiennej losowej

1.3.3. Parametry jednowymiarowej zmiennej losowej

Podstawowymi parametrami jednowymiarowej zmiennej losowej są:

- wartość oczekiwana (wartość przeciętna, nadzieja matematyczna),
- wariancja i odchylenie standardowe.

Wartość oczekiwana dyskretnej zmiennej losowej definiowana jest następująco:

$$E(Y) = \sum_i y_i p_i. \quad (1.46)$$

W przypadku zmiennej losowej ciągłej funkcję prawdopodobieństwa należy zastąpić funkcją gęstości prawdopodobieństwa, a zabieg sumowania należy zastąpić całkowaniem. Mamy więc:

$$E(Y) = \int_{-\infty}^{+\infty} y \cdot f(y) dy. \quad (1.47)$$

Wartość oczekiwana zmiennej losowej dyskretnej i ciągłej istnieje tylko wtedy, gdy szereg lub zastąpiona przez znak sumy całka są bezwzględnie zbieżne. W przeciwnym wypadku wartość oczekiwana zmiennej losowej Y nie istnieje⁶.

Wariancja zmiennej losowej Y jest wartością oczekiwaną następującej zmiennej losowej:

$$[Y - E(Y)]^2. \quad (1.48)$$

Analogicznie jak wyżej, rozróżniamy dwa przypadki, w zależności od tego, czy zmienna losowa jest dyskretna, czy też ciągła. W przypadku zmiennej losowej dyskretnej mamy:

$$D^2(Y) = E[Y - E(Y)]^2 = \sum_i p_i [y_i - E(Y)]^2. \quad (1.49)$$

Przy obliczaniu wariancji wygodnie jest posłużyć się wzorem:

$$D^2(Y) = E(Y^2) - [E(Y)]^2. \quad (1.50)$$

⁶ Przykładem zmiennej losowej, która nie ma wartości oczekiwanej, jest zmienna o rozkładzie Cauchy'ego, określonym gęstością prawdopodobieństwa:

$$f(x) = \frac{1}{\pi} \cdot \frac{\lambda}{\lambda^2 + (x - a)^2},$$

gdzie: a – dowolna liczba rzeczywista, $\lambda > 0$, $x \in \mathbf{R}$.

Jeżeli natomiast zmienna losowa jest ciągła, to wtedy:

$$D^2(Y) = E[Y - E(Y)]^2 = \int_{-\infty}^{+\infty} f(y)[y - E(Y)]^2 dy. \quad (1.51)$$

Odchylenie standardowe zmiennej losowej, niezależnie od tego, czy jest to zmienna losowa dyskretna, czy też ciągła, jest dodatnim pierwiastkiem kwadratowym z wariancji:

$$D(Y) = \sqrt{D^2(Y)}. \quad (1.52)$$

Przykład 1.4

Mając funkcję prawdopodobieństwa zmiennej losowej Y , wyznaczoną w przykładzie 1.2, można obliczyć wartość oczekiwaną tej zmiennej. Na podstawie wzoru (1.46) mamy:

$$E(Y) = 1 \cdot 0,05 + 2 \cdot 0,40 + 3 \cdot 0,55 = 2,5.$$

W celu obliczenia wariancji tej zmiennej losowej należy zastosować wzór (1.49):

$$\begin{aligned} D^2(Y) &= 0,05(1-2,5)^2 + 0,40(2-2,5)^2 + 0,55(3-2,5)^2 = \\ &= 0,1125 + 0,1000 + 0,1375 = 0,3500. \end{aligned}$$

Wartość odchylenia standardowego wynika z następującego rachunku:

$$D(Y) = \sqrt{0,3500} \approx 0,59. \blacksquare$$

Niektóre twierdzenia dotyczące parametrów zmiennej losowej:

1. Jeżeli A jest wielkością stałą, to wówczas jej wartość oczekiwana równa się tej stałej, tj. $E(A) = A$.

2. Wartość oczekiwana sumy dowolnej, skończonej liczby zmiennych losowych równa się sumie wartości oczekiwanych tych zmiennych:

$$E\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n E(X_i). \quad (1.53)$$

3. Wartość oczekiwana iloczynu dowolnej, skończonej liczby wzajemnie niezależnych zmiennych losowych równa się iloczynowi ich wartości oczekiwanych:

$$E\left(\prod_{i=1}^n X_i\right) = \prod_{i=1}^n E(X_i). \quad (1.54)$$

4. Jeżeli A jest wielkością stałą, to wtedy jej wariancja równa się zero, tj. $D^2(A) = 0$.

5. Wariancja sumy dowolnej, skończonej liczby niezależnych zmiennych losowych równa się sumie wariancji tych zmiennych:

$$D^2\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n D^2(X_i). \quad (1.55)$$

1.4. Wybrane rozkłady jednowymiarowych zmiennych losowych

1.4.1. Rozkłady dyskretnych zmiennych losowych

1.4.1.1. Rozkład dwupunktowy

Zmienna losowa X ma rozkład dwupunktowy, jeżeli przyjmuje tylko dwie wartości x_1 i x_2 , z następującymi prawdopodobieństwami:

$$P(X = x_1) = p, \quad P(X = x_2) = q, \quad \text{przy czym } p + q = 1. \quad (1.56)$$

Szczególnym przypadkiem takiego rozkładu jest rozkład dwupunktowy zmiennej zero-jedynkowej. Mamy wtedy:

$$P(X = 1) = p, \quad P(X = 0) = 1 - p = q. \quad (1.57)$$

Wartość oczekiwana i wariancja wynoszą tutaj odpowiednio:

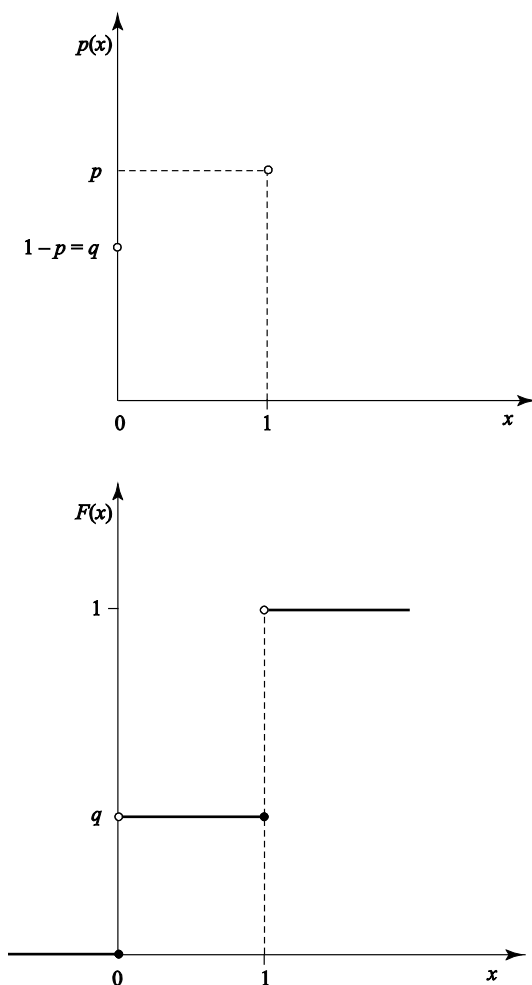
$$E(X) = p \cdot 1 + q \cdot 0 = p, \quad (1.58)$$

$$D^2(X) = p(1 - p)^2 + q(0 - p)^2 = pq^2 + qp^2 = pq(q + p) = pq. \quad (1.59)$$

Dystrybuanta tej zmiennej losowej wyrażona jest następująco:

$$F(x) = \begin{cases} 0 & \text{dla } x \leq 0, \\ q & \text{dla } 0 < x \leq 1, \\ 1 & \text{dla } x > 1. \end{cases} \quad (1.60)$$

Wykresy funkcji prawdopodobieństwa (1.57) oraz dystrybuanty (1.60) pokazano na rys. 1.4.



Rys. 1.4. Funkcja prawdopodobieństwa i dystrybuanta zero-jedynkowej zmiennej losowej

Zero-jedynkowa zmienna losowa ma jeden parametr, a mianowicie p . Parametr ten jednoznacznie określa szczegółową postać dystrybuanty. Fakt, że zmienna losowa X ma dwupunktowy (zero-jedynkowy) rozkład prawdopodobieństwa o parametrze p , będziemy zapisywać następująco: $X \sim D(p)$. Zero-jedynkowe zmienne losowe są szczególnie często wykorzystywane w statystycznej kontroli jakości. Przyjmuje się wówczas, że:

$$X = \begin{cases} 0 & \text{– gdy wyrób spełnia} \\ & \text{wymagania jakościowe} \\ 1 & \text{– gdy wyrób nie spełnia} \\ & \text{wymagań jakościowych.} \end{cases} \quad (1.61)$$

W takiej sytuacji prawdopodobieństwo $p = P(X = 1)$ nazywane jest wadliwością strumienia albo partii (zasobu) wyrobu. Jest to jedna z dwóch podstawowych miar jakości wykonania. Drugą miarą jest liczba wad w jednostce produktu.

1.4.1.2. Rozkład dwumianowy

Niech będzie dana zmienna losowa Z , taka że:

$$Z = X_1 + \dots + X_i + \dots + X_n = \sum_{i=1}^n X_i, \quad (1.62)$$

gdzie X_i ($i = 1, 2, \dots, n$) oznacza zero-jedynkową zmienną losową, przy czym:

$$\bigwedge_i E(X_i) = p_i = p.$$

Mamy więc do czynienia z sumą zero-jedynkowych zmiennych losowych, o takich samych rozkładach prawdopodobieństwa. Tak zdefiniowana zmienna losowa Z może przyjmować wartości: $0, 1, \dots, n$.

Zmienna losowa Z ma funkcję prawdopodobieństwa określoną wzorem:

$$P(Z = z; n, p) = \binom{n}{z} p^z q^{n-z}, \quad (1.63)$$

gdzie:

$$q = 1 - p; \quad 0 \leq p \leq 1; \quad z = 0, 1, \dots, n.$$

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

Obliczane prawdopodobieństwa są składnikami rozwinięcia dwumianu Newtona $(p + q)^n$. Dlatego też omawiany rozkład nazywany jest rozkładem dwumianowym (binomialnym). Fakt, że zmienna losowa Z ma rozkład binomialny o parametrach n i p , zapisujemy następująco: $Z \sim B(n, p)$.

Ponieważ (1.63) jest funkcją prawdopodobieństwa, zatem suma prawdopodobieństw obliczanych według tego wzoru równa jest jedności:

$$\sum_{z=0}^n P(Z = z) = \sum_{z=0}^n \binom{n}{z} p^z q^{n-z} = (p + q)^n = 1.$$

Dystrybuanta zmiennej losowej Z ma postać:

$$F_Z(x) = \begin{cases} 0 & \text{dla } x \leq 0, \\ \sum_{z < x} \binom{n}{z} p^z q^{n-z} & \text{dla } 0 < x \leq n, \\ 1 & \text{dla } x > n. \end{cases} \quad (1.64)$$

Wykresy funkcji prawdopodobieństwa oraz dystrybuanty zmiennej losowej Z przedstawiono na rys. 1.5. Zmienna losowa Z ma wartość oczekiwaną

$$E(Z) = np \quad (1.65)$$

oraz wariancję

$$D^2(Z) = npq. \quad (1.66)$$

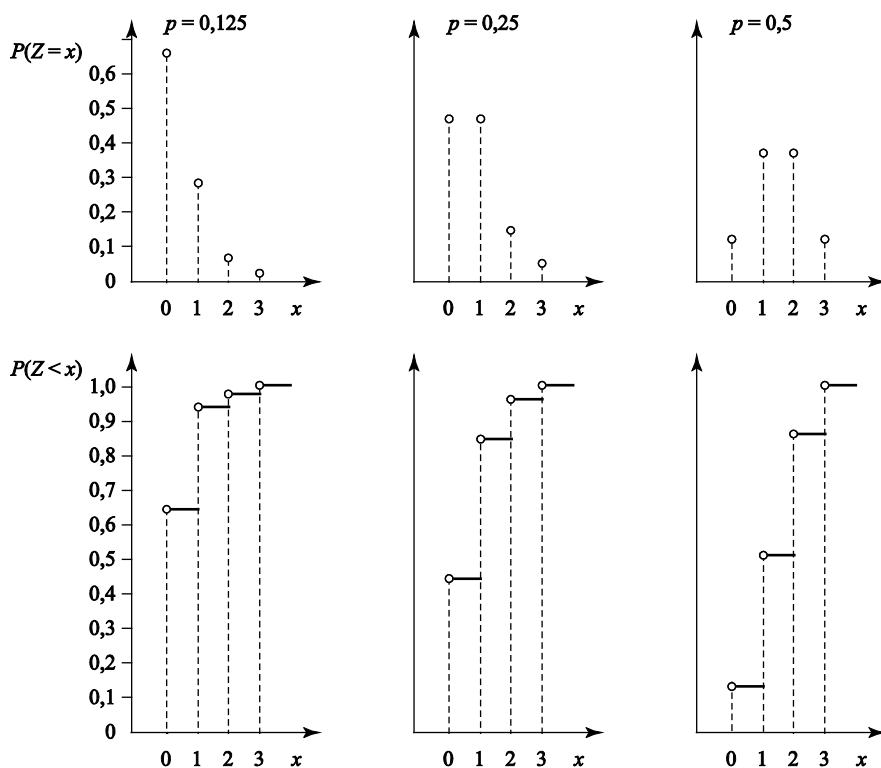
Obydwa wzory wynikają bezpośrednio z faktu, że zmienna losowa Z ma postać (1.62). Wykorzystując wzór (1.53) mamy:

$$E(Z) = E(X_1) + \dots + E(X_i) + \dots + E(X_n) = \sum_{i=1}^n E(X_i) = np.$$

Na mocy zależności (1.55) mamy natomiast:

$$D^2(Z) = D^2(X_1) + \dots + D^2(X_i) + \dots + D^2(X_n) = \sum_{i=1}^n D^2(X_i) = npq.$$

	$n = 3$					
	$p = 0,125$		$p = 0,25$		$p = 0,5$	
	$P(Z = x)$	$P(Z < x)$	$P(Z = x)$	$P(Z < x)$	$P(Z = x)$	$P(Z < x)$
$x < 0$	0	0	0	0	0	0
$x = 0$	0,6699	0	0,4219	0	0,1250	0
$0 < x < 1$	0	0,6699	0	0,4219	0	0,1250
$x = 1$	0,2871	0,6699	0,4219	0,4219	0,3750	0,1250
$1 < x < 2$	0	0,9570	0	0,8438	0	0,5000
$x = 2$	0,0410	0,9570	0,1406	0,8438	0,3750	0,5000
$2 < x < 3$	0	0,9980	0	0,9844	0	0,8750
$x = 3$	0,0020	0,9980	0,0156	0,9844	0,1250	0,8750
$x > 3$	0	1	0	1	0	1



Rys. 1.5. Funkcja prawdopodobieństwa i dystrybuanta zmiennej losowej o rozkładzie dwumianowym

Dwumianowy rozkład prawdopodobieństwa ma również zmienna losowa:

$$W = \frac{X_1}{n} + \dots + \frac{X_i}{n} + \dots + \frac{X_n}{n} = \frac{\sum_{i=1}^n X_i}{n} = \frac{Z}{n}, \quad (1.67)$$

o wartościach: $1/n, \dots, z/n, \dots, 1$.

Realizacje tej zmiennej losowej są frakcjami jedynek w n -elementowych ciągach zer i jedynek, generowanych przy stałym prawdopodobieństwie $p = P(X = 1)$. Funkcja prawdopodobieństwa i dystrybuanta zmiennej losowej W mają postać (1.63), (1.64). Jeśli wartości p i n są ustalone, to $P(Z = z; n, p) = P(W = z/n; n, p)$.

Wartość oczekiwana zmiennej losowej W wynika z odpowiedniego przekształcenia wzoru (1.65). Mamy mianowicie:

$$E(W) = \frac{E(Z)}{n} = \frac{np}{n} = p. \quad (1.68)$$

Przekształcając wyrażenie (1.66) otrzymujemy wzór na wariancję zmiennej losowej W :

$$D^2(W) = \frac{D^2(Z)}{n^2} = \frac{npq}{n^2} = \frac{pq}{n}. \quad (1.69)$$

Fakt, że zmienna losowa W ma rozkład dwumianowy o parametrach p i n , zapisujemy następująco: $W \sim B(n, p)$.

Przykład 1.5

Z partii towaru, o której wiadomo, że zawiera 2% sztuk wadliwych, pobrano do badania $n = 5$ sztuk, stosując przy tym losowanie niezależne. Jakie jest prawdopodobieństwo tego, że wśród badanych sztuk towaru znajdzie się jedna wadliwa? Można też sformułować inne – równoważne – pytanie. Jakie jest prawdopodobieństwo tego, że frakcja sztuk wadliwych wyniesie $1/5$?

$$\begin{aligned} P(Z = 1; p = 0,02; n = 5) &= P(W = 1/5; p = 0,02; n = 5) = \\ &= \binom{5}{1} \cdot 0,02 \cdot 0,98^4 = \frac{5!}{1!4!} \cdot 0,02 \cdot 0,98^4 = 5 \cdot 0,02 \cdot 0,98^4 = 0,092. \blacksquare \end{aligned}$$

1.4.1.3. Rozkład Poissona

Zmienna losowa Z ma rozkład Poissona⁷, jeśli może przyjmować wartości wyrażające się liczbami całkowitymi nieujemnymi ($z = 0, 1, 2, \dots$), z prawdopodobieństwami:

$$P(Z = z; \lambda) = \frac{\lambda^z}{z!} e^{-\lambda}, \quad (1.70)$$

gdzie $\lambda = np$ jest parametrem tego rozkładu.

Rozkład Poissona jest rozkładem granicznym dla rozkładu dwumianowego. Jeśli n rośnie nieograniczenie, a p zmienia się wraz z n w ten sposób, że iloczyn np jest stały, to dla każdego z zachodzi zbieżność rozkładu dwumianowego do rozkładu Poissona. Wzór (1.70) można otrzymać ze wzoru (1.63), dokonując podstawienia $np = \lambda$ i następnie przechodząc do granicy:

$$\begin{aligned} P(Z = z; n, p) &= \binom{n}{z} p^z q^{n-z} = \frac{n!}{z!(n-z)!} \left(\frac{\lambda}{n}\right)^z \left(1 - \frac{\lambda}{n}\right)^{n-z} = \\ &= \frac{\lambda^z}{n!} \cdot \frac{n(n-1)\dots(n-z+1)}{n^z} \cdot \frac{\left(1 - \frac{\lambda}{n}\right)^n}{\left(1 - \frac{\lambda}{n}\right)^z} = \\ &= \frac{\lambda^z}{z!} \cdot \left(1 - \frac{1}{n}\right) \left(1 - \frac{2}{n}\right) \dots \left(1 - \frac{z-1}{n}\right) \cdot \frac{\left(1 - \frac{\lambda}{n}\right)^n}{\left(1 - \frac{\lambda}{n}\right)^z}. \end{aligned}$$

Mamy więc:

$$\lim_{n \rightarrow \infty} \binom{n}{z} p^z q^{n-z} = \frac{\lambda^z}{z!} e^{-\lambda}.$$

⁷ Siméon Denis Poisson (1781–1840), francuski matematyk i fizyk.

Ponieważ (1.70) jest funkcją prawdopodobieństwa, zatem suma wszystkich prawdopodobieństw obliczanych według tego wzoru równa jest jedności. Mamy mianowicie:

$$\sum_{z=0}^{\infty} P(Z = z; \lambda) = e^{-\lambda} \sum_{z=0}^{\infty} \frac{\lambda^z}{z!} = e^{-\lambda} \left(\frac{\lambda^0}{0!} + \frac{\lambda^1}{1!} + \dots \right) = e^{-\lambda} \cdot e^{\lambda} = 1.$$

Dystrybuanta zmiennej losowej Z o rozkładzie Poissona przedstawia się następująco:

$$F_Z(x) = \begin{cases} 0 & \text{dla } x \leq 0, \\ \sum_{z < x} \frac{\lambda^z}{z!} e^{-\lambda} & \text{dla } x > 0. \end{cases} \quad (1.71)$$

Na rys. 1.6 i 1.7 przedstawiono funkcję prawdopodobieństwa i dystrybuantę zmiennej losowej Poissona dla dwóch wybranych wartości λ .

W przypadku zmiennej losowej Z o rozkładzie Poissona mamy:

$$E(Z) = D^2(Z) = \lambda. \quad (1.72)$$

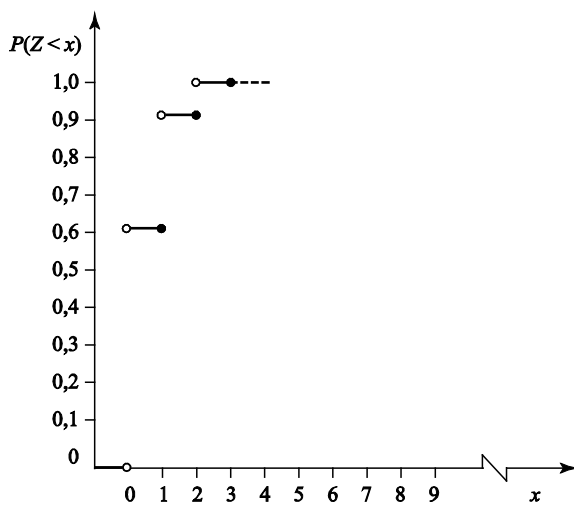
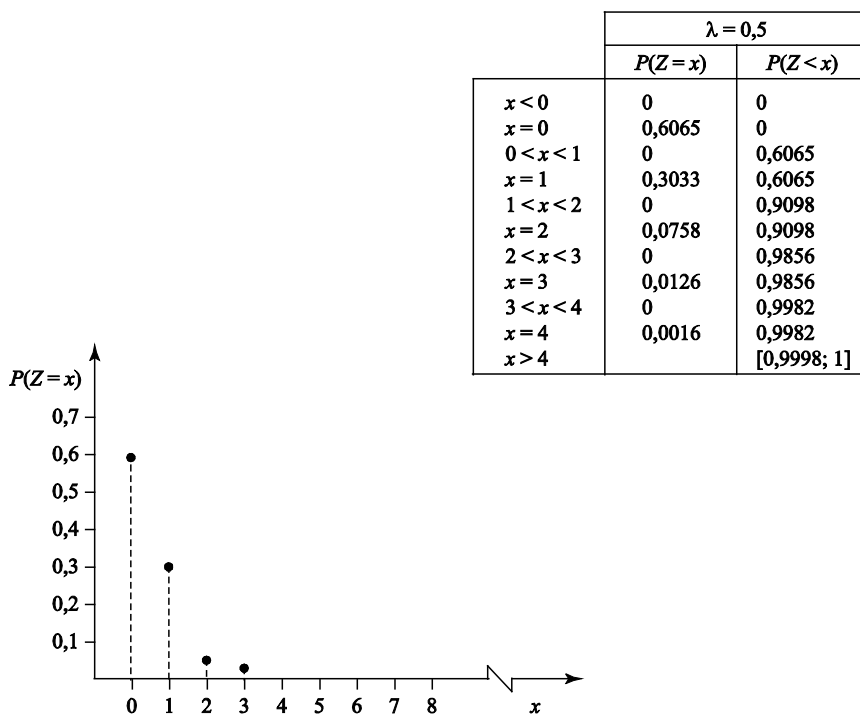
Fakt, że zmienna losowa Z ma rozkład Poissona o parametrze λ , będziemy zapisywać następująco: $Z \sim P(\lambda)$.

Rozkład Poissona jest często nazywany rozkładem zdarzeń rzadkich. Jest on wykorzystywany w statystycznej kontroli jakości, szczególnie wówczas, gdy kontroli podlega liczba wad w jednostce produktu. Rozkład Poissona wykorzystywany jest nie tylko wówczas, gdy jest on merytorycznie uzasadniony, ale również jako przybliżenie rozkładu dwumianowego. Przyjmuje się zwykle, że przybliżenie to jest wystarczająco dobre, gdy $n \geq 20$ i jednocześnie $p \leq 0,2$.

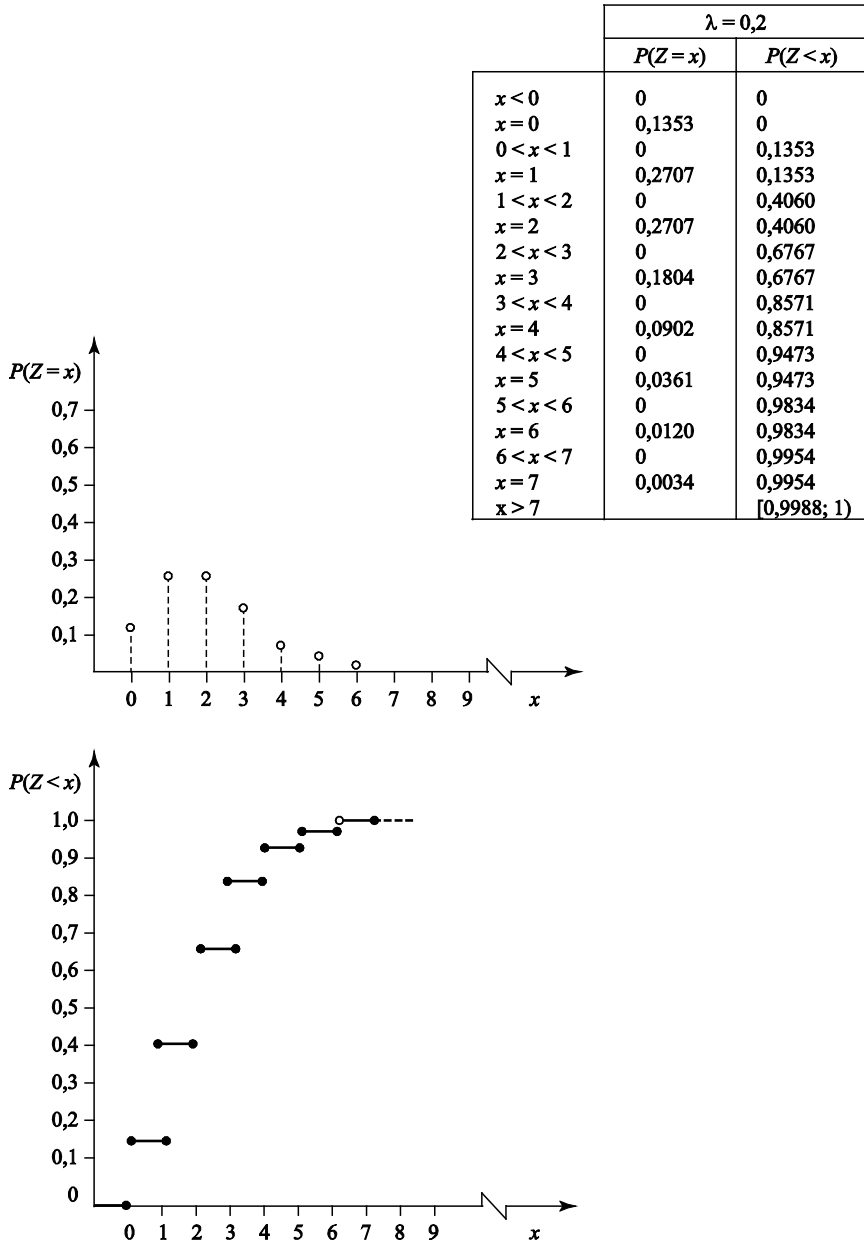
Przykład 1.6

Z partii pewnego produktu, o wadliwości $p = 0,05$ (5%), pobrano do badania $n = 40$ sztuk. Jakie jest prawdopodobieństwo tego, że wśród badanych sztuk znajdziemy co najwyżej jedną wadliwą?

Dokładną wartość tego prawdopodobieństwa można wyznaczyć z rozkładu dwumianowego. Mamy mianowicie:



Rys. 1.6. Funkcja prawdopodobieństwa i dystrybuanta zmiennej losowej o rozkładzie Poissona ($\lambda = 0,5$)



Rys. 1.7. Funkcja prawdopodobieństwa i dystrybuanta zmiennej losowej o rozkładzie Poissona ($\lambda = 2$)

$$\begin{aligned}
 F_Z(1) &= P(Z \leq 1) = P(Z = 0) + P(Z = 1) = \\
 &= \binom{40}{0} 0,05^0 \cdot 0,95^{40} + \binom{40}{1} 0,05^1 \cdot 0,95^{39} = \\
 &= 0,1285 + 40 \cdot 0,05 \cdot 0,1353 = 0,1285 + 0,2706 = 0,3991.
 \end{aligned}$$

Można również dokonać przybliżenia za pomocą rozkładu Poissona, albowiem $n > 20$ i $p < 0,2$. Obliczenia przebiegają następująco:

$$\lambda = np = 40 \cdot 0,05 = 2.$$

$$\begin{aligned}
 F_Z(1) &= P(Z \leq 1) = P(Z = 0) + P(Z = 1) = \\
 &= \frac{2^0}{0!} e^{-2} + \frac{2^1}{1!} e^{-2} = 3 \cdot e^{-2} = 0,4060.
 \end{aligned}$$

Błąd bezwzględny uzyskanego przybliżenia wynosi $0,4060 - 0,3991 = 0,0069$. Obliczenia z wykorzystaniem rozkładu Poissona są znacznie mniej pracochłonne, niż obliczenia oparte na rozkładzie dwumianowym. ■

1.4.1.4. Rozkład hipergeometryczny

Rozważmy następujące doświadczenie losowe:

- niech X oznacza zero-jedynkową zmienną losową,
- urna zawiera N kul, z których M należy do kategorii $X = 1$, natomiast pozostałe (czyli $N - M$) do kategorii $X = 0$,
- z urny losujemy kolejno bez zwracania n kul,
- obserwujemy sumę:

$$z = \sum_{i=1}^n x_i,$$

gdzie x_i oznacza realizację zero-jedynkowej zmiennej losowej X w i -tej kuli.

Otrzymane w taki sposób wartości z są realizacjami zmiennej losowej Z o hipergeometrycznym rozkładzie prawdopodobieństwa.

Funkcja prawdopodobieństwa tej zmiennej ma postać:

$$P(Z = z; N, M, n) = \frac{\binom{M}{z} \binom{N-M}{n-z}}{\binom{N}{n}}, \quad (1.73)$$

gdzie $z = 0, 1, 2, \dots, z'$, przy czym $z' = n$, gdy $n < M$, lub $z' = M$, gdy $n \geq M$.

Ponieważ $p = M/N$, wzór (1.73) można również zapisać następująco:

$$P(Z = z; N, p, n) = \frac{\binom{Np}{z} \binom{Nq}{n-z}}{\binom{N}{n}}. \quad (1.73a)$$

Suma prawdopodobieństw obliczanych według tych wzorów równa jest jedności:

$$\begin{aligned} \sum_{z=0}^n P(Z = z) &= \sum_{z=0}^n \frac{\binom{Np}{z} \binom{Nq}{n-z}}{\binom{N}{n}} = \\ &= \binom{N}{n}^{-1} \sum_{z=0}^n \binom{Np}{z} \binom{Nq}{n-z} = \binom{N}{n}^{-1} \binom{N}{n} = 1. \end{aligned}$$

Dystrybuanta zmiennej losowej Z przedstawia się następująco:

$$F_Z(x) = \begin{cases} 0 & \text{dla } x \leq 0, \\ \sum_{z < x} \frac{\binom{Np}{z} \binom{Nq}{n-z}}{\binom{N}{n}} & \text{dla } 0 < x \leq z', \\ 1 & \text{dla } x > z'. \end{cases} \quad (1.74)$$

Wartość oczekiwana zmiennej losowej $Z \sim H(N, M, n)$ – czyli zmiennej losowej o rozkładzie hipergeometrycznym z parametrami N, M, n – ma postać:

$$E(Z) = \frac{nM}{N} = np, \quad (1.75)$$

gdzie p jest prawdopodobieństwem zdarzenia losowego $X = 1$ podczas pierwszego spośród n losowań. W kolejnych losowaniach wartość p się zmienia. Jest ona zależna od wyników poprzednich losowań.

Wariancja zmiennej losowej $Z \sim H(N, M, n)$ przedstawia się następująco:

$$D^2(Z) = \frac{N-n}{N-1} \cdot n \cdot \frac{M}{N} \cdot \frac{M-N}{N} = \frac{N-n}{N-1} npq. \quad (1.76)$$

Rozkład hipergeometryczny może być aproksymowany za pomocą rozkładu dwumianowego. Przybliżenie to jest tym lepsze, im mniejsza jest wartość ilorazu n/N .

Przykład 1.7

Partia pewnego towaru składa się z $N = 50$ sztuk, wśród których znajduje się $M = 5$ sztuk wadliwych. Jakie jest prawdopodobieństwo tego, że wśród pobranych do badania $n = 3$ sztuk znajdzie się jedna wadliwa?

$$\begin{aligned} P(Z=1) &= \frac{\binom{5}{1} \binom{45}{2}}{\binom{50}{3}} = \frac{5! \cdot 45!}{1!4! \cdot 2!43!} = \frac{5!45!3!47!}{1!4!2!43!50!} = \\ &= \frac{3 \cdot 5 \cdot 44 \cdot 45}{48 \cdot 49 \cdot 50} = 0,253. \blacksquare \end{aligned}$$

1.4.2. Rozkłady ciągłych zmiennych losowych

1.4.2.1. Rozkład jednostajny

Zmienna losowa X ma rozkład jednostajny (równomierny, prostokątny) w przedziale $[a, b]$, jeśli jej funkcja gęstości prawdopodobieństwa określona jest wzorem:

$$f(x) = \begin{cases} 0 & \text{dla } x < a, \\ \frac{1}{b-a} & \text{dla } a \leq x \leq b, \\ 0 & \text{dla } x > b. \end{cases} \quad (1.77)$$

Wykres tej funkcji pokazano na rys. 1.8A. Z rysunku tego wynika bezpośrednio postać funkcji gęstości. Zgodnie ze wzorem (1.35) pole prostokąta na rys. 1.8 A równe jest jedności. Mamy więc:

$$f(x) \cdot (b - a) = 1,$$

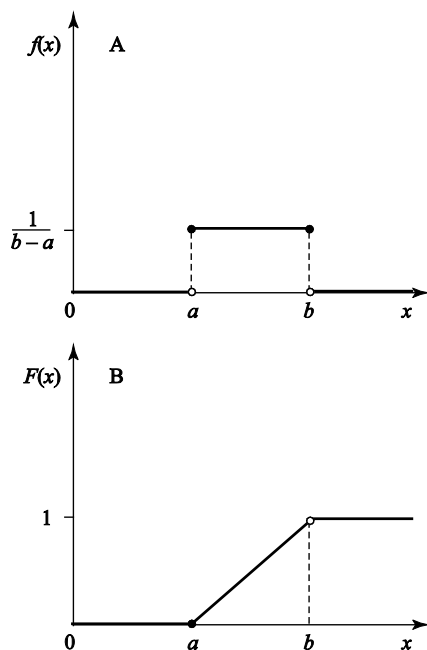
a stąd

$$f(x) = \frac{1}{b - a}.$$

Dystrybuanta zmiennej losowej dana jest wzorem:

$$F(x) = \begin{cases} 0 & \text{dla } x \leq a, \\ \frac{x - a}{b - a} & \text{dla } a < x \leq b, \\ 1 & \text{dla } x > b. \end{cases} \quad (1.78)$$

Wykres tej funkcji pokazano na rys. 1.8 B.



Rys. 1.8. Funkcja gęstości prawdopodobieństwa i dystrybuanta zmiennej losowej o rozkładzie jednostajnym

Wartość oczekiwaną i wariancję zmiennej losowej wyznacza się następująco:

$$E(X) = \int_a^b x \cdot \frac{1}{b-a} dx = \frac{1}{b-a} \int_a^b x dx = \frac{b^2 - a^2}{2(b-a)} = \frac{a+b}{2}. \quad (1.79)$$

$$\begin{aligned} D^2(X) &= E(X)^2 - [E(X)]^2 = \int_a^b x^2 \cdot \frac{1}{b-a} dx - \frac{(b+a)^2}{4} = \\ &= \frac{1}{b-a} \int_a^b x^2 dx - \frac{(b+a)^2}{4} = \frac{b^3 - a^3}{3(b-a)} - \frac{(b+a)^2}{4} = \frac{(b-a)^2}{12}. \end{aligned} \quad (1.80)$$

Parametrami rozkładu jednostajnego są końce przedziału, na którym ten rozkład jest rozpięty, czyli wartości a i b . Fakt, że zmienna losowa X ma rozkład jednostajny na przedziale $[a, b]$, będziemy zapisywać następująco: $X \sim R(a, b)$.

1.4.2.2. Rozkład normalny

Zmienna losowa X ma rozkład normalny⁸, jeśli jej funkcja gęstości prawdopodobieństwa ma postać:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right], \quad (1.81)$$

przy czym $x \in \mathbf{R}$, $\mu \in \mathbf{R}$, natomiast $\sigma \in \mathbf{R}^+$. Parametrami rozkładu normalnego są μ i σ . Fakt, że zmienna losowa X ma rozkład normalny o parametrach μ i σ , będziemy zapisywać następująco: $X \sim N(\mu, \sigma)$. Zachodzą przy tym następujące zależności:

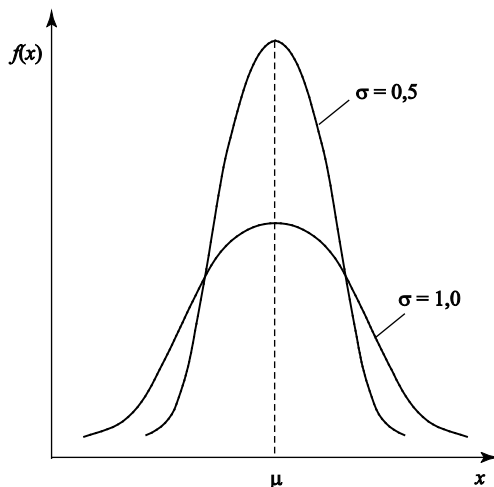
$$E(X) = \mu, \quad (1.82)$$

$$D^2(X) = \sigma^2. \quad (1.83)$$

Parametry μ i σ są niezależne. Parametr μ określa położenie funkcji (1.81) na osi liczb rzeczywistych w tym sensie, że funkcja ta osiąga maksimum w punkcie $x = \mu$. Parametr σ określa kształt „krzywej dzwonowej”, będącej wykresem funk-

⁸ Podstawy teoretyczne rozkładu normalnego są dziełem A. de Moivre'a (1667–1754), P.S. de Laplace'a (1749–1827) i C.F. Gaussa (1777–1855).

cji (1.81). Im większa jest wartość σ , tym bardziej spłaszczona jest owa krzywa. Na rys. 1.9 przedstawiono wykresy funkcji (1.81) dla dwóch różnych wartości σ .



Rys. 1.9. Funkcje gęstości prawdopodobieństwa dwóch normalnych zmiennych losowych

Dystrybuanta zmiennej losowej $X \sim N(\mu, \sigma)$ określona jest wzorem:

$$F(x) = P(X < x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^x \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] dx. \quad (1.84)$$

Rys. 1.10 ilustruje podstawowe własności funkcji gęstości prawdopodobieństwa oraz dystrybuanty zmiennej losowej $X \sim N(\mu, \sigma)$, a także powiązania między tymi charakterystykami rozkładu.

Na uwagę zasługują w szczególności następujące własności rozkładu normalnego:

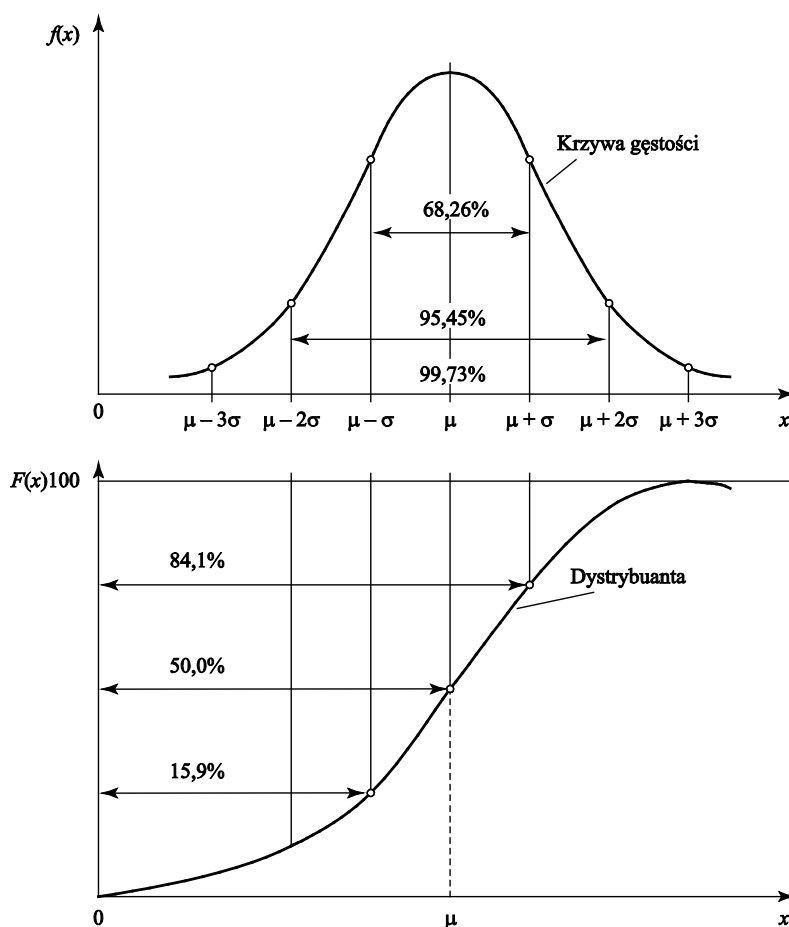
$$P(\mu - \sigma < X < \mu + \sigma) = 0,6826, \quad (1.85)$$

$$P(\mu - 2\sigma < X < \mu + 2\sigma) = 0,9545, \quad (1.86)$$

$$P(\mu - 3\sigma < X < \mu + 3\sigma) = 0,9973. \quad (1.87)$$

Wynika stąd bezpośrednio, że:

$$P(|X - \mu| > \sigma) = 0,3174, \quad (1.88)$$



Rys. 1.10. Funkcja gęstości prawdopodobieństwa i dystrybuanta normalnej zmiennej losowej

$$P(|X - \mu| > 2\sigma) = 0,0455, \quad (1.89)$$

$$P(|X - \mu| > 3\sigma) = 0,0027. \quad (1.90)$$

Relacje wyrażone wzorami (1.87) i (1.90) określane są niekiedy jako tzw. reguła trzech sigm. Reguła ta wyraża jedną z podstawowych własności rozkładu normalnego, polegającą na tym, że niemal wszystkie możliwe realizacje zmiennej losowej $X \sim N(\mu, \sigma)$ należą do przedziału $(\mu - 3\sigma; \mu + 3\sigma)$, mimo że dziedziną funkcji (1.81) jest cały zbiór liczb rzeczywistych \mathbf{R} .

Zarówno w rozważaniach teoretycznych, jak i w zastosowaniach podstawową rolę odgrywa tzw. standaryzowana (albo znormalizowana lub unormowana) normalna zmienna losowa:

$$U = \frac{X - E(X)}{D(X)} \quad (1.91)$$

o wartościach

$$u = \frac{x - \mu}{\sigma}. \quad (1.92)$$

Dokonując standaryzacji, transformujemy dowolną zmienną losową $X \sim N(\mu, \sigma)$ do postaci $U \sim N(0; 1)$. Tak więc:

$$E(U) = 0, \quad (1.93)$$

$$D^2(U) = 1. \quad (1.94)$$

Funkcja gęstości prawdopodobieństwa zmiennej losowej U ma postać:

$$\varphi(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right). \quad (1.95)$$

przy czym $u \in \mathbf{R}$.

Dystrybuanta zmiennej losowej U przedstawia się następująco:

$$\Phi(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^u \exp\left(-\frac{u^2}{2}\right) du. \quad (1.96)$$

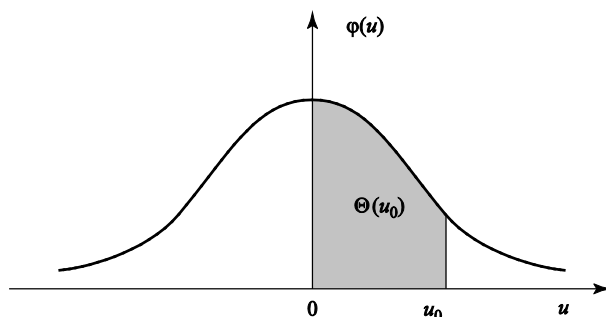
W zbiorach tablic statystycznych, a także w podręcznikach statystyki, publikowane są tablice wartości tej całki; zwykle w przedziale od $u \approx -3$ do $u \approx 3$. Dysponując takimi tablicami nie musimy obliczać wartości całki (1.84) przy rozwiązywaniu problemów rachunkowych związanych z rozkładem normalnym, albowiem dla każdego $x \in \mathbf{R}$ zachodzi równość:

$$F(x) = \Phi\left(\frac{x - \mu}{\sigma}\right) = \Phi(u). \quad (1.97)$$

Ponieważ funkcja gęstości prawdopodobieństwa $\varphi(u)$ jest funkcją parzystą [$\varphi(u) = \varphi(-u)$], zatem dla celów aplikacyjnych dystrybuanta $\Phi(u)$ może być zastąpiona funkcją Laplace'a postaci:

$$\Theta(u) = \frac{1}{\sqrt{2\pi}} \int_0^u \exp\left(-\frac{u^2}{2}\right) du. \quad (1.98)$$

Graficzną ilustracją całki (1.98) jest rys. 1.11, natomiast jej wartości dla $0 \leq u \leq 3,09$ zestawione są w tabelicy I, zamieszczonej w aneksie.



Rys. 1.11. Geometryczna interpretacja całki Laplace'a

Między dystrybuantą $\Phi(u)$ a funkcją $\Theta(u)$ zachodzą następujące związki:

– jeśli wartość u obliczona według wzoru (1.92) jest dodatnia (a więc jeśli $x \geq \mu$), to:

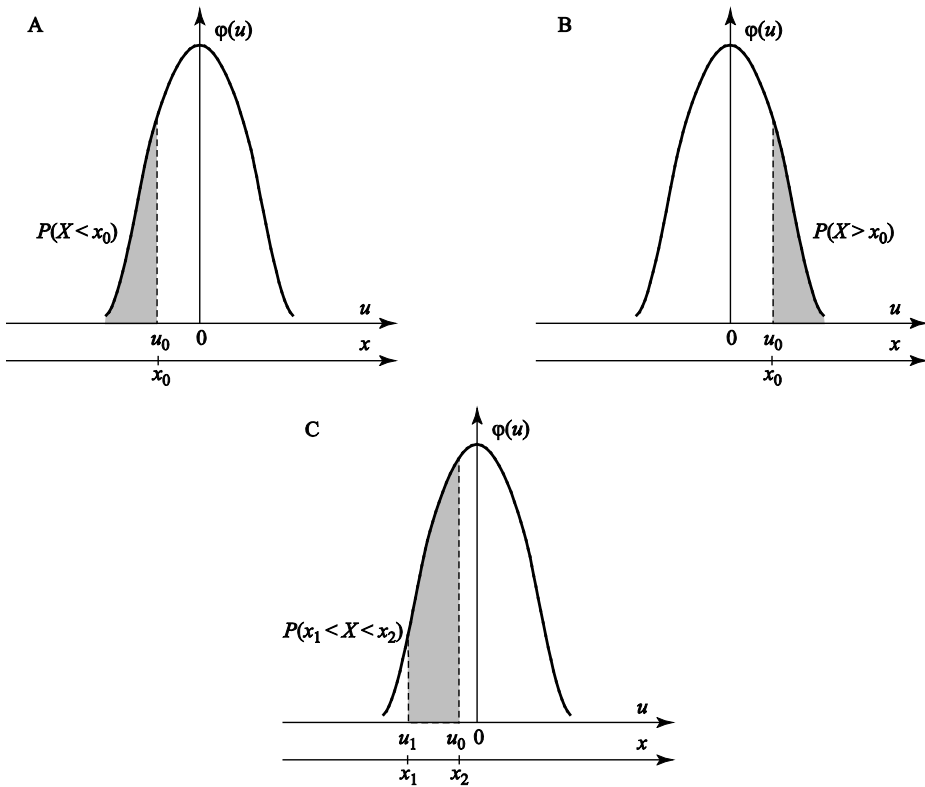
$$\Phi(u) = 0,5 + \Theta(u), \quad (1.99)$$

– jeśli natomiast wartość u otrzymana ze wzoru (1.92) jest ujemna (a więc jeśli $x < \mu$), to:

$$\Phi(u) = 0,5 - \Theta(-u). \quad (1.100)$$

W przypadku gdy $u = 0$, a więc gdy $x = \mu$, mamy $\Phi(0) = 0,5$, bowiem – zgodnie ze wzorem (1.98) – $\Theta(0) = 0$.

Mając do dyspozycji tablicę wartości całki $\Theta(u)$ oraz wzory (1.97), (1.99) i (1.100), można rozwiązać podstawowe zadania związane z rozkładem normalnym. Poniżej rozważymy trzy problemy, najczęściej występujące w zastosowaniach. Graficzną ilustracją tych rozważań jest rys. 1.12.



Rys. 1.12. Obliczanie prawdopodobieństw w rozkładzie normalnym (interpretacja geometryczna)

A. Prawdopodobieństwo zdarzenia losowego A , polegającego na tym, że zmienna losowa $X \sim N(\mu, \sigma)$ przyjmie wartość mniejszą niż x_0 . Wykorzystując definicję dystrybuanty oraz zależność (1.97) możemy napisać:

$$P(A) = P(X < x_0) = F(x_0) = \Phi\left(\frac{x_0 - \mu}{\sigma}\right) = \Phi(u_0). \quad (1.101)$$

Jeśli $u_0 > 0$, to w dalszym postępowaniu należy zastosować wzór (1.99), jeśli natomiast $u_0 < 0$, to właściwy jest wzór (1.100). W przypadku gdy $u_0 = 0$, mamy $P(A) = 0,5$.

B. Prawdopodobieństwo zdarzenia losowego B , polegającego na tym, że zmienna losowa $X \sim N(\mu, \sigma)$ przyjmie wartość większą niż x_0 :

$$P(B) = P(X > x_0) = 1 - F(x_0) = 1 - \Phi\left(\frac{x_0 - \mu}{\sigma}\right) = 1 - \Phi(u_0). \quad (1.102)$$

Dalsze obliczenia przebiegają analogicznie jak powyżej.

C. Prawdopodobieństwo zdarzenia losowego C , polegającego na tym, że zmienna losowa $X \sim N(\mu, \sigma)$ przyjmie wartość z przedziału (x_1, x_2) :

$$\begin{aligned} P(C) &= P(x_1 < X < x_2) = F(x_2) - F(x_1) = \\ &= \Phi\left(\frac{x_2 - \mu}{\sigma}\right) - \Phi\left(\frac{x_1 - \mu}{\sigma}\right) = \Phi(u_2) - \Phi(u_1). \end{aligned} \quad (1.103)$$

W dalszych obliczeniach wykorzystuje się – analogicznie jak poprzednio – wzory (1.101), (1.102), w zależności od znaków wartości u_1 i u_2 .

W przedstawionym poniżej przykładzie rozważmy nieco bardziej złożony problem.

Przykład 1.8

Pewien sypki produkt spożywczy paczkowany jest automatycznie w postaci opakowań jednostkowych o nominalnej masie netto $\mu_0 = 1$ kg. Rzeczywista zawartość produktu w opakowaniach jednostkowych jest zmienną losową $X \sim N(\mu_r; 0,05)$, gdzie μ_r jest wartością zadawaną przez operatora urządzenia paczkującego. Z własności rozkładu normalnego wynika, że jeśli $\mu_r = \mu_0$, to 68,26% wszystkich opakowań mieści się – ze względu na masę netto – w przedziale $(0,95; 1,05)$, 95,45% mieści się w przedziale $(0,90; 1,10)$, natomiast 99,73% mieści się w przedziale $(0,85; 1,15)$; zob. wzory (1.85)–(1.87). Na jakim poziomie należy ustalić μ_r , aby tylko 5% opakowań zawierało mniej niż 1 kg paczkowanego produktu?

Żądamy więc, by $P(X < 1) = 0,05$. Wykorzystując poznane zależności możemy napisać:

$$F(1) = \Phi\left(\frac{1 - \mu_r}{0,05}\right) = 0,05.$$

Ponieważ $\mu_r > 1$, mamy więc:

$$F(1) = \Phi\left(\frac{1 - \mu_r}{0,05}\right) = 0,05 = 0,5 - \Phi\left(-\frac{1 - \mu_r}{0,05}\right).$$

Wynika stąd, że:

$$0,05 = 0,5 - \Theta\left(-\frac{1 - \mu_r}{0,05}\right),$$

$$\Theta\left(-\frac{1 - \mu_r}{0,05}\right) = 0,45.$$

Na podstawie tablicy I, zamieszczonej w aneksie, mamy $\Theta(1,645) = 0,45$, a stąd:

$$-\frac{1 - \mu_r}{0,05} = 1,645$$

i w konsekwencji

$$\mu_r \approx 1,0823.$$

Tak więc jeśli $\mu_r \approx 1,0823$, to w dużych partiach produktu (liczących co najmniej kilka tysięcy sztuk) tylko 5% opakowań będzie zawierać mniej niż 1 kg produktu. ■

1.4.2.3. Rozkład wykładniczy

Zmienna losowa T podlega wykładniczemu rozkładowi prawdopodobieństwa, jeśli jej gęstość prawdopodobieństwa wyraża się wzorem:

$$f(t) = \begin{cases} 0 & \text{dla } t < 0, \\ \lambda e^{-\lambda t} & \text{dla } t \geq 0. \end{cases} \quad (1.104)$$

Parametr λ jest powiązany z wartością oczekiwaną i wariancją zmiennej losowej T , następującymi zależnościami:

$$E(T) = \lambda^{-1}, \quad (1.105)$$

$$D^2(T) = \lambda^{-2}. \quad (1.106)$$

Wynika stąd bezpośrednio, że dla dowolnej wartości λ współczynnik zmienności zmiennej losowej T równy jest jedności. Mamy bowiem:

$$v(T) = \frac{E(T)}{D(T)} = \frac{\lambda^{-1}}{\sqrt{\lambda^{-2}}} = -\frac{\sqrt{\lambda^2}}{\lambda} = 1. \quad (1.107)$$

Dystrybuanta zmiennej losowej T wyraża się wzorem:

$$F(t) = \begin{cases} 0 & \text{dla } t \leq 0, \\ 1 - e^{-\lambda t} & \text{dla } t > 0. \end{cases} \quad (1.108)$$

Na rys. 1.13 pokazano wykresy funkcji gęstości prawdopodobieństwa i dystrybuanty dla dwóch wybranych wartości λ .

Jednym z podstawowych zastosowań rozkładu wykładniczego jest ocena niezawodności różnego rodzaju obiektów technicznych. Funkcja niezawodności $R(t)$ wyraża prawdopodobieństwo zdarzenia losowego polegającego na tym, że czas poprawnej pracy obiektu T nie będzie krótszy niż pewna wyróżniona wartość. Mamy więc:

$$R(t) = P(T \geq t). \quad (1.109)$$

Jak łatwo zauważyć:

$$R(t) = 1 - P(T < t) = 1 - F(t), \quad (1.110)$$

albowiem zdarzenia losowe $T \geq t$ i $T < t$ tworzą zupełny układ zdarzeń.

Jeśli zmienna losowa T ma wykładniczy rozkład prawdopodobieństwa, to wówczas:

$$R(t) = 1 - (1 - e^{-\lambda t}) = e^{-\lambda t}. \quad (1.111)$$

W teorii niezawodności często stosuje się dwuparametrowy rozkład wykładniczy, nazywany też rozkładem wykładniczym z progiem czułości, albo z przesunięciem.

Gęstość prawdopodobieństwa zmiennej losowej T o dwuparametrowym, wykładniczym rozkładzie prawdopodobieństwa wyraża się wzorem:

$$f(t) = \begin{cases} 0 & \text{dla } t < t_0, \\ \lambda e^{-\lambda(t-t_0)} & \text{dla } t \geq t_0. \end{cases} \quad (1.112)$$

Parametr t_0 jest taką wartością na osi czasu pracy obiektu T , dla której:

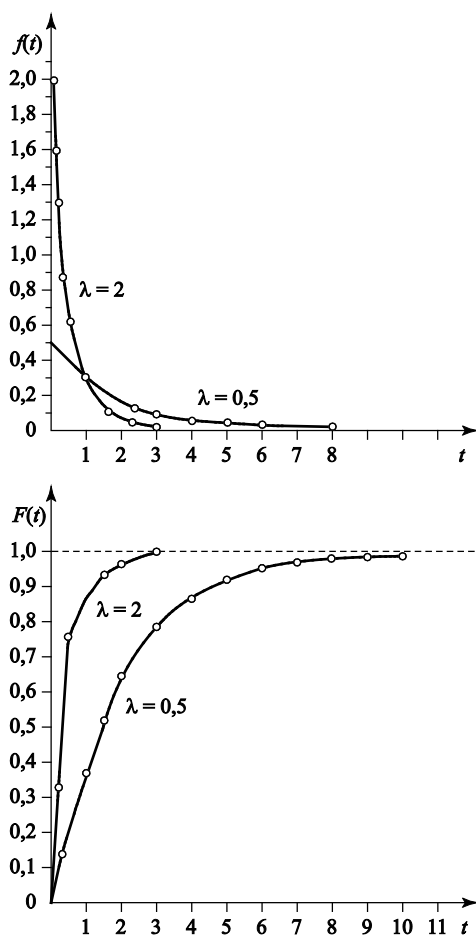
$$P(T \leq t_0) = 0. \quad (1.113)$$

$\lambda = 0,5$

t	0	0,5	1	2	3	4	5	6	7	8	9	10
$f(t)$	0,500	0,389	0,303	0,184	0,112	0,068	0,041	0,025	0,015	0,009	0,006	0,003
$F(t)$	0	0,221	0,393	0,632	0,777	0,865	0,918	0,950	0,970	0,982	0,989	0,993

 $\lambda = 2$

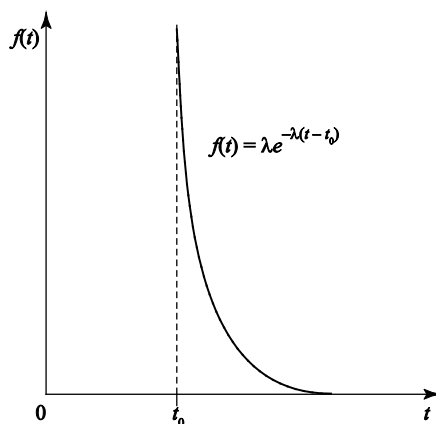
t	0	0,1	0,2	0,4	0,6	0,8	1	1,2	1,4	1,6	1,8	2	3
$f(t)$	2	1,637	1,341	0,899	0,602	0,404	0,271	0,181	0,122	0,082	0,055	0,037	0,005
$F(t)$	0	0,181	0,330	0,551	0,699	0,798	0,865	0,908	0,939	0,959	0,973	0,982	0,998



Rys. 1.13. Funkcje gęstości prawdopodobieństwa i dystrybuanty dwóch zmiennych losowych o rozkładzie wykładniczym

Graficzną ilustrację przesunięcia rozkładu wykładniczego przedstawiono na rys. 1.14. Rozkład ten znajduje zastosowanie wówczas, gdy w czasie pracy obiektu można wyróżnić dwa etapy, a mianowicie:

- etap $(0, t_0]$, w którym badany obiekt jest praktycznie niewrażliwy na działanie narażeń eksploatacyjnych,
- etap $(t_0, t_u]$, w którym narażenia eksploatacyjne doprowadzają do uszkodzenia obiektu.



Rys. 1.14. Funkcja gęstości prawdopodobieństwa zmiennej losowej o rozkładzie wykładniczym z przesunięciem

Dystrybuanta zmiennej losowej T o dwuparametrowym rozkładzie wykładniczym wyraża się wzorem:

$$F(t) = \begin{cases} 0 & \text{dla } t \leq t_0, \\ 1 - e^{-\lambda(t-t_0)} & \text{dla } t > t_0. \end{cases} \quad (1.114)$$

W konsekwencji funkcja niezawodności przybiera postać:

$$R(t) = \begin{cases} 1 & \text{dla } t \leq t_0, \\ e^{-\lambda(t-t_0)} & \text{dla } t > t_0. \end{cases} \quad (1.115)$$

Wynika stąd bezpośrednio, że przedział $(0, t_0]$ może być interpretowany jako okres gwarancyjny dla obiektu, albowiem w przedziale tym poprawna praca obiektu jest zdarzeniem praktycznie pewnym.

Przykład 1.9

Na podstawie długotrwałych obserwacji ustalono, że przeciętny czas świecenia żarówki pewnego typu wynosi 800 h. Jakie jest prawdopodobieństwo zdarzenia losowego polegającego na tym, że losowo wybrana żarówka będzie świecić co najmniej 600 h?

Zakładamy, że czas świecenia żarówki T jest zmienną losową o jednoparametrowym, wykładniczym rozkładzie prawdopodobieństwa. Wykorzystując zależność (1.111) mamy:

$$R(600) = \exp\left(-\frac{1}{800} \cdot 600\right) = \exp(-0,75) = 0,473.$$

Tak więc prawdopodobieństwo tego, że żarówka będzie świecić co najmniej 600 h wynosi 0,473. ■

1.5. Dwuwymiarowe zmienne losowe

Wykorzystując przestrzeń probabilistyczną (Ω, \mathbf{B}, P) można określić układ m funkcji postaci (1.28). Układ taki przyporządkowuje każdemu zdarzeniu losowemu, wyróżnionemu w przestrzeni Ω , m liczb rzeczywistych. W rezultacie otrzymuje się m -wymiarową zmienną losową:

$$(X_1, \dots, X_r, \dots, X_m) \tag{1.116}$$

o wartościach w zbiorze \mathbf{R}^m . Realizacją takiej zmiennej losowej jest wektor postaci:

$$[x_1, \dots, x_r, \dots, x_m], \tag{1.117}$$

czyli punkt w m -wymiarowej przestrzeni.

Dalsze rozważania ograniczymy do zmiennych dwuwymiarowych ($m = 2$). Zmienne te będziemy oznaczać symbolami (X_1, X_2) , (X, Y) itp.

1.5.1. Funkcja prawdopodobieństwa, funkcja gęstości prawdopodobieństwa i dystrybuanta dwuwymiarowej zmienniej losowej

Funkcja prawdopodobieństwa skokowej, dwuwymiarowej zmienniej losowej (X, Y) wyraża się wzorem:

$$P(X = x_i, Y = y_j) = p_{ij}, \quad (1.118)$$

$$i = 1, 2, 3, \dots,$$

$$j = 1, 2, 3, \dots,$$

przy czym:

$$\sum_i \sum_j p_{ij} = 1. \quad (1.119)$$

Zbiór wartości (x_i, y_j) skokowej zmienniej losowej (X, Y) jest skończony lub co najwyżej przeliczalny.

W przypadku ciągłej zmienniej losowej (X, Y) zbiór wartości jest nieprzeliczalny. W konsekwencji, analogicznie jak w przypadku jednowymiarowej zmienniej losowej, funkcja prawdopodobieństwa powinna być zastąpiona funkcją gęstości prawdopodobieństwa:

$$f(x, y) = \lim_{\substack{\Delta x \rightarrow 0 \\ \Delta y \rightarrow 0}} \frac{P(x \leq X < x + \Delta x, y \leq Y < y + \Delta y)}{\Delta x \Delta y}, \quad (1.120)$$

przy czym:

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy = 1, \quad (1.121)$$

gdzie:

$$f(x, y) \geq 0.$$

Podane powyżej wzory (1.118)–(1.121) są odpowiednikami wzorów (1.30), (1.31) i (1.34), (1.35), dotyczących jednowymiarowych zmiennych losowych.

Dystrybuanta dwuwymiarowej zmienniej losowej (X, Y) definiowana jest jednym z następujących wzorów:

$$F(x, y) = P(X < x, Y < y) \quad (1.122)$$

lub

$$P(x, y) = P(X \leq x, Y \leq y). \quad (1.123)$$

Są one odpowiednikami wzorów (1.36) i (1.37), dotyczących jednowymiarowych zmiennych losowych. W przypadku ciągłych zmiennych losowych, wobec nieprzeliczalności zbiorów ich wartości, obydwie konwencje są równoważne w sensie wartości prawdopodobieństw. Przejście od konwencji (1.22) do (1.23) oznacza tylko zmianę charakteru ciągłości dystrybuanty względem każdego z argumentów (x, y) , z przynajmniej lewostronnej na przynajmniej prawostronną. W przypadku dyskretnych zmiennych losowych zmiana konwencji prowadzi do zmiany wartości odpowiednich prawdopodobieństw. Konwencja (1.123) jest niekiedy wygodniejsza w zastosowaniach.

W odniesieniu do dyskretnych, dwuwymiarowych zmiennych losowych (X, Y) wzór (1.22) przybiera następującą postać:

$$F(x, y) = \sum_{x_i < x} \sum_{y_j < y} p_{ij}. \quad (1.124)$$

W przypadku ciągłych zmiennych losowych (X, Y) sumowanie wartości funkcji prawdopodobieństwa należy zastąpić całkowaniem funkcji gęstości prawdopodobieństwa:

$$F(x, y) = \int_{-\infty}^x \int_{-\infty}^y f(x, y) dx dy. \quad (1.125)$$

1.5.2. Rozkłady brzegowe i rozkłady warunkowe

W tablicy pokazanej na rys. 1.15 przedstawiono rozkład prawdopodobieństwa dwuwymiarowej, skokowej zmiennej losowej (X, Y) . Sumując wartości p_{ij} w poszczególnych kolumnach tablicy, otrzymujemy wartości:

$$p_{.j} = \sum_i p_{ij}, \quad (1.126)$$

będące prawdopodobieństwami odpowiednich wartości zmiennej losowej Y : $y_1, \dots, y_j, \dots, y_m$. Mamy więc następującą funkcję prawdopodobieństwa:

$$P(Y = y_j) = p_{.j}, \quad (1.127)$$

$$j = 1, 2, \dots$$

Analogicznie, sumując wartości p_{ij} w kolejnych wierszach tablicy otrzymujemy:

$$P_{i\cdot} = \sum_j P_{ij}, \quad (1.128)$$

$$P(X = x_i) = p_{i\cdot}. \quad (1.129)$$

		Y					Rozkład brzegowy X	
		y ₁	y ₂	...	y _j	...		y _m
X	x ₁	P ₁₁	P ₁₂	...	P _{1j}	...	P _{1m}	$P_{1\cdot} = \sum_{j=1}^m p_{1j}$
	x ₂	P ₂₁	P ₂₂	...	P _{2j}	...	P _{2m}	$P_{2\cdot} = \sum_{j=1}^m p_{2j}$
	⋮	⋮	⋮		⋮		⋮	⋮
	x _i	P _{i1}	P _{i2}	...	P _{ij}	...	P _{im}	$P_{i\cdot} = \sum_{j=1}^m p_{ij}$
	⋮	⋮	⋮		⋮		⋮	⋮
	x _n	P _{n1}	P _{n2}	...	P _{nj}	...	P _{nm}	$P_{n\cdot} = \sum_{j=1}^m p_{nj}$
Rozkład brzegowy Y		$P_{\cdot 1} = \sum_{i=1}^n p_{i1}$	$P_{\cdot 2} = \sum_{i=1}^n p_{i2}$		$P_{\cdot j} = \sum_{i=1}^n p_{ij}$		$P_{\cdot m} = \sum_{i=1}^n p_{im}$	$\sum_i p_{i\cdot} = \sum_j p_{\cdot j} = 1$

Rys. 1.15. Rozkład prawdopodobieństwa dwuwymiarowej, dyskretnej zmiennej losowej

Wzory (1.127) i (1.129) przedstawiają funkcje prawdopodobieństwa brzegowych rozkładów prawdopodobieństwa zmiennych losowych X i Y . Wartości tych funkcji informują o prawdopodobieństwach zrealizowania się poszczególnych wartości x_i i y_j , niezależnie od tego, z którą wartością drugiej zmiennej kojarzy się x_i (albo y_j) w konkretnej realizacji dwuwymiarowej zmiennej losowej (X, Y) . Na podstawie podanych w tablicy wartości p_{ij} ($i = 1, 2, \dots; j = 1, 2, \dots$) oraz sformułowanych powyżej funkcji prawdopodobieństwa rozkładów brzegowych, można wyznaczyć następujące prawdopodobieństwa warunkowe:

$$P(X = x_i | Y = y_j) = \frac{p_{ij}}{p_{.j}}, \quad (1.130)$$

$$P(Y = y_j | X = x_i) = \frac{p_{ij}}{p_{i.}}. \quad (1.131)$$

Są to funkcje prawdopodobieństwa warunkowych (albo względnych) rozkładów prawdopodobieństwa zmiennych losowych X i Y . Informują one o prawdopodobieństwie zrealizowania się poszczególnych wartości jednej zmiennej, przy ustalonych wartościach drugiej zmiennej. Można przy tym łatwo wykazać, że dla każdego y_j suma prawdopodobieństw wyrażonych równaniem (1.130) oraz dla każdego x_i suma prawdopodobieństw wyrażonych równaniem (1.131) równa jest jedności. Mamy bowiem:

$$\sum_i P(X = x_i | Y = y_j) = \frac{\sum_i p_{ij}}{p_{.j}} = \frac{p_{.j}}{p_{.j}} = 1, \quad (1.132)$$

$$\sum_j P(Y = y_j | X = x_i) = \frac{\sum_j p_{ij}}{p_{i.}} = \frac{p_{i.}}{p_{i.}} = 1. \quad (1.133)$$

Podane powyżej rozważania można rozszerzyć na ciągle, dwuwymiarowe zmienne losowe (X, Y) .

Brzegowe rozkłady prawdopodobieństwa zmiennych losowych X i Y posiadają następujące gęstości prawdopodobieństwa:

$$f_X(x) = \int_{-\infty}^{+\infty} f(x, y) dx, \quad (1.134)$$

$$f_Y(y) = \int_{-\infty}^{+\infty} f(x, y) dy. \quad (1.135)$$

Zachodzą przy tym następujące zależności:

$$\int_{-\infty}^{+\infty} f_X(x) dx = \int_{-\infty}^{+\infty} \left\{ \int_{-\infty}^{+\infty} f(x, y) dy \right\} dx = 1, \quad (1.136)$$

$$\int_{-\infty}^{+\infty} f_Y(y) dy = \int_{-\infty}^{+\infty} \left\{ \int_{-\infty}^{+\infty} f(x, y) dx \right\} dy = 1. \quad (1.137)$$

Gęstość prawdopodobieństwa warunkowych rozkładów zmiennych losowych X i Y można wyrazić następująco:

$$f(x|y) = \frac{f(x, y)}{f_Y(y)}, \quad (1.138)$$

$$f(y|x) = \frac{f(x, y)}{f_X(x)}, \quad (1.139)$$

przy czym, analogicznie jak w przypadku dyskretnych zmiennych losowych, zachodzą następujące zależności:

$$\int_{-\infty}^{+\infty} f(x|y) dx = \int_{-\infty}^{+\infty} \frac{f(x, y)}{f_Y(y)} dx = 1, \quad (1.140)$$

$$\int_{-\infty}^{+\infty} f(y|x) dy = \int_{-\infty}^{+\infty} \frac{f(x, y)}{f_X(x)} dy = 1. \quad (1.141)$$

Wykorzystując określenie dystrybuanty dwuwymiarowej zmiennej losowej oraz dystrybuanty rozkładów brzegowych zmiennych losowych można zdefiniować pojęcie niezależności zmiennych losowych X, Y . Zmienne losowe X, Y są niezależne, gdy dla każdej pary wartości x, y spełnione jest następujące równanie:

$$F(x, y) = F_1(x) \cdot F_2(y), \quad (1.142)$$

gdzie:

$F(x, y)$ – dystrybuanta dwuwymiarowej zmiennej losowej X, Y ,

$F_1(x), F_2(y)$ – dystrybuanty rozkładów brzegowych zmiennych X, Y .

1.5.3. Dwuwymiarowy rozkład normalny

W punkcie 1.4.2 określono pojęcie funkcji gęstości prawdopodobieństwa dwuwymiarowych zmiennych losowych. Dla zmiennych tych istnieją różne postaci rozkładów prawdopodobieństwa. Jedną z nich jest dwuwymiarowy rozkład normalny.

Zmienna losowa (X, Y) ma dwuwymiarowy rozkład normalny, jeżeli funkcja gęstości prawdopodobieństwa jej dwuwymiarowego rozkładu wyrażona jest wzorem:

$$f(x, y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp \left\{ -\frac{1}{2(1-\rho^2)} \left[\frac{(x-\mu_x)^2}{\sigma_x^2} - 2\rho \frac{(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y} + \frac{(y-\mu_y)^2}{\sigma_y^2} \right] \right\} \quad (x, y) \in \mathbf{R} \quad (1.143)$$

gdzie:

- μ_x, μ_y – wartości oczekiwane zmiennych losowych X i Y ,
- σ_x, σ_y – ich odchylenia standardowe,
- ρ – wartość współczynnika korelacji.

Dwuwymiarowy rozkład normalny $N(\mu_x, \mu_y, \sigma_x, \sigma_y, \rho)$ zależy więc od pięciu parametrów.

Wykres dwuwymiarowego rozkładu normalnego – w przestrzeni trójwymiarowej – przedstawiono na rys. 1.16.

Współczynnik korelacji ρ jest miarą liniowej współzależności między zmiennymi X i Y . Zauważmy przy tym, że:

1. Jeżeli $|\rho| = 1$, to funkcja gęstości prawdopodobieństwa nie jest określona; między badanymi zmiennymi losowymi zachodzi zależność funkcyjna. Otrzymany rozkład nosi nazwę osobliwego dwuwymiarowego rozkładu normalnego.

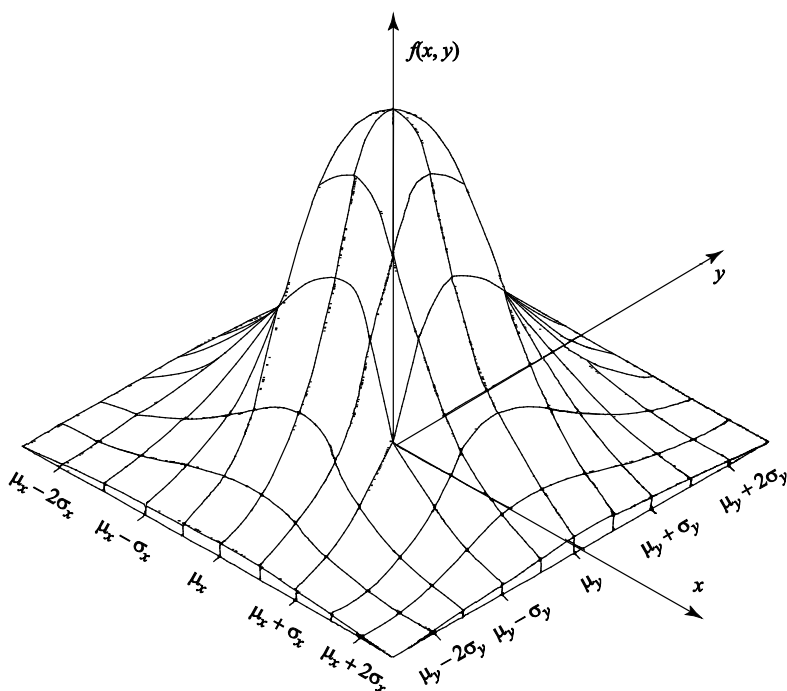
2. Jeżeli $\rho = 0$, to zmienne losowe (X, Y) nie są skorelowane. W przypadku rozkładu normalnego są one jednocześnie niezależne:

$$f(x, y) = \frac{1}{2\pi\sigma_x\sigma_y} \exp \left\{ -\frac{1}{2} \left[\frac{(x-\mu_x)^2}{\sigma_x^2} + \frac{(y-\mu_y)^2}{\sigma_y^2} \right] \right\} = \quad (1.144)$$

$$= f_1(x) \cdot f_2(y),$$

gdzie:

$f_1(x), f_2(y)$ – brzegowe prawdopodobieństwa zmiennych losowych (X, Y) .



Rys. 1.16. Dwuwymiarowy rozkład normalny

Rozkład dwuwymiarowy można standaryzować. W efekcie tego zabiegu nowe zmienne losowe spełniają następujące warunki: $\mu_x = 0$; $\mu_y = 0$; $\sigma_x^2 = 1$; $\sigma_y^2 = 1$. Zatem funkcja gęstości prawdopodobieństwa w dwuwymiarowym, standaryzowanym rozkładzie normalnym zależy tylko od jednego parametru, tzn. od wartości współczynnika korelacji ρ .

Jeżeli zmienne losowe (X, Y) mają dwuwymiarowy rozkład normalny, to wówczas rozkłady warunkowe tych zmiennych losowych są rozkładami normalnymi:

$$f(y|x) = \frac{1}{\sigma_y \sqrt{2\pi} \sqrt{1-\rho^2}} \exp \left\{ -\frac{1}{2\sigma_y^2(1-\rho^2)} [(y-\mu_y) - \rho \frac{\sigma_y}{\sigma_x} (x-\mu_x)]^2 \right\}, \quad (1.145)$$

$$f(x|y) = \frac{1}{\sigma_x \sqrt{2\pi} \sqrt{1-\rho^2}} \exp \left\{ -\frac{1}{2\sigma_x^2(1-\rho^2)} [(x-\mu_x) - \rho \frac{\sigma_x}{\sigma_y} (y-\mu_y)]^2 \right\}. \quad (1.146)$$

Warunkowe wartości oczekiwane $\mu_x(y) = E(X | Y = y)$ i $\mu_y(x) = E(Y | X = x)$ są określone wzorami:

$$\mu_x(y) = \mu_x + \rho \frac{\sigma_x}{\sigma_y} (y - \mu_y), \quad (1.147)$$

$$\mu_y(x) = \mu_y + \rho \frac{\sigma_y}{\sigma_x} (x - \mu_x). \quad (1.148)$$

W przypadku wzoru (1.147) zmienną objaśniającą jest Y , natomiast X jest zmienną objaśnianą. We wzorze (1.148) role zmiennych są odwrotne.

Warunkowe odchylenia standardowe wynikają ze wzorów:

$$\sigma_x(y) = \sigma_x \sqrt{1-\rho^2}, \quad (1.149)$$

$$\sigma_y(x) = \sigma_y \sqrt{1-\rho^2}. \quad (1.150)$$

Zauważmy, że warunkowe odchylenia standardowe są wielkościami stałymi. Nie zależą one od zmiennych objaśniających.

1.5.4. Współczynnik korelacji liniowej

Współczynnik korelacji liniowej (ρ) jest kolejną charakterystyką rozkładu prawdopodobieństwa dwuwymiarowej zmiennej losowej (X, Y) . Jest on miarą współzależności między zmiennymi X, Y i jest definiowany następująco:

$$\rho_{X,Y} = \frac{\text{cov}(X, Y)}{D(X) \cdot D(Y)}, \quad (1.151)$$

gdzie:

$$\text{cov}(X, Y) = E(XY) - E(X)E(Y). \quad (1.152)$$

Zdefiniowana w ten sposób kowariancja (łączna wariancja) zmiennych losowych X i Y jest miarą współzależności tych zmiennych. W przypadku gdy zmienne X i Y są niezależne:

$$E(XY) = E(X) \cdot E(Y) \quad (1.153)$$

i w konsekwencji $\text{cov}(X, Y) = 0$. Im silniejsze jest powiązanie zmiennych X i Y , tym bardziej $\text{cov}(X, Y)$ różni się od zera.

Kowariancja jest jednak niewygodną miarą współzależności, albowiem wyraża się w jednostkach mianowanych, a jej przedział zmienności nie jest ograniczony. W celu uniknięcia tych niedogodności dokonuje się zabiegu standaryzacji, dzieląc $\text{cov}(X, Y)$ przez iloczyn odchyłeń standardowych zmiennych X i Y . Skonstruowany w ten sposób współczynnik korelacji ρ przybiera wartości z przedziału $[-1; 0]$ albo $[0; +1]$, przy czym:

a) $\rho = 0$, gdy $\text{cov}(X, Y) = 0$, a więc wówczas, gdy zmienne X i Y nie są skorelowane;

b) $\rho = -1$, gdy zmienne losowe X i Y powiązane są funkcyjnie, a zależność między nimi ma charakter funkcji malejącej;

c) $\rho = +1$, gdy zmienne losowe X i Y powiązane są funkcyjnie, a zależność między nimi ma charakter funkcji rosnącej;

d) $-1 < \rho < 0$ albo $0 < \rho < 1$, gdy istnieje współzależność między zmiennymi X i Y , ale nie ma ona charakteru funkcyjnego, przy czym ρ tym bardziej odbiega od zera, im współzależność ta jest silniejsza.

1.6. Twierdzenia graniczne

Pod nazwą twierdzeń granicznych należy rozumieć zespół twierdzeń, w których rozważa się ciągi niezależnych zmiennych losowych i bada się graniczne

własności ich rozkładów. W twierdzeniach tych używa się pojęcia zbieżności stochastycznej, czyli zbieżności według prawdopodobieństwa.

1.6.1. Prawo wielkich liczb Bernoulliego

Twierdzenie graniczne, znane pod nazwą prawa wielkich liczb Bernoulliego⁹, stanowi teoretyczną podbudowę częstościowych definicji prawdopodobieństwa.

Przypomnijmy, że zmienna losowa w postaci (1.167) ma dwumianowy rozkład prawdopodobieństwa o parametrach p i n oraz że może ona przyjmować wartości: $0, 1/n, \dots, 2/n, \dots, z/n, \dots, 1$. Zmienną tę oznaczymy obecnie symbolem W_n , gdzie n – analogicznie jak dotychczas – oznacza liczbę składników w sumie (1.62). Ciąg zmiennych losowych W_n jest zbieżny stochastycznie – czyli według prawdopodobieństwa – do p . Mamy mianowicie:

$$\lim_{n \rightarrow \infty} P(|W_n - p| > \varepsilon) = 0 \quad (1.154)$$

albo

$$\lim_{n \rightarrow \infty} P(|W_n - p| \leq \varepsilon) = 1, \quad (1.155)$$

gdzie $\varepsilon > 0$ jest dowolnie małą liczbą rzeczywistą.

Wynika stąd, że im większe jest n – a więc im większa jest liczba składników w sumie (1.62) – tym większe jest prawdopodobieństwo zdarzenia losowego $-\varepsilon < W_n - p < \varepsilon$. Fakt ten stanowi formalną podstawę częstościowych definicji prawdopodobieństwa, zgodnie z którymi częstość względna (1.9) – będąca realizacją zmiennej losowej W_n – przyjmuje się jako przybliżenie prawdopodobieństwa.

Przykład 1.10

Rozważmy następujące zmienne losowe: $W_5 \sim B(n = 5, p = 0,5)$, $W_{10} \sim B(n = 10, p = 0,5)$, $W_{20} \sim B(n = 20, p = 0,5)$. Wartości $w_n = z_n/n$, będące realizacjami tych zmiennych losowych powinny – zgodnie z prawem wielkich liczb – tym wyraźniej grupować się wokół punktu $p = 0,5$, im większa jest wartość n .

Zbadamy to zjawisko, przyjmując $\varepsilon = 0,1$. Za otoczenie punktu $p = 0,5$ będziemy więc uważać przedział $[0,4; 0,6]$. Zbiór wartości zmiennej losowej W_5 ma postać:

$$\{0; 0,2; 0,4; 0,6; 0,8; 1\}.$$

⁹ Jakob Bernoulli (1654–1705), szwajcarski matematyk i fizyk.

Do otoczenia punktu p należą wartości 0,4 i 0,6. Mamy więc:

$$P(|W_5 - 0,5| \leq 0,1) = P(W_5 = 0,4) + P(W_5 = 0,6) = 0,624.$$

Zbiór wartości zmiennej losowej W_{10} ma postać:

$$\{0; 0,1; 0,2; 0,3; 0,4; 0,5; 0,6; 0,7; 0,8; 0,9; 1\}.$$

Do otoczenia punktu p należą trzy wartości, a mianowicie 0,4; 0,5 i 0,6. Po wykonaniu niezbędnych obliczeń mamy:

$$P(|W_{10} - 0,5| \leq 0,1) = P(W_{10} = 0,4) + P(W_{10} = 0,5) + P(W_{10} = 0,6) = 0,656.$$

W przypadku zmiennej losowej W_{20} mamy:

$$\{0; 0,05; 0,1; 0,15; 0,2; 0,25; 0,3; 0,35; 0,4; 0,45; 0,5; 0,55; 0,6; 0,65; 0,7; 0,75; 0,8; 0,85; 0,9; 0,95; 1\}.$$

Do otoczenia punktu $p = 0,5$ należy pięć wartości, a mianowicie: 0,4; 0,45; 0,5; 0,55 i 0,6. Sumując prawdopodobieństwa odpowiadające tym wartościom otrzymujemy:

$$P(|W_{20} - 0,5| \leq 0,1) = P(W_{20} = 0,4) + P(W_{20} = 0,45) + P(W_{20} = 0,5) + P(W_{20} = 0,55) + P(W_{20} = 0,6) = 0,736.$$

Należy zauważyć, że wyniki przeprowadzonych obliczeń potwierdzają wnioski wynikające z prawa wielkich liczb Bernoulliego. ■

1.6.2. Twierdzenie Moivre'a-Laplace'a

Rozważmy powtórnie ciąg zero-jedynkowych zmiennych losowych X_i ($i = 1, 2, \dots, n$), o jednakowych rozkładach prawdopodobieństwa. Przypomnijmy, że suma takich zmiennych losowych – wyrażona wzorem (1.62) – ma dwumianowy (binomialny) rozkład prawdopodobieństwa o parametrach p i n .

Zdefiniujemy obecnie następującą zmienną losową:

$$U_n = \frac{\sum_{i=1}^n X_i - np}{\sqrt{np(1-p)}} = \frac{Z_n - np}{\sqrt{np(1-p)}}. \quad (1.156)$$

Dystrybuantę tej zmiennej losowej oznaczmy następująco:

$$F_n(u) = P(U_n < u), \quad u \in \mathbf{R}. \quad (1.157)$$

Tak zwane integralne twierdzenie Moivre'a-Laplace'a głosi, że ciąg dystrybuant $F_n(u)$ jest zbieżny do dystrybuanty standaryzowanego rozkładu normalnego $\Phi(u)$. Mamy więc:

$$\lim_{n \rightarrow \infty} F_n(u) = \Phi(u) = \int_{-\infty}^u \varphi(u) du = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^u \exp\left(-\frac{1}{2}u^2\right) du \quad (1.158)$$

Wynika stąd, że:

$$\lim_{n \rightarrow \infty} P\left(a \leq \frac{Z_n - np}{\sqrt{np(1-p)}} < b\right) = \int_a^b \varphi(u) du. \quad (1.159)$$

Pozwala to stosować następujące przybliżenie rachunkowe:

$$P\left(a \leq \frac{Z_n - np}{\sqrt{np(1-p)}} < b\right) \approx \Phi(b) - \Phi(a). \quad (1.160)$$

$$P\left(np + a\sqrt{np(1-p)} \leq Z_n < np + b\sqrt{np(1-p)}\right) \approx \Phi(b) - \Phi(a). \quad (1.161)$$

Przykład 1.11

Załóżmy, że $n = 100$ i $p = 0,01$. Obliczymy prawdopodobieństwo zdarzenia losowego polegającego na tym, że $0 \leq Z_{100} < 1$. Zauważmy, że $P(0 \leq Z_{100} < 1) = P(Z_{100} = 0)$. Wykorzystując wzór (1.63) otrzymujemy:

$$P(Z_{100} = 0) = \binom{100}{0} 0,01^0 0,99^{100} \approx 0,366.$$

Zastosowanie wzorów (1.160) i (1.161) wymaga obliczenia wartości a i b . Wykorzystując lewą stronę wzoru (1.161) mamy:

$$\begin{aligned} np + a\sqrt{np(1-p)} &= 0 \\ 100 \cdot 0,01 + a\sqrt{100 \cdot 0,01 \cdot 0,99} &= 0 \\ a &= -\frac{1}{0,995} \approx -1,01 \\ np + b\sqrt{np(1-p)} &= 1 \\ 100 \cdot 0,01 + b\sqrt{100 \cdot 0,01 \cdot 0,99} &= 1 \\ b &= 0. \end{aligned}$$

Na podstawie wzoru (1.161) otrzymujemy:

$$P(0 \leq Z_{100} < 1) \approx \Phi(0) - \Phi(-1,01) = 0,5 - [0,5 - \Theta(1,01)] = \Theta(1,01) = 0,344.$$

Błąd przybliżenia wynosi więc $0,344 - 0,366 = -0,022$. ■

1.6.3. Nierówność Czebyszewa

W statystycznej kontroli jakości istotną rolę odgrywają tzw. statystyczne przedziały tolerancji. Są to przedziały, które z określonym prawdopodobieństwem pokrywają określoną frakcję realizacji obserwowanej zmiennej losowej. Porównanie takiego przedziału z technicznym przedziałem tolerancji, wyznaczającym dopuszczalne granice zmienności badanej cechy, pozwala najlepiej ocenić jakość wykonania danego wyrobu. Wyznaczenie statystycznego przedziału tolerancji uwarunkowane jest znajomością rozkładu prawdopodobieństwa zmiennej losowej. Jeśli rozkład ten nie jest znany, to przybliżone oceny można uzyskać za pomocą nierówności Czebyszewa¹⁰. Odpowiednie twierdzenie formułowane jest następująco: dla każdej zmiennej losowej X , o wartości oczekiwanej $E(X) = \mu$ i wariancji $D^2(X) = \sigma^2$, przy dowolnej liczbie rzeczywistej $\varepsilon > 0$ zachodzi nierówność:

$$P(|X - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2}. \quad (1.162)$$

Wynika stąd bezpośrednio, że:

$$P(|X - \mu| < \varepsilon) \geq 1 - \frac{\sigma^2}{\varepsilon^2}, \quad (1.163)$$

albowiem

$$P(|X - \mu| \geq \varepsilon) + P(|X - \mu| < \varepsilon) = 1.$$

Przekształcając wyrażenie (1.163) otrzymujemy:

$$P(-\varepsilon < X - \mu < \varepsilon) \geq 1 - \frac{\sigma^2}{\varepsilon^2}, \quad (1.164)$$

$$P(\mu - \varepsilon < X < \mu + \varepsilon) \geq 1 - \frac{\sigma^2}{\varepsilon^2}. \quad (1.165)$$

¹⁰ Pafnutij L. Czebyszew (1821–1894), matematyk rosyjski.

Jeśli przyjąć:

$$\varepsilon = k\sigma,$$

gdzie $k > 0$, to zależności (1.162)–(1.165) można zapisać następująco:

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}, \quad (1.166)$$

$$P(|X - \mu| < k\sigma) \geq 1 - \frac{1}{k^2} = \frac{k^2 - 1}{k^2}, \quad (1.167)$$

$$P(-k\sigma < X - \mu < k\sigma) \geq \frac{k^2 - 1}{k^2}, \quad (1.168)$$

$$P(\mu - k\sigma < X < \mu + k\sigma) \geq \frac{k^2 - 1}{k^2}. \quad (1.169)$$

Przykład 1.12

Załóżmy, że $X \sim N(\mu = 10, \sigma = 2)$. Z własności rozkładu normalnego (zob. punkt 1.4.2.2) wynika, że:

$$P(8 < X < 12) = 0,6826, \quad (i)$$

$$P(6 < X < 14) = 0,9545, \quad (ii)$$

$$P(4 < X < 16) = 0,9973. \quad (iii)$$

Stosując (1.169) mamy:

$$P(8 < X < 12) \geq 0, \quad (i)$$

$$P(6 < X < 14) \geq 0,75, \quad (ii)$$

$$P(4 < X < 16) \geq 0,89. \quad (iii)$$

Nietrudno zauważyć, że w przypadku (i) przybliżenie uzyskane za pomocą nierówności Czebyszewa nie ma wartości aplikacyjnej. W przypadku (ii) oraz (iii) oceny uzyskane za pomocą nierówności Czebyszewa mają charakter „grubych” przybliżeń. Podkreślmy jednak, że zostały one uzyskane bez jakichkolwiek założeń w odniesieniu do typu rozkładu obserwowanej zmiennej losowej. ■

Rozdział 2

METODOLOGIA BADAŃ STATYSTYCZNYCH

2.1. Wprowadzenie

W naukach ekonomicznych, podobnie jak w innych dziedzinach wiedzy, odwołujemy się często do badań empirycznych. Są one podejmowane bądź to w celu zweryfikowania (sprawdzenia) hipotez sformułowanych na podstawie rozważań teoretycznych, bądź też w celu pozyskania informacji niezbędnych do rozwiązania jakiegoś problemu. Badania empiryczne polegają na przeprowadzaniu odpowiednio zaplanowanych doświadczeń (eksperymentów). Mogą to być eksperymenty czynne albo bierne. W doświadczeniu czynnym eksperymentator sam kreuje sytuację, które pozwalają badać interesujący go obiekt lub zjawisko. W doświadczeniu biernym natomiast prowadzący badanie jest tylko obserwatorem. Nie ma on możliwości kreowania sytuacji doświadczalnych, a jego rola polega przede wszystkim na wyborze odpowiedniej metody badania. W naukach ekonomicznych, w odróżnieniu od nauk przyrodniczych i technicznych, dominują badania o charakterze eksperymentów biernych.

Oto przykłady takich badań:

- (i) Śledzenie liczby samochodów osobowych eksploatowanych w wybranym regionie albo w całym kraju.
- (ii) Analiza przyczyn zgonów w określonym przedziale czasu w wybranym regionie albo w całym kraju.
- (iii) Śledzenie i analiza kursów akcji wybranej spółki na Gieldzie Papierów Wartościowych w Warszawie.
- (iv) Śledzenie i analiza cen nieruchomości w określonym regionie.

W żadnym z wymienionych przykładów, ani też w żadnej innej podobnej sytuacji, prowadzący badania nie ma wpływu na przebieg obserwowanego zjawiska. Nie ma on też możliwości odtworzenia jakiegokolwiek fragmentu obserwowanej rzeczywistości w sztucznie spreparowanych warunkach. Dlatego właśnie mamy tu do czynienia z doświadczeniem biernym. W badaniach takich

podstawowym problemem jest zorganizowanie procesu pozyskiwania pierwotnych informacji o obserwowanych faktach. Jest to problem o znaczeniu podstawowym przede wszystkim dlatego, że od jakości owych pierwotnych (źródłowych) informacji zależy celowość podejmowania badań statystycznych. Należy wszak pamiętać, że nawet najbardziej wyrafinowane metody analizy statystycznej nie wzbogacają materiału źródłowego¹. Ich zastosowanie pozwala tylko ujawnić te informacje, które tkwią w wynikach przeprowadzonych obserwacji lub pomiarów. Dodajmy, że zastosowanie niewłaściwych albo zbyt prymitywnych metod analizy statystycznej jest często przyczyną utraty znacznej części owych źródłowych informacji. Oznacza to marnotrawstwo środków wydanych na pozyskanie informacji o badanym obiekcie lub zjawisku. W empirycznych badaniach zjawisk, które należą do kręgu zawodowych i naukowych zainteresowań ekonomisty, bardzo często można korzystać z informacji, które – z mocy prawa – gromadzone są przez wyspecjalizowane agendy państwowe i samorządowe. Podstawową rolę odgrywa tu Główny Urząd Statystyczny (GUS)², a w skali regionalnej urzędy statystyczne. Instytucje te kontynuują pierwotne funkcje statystyki, polegające na pozyskiwaniu, gromadzeniu i odpowiednim przetwarzaniu informacji o stanie państwa³.

Zauważmy, że na podstawie bieżąco rejestrowanych, gromadzonych i publikowanych informacji można podjąć badania statystyczne w zakresie trzech spośród czterech wymienionych powyżej przykładów.

Tablica 2.1

Zarejestrowane w Polsce samochody osobowe

Rok	Liczba zarejestrowanych samochodów (tys. szt.)	Rok	Liczba zarejestrowanych samochodów (tys. szt.)
1980	2383,0	1986	3964,0
1981	2643,3	1987	4231,7
1982	2881,7	1988	4519,1
1983	3178,9	1989	4846,4
1984	3425,8	1990	5260,6
1985	3671,4	1991	6112,2

Źródło: Rocznik Statystyczny GUS, 1992.

¹ Istotę poruszanego problemu najlepiej oddaje lapidarne powiedzenie, popularne wśród statystyków amerykańskich: „rubbish in, rubbish out”, czyli „jeśli włożysz śmiecie, to wyjmiesz śmiecie”.

² Główny Urząd Statystyczny (GUS) został utworzony w 1918 r. jako jedna z pierwszych wyspecjalizowanych agend rządowych w odrodzonej Polsce.

³ Przypomnijmy tu, że słowo „statystyka” wywodzi się od łacińskiego słowa „status”, czyli państwo.

Przykład (i): Najpewniejszym źródłem informacji o liczbie eksploatowanych samochodów osobowych są rejestry prowadzone przez wyspecjalizowane agendy samorządu terytorialnego. Pozyskiwane w ten sposób informacje są gromadzone i przetwarzane przez GUS, a następnie publikowane w rocznikach statystycznych. Przykład tego typu informacji przedstawiono w tabelicy 2.1.

Przykład (ii): Informacje dotyczące naturalnego ruchu ludności (urodziny, zgony) są bieżąco rejestrowane i gromadzone przez urzędy stanu cywilnego, a następnie opracowywane i udostępniane dla celów badawczych przez GUS. Informacje tego typu udostępniane są również organizacjom międzynarodowym. Ułatwia to różnego rodzaju badania porównawcze. W tabelicy 2.2 przedstawiono informacje opublikowane przez Światową Organizację Zdrowia (World Health Organization; WHO), dotyczące zgonów spowodowanych chorobami nowotworowymi w Polsce.

Tabela 2.2

Zgony w Polsce spowodowane chorobami nowotworowymi

Rok	Liczba zgonów na raka na tysiąc zgonów ogółem	Rok	Liczba zgonów na raka na tysiąc zgonów ogółem	Rok	Liczba zgonów na raka na tysiąc zgonów ogółem
1961	97,40	1971	143,75	1981	170,10
1962	103,90	1972	148,80	1982	171,00
1963	112,45	1973	152,40	1983	175,80
1964	119,35	1974	154,20	1984	179,60
1965	123,85	1975	157,05	1985	180,75
1966	125,60	1976	160,75	1986	182,50
1967	128,80	1977	162,15	1987	185,00
1968	133,20	1978	164,55	1988	188,40
1969	135,40	1979	166,35	1989	189,20
1970	138,00	1980	168,40	1990	192,15

Źródło: World Health Organization.

Przykład (iii): Najprostszą, a jednocześnie wystarczająco pewną metodą pozyskiwania informacji o kursach akcji spółek notowanych na Giełdzie Papierów Wartościowych w Warszawie jest korzystanie z tablic kursów, publikowanych przez biura maklerskie w prasie fachowej i codziennej. Zarząd Giełdy ma ustawowy obowiązek publikowania wyników każdej sesji giełdowej.

W przykładzie (iv) problem pozyskiwania informacji źródłowych jest bardziej skomplikowany, albowiem nie tylko nie ma obowiązku publikowania informacji o transakcyjnych cenach nieruchomości, ale pewne informacje o zawieranych transakcjach chronione są tajemnicą. W takiej, a także w wielu innych podobnych sytuacjach, podstawową metodą pozyskiwania informacji jest

ankieta. Jest to zawsze metoda mniej pewna, niż bezpośrednio wykorzystywanie informacji pochodzących z bieżącej rejestracji. Jest to też metoda kosztowniejsza i wymagająca pewnej wiedzy w zakresie metody reprezentacyjnej i teorii społecznego komunikowania⁴.

Nie wszystkie badania empiryczne podejmowane w naukach ekonomicznych mają charakter eksperymentów biernych. Podejmowane są także, aczkolwiek znacznie rzadziej niż w naukach przyrodniczych i technicznych, badania o charakterze eksperymentów czynnych. Oto przykłady:

- (v) Badanie reakcji konsumentów na nowe formy sprzedaży pewnego produktu, wprowadzone w celu zwiększenia utargu.
- (vi) Poszukiwanie najbardziej efektywnej formy organizacji procesu wytwarzania pewnego produktu.

W przykładzie (v) mamy do czynienia z przedsięwzięciem marketingowym podejmowanym w celu aktywizacji sprzedaży pewnego produktu. Jeśli nie są to działania rutynowe, to – niezależnie od wszystkich aspektów praktycznych – powinny być one potraktowane w kategoriach eksperymentu. Pozwala to racjonalnie zorganizować podejmowane działania, z uwzględnieniem teorii planowania doświadczeń⁵. Z bardzo podobną sytuacją mamy do czynienia w przykładzie (vi). Organizator procesu produkcyjnego stoi tu jednak wobec nieco łatwiejszego zadania niż szef służb marketingowych w przykładzie (v). Zadanie jest łatwiejsze, albowiem w mniejszym stopniu zależy od otoczenia rynkowego, na które nie mamy wpływu.

Każde badanie empiryczne – zarówno o charakterze eksperymentu czynnego, jak i biernego – należy interpretować jako doświadczenie losowe. Przypomnijmy (zob. rozdział 1), że jest to takie przedsięwzięcie empiryczne, którego wynik jest zdarzeniem losowym. Ma to ten skutek, że wyniki badań empirycznych mogą być racjonalnie opisywane i interpretowane tylko za pomocą metod statystycznych. Metody te pozwalają odróżniać skutki działania czynników systematycznych od skutków losowej (przypadkowej) zmienności obserwowanych układów doświadczalnych. Wykrycie, wyodrębnienie i scharakteryzowanie owej systematycznej składowej w ogólnej zmienności badanego zjawiska jest zawsze jednym z podstawowych celów podejmowania badań empirycznych.

⁴ Zainteresowanego Czytelnika odsyłamy do książki J. Steczkowskiego: *Metoda reprezentacyjna w badaniach zjawisk ekonomiczno-społecznych*, Wydawnictwo Naukowe PWN, Warszawa-Kraków 1995.

⁵ Zainteresowanego Czytelnika odsyłamy do książek: Z. Polański, *Planowanie doświadczeń w technice*, PWN, Warszawa 1984 oraz D. Rasch, G. Herrendörfer, *Statystyczne planowanie doświadczeń*, Wydawnictwo Naukowe PWN, Warszawa 1991.

2.2. Zbiorowość generalna

Zbiorowość generalna, albo populacja generalna, jest to zbiór obiektów materialnych lub potencjalnych powtórzeń zjawiska, który jest przedmiotem zainteresowania badacza. Zbiorowość generalną będziemy oznaczać symbolem \mathbf{A} , natomiast jej elementy – czyli pojedyncze obiekty materialne lub pojedyncze powtórzenia zjawiska – symbolem A , zaopatrzonym (w miarę potrzeby) w odpowiednie indeksy. Poddawana badaniu zbiorowość generalna \mathbf{A} musi być zbiorem jednorodnym ze względu na cechy kwalifikujące, należące do zbioru:

$$\mathbf{K} = \{K_r; r = 1, 2, \dots, q\}. \quad (2.1)$$

Zawartość tego zbioru powinna w sposób klarowny wynikać z ogólnego planu przedsięwzięcia badawczego. W planie takim powinien też być zdefiniowany zbiór cech badanych:

$$\mathbf{B} = \{B_s; s = 1, 2, \dots, m\}, \quad (2.2)$$

czyli zbiór tych cech B_s , ze względu na które – zgodnie z celem podejmowanych badań – będą badane elementy $A \in \mathbf{A}$. Zbiory \mathbf{K} i \mathbf{B} powinny być rozłączne:

$$\mathbf{K} \cap \mathbf{B} = \emptyset. \quad (2.3)$$

Zdefiniowanie zbiorów \mathbf{K} i \mathbf{B} jest warunkiem wstępnym podjęcia badań empirycznych, albowiem zanim rozpocznie się takie badanie, powinno się dokładnie wiedzieć, co i ze względu na co chce się badać.

Każdej cesze kwalifikującej $K_r \in \mathbf{K}$ odpowiada pewien zbiór rozróżnialnych stanów \mathbf{K}_r° . Szczegółowa postać tego zbioru zależy od zastosowanej metody badania cechy K_r , a ta z kolei uzależniona jest od potrzeb badania, określonych przez cechy $B_s \in \mathbf{B}$. Podczas projektowania doświadczenia, w każdym zbiorze \mathbf{K}_r° wyróżnia się pewien podzbiór \mathbf{K}_r^+ , stanowiący kryterium przynależności jednostek A do zbiorowości generalnej \mathbf{A} . Mając te podzbiory ($r = 1, 2, \dots, q$) można sformułować warunek zakwalifikowania jednostki A , należącej do pewnego nadzbioru $\bar{\mathbf{A}}$, do zbiorowości generalnej \mathbf{A} . Mamy mianowicie:

$$A \in \bar{\mathbf{A}} \Leftrightarrow \bigwedge_r K_r(A) \in \mathbf{K}_r^+, \quad (2.4)$$

gdzie $K_r(A)$ oznacza stan elementu A ze względu na cechę $K_r \in \mathbf{K}$.

Z równoważności (2.4) wynika bezpośrednio, że o szczegółowej postaci zbiorowości generalnej $\bar{\mathbf{A}}$ decyduje to, w jaki sposób zostaną określone poszczególne

gólne podzbiory K_r^+ ($r = 1, 2, \dots, q$). Należy podkreślić, że przy konstruowaniu tych podzbiorów powinny decydować przesłanki merytoryczne, a nie formalne. Reguła ta powinna obowiązywać zwłaszcza na wstępnym etapie badań, albowiem przede wszystkim względy merytoryczne powinny decydować o tym, co należy, a co nie należy do obszaru badań.

Przykład 2.1

Załóżmy, że producent pewnego typu samochodów osobowych, począwszy od pojazdu o numerze fabrycznym y_0 , zastosował ulepszone zabezpieczenie antykorozyjne karoserii. Po upływie trzech lat chce on sprawdzić skuteczność zastosowanego ulepszenia. Zbiorowość generalną stanowią więc wszystkie samochody osobowe danej marki i danego typu, o numerach fabrycznych nie niższych niż y_0 i nie wyższych niż y_1 . Numer fabryczny y_1 został tak dobrany, by do populacji generalnej włączyć samochody eksploatowane co najmniej przez jeden rok. Tak zdefiniowana zbiorowość generalna wynika z przedstawionego poniżej rozumowania.

- Nadzbiór \bar{A} jest zbiorem wszystkich eksploatowanych samochodów osobowych. Z tego nadzbioru należy wyodrębnić zbiór A interesujący producenta.
- Zbiór cech kwalifikujących jest trójelementowy i ma postać:

$$K = \left\{ \begin{array}{l} K_1 - \text{przynależność do marki} \\ K_2 - \text{przynależność do typu} \\ K_3 - \text{numer fabryczny} \end{array} \right\} \quad (\text{i})$$

- Dla potrzeb zamierzonego badania zbiory rozróżnialnych stanów tych cech można zdefiniować następująco:

$$K_1^\circ = \left\{ \begin{array}{l} k_{1.1} - \text{spełnia kryterium} \\ \text{przynależności do marki} \\ k_{1.2} - \text{nie spełnia kryterium} \\ \text{przynależności do marki} \end{array} \right\} \quad (\text{ii})$$

$$K_2^\circ = \left\{ \begin{array}{l} k_{2.1} - \text{spełnia kryterium} \\ \text{przynależności do typu} \\ k_{2.2} - \text{nie spełnia kryterium} \\ \text{przynależności do typu} \end{array} \right\} \quad (\text{iii})$$

$$\mathbf{K}_3^\circ = \left\{ \begin{array}{l} k_{3,1}: y \in [y_0; y_1] \\ k_{3,2}: y \notin [y_0; y_1] \end{array} \right\} \quad (\text{iv})$$

gdzie y oznacza numer fabryczny konkretnego samochodu.

– Podzbiory stanów wyróżnionych (pożądanych):

$$\mathbf{K}_1^+ = \{k_{1,1} - \text{spełnia kryterium przynależności do marki}\} \quad (\text{v})$$

$$\mathbf{K}_2^+ = \{k_{2,1} - \text{spełnia kryterium przynależności do typu}\} \quad (\text{vi})$$

$$\mathbf{K}_3^+ = \{k_{3,1} : y \in [y_0; y_1]\} \quad (\text{vii})$$

– Tak więc, zgodnie ze wzorem (2.4), mamy:

$$A \in \mathbf{A} \Leftrightarrow K_1(A) \in \mathbf{K}_1^+ \wedge K_2(A) \in \mathbf{K}_2^+ \wedge K_3(A) \in \mathbf{K}_3^+ \quad (\text{viii})$$

– Zbiorowość generalna \mathbf{A} przedstawia się więc następująco:

$$\mathbf{A} = \{A: K_1(A) \in \mathbf{K}_1^+ \wedge K_2(A) \in \mathbf{K}_2^+ \wedge K_3(A) \in \mathbf{K}_3^+\} \quad (\text{ix})$$

gdzie A oznacza samochód osobowy, natomiast $K_1(A)$, $K_2(A)$ i $K_3(A)$ oznaczają stan tego samochodu ze względu na cechy kwalifikujące K_1 , K_2 i K_3 .

Wzór (ix) definiuje zbiorowość generalną niezależnie od szczegółowego planu badania. Zbiór \mathbf{B} musi zawierać cechę „stopień skorodowania karoserii”, albowiem wynika to z celu podejmowanych badań. Do zbioru \mathbf{B} mogą być jednak włączone również inne cechy, takie jak intensywność eksploatacji, skażenie środowiska w danym regionie, przebieg itp. Włączenie tych cech do zbioru \mathbf{B} pozwala wyjaśnić zróżnicowanie badanych obiektów ze względu na cechę podstawową. ■

Wyróżnia się zbiorowości generalne przedmiotowe i zdarzeniowe. Elementami przedmiotowych zbiorowości generalnych są obiekty materialne (przedmioty), natomiast zdarzeniowe zbiorowości generalne składają się z powtórzeń (albo potencjalnych powtórzeń) badanego zjawiska. Przedmiotowe zbiorowości generalne są najczęściej zbiorami skończonymi:

$$\mathbf{A} = \{A_z: \bigwedge_z \bigwedge_r K_r(A_z) \in \mathbf{K}_r^+; z = 1, 2, \dots, N\}, \quad (2.5)$$

gdzie N jest licznością zbiorowości generalnej. Tego typu zbiorowość generalną przedstawiono w przykładzie 2.1. Do tego typu zbiorowości należy również – na przykład – partia produktu poddawanej ocenie jakości albo zbiór konsumentów tworzących segment rynku.

W pewnych sytuacjach przedmiotową zbiorowość generalną wygodnie jest traktować jako zbiór przeliczalny postaci:

$$A = \{A_z: \bigwedge_z \bigwedge_r K_r(A_z) \in \mathbf{K}_r^+; z = 1, 2, \dots\}. \quad (2.6)$$

Z sytuacją taką mamy do czynienia na przykład wówczas, gdy obserwowanym obiektem jest nie partia (zasób), a strumień wyrobu generowany przez agregat produkcyjny. Zauważmy jednak, że żaden rzeczywisty proces produkcyjny nie może trwać nieskończenie długo, a ponadto, nawet w przypadku agregatów produkcyjnych funkcjonujących w ruchu ciągłym, ocenę jakości odnosi się zawsze do pewnych ustalonych i niezbyt długich okresów rozliczeniowych. Tak więc zbiorowość generalna jest w takich przypadkach zbiorem skończonym, a za modelem (2.6) przemawia przede wszystkim to, że przed zakończeniem procesu kumulowania strumienia nie jest znana liczność (N) zbiorowości generalnej. Ma to wpływ na technikę pobierania próby do badania (zob. punkt 2.3).

Zdarzeniowe zbiorowości generalne traktowane są najczęściej jako zbiory nieskończone. Mogą to być zbiory przeliczalne postaci (2.6) albo zbiory nieprzeliczalne:

$$A = \{A: \bigwedge_r K_r(A) \in \mathbf{K}_r^+\}. \quad (2.7)$$

Podkreślmy jednak, że zbiorem nieskończonym może być tylko zbiór potencjalnych powtórzeń obserwowanego zjawiska, i to tylko wówczas, gdy generowanie tych powtórzeń nie jest związane z funkcjonowaniem jakiegoś urządzenia technicznego. Jeśli urządzenie takie jest niezbędnym elementem układu doświadczalnego, to zbiór potencjalnych powtórzeń zjawiska jest skończony, albowiem trwałość każdego urządzenia technicznego jest ograniczona. Nie zmienia to jednak faktu, że zbiory postaci (2.6) i (2.7) są najczęściej stosowanymi modelami zdarzeniowych zbiorowości generalnych. Nawet wówczas, gdy są to zbiory w istocie skończone, ich liczność jest zwykle bardzo duża i trudna do określenia.

Badania zbiorowości generalnej A ze względu na cechy $B_s \in \mathbf{B}$ mogą być badaniami wyczerpującymi (całkowitymi) albo badaniami niewyczerpującymi (częściowymi, wrywkowymi). W badaniach wyczerpujących badaniu poddawane są wszystkie elementy $A \in A$, w badaniach niewyczerpujących natomiast

badaniu podlega tylko pewien podzbiór $\underline{A} \subset A$, a wnioski odnoszone są do całego zbioru A . Podzbiór nosi nazwę próby (zob. punkt 2.3).

Badania wyczerpujące możliwe są wówczas, gdy:

- (i) zbiorowość generalna jest zbiorem skończonym postaci (2.5),
- (ii) metoda badania każdej cechy $B_s \in \mathbf{B}$ ma charakter nieniszczący,
- (iii) koszty związane z badaniem nie przewyższają pewnej wartości, ustalonej na podstawie przesłanek ekonomicznych.

Jeśli nie jest spełniony warunek (i), to badania wyczerpujące są niemożliwe *ex definitione*. Jeśli warunek (i) jest spełniony, a jednocześnie nie jest spełniony warunek (ii), to badania wyczerpujące są formalnie możliwe, ale są one pozbawione sensu. Rezultatem takich badań byłoby bowiem zniszczenie całej zbiorowości generalnej. Jeśli warunki (i), (ii) są spełnione, to o możliwości prowadzenia badań wyczerpujących rozstrzyga ich ekonomiczna opłacalność (warunek (iii)). Należy przy tym brać pod uwagę nie tylko koszty samego badania, ale również koszty związane z czasem oczekiwania na wyniki badań. Jeśli ten okres oczekiwania jest długi, to wyniki badań wyczerpujących mogą być nieaktualne lub nieprzydatne już w chwili ich uzyskania. Dlatego też w praktyce podstawowe znaczenie mają badania niewyczerpujące (częściowe). Badania takie są zawsze mniej kosztowne, a wiarygodność wniosków formułowanych w odniesieniu do zbiorowości generalnej nie jest niższa niż w przypadku badań wyczerpujących. W pewnych sytuacjach badania wyczerpujące są jednak niezastąpione. Można tu wymienić powszechne spisy ludności, a także spisy rolne. Spisy te nie tylko dostarczają szczegółowych informacji, ale pozwalają również weryfikować poprawność procedur badań niewyczerpujących, stosowanych w okresach między spisami.

2.3. Próba

Próba, nazywana też próbką albo populacją lub zbiorowością próbną, jest podzbiorem zbiorowości generalnej, który poddawany jest badaniu i który stanowi podstawę do formułowania wniosków o zbiorowości generalnej. Próbę pobraną ze zbiorowości generalnej A będziemy oznaczać symbolem \underline{A} . Próba jest zawsze zbiorem skończonym, niezależnie od tego, czy zbiorowość generalna jest zbiorem skończonym czy nieskończonym. Mamy więc:

$$\underline{A} = \{A_i: \bigwedge_i A_i \in A; \quad i = 1, 2, \dots, n\}, \quad (2.8)$$

gdzie n jest licznością próby.

Elementy zbiorowości generalnej mogą być wybierane do próby w sposób celowy lub losowy (przypadkowy). Próbką celową stosowana jest przede wszystkim wówczas, gdy badane cechy $B_s \in \mathbf{B}$ są skorelowane z pewnymi cechami kwalifikującymi, co umożliwia przeprowadzenie wnioskowania na podstawie wartości ekstremalnych, czyli tzw. wnioskowania z nadmiarem.

Przykład 2.2

Załóżmy, że zbiorowością generalną A jest partia pewnego tłuszczu spożywczego złożona z N opakowań jednostkowych. Załóżmy też, że partia ta była przechowywana przez pewien czas w nieodpowiednich warunkach i zachodzi podejrzenie, iż stężenie produktów jęlczenia jest większe niż dopuszczalne, przynajmniej w niektórych opakowaniach jednostkowych. W celu rozstrzygnięcia tego problemu należy posłużyć się próbką celową. Należy do niej zaliczyć te opakowania jednostkowe, które wykazują najbardziej zaawansowane zewnętrzne objawy jęlczenia (zmiany barwy, zapachu). Jeśli badanie tak skonstruowanej próbki nie potwierdzi podejrzeń, to prawie na pewno partia produktu odpowiada wymaganiom jakościowym. ■

Przykład 2.3

Zarząd supermarketu, działając na zlecenie producenta, chce poznać opinię swoich klientów na temat:

- walorów smakowych pewnej odmiany margaryny,
- właściwości użytkowych opakowań jednostkowych, w których ta margaryna jest sprzedawana.

Wyniki tych badań mają posłużyć do opracowania ulepszonej koncepcji produktu, a w szczególności do zrjonalizowania opakowania jednostkowego. Na podstawie posiadanego doświadczenia handlowego zarząd supermarketu ocenia, że ponad 90% zakupów tego produktu dokonują kobiety w młodym i średnim wieku. Uznano więc, że próbkę należy wylosować spośród tej grupy klientów, gdyż dostosowanie produktu do ich wymagań może mieć największy wpływ na zwiększenie zakupów. Będzie to próbka celowo-losowa. Dodajmy, że zastosowanie celowego doboru próbki ma tu bardzo racjonalne uzasadnienie. Pobieranie próbki z całej populacji klientów zmniejsza wartość informacyjną uzyskanych rezultatów. Wyniki te będą bowiem obciążone opiniami osób, które nie realizują tych opinii w formie zakupów. ■

Próbki celowe, mimo że bardzo użyteczne w pewnych sytuacjach doświadczalnych, odgrywają drugorzędną rolę zarówno w praktyce jak i rozważaniach teoretycznych. Podstawowe znaczenie mają próbki losowe. Próbki takie otrzy-

muje się na drodze losowania, czyli takiego postępowania, w którym o wyborze danego elementu zbiorowości generalnej do próby decyduje jedynie przypadek. Przyjmuje się, że próbka losowa o dostatecznie dużej liczności jest próbką reprezentatywną, czyli zachowującą wszystkie cechy strukturalne zbiorowości generalnej.

Istnieją dwie podstawowe techniki losowania, a mianowicie:

- losowanie ze zwracaniem,
- losowanie bez zwracania.

Losowanie ze zwracaniem może być stosowane tylko w przypadku przedmiotowych zbiorowości generalnych, i to pod warunkiem, że metody badania wszystkich cech $B_s \in \mathbf{B}$ są metodami nieniszczącymi. We wszystkich pozostałych przypadkach może być stosowane tylko losowanie bez zwracania.

W przypadku nieskończonej licznych zbiorowości generalnych postaci (2.6) i (2.7) obydwie techniki losowania są równoważne, zarówno formalnie jak i praktycznie. Jeśli natomiast mamy do czynienia ze skończoną zbiorowością generalną postaci (2.5), to losowanie ze zwracaniem jest losowaniem niezależnym, podczas gdy losowanie bez zwracania jest losowaniem zależnym⁶. W rozważanym przypadku losowanie ze zwracaniem zapewnia niezmienną stochastyczną strukturę zbiorowości generalnej w całym procesie losowania. Oznacza to, że prawdopodobieństwo wylosowania danego elementu zbiorowości generalnej, w pojedynczym akcie losowania, nie zależy od wyników poprzednich losowań. W przypadku losowania bez zwracania zależność taka istnieje, albowiem każdy akt losowania zmienia stochastyczną strukturę zbiorowości generalnej. Zmiany te mogą być mniej lub bardziej znaczące, w zależności od stosunku liczności próbki (n) do liczności zbiorowości generalnej (N). Do oceny tego zjawiska można wykorzystać następujące wyrażenie:

$$\varepsilon = \frac{N - n}{N - 1}. \quad (2.9)$$

Im bliższa jedności jest wartość ε , tym mniej znaczące jest naruszenie niezależności losowania, mimo zastosowania losowania bez zwracania. Zauważmy, że jeśli $n = 1$, to $\varepsilon = 1$, niezależnie od wartości N . Potwierdza to intuicyjnie oczywisty wniosek, że jeśli próbka jest jednoelementowa, to losowanie bez zwracania jest całkowicie równoważne losowaniu ze zwracaniem.

Obydwie wyróżnione techniki losowania (ze zwracaniem i bez zwracania) mogą być wykorzystane w różnych schematach pobierania próbek losowych. Należy tu wymienić przede wszystkim *losowanie indywidualne nieograniczone*,

⁶ Przypomnijmy, że z problemem losowania niezależnego i zależnego zetknęliśmy się w rozdziale 1, omawiając rozkład dwumianowy i hipergeometryczny.

które stanowi swego rodzaju „wzorzec” i układ odniesienia dla wszystkich pozostałych schematów losowania. W losowaniu tym pojedyncze elementy $A \in A$ pobierane są bez ograniczeń z całej zbiorowości generalnej. Jeśli jest to losowanie niezależne, to w rezultacie uzyskujemy tzw. *prostą próbkę losową*, odgrywającą podstawową rolę zarówno w rozważaniach teoretycznych, jak i w praktyce. *Losowanie zespołowe* stosowane jest wówczas, gdy elementy zbiorowości generalnej połączone są w pewne zespoły, których dezagregacja – w procesie losowania – jest niemożliwa lub niecelowa, na przykład z powodu zbyt wysokich kosztów. Jednostką losowania jest wówczas cały zespół (agregat), a nie pojedynczy element. Jeśli w dalszych etapach postępowania owe agregaty redukowane są do zespołów o stopniowo malejących liczebnościach, to mamy do czynienia z *losowaniem wielostopniowym*. Osobną grupę schematów losowania stanowią procedury pobierania próbek z rozwarstwionych zbiorowości generalnych. Są to takie zbiorowości, które mogą być podzielone (rozcięte) na bardziej jednorodne podzbiory, a które – z jakichkolwiek powodów – muszą być traktowane łącznie. Pobieranie próbek z takich zbiorowości generalnych realizowane jest w formie *losowania warstwowego*⁷. Dodajmy, że jeśli liczebność próbki jest dostatecznie duża, to prosta próbka losowa jest tak samo efektywna jak próbka uzyskana w rezultacie losowania warstwowego.

Pobieranie próbki losowej wymaga zastosowania pewnych technik zapewniających całkowitą beztendencyjność procesowi kwalifikowania elementów zbiorowości generalnej do próby. W przypadku skończonych zbiorowości generalnych postaci (2.5) można do tego celu wykorzystać urnę z odpowiednio ponumerowanymi, a jednocześnie nierozróżnialnymi na podstawie innych cech kulami albo karteczkami (losami). Podstawowe znaczenie mają jednak tablice liczb losowych, a także generatory liczb pseudolosowych, w jakie wyposażone są współczesne komputery, a nawet niektóre kalkulatory kieszonkowe. Fragment tablicy liczb losowych, zaczerpnięty z normy PN/N-03010 „Statystyczna kontrola jakości. Losowy wybór sztuk do próbek”, zamieszczono w aneksie (tablica VII). W tablicy tej cyfry losowe zestawiono w czteroelementowe grupy, a w każdej takiej grupie wyróżnione są z kolei dwie podgrupy dwuelementowe. Grupy te można traktować jako czterocyfrowe i dwucyfrowe liczby losowe. Skreślając – na przykład – ostatnią cyfrę w grupie czterocyfrowej otrzymujemy liczbę trzycyfrową itd. Tablice cyfr losowych mogą być stosowane wówczas, gdy istnieje możliwość ponumerowania wszystkich elementów zbiorowości generalnej. Chodzi przy tym, co wydaje się oczywiste, nie o fizyczne etykietowanie tych elementów, ale o możliwość jednoznacznej identyfikacji każdego elementu $A \in A$ za pomocą liczby naturalnej ze zbioru $\{1, 2, 3, \dots, N\}$.

⁷ Czytelnicy zainteresowani tymi problemami znajdą szczegółowe informacje w podręcznikach z zakresu metody reprezentacyjnej, na przykład w cytowanej już książce J. Steczkowskiego.

Początek odczytywania sekwencji liczb losowych wybiera się w sposób przypadkowy. Następnie, przesuwając się wzdłuż wierszy (od strony lewej ku prawej lub odwrotnie) albo wzdłuż kolumn (od góry ku dołowi lub odwrotnie), odczytuje się kolejne grupy cyfr losowych o ustalonej długości, czyli kolejne liczby losowe. W przypadku losowania ze zwracaniem odrzuca się tylko liczby większe od N . Jeśli natomiast realizowane jest losowanie bez zwracania, to odrzuca się również już raz otrzymane liczby. Postępowanie kończy się z chwilą uzyskania zbioru liczb losowych o liczności n .

Przykład 2.4

Do odbioru jakościowego przedstawiono partię tafli szklanych o liczności $N = 100$. Do badania postanowiono pobrać próbkę losową o liczności $n = 10$. Tafle te były zapakowane w skrzynię drewniane po 10 sztuk. Ponieważ program badań przewidywał stosowanie tylko nieniszczących metod oceny jakości, zatem można było zastosować losowanie ze zwracaniem. Numerowanie wszystkich tafli szklanych wymagałoby bardzo dużego, i całkowicie zbędnego, nakładu czasu i pracy, a ponadto wiązałoby się z ryzykiem uszkodzenia tafli podczas manipulacji.

Zastosowano następujące postępowanie:

- skrzynie oznakowano kolejnymi cyframi od 0 do 9,
- numer każdej skrzyni umieszczono na bocznej krawędzi, od której biegła numeracja tafli w danej skrzyni.

Ponumerowano w ten sposób jednoznacznie wszystkie tafle, bez otwierania skrzyń. Chcąc poddać badaniu – na przykład – taflę o numerze 24, należało wziąć czwartą taflę (licząc od oznakowanej krawędzi) ze skrzyni o numerze 2 itd. Ponieważ $N = 100$, należało więc zastosować dwucyfrowe liczby losowe, przyjmując dodatkowo, że numerowi „100” odpowiada liczba losowa „00”. Za początek odczytywania przyjęto grupę cyfr losowych na przecięciu wiersza 7 i kolumny 4, w bloku 1 tablicy VII zamieszczonej w aneksie. Odczytując kolejne grupy dwucyfrowe w wierszu 7 otrzymujemy:

27, 66, 89, 72, 21, 17, 71, 69, 95, 17.

Zauważmy, że taflę szklaną o numerze 17 należało zbadać dwukrotnie. ■

Zwróćmy obecnie uwagę na przypadek, gdy zbiorowością generalną jest strumień wyrobu sztukowego, generowany przez agregat produkcyjny w pewnym ustalonym przedziale czasu τ . Jeżeli natężenie takiego strumienia nie jest zbyt duże, to produkowane elementy można jednoznacznie uporządkować według kolejności ich otrzymywania. W konsekwencji, do pobierania próby można

wykorzystać tablice liczb losowych. Należy w tym celu oszacować licznosc zbiorowosci generalnej za pomoca wzoru:

$$N \cong \gamma \Delta \tau, \quad (2.10)$$

w którym γ oznacza przeciętne natężenie strumienia produktu (czyli przeciętną liczbę sztuk wyrobu przypadającą na jednostkę czasu), natomiast $\Delta \tau$ jest długością przedziału czasu τ (okresu rozliczeniowego). W kolejnym etapie postępowania należy, posługując się tablicami liczb losowych, wygenerować n -elementowy zbiór liczb losowych, nie większych od N . Zbiór ten należy uporządkować za pomocą relacji „ \leq ”, a następnie można przystąpić do losowania elementów zbiorowości generalnej do próby.

Opisana procedura konstruowania próby nie może być zastosowana wówczas, gdy elementów strumienia wyrobu nie można uporządkować w kolejności ich otrzymywania. Z sytuacją taką mamy do czynienia w przypadku strumienia o dużych natężeniach γ . W takich przypadkach jednostkami losowania nie mogą być sztuki wyrobu, lecz chwile, w których należy pobierać elementy zbiorowości generalnej do próby. Do losowego wyboru owych chwil można wykorzystać tablice liczb losowych albo tablice tzw. liczb złotych, zaproponowanych przez H. Steinhausa⁸. Fragment tablicy liczb złotych zamieszczono w aneksie (tablica VIII). Sposób posługiwania się tymi tablicami ilustruje podany niżej przykład.

Przykład 2.5

Z obserwowanego strumienia wyrobu należało pobrać próbkę losową o liczności $n = 10$, w przedziale τ od godziny 10.00 do 14.00. W tym celu z tablicy VIII (aneks) odczytano 10 kolejnych liczb złotych. Ułamkowe części tych liczb (mantysy) zestawiono w kolumnie 2 tablicy 2.3. Wartości przedstawione w kolumnie 3 są iloczynami postaci $w \Delta \tau = w 4$, gdzie w jest mantysą liczby złotej, natomiast $\Delta \tau = 14.00 - 10.00 = 4$ oznacza długość przedziału czasu, w którym obserwowano proces produkcyjny. W kolumnach 5 i 6 tablicy 2.3 podano punkty na osi czasu zegarowego, w których należało pobrać kolejne elementy do próbki. ■

W przypadku gdy zbiorowość generalna ma postać strumienia kolejno generowanych elementów, przy konstruowaniu próbki można zastosować schemat losowania znany jako *losowanie systematyczne*. Polega ono na tym, że losowo wyznacza się początek sekwencji, a następnie w regularnych odstępach pobiera

⁸ Hugo Dionizy Steinhaus (1887–1972), matematyk i statystyk polski.

Tablica 2.3

Wyznaczanie za pomocą liczb złotych punktów pobierania elementów do próbki

i	Mantysa liczby złotej	Punkt na osi czasu w przedziale (0, 4)	(i)	Czas pobierania elementów	
				godziny	minuty
1	2	3	4	5	6
1	0,6180	2,4720	1	10	22
2	0,2361	0,9444	2	10	43
3	0,8541	3,4164	3	10	57
4	0,4721	1,8884	4	11	18
5	0,0902	0,3608	5	11	53
6	0,7082	3,8328	6	12	15
7	0,3262	1,3048	7	12	28
8	0,9443	3,7772	8	12	50
9	0,5623	2,2492	9	13	25
10	0,1803	0,7212	10	13	47

Źródło: opracowanie własne.

się kolejne elementy zbiorowości generalnej do próby. Jeśli elementy strumienia mogą być ponumerowane, to do próbki trafiają elementy o numerach:

$$z = z_0 + ck, \quad (2.11)$$

gdzie z_0 oznacza losowy początek sekwencji, k jest tzw. interwałem losowania, czyli stałym odstępem między elementami, natomiast $c = 0, 1, 2, \dots$. Jeśli elementy strumienia nie mogą być ponumerowane, to kolejne jednostki do próby pobiera się w chwilach wynikających ze wzoru:

$$t = t_0 + ck, \quad (2.12)$$

w którym t_0 oznacza losowy początek sekwencji punktów na osi czasu. Przedstawiona procedura pozwala zapewnić reprezentatywność próbki tylko wówczas, gdy mechanizm generujący elementy strumienia jest czysto losowy.

2.4. Zmienne losowe w badaniach empirycznych

W celu uzyskania możliwości operowania ilościowymi charakterystykami właściwości badanych obiektów i zjawisk, a także dla potrzeb numerycznych analiz statystycznych, cechy należące do zbioru \mathbf{B} (zob. wzór (2.2)) przekształca się w odpowiednie zmienne, czyli ich liczbowe obrazy. Zauważmy, że cecha jest pojęciem bardziej pierwotnym niż zmienna. Jednej cesze można przyporządkować

tyle zmiennych, ile różnych metod można zastosować do badania cechy B i ile różnych funkcji pomiarowych można zdefiniować w konkretnym przypadku⁹.

Każdej metodzie badania odpowiada pewien, charakterystyczny dla tej metody, zbiór rozróżnialnych stanów cechy B . Jest to zbiór skończony postaci:

$$\mathbf{B}^\circ = \{b_j^\circ; j = 1, 2, \dots, k\} \quad (2.13)$$

albo zbiór przeliczalny:

$$\mathbf{B}^\circ = \{b_j^\circ; j = 1, 2, \dots\}. \quad (2.14)$$

O szczegółowej postaci tego zbioru decyduje zarówno natura cechy B , jak i zastosowana metoda jej badania.

Odwzorowując zbiór rozróżnialnych stanów (\mathbf{B}°) w zbiór liczb rzeczywistych, za pomocą funkcji pomiarowej:

$$x_j^\circ = f(b_j^\circ), \quad (2.15)$$

otrzymujemy zbiór wartości odpowiedniej zmiennej losowej. Zbiór ten będziemy oznaczać symbolem \mathbf{X}° . Może to być – analogicznie jak \mathbf{B}° – albo zbiór skończony:

$$\mathbf{X}^\circ = \{x_j^\circ; j = 1, 2, \dots, k\} \quad (2.16)$$

albo zbiór przeliczalny:

$$\mathbf{X}^\circ = \{x_j^\circ; j = 1, 2, \dots\}. \quad (2.17)$$

Jeśli zastosowana metoda badania jest bardzo czuła, a więc jeśli ma ona zdolność rozróżniania bardzo mało różniących się stanów obiektów lub zjawisk, to \mathbf{X}° można interpretować jako przedział na osi liczb rzeczywistych. Zbiór (2.16) można wówczas zapisać następująco:

$$\mathbf{X}^\circ = \{x: x_d \leq x \leq x_g\}, \quad (2.18)$$

natomiast w przypadku zbioru (2.17) mamy:

$$\mathbf{X}^\circ = \{x: x \leq x_g\} \quad (2.19)$$

albo

$$\mathbf{X}^\circ = \{x: x \geq x_d\}. \quad (2.20)$$

⁹ Obszerniejsze omówienie tych problemów zainteresowany Czytelnik znajdzie w pracy A. Iwasiewicza: *Problemy niepełnej sprawności diagnostycznej w statystycznej kontroli jakości; studium metodologiczne*, Zeszyty Naukowe AE w Krakowie, Seria specjalna: Monografie, nr 80, Kraków 1987.

Podkreślmy, że traktowanie zbioru X° jako przedziału na osi liczb rzeczywistych jest zawsze wybiegiem formalnym, mającym na celu uzyskanie możliwości zastosowania ciągłego modelu matematycznego. Zbiory X° są z natury co najwyżej przeliczalne, albowiem każda metoda badania cechy B ma ograniczoną zdolność rozróżniania stanów tej cechy.

Przykład 2.6

Badano czas świecenia (T) żarówek pewnego typu (zob. rozdział 1, przykład 1.9). Zastosowana aparatura pomiarowa rejestrowała stan każdej żarówki (świeci, nie świeci) w odstępach 15-minutowych (0,25 h). Tak więc zbiór wartości (T°) obserwowanej zmiennej losowej T przedstawiał się następująco:

$$T^\circ = \{0; 0,25; 0,50; 0,75; \dots\}. \quad (i)$$

Ponieważ jednak $\Delta t = 0,25$ h jest wartością bardzo małą w porównaniu z przeciętnym czasem świecenia (w przykładzie 1.9 przyjęto, że wynosi on 800 h), zatem w celu uzyskania możliwości zastosowania rozkładu wykładniczego przyjęto, że:

$$T^\circ = \{t: t \geq 0\}. \quad (ii)$$

Nie zmienia to oczywiście faktu, że pewne wartości należące do przedziału $[0; +\infty)$ nie mogą się zrealizować. ■

W przypadku wielocechowego opisu badanych obiektów lub powtórzeń zjawiska, w rezultacie kwantyfikacji cech otrzymujemy wielowymiarowe zmienne losowe. Zwróćmy uwagę na przypadek, gdy każdy obiekt, lub każde powtórzenie zjawiska, badane jest ze względu na dwie cechy (B_1 i B_2). Każdej z tych cech odpowiada zbiór rozróżnialnych stanów (B_1°, B_2°) charakterystyczny dla zastosowanej metody badania. Dokonując kwantyfikacji każdej z tych cech otrzymujemy dwuwymiarową zmienną losową (X_1, X_2), taką że:

$$x_{1i}^\circ = f_1(b_{1i}^\circ), \quad (2.21)$$

$$x_{2j}^\circ = f_2(b_{2j}^\circ), \quad (2.22)$$

gdzie $b_{1i}^\circ \in B_1^\circ$ i $b_{2j}^\circ \in B_2^\circ$.

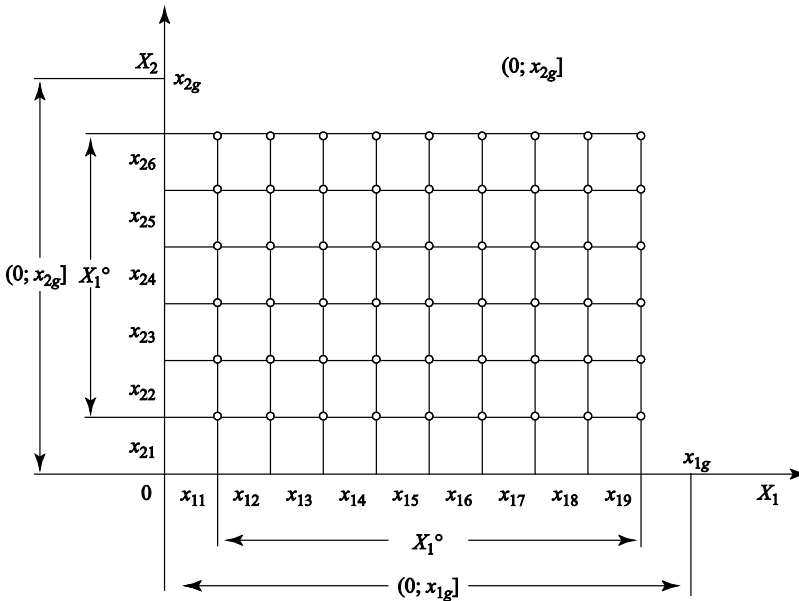
Zbiór wartości zmiennej losowej (X_1, X_2) jest iloczynem kartezjańskim zbiorów \mathbf{X}_1° i \mathbf{X}_2° . Mamy więc:

$$(\mathbf{X}_1^\circ, \mathbf{X}_2^\circ) = \mathbf{X}_1^\circ \times \mathbf{X}_2^\circ = \{(x_{1i}^\circ, x_{2j}^\circ) : x_{1i}^\circ \in \mathbf{X}_1^\circ, x_{2j}^\circ \in \mathbf{X}_2^\circ\}. \quad (2.23)$$

Przykład 2.7

Przy badaniu fizyko-mechanicznych właściwości dowolnego materiału konstrukcyjnego wykonuje się zwykle próbę statycznego rozciągania. Celem takiego badania jest wyznaczenie wartości kilku charakterystyk materiału, przede wszystkim wytrzymałości na rozciąganie (B_1) i wydłużenia względnego przy zerwaniu (B_2). W rezultacie kwantyfikacji obu tych cech otrzymujemy dwuwymiarową zmienną losową (X_1, X_2) . Zbiór wartości tej zmiennej przedstawiono schematycznie na rys. 2.1. Zauważmy, że teoretycznie zmienne X_1 i X_2 są ciągłe w przedziałach $(0; x_{1g})$ i $(0; x_{2g})$, gdzie x_{1g} i x_{2g} są największymi wartościami zmiennych X_1 i X_2 , charakterystycznymi dla badanego materiału. W konsekwencji, teoretycznie zbiorem wartości zmiennej losowej (X_1, X_2) jest iloczyn kartezjański:

$$(0; x_{1g}) \times (0; x_{2g}).$$



Rys. 2.1. Zbiór wartości dwuwymiarowej zmiennej losowej

Praktycznie jednak, skutkiem ograniczeń wnoszonych przez instrumentarium pomiarowe, zmienne X_1 i X_2 są quasi-ciągłe. W konsekwencji, zbiór wartości zmiennej losowej (X_1, X_2) redukuje się do postaci, którą schematycznie pokazano na rys. 2.1. ■

Rezultatem doświadczenia jest zbiór realizacji obserwowanej zmiennej losowej. W przypadku wyczerpujących badań zbiorowości generalnej A o licznosci N , ze względu na jedną cechę, otrzymujemy:

$$\mathbf{X}_N = \{x_j; j = 1, 2, \dots, N\}. \quad (2.24)$$

Jeśli badania mają charakter niewyczerpujący, to rezultatem doświadczenia jest zbiór realizacji postaci:

$$\mathbf{X}_n = \{x_i; i = 1, 2, \dots, n\}, \quad (2.25)$$

gdzie n oznacza licznosc próbki poddanej badaniu.

W przypadku dwuwymiarowej zmiennej losowej (X_1, X_2) mamy:

$$(\mathbf{X}_1, \mathbf{X}_2)_N = \{(x_{1j}, x_{2j}); j = 1, 2, \dots, N\} \quad (2.26)$$

albo

$$(\mathbf{X}_1, \mathbf{X}_2)_n = \{(x_{1i}, x_{2i}); i = 1, 2, \dots, n\}. \quad (2.27)$$

Elementami tych zbiorów są uporządkowane pary liczb otrzymane w rezultacie badania kolejnych elementów zbiorowości generalnej (A) albo próby (\underline{A}). Podane wzory można łatwo transformować do postaci odpowiadającej k -wymiarowej zmiennej losowej.

Każdy element zbioru realizacji zmiennej losowej, zarówno jednowymiarowej jak i wielowymiarowej, można identyfikować jako element odpowiedniego zbioru wartości. Jeśli plan doświadczenia jest logicznie spójny, to w rezultacie badania dowolnego elementu zbiorowości generalnej lub próby nie można uzyskać wyniku spoza zbioru wartości i jednocześnie zbiór ten nie może zawierać elementów, których uzyskanie w rezultacie doświadczenia byłoby niemożliwe.

2.5. Skale pomiarowe

Pomiar jest procesem empirycznym, polegającym na przyporządkowywaniu wartości liczbowych stanom obserwowanych cech, zrealizowanych w badanych obiektach lub powtórzeniach zjawiska. Stosowane w tym postępowaniu funkcje pomiarowe są więc funkcjami zbioro-liczbowymi. Dziedziną takiej funkcji jest

zbiór obiektów (powtórzeń zjawiska) lub ich stanów ze względu na badaną cechę, natomiast przeciwdziedzina jest pewien podzbiór zbioru liczb rzeczywistych. Funkcje pomiarowe muszą być tak skonstruowane, by zapewniały izomorfizm obiektowego systemu relacyjnego i liczbowego systemu relacyjnego. Relacje między liczbami uzyskiwanymi w rezultacie pomiaru muszą odwzorowywać relacje między badanymi obiektami lub powtórzeniami zjawiska. Jeśli warunek ten nie jest spełniony, to proces przyporządkowywania liczb nie jest pomiarem. Wynika stąd bezpośrednio, że jeśli system pomiarowy wykorzystywany w procesie pozyskiwania informacji nie jest właściwie skonstruowany, to zbiory realizacji zmiennych losowych (zob. punkt 2.4) nie dają poprawnego obrazu obiektu lub zjawiska¹⁰.

Wyróżnia się pięć typów skal pomiarowych, a mianowicie:

- skale nominalne,
- skale porządkowe,
- skale przedziałowe,
- skale ilorazowe,
- skale absolutne.

Skala nominalna stosowana jest wówczas, gdy stany $b_j^\circ \in \mathbf{B}^\circ$, rozróżniane przez zastosowaną metodę badania obiektu (zjawiska) ze względu na cechę B są tylko rozłącznymi kategoriami jakościowymi. W takiej sytuacji izomorfizm między obiektowym systemem relacyjnym i liczbowym systemem relacyjnym zostanie zachowany, jeśli funkcja pomiarowa zostanie tak skonstruowana, by:

- jednakowym obiektom lub powtórzeniom zjawiska przyporządkowywała jednakowe wartości liczbowe i jednocześnie by
- różnym obiektom (powtórzeniom zjawiska) przyporządkowywała różne wartości liczbowe.

Skala porządkowa stosowana jest wówczas, gdy stany $b_j^\circ \in \mathbf{B}^\circ$ są nie tylko rozłącznymi kategoriami jakościowymi, ale gdy są one także uporządkowane według rosnącej lub malejącej preferencji. Oznacza to, że między dowolnymi dwoma stanami $b_i^\circ, b_j^\circ \in \mathbf{B}^\circ$ zachodzi relacja poprzedzania $b_i^\circ < b_j^\circ$, albo $b_i^\circ > b_j^\circ$. Potrafimy więc zawsze wskazać, który stan jest bardziej pożądanym. Oceny szkolne są pomiarem wiedzy i umiejętności ucznia na skali porządkowej, przy czym w polskim systemie szkolnym funkcja pomiarowa jest tak skonstruowana, że stanowi bardziej preferowanemu przyporządkowuje się większą liczbę. Miejsca zajmowane przez zawodników na mecie wyścigu kolarskiego są

¹⁰ Obszerniejsze informacje na temat skal pomiarowych i ich zastosowań zainteresowany Czytelnik znajdzie w opracowaniach: A. Góralski, *Metody opisu i wnioskowania statystycznego w psychologii i pedagogice*, PWN, Warszawa 1987 oraz M. Walesiak, *Metody analizy danych marketingowych*, Wydawnictwo Naukowe PWN, Warszawa 1996.

również pomiarem ich wydolności i umiejętności na skali porządkowej. Obowiązuje tu jednak konwencja, że stanowi bardziej preferowanemu przyporządkowuje się mniejszą liczbę (numer zajętego miejsca).

Skala przedziałowa zachowuje wszystkie możliwości pomiarowe omówionych wyżej skal (nominalnej i porządkowej), a jednocześnie umożliwia pomiar dystansu między dowolnymi dwoma stanami $b_i^\circ, b_j^\circ \in \mathbf{B}^\circ$. Jest to możliwe dzięki wyposażeniu skali w stałą jednostkę miary i umowne zero. Fakt, że jest to zero umowne (a nie absolutne), decyduje o tym, że wyniki pomiarów na skali przedziałowej nie mogą stanowić podstawy do oceny stosunku mierzonych wielkości. Nie jest to możliwe, albowiem zmiana położenia owego umownego zera na osi badanej zmiennej zmienia stosunki między liczbami otrzymywanymi w rezultacie pomiaru, mimo że relacje między odpowiednimi stanami badanej cechy pozostają nie zmienione.

Przykład 2.8

Skala termometryczna Celsjusza jest skalą przedziałową. Dlatego też jeśli temperatura jakiegoś ciała wzrosła z 20°C do 40°C , oznacza to tylko, że wzrosła ona o 20°C . Nie jest natomiast prawdą, że temperatura tego ciała wzrosła dwukrotnie, mimo że $40/20 = 2$. Zauważmy, że gdybyśmy przesunęli umowne zero do punktu oznaczającego 10°C , to doszlibyśmy do wniosku, że temperatura wzrosła trzykrotnie, mimo że fizyczny (obiektywny) układ relacyjny pozostałby nie zmieniony. Gdybyśmy byli zainteresowani oceną stosunku przytoczonych powyżej poziomów temperatury, to powinniśmy posłużyć się skalą termometryczną Kelvina. Mielibyśmy wówczas:

$$(273^\circ\text{C} + 40^\circ\text{C}) / (273^\circ\text{C} + 20^\circ\text{C}) = 313\text{K} / 293\text{K} \cong 1,07.$$

Temperatura badanego ciała wzrosła więc tylko 1,07 raza, czyli o 7%. ■

Skala ilorazowa zachowuje wszystkie możliwości pomiarowe skali przedziałowej, a dodatkowo – dzięki wyposażeniu jej w absolutne zero – umożliwia ocenę stosunków między mierzonymi wielkościami. Skalą ilorazową jest – na przykład – skala termometryczna Kelvina.

Przedstawione powyżej skale pomiarowe charakteryzują się różną mocą. Zależy ona od tego, jak liczna jest klasa dopuszczalnych transformacji funkcji pomiarowej, czyli takich przekształceń tej funkcji, które nie niszczą izomorfizmu między obiektywnym systemem relacyjnym i liczbowym systemem relacyjnym. Tak pojmowana moc skali jest najniższa w przypadku skali nominalnej, albowiem dana funkcja pomiarowa może tu być zastąpiona dowolną inną funkcją

różnowartościową. Skala porządkowa jest mocniejsza niż skala nominalna, gdyż dopuszczalne transformacje funkcji pomiarowej muszą spełniać dodatkowe warunki, wynikające z relacji porządku ($<$) w zbiorze obiektów (powtórzeń zjawiska) i konieczności zachowania tego porządku ($<$, $>$) w odpowiednim zbiorze liczb. W przypadku skali przedziałowej klasa dopuszczalnych przekształceń funkcji pomiarowej ulega dalszemu ograniczeniu. Dopuszczalne są tylko transformacje postaci $x' = a + cx$, gdzie x jest wynikiem pomiaru na skali pierwotnej, natomiast x' oznacza wynik pomiaru na skali przekształconej, przy czym $a \in \mathbf{R}$ i $c \in \mathbf{R}^+$. Skala przedziałowa ma więc większą moc niż skala porządkowa, a tym bardziej niż skala nominalna. W przypadku skali ilorazowej klasa dopuszczalnych transformacji ograniczona jest do przekształceń postaci $x' = cx$, $c \in \mathbf{R}^+$. Jest to więc skala najmocniejsza spośród dotychczas omówionych. Zauważmy, że klasę transformacji charakterystyczną dla skali ilorazowej otrzymujemy przez podstawienie $a = 0$ w równaniu $x' = a + cx$, odpowiadającemu skali przedziałowej. Dalsze ograniczenie klasy dopuszczalnych przekształceń można osiągnąć przyjmując $c = 1$. Po dokonaniu takiego podstawienia mamy $x' = x$, czyli tzw. przekształcenie identycznościowe, charakterystyczne dla skali absolutnej. Wynik pomiaru na takiej skali nie może podlegać żadnym przekształceniom, jeśli ma zostać zachowana informacja o badanym obiekcie lub zjawisku. Przykładem tego typu pomiaru jest zliczanie głosów w wyborach, zliczanie klientów wchodzących do sklepu w określonym przedziale czasu, zliczanie wad w jednostce produktu itp. Jakakolwiek transformacja uzyskanego wyniku niszczy izomorfizm między obiektywnym systemem relacyjnym a odpowiednim liczbowym systemem relacyjnym.

Rozdział 3

CHARAKTERYSTYKI OPISOWE ZMIENNYCH LOSOWYCH W BADANIACH WYCZERPUJĄCYCH

3.1. Wprowadzenie

W rozważaniach przedstawionych w tym rozdziale będziemy zakładać, że skończona zbiorowość generalna A (o elementach A_i ; $i = 1, 2, \dots, N$) poddawana jest badaniu wyczerpującemu. Omówimy przede wszystkim charakterystyki opisowe jednowymiarowych zmiennych losowych, a spośród zmiennych wielowymiarowych najwięcej uwagi poświęcimy zmiennym losowym dwuwymiarowym, a także trójwymiarowym.

W przypadku jednowymiarowej zmiennej losowej X , w rezultacie doświadczenia otrzymujemy zbiór realizacji postaci (2.26). Zbiór ten zapiszemy obecnie następująco:

$$\mathbf{X}_N = \{x_i; \quad i = 1, 2, \dots, N\}.$$

Otrzymany w rezultacie doświadczenia zbiór realizacji k -wymiarowej zmiennej losowej (X_1, X_2, \dots, X_k) ma postać:

$$(\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_k)_N = \{(x_{1i}, x_{2i}, \dots, x_{ki}); \quad i = 1, 2, \dots, N\},$$

przy czym wartości $x_{1i}, x_{2i}, \dots, x_{ki}$ są realizacjami zmiennych losowych X_1, X_2, \dots, X_k w tym samym elemencie $A_i \in A$.

Jeśli obserwowana zmienna losowa jest dwuwymiarowa ($k = 2$), to zbiór realizacji przybiera postać (2.34). Tak więc dwuwymiarowej zmiennej losowej (Y, X) odpowiada zbiór realizacji:

$$(\mathbf{Y}, \mathbf{X})_N = \{(y_i, x_i); \quad i = 1, 2, \dots, N\}.$$

W przypadku trójwymiarowej zmiennej losowej (Y, X_1, X_2) mamy natomiast:

$$(\mathbf{Y}, \mathbf{X}_1\mathbf{X}_2)_N = \{(y_i, x_{1i}, x_{2i}); \quad i = 1, 2, \dots, N\}.$$

Informacje zawarte w tych zbiorach nie są obciążone błędami związanymi z pobieraniem próbki, albowiem – zgodnie z przyjętym założeniem – badaniu podlega cała zbiorowość A . Wyczerpujący charakter badań nie uwalnia natomiast podanych powyżej zbiorów od błędów generowanych w procesie pozyskiwania informacji, a więc od błędów pojawiających się podczas rozpoznawania stanu kolejnych elementów $A_i \in A$, a także podczas przyporządkowywania tym elementom wartości liczbowych. Ma to ten skutek, że przy wielokrotnym powtarzaniu wyczerpującego badania zbiorowości A , w kolejnych krokach postępowania nie należy oczekiwać takich samych zbiorów realizacji postaci X_N albo $(X_1, X_2)_N$. Zjawisko to nie występuje tylko w najprostszych układach doświadczalnych. Jeśli – na przykład – doświadczenie polegałoby na rozsegregowaniu kul białych i czarnych, to można oczekiwać, że wszystkie kolejne powtórzenia badania zakończą się takim samym rezultatem. Jeśli jednak segregowane kule mało różniłyby się kolorem (na przykład dwa odcienie szarości), to jest rzeczą niemal pewną, że w kolejnych powtórzeniach doświadczenia otrzymywalibyśmy różne zbiory realizacji, ogólnej postaci X_N . Tak więc, jeśli wyłączyć z pola rozważań pewne ekstremalnie uproszczone układy doświadczalne, to należy przyjąć, że badanie wyczerpujące dostarcza informacji o *empirycznym rozkładzie* obserwowanej zmiennej losowej w przeprowadzonym doświadczeniu. Wynika stąd bezpośrednio, że rola metod statystycznych w badaniach wyczerpujących sprowadza się do syntetycznego opisu wyników eksperymentu.

3.2. Jednowymiarowe zmienne losowe

3.2.1. Rozkład empiryczny

Pod pojęciem rozkładu empirycznego, albo rozkładu liczebności lub rozkładu częstości, będziemy rozumieć funkcję odwzorowującą zbiór realizacji X_N w odpowiedni zbiór wartości X° . Funkcja ta opisuje regułę rządzącą procesem rozdziału (dystrybucji) realizacji $x_i \in X_N$ pomiędzy poszczególne wartości $x_j^\circ \in X^\circ$ albo pomiędzy wyróżnione podzbiory zbioru X° . Do opisu tej reguły wykorzystuje się dwojakiemu rodzaju charakterystyki, a mianowicie szereg rozdzielczy oraz szereg kumulacyjny. Pojęciowo odpowiadają one charakterystykom rozkładu prawdopodobieństwa zmiennej losowej, omówionym w rozdziale 1. Szereg rozdzielczy jest empirycznym odpowiednikiem funkcji prawdopodobieństwa i funkcji gęstości prawdopodobieństwa, natomiast szereg kumulacyjny jest empirycznym analogiem dystrybuanty.

3.2.1.1. Szereg rozdzielczy

Zbiór realizacji \mathbf{X}_N nazywany jest w statystyce opisowej szczegółowym szeregiem statystycznym¹. Każdy element tego szeregu można identyfikować jako element zbioru wartości \mathbf{X}° , albowiem:

$$\bigwedge_i \bigvee_j x_i = x_j^\circ, \quad x_i \in \mathbf{X}_N, x_j^\circ \in \mathbf{X}^\circ. \quad (3.1)$$

W konsekwencji, w przypadku dyskretnej zmiennej losowej szczegółowy szereg statystyczny \mathbf{X}_N można transformować do postaci:

$$\begin{array}{l|l} x_1^\circ & f_1 = f(x_1^\circ) \\ x_2^\circ & f_2 = f(x_2^\circ) \\ \vdots & \vdots \\ x_j^\circ & f_j = f(x_j^\circ) \\ \vdots & \vdots \\ x_k^\circ & f_k = f(x_k^\circ) \end{array} \quad (3.2)$$

gdzie $f_j = f(x_j^\circ)$ oznacza liczbę realizacji zmiennej losowej X w zbiorze \mathbf{X}_N o wartościach równych $x_j^\circ; j = 1, 2, \dots, k$.

Zauważmy, że zachodzi następująca zależność:

$$\sum_{j=1}^k f_j = N. \quad (3.3)$$

Liczebności f_j (nazywane też frekwencjami) można zastąpić w szeregu (3.2) odpowiednimi częstościami względnymi:

$$v(x_j^\circ) = v_j = \frac{f_j}{\sum_{j=1}^k f_j} = \frac{f_j}{N}, \quad j = 1, 2, \dots, k, \quad (3.4)$$

takimi, że:

$$\sum_{j=1}^k v_j = 1. \quad (3.5)$$

¹ Zauważmy, że określenie „szereg” ma w statystyce inne znaczenie niż w analizie matematycznej.

Szereg rozdzielczy (3.2) można więc zapisać następująco:

$$\begin{array}{c|c}
 x_1^\circ & v_1 = v(x_1^\circ) \\
 x_2^\circ & v_2 = v(x_2^\circ) \\
 \vdots & \vdots \\
 x_j^\circ & v_j = v(x_j^\circ) \\
 \vdots & \vdots \\
 x_k^\circ & v_k = v(x_k^\circ)
 \end{array} \tag{3.6}$$

Przykład 3.1

Przedmiotem badania była partia pewnego produktu przemysłowego o liczności $N = 40$. Każdą sztukę poddano badaniu, którego celem było wykrycie wad technologicznych. Liczba wad w danej sztuce jest realizacją obserwowanej zmiennej losowej X . Wyniki badań – czyli zbiór realizacji \mathbf{X}_N – przedstawiono w tabelicy 3.1.

Tablica 3.1

Wyniki badania liczby wad technologicznych w jednostce produktu

i	x_i	i	x_i	i	x_i	i	x_i
1	2	1	2	1	2	1	2
1	0	11	1	21	0	31	1
2	0	12	0	22	0	32	2
3	2	13	0	23	0	33	0
4	1	14	1	24	1	34	0
5	0	15	0	25	1	35	0
6	0	16	3	26	1	36	0
7	0	17	0	27	0	37	1
8	1	18	0	28	0	38	0
9	0	19	1	29	1	39	0
10	0	20	0	30	1	40	2

Źródło: opracowanie własne.

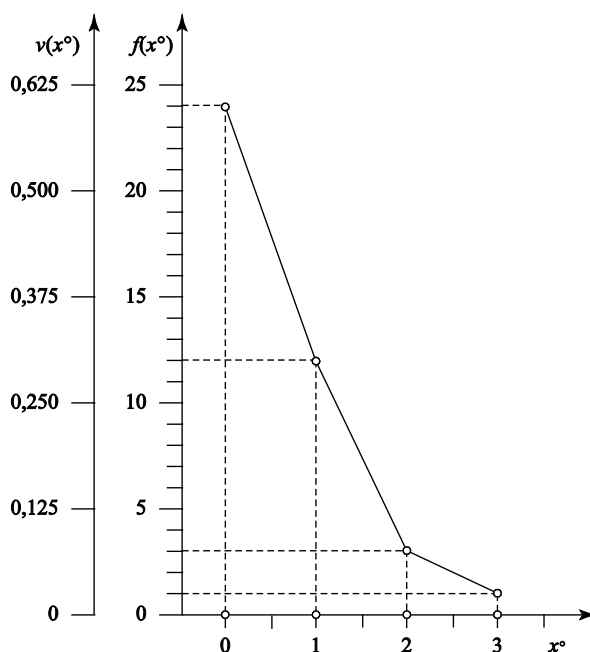
Zbiór wartości ma postać $X^\circ = \{0, 1, 2, 3\}$. Zauważmy, że zbiór ten można określić dopiero po zakończeniu badań, albowiem nie ma żadnych teoretycznych przesłanek, pozwalających wskazać największą liczbę wad technologicznych w jednostce produktu. Identyfikując kolejne wartości x_i z tabelicy 3.1, jako elementy zbioru X° , sporządzono tabelicę 3.2. Kolumny 2 i 3 przedstawiają szereg rozdzielczy postaci (3.2), natomiast kolumny 2 i 4 przedstawiają szereg roz-

dzielnicy (3.6). Graficzną ilustracją tych obu szeregów jest rys. 3.1. Wykres pokazany na tym rysunku nazywany jest wielobokiem częstości względnych. ■

Tablica 3.2
Szeregi rozdzielcze liczby wad technologicznych w jednostce produktu;
szereg szczegółowy w tablicy 3.1

j	x_j°	f_j	v_j
1	2	3	4
1	0	24	0,600
2	1	12	0,300
3	2	3	0,075
4	3	1	0,025
Σ	*	40	1,000

Źródło: opracowanie własne.



Rys. 3.1. Rozkład liczebności i rozkład częstości względnych dyskretnej zmiennej losowej

W przypadku gdy zbiór X° ma postać (2.18), a więc gdy obserwowana zmienna losowa X jest ciągła w przedziale $[x_d, x_g]$, przedstawione powyżej postępowanie nie może być zrealizowane, albowiem każdy przedział $[x_d, x_g] \subset \mathbf{R}$ jest zbiorem nieprzeliczalnym. W takiej sytuacji konstruowanie szeregu rozdzielczego przebiega według podanego poniżej schematu.

– Należy znaleźć wartość najmniejszą (x_{\min}) oraz wartość największą (x_{\max}) w zbiorze realizacji X_N . Wartości te wyznaczają przedział zmienności $[x_{\min}, x_{\max}] \subset [x_d, x_g]$, o długości $x_{\max} - x_{\min}$.

– Przedział zmienności należy rozciąć na k przedziałów klasowych o długości l . Wartość k należy wyznaczyć według następujących wzorów²:

$$k' = 1 + 3,3 \lg_{10} N, \quad (3.7)$$

$$0,5\sqrt{N} \leq k \leq \sqrt{N}. \quad (3.8)$$

W przypadku wzoru (3.7) wartość k uzyskuje się przez zaokrąglenie k' do najbliższej liczby całkowitej.

– Jeśli iloraz:

$$\frac{x_{\max} - x_{\min}}{k} \quad (3.9)$$

ma wartość utrudniającą dalsze operacje numeryczne, to należy skorygować końce przedziału zmienności. Kres dolny można skorygować do poziomu x'_{\min} , takiego że $x_d < x'_{\min} < x_{\min}$, natomiast górny do poziomu x'_{\max} , spełniającego nierówność $x_{\max} < x'_{\max} < x_g$. Korekta ta powinna być możliwie najmniejsza. Można również korygować tylko kres dolny albo tylko kres górny.

– Mając wartości x'_{\min} i x'_{\max} można wyznaczyć długość przedziału klasowego według wzoru:

$$l = \frac{x'_{\max} - x'_{\min}}{k}. \quad (3.10)$$

² Wzór (3.7) zaczerpnięto z książki: A.K. Govil, *Definitions and Formulae in Statistics*, 2. ed., Macmillan Press, London 1984, s. 2. W książce: S.C. Puri, D. Ennis, K. Mullen, *Statistical Quality Control for Food and Agricultural Scientists*, G.K. Hall and Co., Boston, Massachusetts 1979, s. 21, zaproponowano zależność $k = \sqrt{N}$. Wzór (3.8) stanowi rozwiniętą formę tej propozycji. Wzór ten zaczerpnięto z książki: D. Bobrowski, *Probabilistyka w zastosowaniach technicznych*, WNT, Warszawa 1980, s. 41.

– Kolejne przedziały klasowe mają postać:

$$(x_{dj}; x_{gj}], \quad (3.11)$$

przy czym: $j = 1, 2, \dots, k$; $x_{d,1} = x'_{\min}$; $x_{g,k} = x'_{\max}$; $x_{g,j} - x_{d,j} = 1$.

– Kolejne realizacje $x_i \in X_N$ zalicza się do odpowiednich przedziałów klasowych. Realizacja x_i należy do j -tego przedziału klasowego, gdy spełniona jest podwójna nierówność:

$$x_{dj} < x_i \leq x_{gj}. \quad (3.12)$$

W rezultacie takiego postępowania otrzymuje się szereg rozdzielczy postaci:

$(x_{d,1}; x_{g,1}]$	f_1	(3.13)
$(x_{d,2}; x_{g,2}]$	f_2	
\vdots	\vdots	
$(x_{d,j}; x_{g,j}]$	f_j	
\vdots	\vdots	
$(x_{d,k}; x_{g,k}]$	f_k	

Zastępując liczebności f_j odpowiednimi częstościami względnymi v_j mamy:

$(x_{d,1}; x_{g,1}]$	v_1	(3.14)
$(x_{d,2}; x_{g,2}]$	v_2	
\vdots	\vdots	
$(x_{d,j}; x_{g,j}]$	v_j	
\vdots	\vdots	
$(x_{d,k}; x_{g,k}]$	v_k	

gdzie v_j wynika ze wzoru (3.4).

Przykład 3.2

Badaniu poddano $N = 36$ plastyfikatów polichlorowinyłowych, wykonanych metodą odlewania i żelatynizacji bezcisnieniowej, w temperaturze 443 K, w czasie 30 minut. Badano doraźną wytrzymałość na rozciąganie (R_r). Otrzymane wyniki pomiarów (kG/cm^2) przedstawiono w tablicy 3.3.

Tablica 3.3

Wyniki pomiarów doraźnej wytrzymałości na rozciąganie plastyfikatów polichlorowinyowych

i	x_i	i	x_i	i	x_i
	R_r (kG/cm ²)		R_r (kG/cm ²)		R_r (kG/cm ²)
	2	1	2	1	2
1	73,70	13	77,43	25	84,35
2	61,64	14	76,47	26	78,70
3	78,24	15	72,22	27	67,88
4	67,68	16	77,03	28	71,20
5	87,28	17	75,36	29	75,48
6	86,39	18	68,12	30	79,64
7	81,98	19	77,04	31	77,65
8	73,41	20	74,20	32	75,57
9	86,43	21	77,85	33	82,49
10	76,00	22	88,21	34	87,50
11	83,82	23	77,27	35	77,40
12	72,34	24	82,60	36	78,81

Źródło: badania własne.

Tablica 3.4

Konstruowanie szeregu rozdzielczego ciągłej zmiennej losowej

j	$(x_{d,j}; x_{g,j}]$	Zaliczanie obserwacji do przedziałów	f_j	v_j
1	2	3	4	5
1	(61,50; 66,00]		1	0,028
2	(66,00; 70,50]	┌	3	0,083
3	(70,50; 75,00]	▣	6	0,167
4	(75,00; 79,50]	▣ ▣ ▣	15	0,416
5	(79,50; 84,00]	▣	5	0,139
6	(84,00; 88,50]	▣	6	0,167
Σ	*	*	36	1,000

Źródło: opracowanie własne.

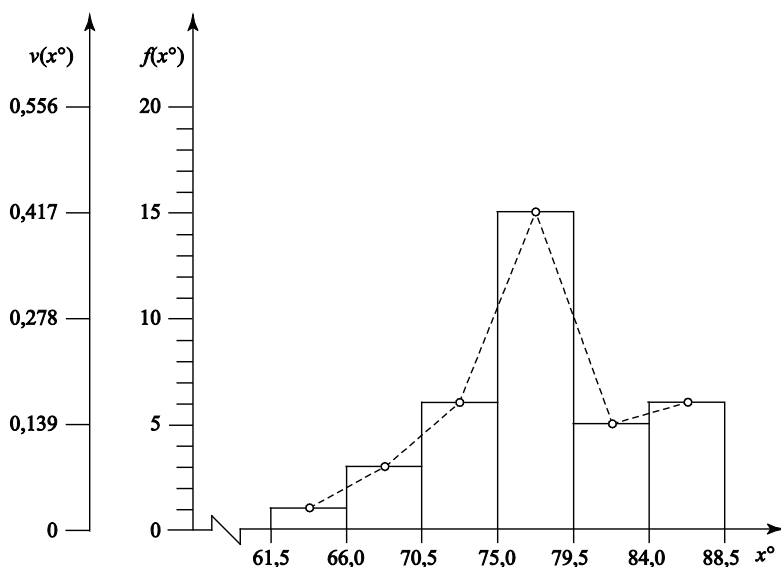
Badaną zmienną losową oznaczono symbolem X , a jej realizacje w poszczególnych plastyfikatach symbolem x_i , przy czym $i = 1, 2, \dots, 36$. Tablica 3.3 przedstawia zbiór X_N , czyli szczegółowy szereg statystyczny. Z tablicy wynika,

że $x_{\min} = 61,64$ oraz $x_{\max} = 88,21$. Wartość x_{\min} zaokrąglamy do $x'_{\min} = 61,50$, natomiast wartość x_{\max} do $x'_{\max} = 88,50$. Wykorzystując wzory (3.7) i (3.10) mamy:

$$k = 1 + 3,3 \lg_{10} 36 \approx 6,1358 \approx 6,$$

$$l = \frac{88,50 - 61,50}{6} = 4,50.$$

Technikę konstruowania szeregu rozdzielczego ciągłej zmiennej losowej X pokazano w tablicy 3.4. Kolumny 2 i 4 przedstawiają szereg rozdzielczy postaci (3.13), natomiast kolumny 2 i 5 to szereg rozdzielczy postaci (3.14). Na rys. 3.2 pokazano histogram (wykres słupkowy), będący graficznym obrazem obu szeregów rozdzielczych. Przerywaną linią zaznaczono też odpowiedni wielobok liczebności (wielobok częstości względnych).



Rys. 3.2. Rozkład liczebności i rozkład częstości względnych ciągłej zmiennej losowej

3.2.1.2. Szereg kumulacyjny

Szereg kumulacyjny (albo skumulowany) można skonstruować zarówno w oparciu o szereg szczegółowy, jak i szereg rozdzielczy. Jeśli punktem wyjścia jest szereg rozdzielczy dyskretnej zmiennej losowej, to szereg kumulacyjny ma postać:

x_1°	f_1	(3.15)
x_2°	$f_1 + f_2$	
\vdots	\vdots	
x_j°	$f_1 + f_2 + \dots + f_j$	
\vdots	\vdots	
x_k°	$f_1 + f_2 + \dots + f_j + \dots + f_k = N$	

albo

x_1°	v_1	(3.16)
x_2°	$v_1 + v_2$	
\vdots	\vdots	
x_j°	$v_1 + v_2 + \dots + v_j$	
\vdots	\vdots	
x_k°	$v_1 + v_2 + \dots + v_j + \dots + v_k = 1$	

Każdy element szeregu (3.15) informuje, ile realizacji obserwowanej zmiennej losowej X ma wartości nie przekraczające x_j° , natomiast każdy element szeregu (3.16) odpowiada na pytanie, jaka frakcja realizacji nie przekracza wartości x_j° . W przypadku ciągłej zmiennej losowej mamy:

$(x_{d,1}; x_{g,1}]$	f_1	(3.17)
$(x_{d,2}; x_{g,2}]$	$f_1 + f_2$	
\vdots	\vdots	
$(x_{d,j}; x_{g,j}]$	$f_1 + f_2 + \dots + f_j$	
\vdots	\vdots	
$(x_{d,k}; x_{g,k}]$	$f_1 + f_2 + \dots + f_j + \dots + f_k = N$	

$(x_{d,1}; x_{g,1}]$	v_1	(3.18)
$(x_{d,2}; x_{g,2}]$	$v_1 + v_2$	
\vdots	\vdots	
$(x_{d,j}; x_{g,j}]$	$v_1 + v_2 + \dots + v_j$	
\vdots	\vdots	
$(x_{d,k}; x_{g,k}]$	$v_1 + v_2 + \dots + v_j + \dots + v_k = 1$	

Kolejne elementy szeregu (3.17) informują, ile realizacji obserwowanej zmiennej losowej X nie przewyższa górnego kresu j -tego przedziału klasowego, czyli wartości $x_{g,j}$; $j = 1, 2, \dots, k$. Kolejne elementy szeregu (3.18) odpowiadają na pytanie, jaka frakcja realizacji nie przewyższa wartości $x_{g,j}$. Jeśli szereg kumulacyjny konstruowany jest w oparciu o szczegółowy szereg statystyczny, to postępowanie jest dwuetapowe. W pierwszym etapie porządkuje się zbiór realizacji X_N za pomocą relacji „ \leq ”. W rezultacie otrzymuje się tzw. szereg pozycyjny, czyli zbiór:

$$\{x_{(i)}; (i) = 1, 2, \dots, N\}, \quad (3.19)$$

taki że:

$$x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(i)} \leq \dots \leq x_{(N)}. \quad (3.20)$$

Szereg ten rozcina się na ciągi realizacji $x_{(i)}$, o takich samych wartościach. W każdym takim ciągu wyznacza się największą wartość wskaźnika (i) . Wartość tę oznaczmy symbolem (\underline{i}) . Następnie konstruuje się zbiór postaci:

$$\{[x_{(\underline{i})}, (\underline{i})]; (\underline{i}) \in \mathbf{I}\} \quad (3.21)$$

albo zbiór:

$$\{[x_{(\underline{i})}, \frac{(\underline{i})}{N}]; (\underline{i}) \in \mathbf{I}\}, \quad (3.22)$$

gdzie \mathbf{I} jest zbiorem wartości wskaźnika i . Jeśli w szeregu pozycyjnym (3.19) nie ma ciągów realizacji $x_{(i)}$ o takich samych wartościach, to wszystkie wartości wskaźnika (i) oraz (\underline{i}) są sobie równe. Szereg kumulacyjny (3.22) można stransformować do postaci określonej jako dystrybuanta empiryczna. Jeśli przyjąć definicję dystrybuanty wyrażoną wzorem (1.37), to wówczas dystrybuanta empiryczna ma postać:

$$F_N(x) = \begin{cases} 0 & \text{dla } x < x_{(i_1)} \\ \frac{i_r}{N} & \text{dla } x_{(i_r)} \leq x < x_{(i_{r+1})}, \\ 1 & \text{dla } x \geq x_{(i_k)} \end{cases} \quad (3.23)$$

gdzie (i_r) oznacza największą wartość wskaźnika (i) w r -tym ciągu realizacji $x_{(i)}$ o takich samych wartościach, natomiast $r = 1, 2, \dots, k$.

Jeśli w szeregu pozycyjnym (3.19) nie występują ciągi realizacji $x_{(i)}$ o takich samych wartościach, to dystrybuanta empiryczna ma postać:

$$F_N(x) = \begin{cases} 0 & \text{dla } x < x_{(1)} \\ \frac{(i)}{N} & \text{dla } x_{(i)} \leq x < x_{(i+1)}, \\ 1 & \text{dla } x \geq x_{(N)} \end{cases} \quad (3.24)$$

przy czym $(i) = 1, 2, \dots, N$.

Przykład 3.3

Skonstruujemy szeregi kumulacyjne i dystrybuantę empiryczną w oparciu o pierwotne wyniki obserwacji zestawione w tablicy 3.1. W rezultacie uporządkowania zbioru X_N (tablica 3.1) za pomocą relacji „ \leq ” otrzymano szereg pozycyjny przedstawiony w tablicy 3.5. Stosując wzory (3.21) i (3.22) otrzymujemy:

$$\{(0; 24), (1; 36), (2; 39) (3; 40)\},$$

$$\{(0; 24/40), (1; 36/40), (2; 39/40), (3; 40/40)\}.$$

Przy konstruowaniu tych szeregów kumulacyjnych uwzględniono fakt, że w tablicy 3.5:

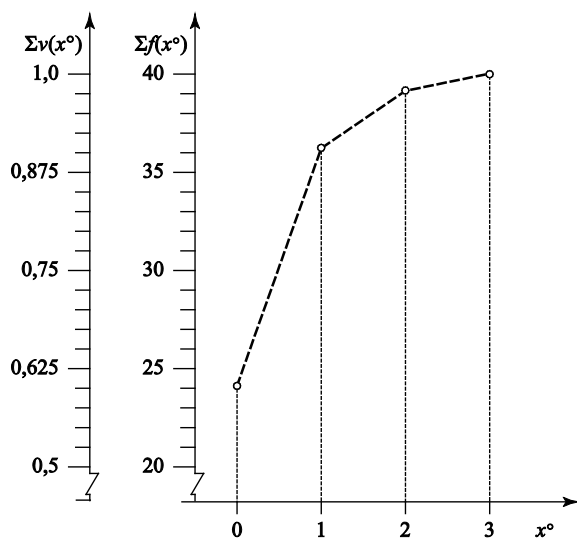
- najwyższa wartość (i) w ciągu par $[0; (i)]$ wynosi $(\underline{i}) = 24$,
- najwyższa wartość (i) w ciągu par $[1; (i)]$ wynosi $(\underline{i}) = 36$,
- najwyższa wartość (i) w ciągu par $[2; (i)]$ wynosi $(\underline{i}) = 39$,
- podzbiór par $[3; (i)]$ jest jednoelementowy i w konsekwencji mamy $(i) = (\underline{i}) = 40$.

Graficznym obrazem skonstruowanych szeregów kumulacyjnych jest wykres przedstawiony na rys. 3.3. Wykorzystując wzór (3.23) otrzymujemy następującą postać dystrybuanty empirycznej:

Tablica 3.5
Szereg pozycyjny dyskretnej zmiennej losowej; dane pierwotne w tablicy 3.1

(i)	i	$x_{(i)}$	(i)	i	$x_{(i)}$	(i)	i	$x_{(i)}$	(i)	i	$x_{(i)}$
1	2	3	1	2	3	1	2	3	1	2	3
1	1	0	11	17	0	21	35	0	31	25	1
2	2	0	12	18	0	22	36	0	32	26	1
3	5	0	13	20	0	23	38	0	33	29	1
4	6	0	14	21	0	24	39	0	34	30	1
5	7	0	15	22	0	25	4	1	35	31	1
6	9	0	16	23	0	26	8	1	36	37	1
7	10	0	17	27	0	27	11	1	37	3	2
8	12	0	18	28	0	28	14	1	38	32	2
9	13	0	19	33	0	29	19	1	39	40	2
10	15	0	20	34	0	30	24	1	40	16	3

Źródło: opracowanie własne.

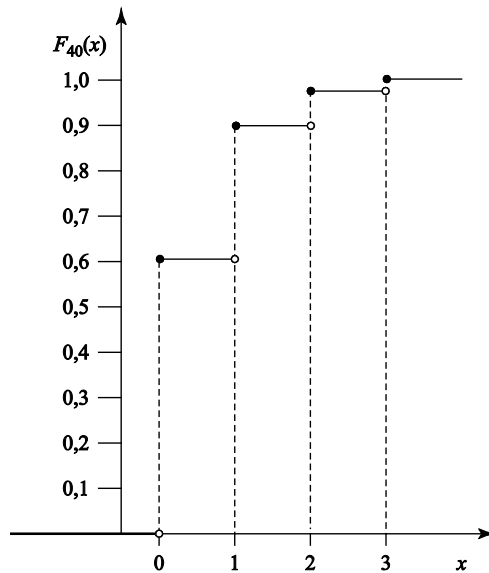


Rys. 3.3. Szeregi kumulacyjne dyskretnej zmiennej losowej

$$F_{40}(x) = \begin{cases} 0 & \text{dla } x < 0 \\ 24/40 = 0,6 & \text{dla } 0 \leq x < 1 \\ 36/40 = 0,9 & \text{dla } 1 \leq x < 2 \\ 39/40 = 0,975 & \text{dla } 2 \leq x < 3 \\ 40/40 = 1 & \text{dla } x \geq 3 \end{cases}$$

W tabelicy 3.5 występują $k = 4$ ciągi realizacji $x_{(i)}$ o takich samych wartościach, a mianowicie:

- $r = 1$; 24-elementowy ciąg zer; $\underline{i}_1 = 24$,
- $r = 2$; 12-elementowy ciąg jedynek; $\underline{i}_2 = 36$,
- $r = 3$; 3-elementowy ciąg dwójek; $\underline{i}_3 = 39$,
- $r = 4$; 4-elementowy ciąg trójek; $\underline{i}_4 = 40$.



Rys. 3.4. Dystrybuanta empiryczna

Wykres dystrybuanty empirycznej pokazano na rys. 3.4. ■

Przykład 3.4

Skonstruujemy obecnie szeregi kumulacyjne ciągłej zmiennej losowej. Wykorzystamy do tego celu zbiór X_N przedstawiony w tabelicy 3.3.

Porządkując zbiór X_N za pomocą relacji „ \leq ” otrzymujemy szereg pozycyjny przedstawiony w tabelicy 3.6. W szeregu tym nie ma realizacji o powtarzających się wartościach. Jest to skutek ciągłości badanej zmiennej losowej. Stosując wzory (3.21) i (3.22) mamy więc:

$$\{(61,64; 1), (67,68; 2), (67,88; 3), \dots, (88,21; 36)\}$$

oraz

$$\{(61,64; 1/36), (67,68; 2/36), (67,88; 3/36), \dots, (88,21; 36/36)\}.$$

Tablica 3.6

Szereg pozycyjny ciągłej zmiennej losowej; dane pierwotne w tabelicy 3.3

(i)	i	$x_{(i)}$	(i)	i	$x_{(i)}$	(i)	i	$x_{(i)}$
1	2	3	1	2	3	1	2	3
1	2	61,64	13	32	75,57	25	36	78,81
2	4	67,68	14	10	76,00	26	30	79,64
3	27	67,88	15	14	76,47	27	7	81,98
4	18	68,12	16	16	77,03	28	33	82,49
5	28	71,20	17	19	77,04	29	11	83,82
6	15	72,22	18	23	77,27	30	25	84,35
7	12	72,34	19	35	77,40	31	24	84,60
8	8	73,41	20	13	77,43	32	6	86,39
9	1	73,70	21	31	77,65	33	9	86,43
10	20	74,20	22	21	77,85	34	5	87,28
11	17	75,36	23	3	78,24	35	34	87,50
12	29	75,48	24	26	78,70	36	22	88,21

Źródło: opracowanie własne.

W tabelicy 3.7 pokazano proces konstruowania szeregu kumulacyjnego ciągłej zmiennej losowej w oparciu o szereg rozdzielczy. Uzyskane rozwiązania odbiegają oczywiście od szeregów kumulacyjnych skonstruowanych na podstawie danych pierwotnych. Jest to konsekwencja grupowania realizacji obserwowanej zmiennej losowej podczas konstruowania szeregu rozdzielczego. Tak więc szeregi kumulacyjne przedstawione w tabelicy 3.7 należy traktować jako rozwiązania przybliżone. Kryterium poprawności takich rozwiązań stanowi ich

niesprzeczność z rozwiązaniem dokładnym. W rozważanym przypadku warunek ten jest spełniony. Graficzną ilustracją szeregów kumulacyjnych przedstawionych w tablicy 3.7 jest rys. 3.5.

Tablica 3.7

Konstruowanie szeregu kumulacyjnego ciągłej zmiennej losowej
na podstawie szeregu rozdzielczego

j	s	$(x_{d,j}; x_{g,j}]$	f_j	$\sum_{j=1}^s f_j$	v_j	$\sum_{j=1}^s v_j$
1	2	3	4	5	6	7
1	1	(61,50; 66,00]	1	1	$\frac{1}{36}$	$\frac{1}{36}$
2	2	(66,00; 70,50]	1	4	$\frac{3}{36}$	$\frac{4}{36}$
3	3	(70,50; 75,00]	6	10	$\frac{6}{36}$	$\frac{10}{36}$
4	4	(75,00; 79,50]	15	25	$\frac{15}{36}$	$\frac{25}{36}$
5	5	(79,50; 84,00]	5	30	$\frac{5}{36}$	$\frac{30}{36}$
6	6	(84,00; 88,50]	6	36	$\frac{6}{36}$	1
Σ		*	36	*	1	*

Źródło: opracowanie własne.

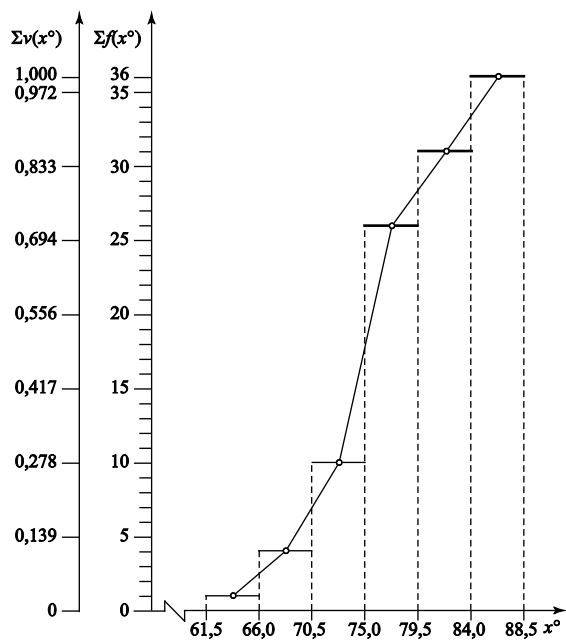
Do kumulacyjnych charakterystyk rozkładu empirycznego zaliczana jest także funkcja nazywana ogiwą. Dziedziną tej funkcji jest zbiór:

$$\{(i); i = 1, 2, \dots, N\},$$

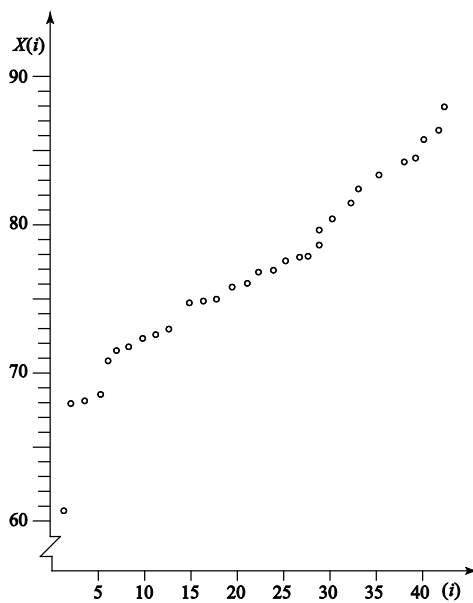
natomiast przeciwdziedziną jest zbiór:

$$\{x_{(i)}; i = 1, 2, \dots, N\}.$$

Ogiwa opisuje więc proces narastania wartości obserwowanej zmiennej losowej w uporządkowanym zbiorze realizacji tej zmiennej. Na rys. 3.6 przedstawiono wykres ogiwy sporządzony w oparciu o szereg pozycyjny z tablicy 3.6. ■



Rys. 3.5. Szeregi kumulacyjne ciągłej zmiennej losowej



Rys. 3.6. Ogiwa

3.2.2. Miary położenia

Omówione powyżej charakterystyki rozkładu empirycznego jednowymiarowej zmiennej losowej są bardzo ważnymi narzędziami analizy statystycznej. Pozwalają one oceniać w szczególności pewne cechy strukturalne zbioru realizacji \mathbf{X}_N . Owe cechy strukturalne nie wyczerpują jednak formułowanych zwykle pytań. Szczególnie często przedmiotem zainteresowania badacza jest ten poziom badanej cechy, o którym można zakładać, że został ukształtowany działaniem czynników systematycznych, kontrolowanych poprzez cechy kwalifikujące $K_r \in \mathbf{K}$. Do wyodrębnienia tego składnika systematycznego służą wartości przeciętne, nazywane też miarami położenia rozkładu empirycznego na osi liczb rzeczywistych. Poniżej omówimy dwa typy miar położenia, a mianowicie średnie oraz tzw. przeciętne pozycyjne.

3.2.2.1. Średnia arytmetyczna

Jeśli funkcjonujące w doświadczeniu czynniki systematyczne i losowe mają charakter addytywny, to każdą realizację $x_i \in \mathbf{X}_N$ można zapisać w formie następującej sumy:

$$x_i = \bar{x} + e_i, \quad (3.25)$$

w której \bar{x} wyraża efekt działania czynników systematycznych, natomiast e_i jest efektem działania czynników losowych. Wartość \bar{x} pozostaje stała w całym doświadczeniu, natomiast wartości e_i zmieniają się w kolejnych powtórzeniach badania. Zapisując każdą realizację $x_i \in \mathbf{X}_N$ w postaci równania (3.25) otrzymujemy następujący układ równań:

$$\left. \begin{array}{l} x_1 = \bar{x} + e_1 \\ x_2 = \bar{x} + e_2 \\ \vdots \\ x_i = \bar{x} + e_i \\ \vdots \\ x_N = \bar{x} + e_N \end{array} \right\} \cdot \quad (3.26)$$

Nie można podać jednoznacznego rozwiązania tego układu. Można natomiast zażądać – zgodnie z ideą metody najmniejszych kwadratów – aby:

$$\sum_{i=1}^N (x_i - \bar{x})^2 = \sum_{i=1}^N e_i^2 = \text{minimum.} \quad (3.27)$$

Sumę tę będziemy rozpatrywać jako funkcję \bar{x} . Aby znaleźć wartość \bar{x} , przy której funkcja:

$$f(\bar{x}) = \sum_{i=1}^N (x_i - \bar{x})^2 \quad (3.28)$$

osiąga minimum, wystarczy wyznaczyć pierwszą pochodną względem \bar{x} , a następnie przyrównać ją do zera. Mamy więc:

$$\frac{df(\bar{x})}{d\bar{x}} = \frac{d \sum_{i=1}^N (x_i - \bar{x})^2}{d\bar{x}} = 2(x_1 - \bar{x}) + 2(x_2 - \bar{x}) + \dots \quad (3.29)$$

$$\dots + 2(x_i - \bar{x}) + \dots + 2(x_N - \bar{x}) = 2 \sum_{i=1}^N x_i - 2N\bar{x}.$$

Po przyrównaniu do zera otrzymujemy:

$$2 \sum_{i=1}^N x_i - 2N\bar{x} = 0,$$

$$\bar{x} = \frac{\sum_{i=1}^N x_i}{N}. \quad (3.30)$$

Okazuje się więc, że poszukiwany składnik systematyczny w równaniach (3.26) jest średnią arytmetyczną realizacji $x_i \in \mathbf{X}_N$.

Wzór (3.30) znajduje zastosowanie wówczas, gdy realizacje obserwowanej zmiennej losowej X dane są w postaci szczegółowego szeregu statystycznego (pierwotnego lub pozycyjnego). W przypadku szeregu rozdzielczego mamy natomiast:

$$\bar{x} = \frac{\sum_{j=1}^k x_j^{\circ} f_j}{\sum_{j=1}^k f_j} = \frac{\sum_{j=1}^k x_j^{\circ} f_j}{N}. \quad (3.31)$$

Jeśli badana zmienna losowa X jest ciągła, to x_j° należy interpretować jako środek j -tego przedziału klasowego. Mamy wówczas:

$$x_j^{\circ} = \frac{x_{d,j} + x_{g,j}}{2}. \quad (3.32)$$

Wykorzystując definicję częstości względnej (3.4), wzór (3.31) można stransformować do postaci:

$$\bar{x} = \sum_{j=1}^k x_j^{\circ} v_j. \quad (3.33)$$

Średnia arytmetyczna obliczana według wzorów (3.31) i (3.33) nazywana jest średnią arytmetyczną ważoną. Podkreśla się w ten sposób, że poszczególnym wartościom $x_j^{\circ} \in X^{\circ}$ nadano różne wagi, proporcjonalne do częstości ich występowania w zbiorze realizacji \mathbf{X}_N . Niezależnie od tego, w jaki sposób obliczono wartość \bar{x} , średnia arytmetyczna ma pewne własności formalne, których znajomość może ułatwić analizę materiału statystycznego. Zauważmy przede wszystkim, że:

$$x_{\min} \leq \bar{x} \leq x_{\max}, \quad (3.34)$$

$$x_{(1)} \leq \bar{x} \leq x_{(N)}, \quad (3.35)$$

gdzie $x_{\min} = x_{(1)}$ jest kresem dolnym, natomiast $x_{\max} = x_{(N)}$ jest kresem górnym zbioru realizacji \mathbf{X}_N . Zauważmy też, że:

$$\bar{x}N = \sum_{i=1}^N x_i, \quad (3.36)$$

$$\sum_{i=1}^N (x_i - \bar{x}) = 0, \quad (3.37)$$

$$\sum_{i=1}^N (x_i - \bar{x})^2 = \text{minimum} \quad (3.38)$$

i w konsekwencji

$$\sum_{i=1}^N (x_i - \bar{x})^2 \leq \sum_{i=1}^N (x_i - a)^2, \quad (3.39)$$

gdzie a jest dowolną liczbą rzeczywistą.

W przypadku szeregu rozdzielczego te same własności zapiszemy następująco:

$$\bar{x} \sum_{j=1}^k f_j = \sum_{j=1}^k x_j^{\circ} f_j, \quad (3.36a)$$

$$\sum_{j=1}^k f_j (x_j^{\circ} - \bar{x}) = 0, \quad (3.37a)$$

$$\sum_{j=1}^k f_j (x_j^{\circ} - \bar{x})^2 = \text{minimum}, \quad (3.38a)$$

$$\sum_{j=1}^k f_j (x_j^{\circ} - \bar{x})^2 \leq \sum_{j=1}^k f_j (x_j^{\circ} - a)^2. \quad (3.39a)$$

Przykład 3.5

W tabelicy 3.8 przedstawiono obliczenia związane z wyznaczaniem średniej arytmetycznej według wzoru (3.31). Kolumny 1–3 pochodzą z tabelicy 3.4. Ze-stawione w kolumnie 4 środki przedziałów klasowych (x_j°) obliczono według wzoru (3.32). Podstawiając sumy kolumn 3 i 5 do wzoru (3.31) otrzymujemy:

$$\bar{x} = \frac{2790}{36} = 77,50.$$

Tablica 3.8

Obliczanie średniej arytmetycznej z szeregu rozdzielczego ciągłej zmiennej losowej X

j	$(x_{d,j}; x_{g,j}]$	f_j	x_j°	$x_j^{\circ} \cdot f_j$
1	2	3	4	5
1	(61,50; 66,00]	1	63,75	63,75
2	(66,00; 70,50]	3	68,25	204,75
3	(70,50; 75,00]	6	72,75	436,50
4	(75,00; 79,50]	15	77,25	1158,75
5	(79,50; 84,00]	5	81,75	408,75
6	(84,00; 88,50]	6	86,25	517,50
Σ	*	36	*	2790,00

Źródło: opracowanie własne.

Tablica 3.9

Obliczanie średniej arytmetycznej z szeregu rozdzielczego ciągłej zmiennej losowej X

j	$(x_{d,j}; x_{g,j}]$	v_j	x_j°	$x_j^{\circ} \cdot v_j$
1	2	3	4	5
1	(61,50; 66,00]	0,028	63,75	1,78500
2	(66,00; 70,50]	0,083	68,25	5,66475
3	(70,50; 75,00]	0,167	72,75	12,14925
4	(75,00; 79,50]	0,416	77,25	32,13600
5	(79,50; 84,00]	0,139	81,75	11,36325
6	(84,00; 88,50]	0,167	86,25	14,40375
Σ	*	1,000	*	77,50200

Źródło: opracowanie własne.

W tablicy 3.9 pokazano obliczenia niezbędne dla wyznaczenia średniej arytmetycznej według wzoru (3.33). Suma kolumny 5 jest poszukiwaną średnią arytmetyczną. Mamy mianowicie: $\bar{x} = 77,50200 \approx 77,50$. Średnia arytmetyczna obliczona według wzoru (3.30) na podstawie szczegółowego szeregu statystycznego (tablica 3.3) wynosi:

$$\bar{x} = \frac{2789,38}{36} = 77,48.$$

Zatem średnie arytmetyczne obliczone z szeregów rozdzielczych obciążone są niewielkim błędem grupowania, wynoszącym $77,50 - 77,48 = 0,02$. ■

3.2.2.2. Średnia harmoniczna

Jeśli wyniki doświadczenia dane są w formie szczegółowego szeregu statystycznego (pierwotnego lub pozycyjnego), to średnią harmoniczną (h_x) oblicza się ze wzoru:

$$h_x = \frac{N}{\sum_{i=1}^N \frac{1}{x_i}}. \quad (3.40)$$

W przypadku szeregu rozdzielczego stosuje się natomiast wzór:

$$h_x = \frac{\sum_{j=1}^k f_j}{\sum_{j=1}^k \frac{f_j}{x_j^\circ}} = \frac{N}{\sum_{j=1}^k \frac{f_j}{x_j^\circ}} \quad (3.41)$$

albo

$$h_x = \frac{1}{\sum_{j=1}^k \frac{v_j}{x_j^\circ}}. \quad (3.42)$$

Przykład 3.6

Obserwowano proces produkcji 500 szt. pewnego wyrobu. Mierzono czas (t) potrzebny do wyprodukowania kolejnych serii produktu o stałej liczności $m = 100$ szt. Czas ten mierzono w godzinach [h]. Następnie obliczono wydajności pracy ($x = m/t = 100/t$) w kolejnych etapach produkcji ($i = 1, 2, \dots, 5$). Wyniki badania zestawiono w tabelicy 3.10. Należy obliczyć przeciętną wydajność pracy.

Stosując średnią arytmetyczną otrzymujemy:

$$\bar{x} = \frac{190}{5} = 38 \text{ [szt./h]}.$$

Jest to wynik błędny. Nietrudno wszak zauważyć, że skoro wyprodukowano 500 szt., a produkcja trwała 14,5 h, to przeciętna wydajność pracy wynosi:

$$\frac{500}{14,5} \approx 34,48 \text{ [szt./h]}.$$

Taki właśnie wynik otrzymamy, jeśli do obliczenia przeciętnej wydajności pracy wykorzystamy średnią harmoniczną. Stosując wzory (3.40) i (3.41) otrzymujemy:

$$h_x = \frac{5}{1/50 + 1/25 + 1/40 + 1/50 + 1/25} = \frac{5}{2/50 + 1/40 + 2/25} \approx 34,48 \text{ [szt./h]}.$$

Tablica 3.10

Dane liczbowe do przykładu 3.6

i	m_i [szt.]	t_i [h]	x_i [szt./h]
1	2	3	4
1	100	2	50
2	100	4	25
3	100	2,5	40
4	100	2	50
5	100	4	25
Σ	500	14,5	190

Źródło: opracowanie własne.

Zauważmy, że badana charakterystyka procesu ma miano szt./h. Zastosowanie średniej arytmetycznej byłoby właściwe tylko wówczas, gdyby mianownik wyrażenia [szt./h] był stały, a więc gdyby mierzono liczbę sztuk wyprodukowanych w jednakowych przedziałach czasu. W rozważanym przypadku czas (t) jest wielkością zmienną, natomiast stały jest licznik wyrażenia [szt./h]. To właśnie przesądza o konieczności zastosowania średniej harmoniczej. ■

3.2.2.3. Średnie potęgowe

Zarówno średnia arytmetyczna (\bar{x}), jak i średnia harmoniczna (h_x) są szczególnymi przypadkami średniej potęgowej m_x^r . W symbolu tym r oznacza stopień średniej, przy czym $r \neq 0$.

W przypadku szczegółowego szeregu statystycznego mamy:

$$m_x^r = \left(\frac{1}{N} \sum_{i=1}^N x_i^r \right)^{\frac{1}{r}}. \quad (3.43)$$

Jeśli wyniki doświadczenia dane są w formie szeregu rozdzielczego, to:

$$m_x^r = \left(\frac{1}{N} \sum_{j=1}^k (x_j^\circ)^r f_j \right)^{\frac{1}{r}}, \quad (3.44)$$

przy czym $N = \sum_{j=1}^k f_j$.

Biorąc $r = 1$ otrzymujemy $m_x^1 = \bar{x}$, czyli średnią arytmetyczną. Podstawiając natomiast $r = -1$ otrzymujemy średnią harmoniczną. Mamy mianowicie:

$$m_x^{-1} = \left(\frac{1}{N} \sum_{i=1}^N x_i^{-1} \right)^{-1} = \frac{N}{\sum_{i=1}^N \frac{1}{x_i}} = h_x \quad (3.45)$$

oraz

$$m_x^{-1} = \left(\frac{1}{N} \sum_{j=1}^k x_j^\circ f_j \right)^{-1} = \frac{N}{\sum_{j=1}^k f_j \frac{1}{x_j^\circ}} = h_x. \quad (3.46)$$

Dodajmy, że w oparciu o wzory (3.43) i (3.44) można zdefiniować całą rodzinę średnich. Ważną rolę w dalszych rozważaniach odegrają tzw. średnie kwadratowe:

$$m_x^2 = \left(\frac{1}{N} \sum_{i=1}^N x_i^2 \right)^{\frac{1}{2}} = \sqrt{\frac{1}{N} \sum_{i=1}^N x_i^2}, \quad (3.47)$$

$$m_x^2 = \left(\frac{1}{N} \sum_{j=1}^k (x_j^\circ)^2 f_j \right)^{\frac{1}{2}} = \sqrt{\frac{1}{N} \sum_{j=1}^k (x_j^\circ)^2 f_j}. \quad (3.48)$$

Średnie te wykorzystamy przy definiowaniu miar wewnętrznego zróżnicowania zbiorów realizacji (X_N) obserwowanych zmiennych losowych X .

3.2.2.4. Średnia geometryczna

Średnia geometryczna znajduje zastosowanie przede wszystkim wówczas, gdy zachodzi potrzeba uśrednienia wartości charakteryzujących stosunki między dwiema wielkościami. Z sytuacją taką mamy w szczególności do czynienia przy badaniu dynamiki zjawisk.

Jeśli elementami ciągu:

$$(y_1, y_2, \dots, y_t, \dots, y_r)$$

są realizacje zmiennej losowej Y w kolejnych chwilach t , to dynamikę obserwowanego zjawiska można badać śledząc wartości wskaźnika łańcuchowego:

$$x_1 = y_2/y_1; \quad x_2 = y_3/y_2, \dots; \quad x_t = y_{t+1}/y_t, \dots; \quad x_{r-1} = y_r/y_{r-1}.$$

Średnia wartość (g_x) wskaźnika x powinna mieć tę własność, że:

$$y_r = y_1 g_x^{r-1}.$$

Własności tej nie posiada średnia arytmetyczna. Ma ją natomiast – jak się przekonamy poniżej – średnia geometryczna, którą oznaczymy symbolem g_x . W przypadku szczegółowego szeregu statystycznego średnią geometryczną wyznacza się ze wzoru:

$$g_x = \sqrt[N]{x_1 \cdot x_2 \cdot \dots \cdot x_i \cdot \dots \cdot x_N} = \sqrt[N]{\prod_{i=1}^N x_i} = \left(\prod_{i=1}^N x_i \right)^{\frac{1}{N}}. \quad (3.49)$$

Jeśli mamy do czynienia z szeregiem rozdzielczym, to wówczas:

$$\begin{aligned} g_x &= \sqrt[N]{(x_1^{\circ})^{f_1} \cdot (x_2^{\circ})^{f_2} \cdot \dots \cdot (x_j^{\circ})^{f_j} \cdot \dots \cdot (x_k^{\circ})^{f_k}} = \\ &= \sqrt[N]{\prod_{j=1}^k (x_j^{\circ})^{f_j}} = \left(\prod_{j=1}^k (x_j^{\circ})^{f_j} \right)^{\frac{1}{N}} \end{aligned} \quad (3.50)$$

przy $N = \sum_{j=1}^k f_j$.

Wzory te można przedstawić w postaci logarytmicznej. Mamy mianowicie:

$$\lg g_x = \frac{\sum_{i=1}^N \lg x_i}{N}, \quad (3.51)$$

$$\lg g_x = \frac{\sum_{j=1}^k f_j \lg x_j^\circ}{N}. \quad (3.52)$$

Wzory te upraszczają obliczenia związane z wyznaczaniem średniej geometrycznej. Jeśli zbiór realizacji X_N jest ustalony, to:

$$g_x \leq \bar{x}. \quad (3.53)$$

Przykład 3.7

W kolumnach 1 i 2 tabeli 3.11 przedstawiono informacje o kształtowaniu się jakości wykonania pewnego wyrobu w kolejnych sześciu okresach rozliczeniowych. Miarą jakości wykonania jest tu frakcja poprawnie wykonanych jednostek wyrobu y_t ($t = 1, 2, \dots, 6$), wyrażona w procentach. Kolumna 3 zawiera ciąg wartości wskaźnika łańcuchowego $x_t = y_t/y_{t-1}$ ($t = 2, 3, \dots, 6$), obrazującego dynamikę zjawiska. Należy obliczyć średnią wartość wskaźnika dynamiki x .

Tabela 3.11

Badanie dynamiki jakości wykonania pewnego wyrobu

t	y_t	$x_t = \frac{y_t}{y_{t-1}}$
1	2	3
1	95,4	*
2	96,5	1,0115
3	93,4	0,9679
4	94,5	1,0118
5	93,2	0,9862
6	96,0	1,0300

Źródło: opracowanie własne.

Stosując wzór (3.49) otrzymujemy:

$$g_x = \sqrt[5]{1,0115 \cdot 0,9679 \cdot 1,0118 \cdot 0,9862 \cdot 1,0300} \approx 1,0011789.$$

Wykorzystując tę wartość mamy:

$$\tilde{y}_6 = y_1 g_x^5 = 95,4 \cdot (1,0011789)^5 \approx 95,9637.$$

Uzyskana ocena (\tilde{y}_6) jest więc obciążona błędem:

$$\Delta y_6 = \tilde{y}_6 - y_6 = 95,9637 - 96,0 = -0,0363.$$

Biorąc średnią arytmetyczną wartości x_t ($t = 2, 3, \dots, 6$), zestawionych w kolumnie 3 tablicy 3.11, uzyskamy ocenę (\tilde{y}_6) obciążoną znacznie większym błędem. Mamy mianowicie:

$$\bar{x} = \frac{1,0115 + 0,9679 + 1,0118 + 0,9862 + 1,0300}{5} \approx 1,00148.$$

$$\tilde{y}_6 = y_1 \bar{x}^5 = 95,4 (1,00148)^5 \approx 96,1081$$

i w konsekwencji

$$\Delta y_6 = \tilde{y}_6 - y_6 = 96,1081 - 96,0 = 0,1081.$$

Średnia arytmetyczna jest więc wyraźnie gorszą oceną średniego poziomu dynamiki badanego zjawiska. ■

3.2.2.5. Mediana

Mediana należy do grupy charakterystyk pozycyjnych nazywanych kwantylami (albo fraktylami). Kwantylem stopnia q nazywamy taką liczbę x_q , która rozcina N -elementowy zbiór realizacji X_N na Nq -elementowy podzbiór wartości $x_i \leq x_q$ oraz na $N(1-q)$ -elementowy podzbiór wartości $x_i > x_q$, przy czym $0 < q < 1$. Podstawowe kwantyle to:

- kwartyle; $q = 0,25; 0,5; 0,75$,
- decyle; $q = 0,1; (0,1); 0,9$,
- centyle; $q = 0,01; (0,01); 0,99$.

Mediana (x_{me}) jest drugim kwantylem. Jest to więc taka wartość, iż połowa realizacji $x_i \in X_N$ ma wartości nie przekraczające x_{me} , natomiast połowa ma – w konsekwencji – wartości wyższe niż x_{me} . Jeśli mediana jest wyznaczana na podstawie szczegółowego szeregu statystycznego, to w pierwszym kroku postępowania należy uporządkować zbiór X_N za pomocą relacji „ \leq ”. W rezultacie otrzymuje się pozycyjny szereg statystyczny postaci (3.19).

W drugim kroku postępowania stosuje się jeden z podanych poniżej wzorów. Jeśli N jest liczbą nieparzystą, to:

$$x_{me} = x_{(\frac{N+1}{2})}. \quad (3.54)$$

Jeśli natomiast N jest liczbą parzystą, to:

$$x_{me} = \frac{1}{2} \left[x_{(\frac{N}{2})} + x_{(\frac{N}{2}+1)} \right]. \quad (3.55)$$

Przykład 3.8

W tabelicy 3.6 przedstawiono szereg pozycyjny otrzymany w rezultacie uporządkowania zbioru realizacji X_N pokazanego w tabelicy 3.3. Należy wyznaczyć medianę. Ponieważ $N = 36$, zatem należy zastosować wzór (3.55). Po podstawieniu wartości z tabelicy 3.6 mamy:

$$x_{me} = \frac{x_{(18)} + x_{(19)}}{2} = \frac{77,27 + 77,40}{2} = 77,335.$$

Realizacje o numerach (1)–(18) mają wartości niższe od 77,335, natomiast realizacje o numerach (19)–(36) mają wartości wyższe niż 77,335. ■

Wyznaczenie mediany na podstawie szeregu rozdzielczego ciągłej zmiennej losowej odbywa się w trzech etapach.

– Należy skonstruować szereg skumulowany postaci (3.17), (3.18); zob. tablica 3.7.

– W szeregu tym należy znaleźć przedział klasowy o najniższym numerze $j = r$, w którym:

$$\sum_{j=1}^r f_j > \frac{N}{2} \quad (3.56)$$

albo

$$\sum_{j=1}^r v_j > \frac{1}{2}. \quad (3.57)$$

W przedziale tym znajduje się mediana.

– Wartość mediany wyznacza się ze wzoru interpolacyjnego:

$$x_{me} = x_{d,r} + \frac{\left(\frac{N}{2} - \sum_{j=1}^{r-1} f_j \right) l}{f_r}, \quad (3.58)$$

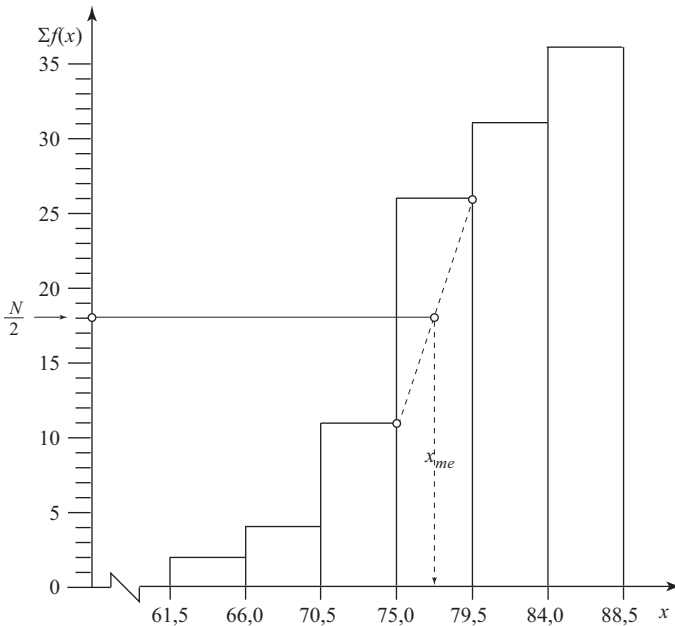
gdzie:

- $x_{d,r}$ – dolny kres przedziału klasowego, w którym znajduje się mediana,
- f_r – liczebność w przedziale klasowym o numerze $j = r$,
- l – długość przedziału klasowego.

Przykład 3.9

Wyznamy medianę z szeregu rozdzielczego, przedstawionego w tabelicy 3.4. Odpowiedni szereg kumulacyjny pokazano w tabelicy 3.7. Analiza kolumny 5 albo 7 prowadzi do wniosku, że mediana znajduje się w przedziale klasowym o numerze $j = r = 4$. Jest to przedział o najniższym numerze (j), w którym spełnione są nierówności (3.56), (3.57). Mamy więc $x_{d,r} = x_{d,4} = 75,00$; $f_r = f_4 = 15$ oraz

$$\sum_{j=1}^{r-1} f_j = \sum_{j=1}^3 f_j = 10.$$



Rys. 3.7. Graficzne wyznaczanie mediany w szeregu rozdzielczym ciągłej zmiennej losowej

Pamiętamy też, że $l = 4,50$ (zob. przykład 3.2). Po podstawieniu tych wartości do wzoru (3.59) otrzymujemy:

$$x_{me} = 75 + \frac{4,5(18-10)}{15} = 75 + 2,40 = 77,40.$$

Mediana wyznaczona ze szczegółowego szeregu statystycznego wynosi $x_{me} = 77,335$ (zob. przykład 3.8). Różnica $77,40 - 77,335 = 0,065$ jest błędem grupowania.

Na rys. 3.7 pokazano graficzną metodę wyznaczania mediany z szeregu rozdzielczego ciągłej zmiennej losowej. ■

3.2.2.6. Modalna

Modalna (albo dominanta, moda) jest wartością najczęstszą w szeregu statystycznym. Charakterystykę tę będziemy oznaczać symbolem x_{mo} . Koniecznym etapem przy jej wyznaczaniu jest skonstruowanie szeregu rozdzielczego.

W przypadku dyskretnej zmiennej losowej wyznaczenie wartości x_{mo} sprowadza się do znalezienia takiej wartości $x_j^o \in X^o$, której odpowiada największa liczebność f_j .

W przypadku ciągłej zmiennej losowej takie postępowanie pozwala wyznaczyć tylko przedział klasowy, w którym znajduje się modalna. Wartość x_{mo} wyznacza się ze wzoru interpolacyjnego:

$$x_{mo} = x_{d,r} + \frac{(f_r - f_{r-1})l}{(f_r - f_{r-1}) + (f_r - f_{r+1})}, \quad (3.59)$$

gdzie:

- $x_{d,r}$ – dolny kres przedziału klasowego, w którym znajduje się modalna,
- l – długość przedziału klasowego,
- f_r, f_{r-1}, f_{r+1} – liczebności w przedziałach klasowych o numerach $j = r, j = r - 1$ oraz $j = r + 1$.

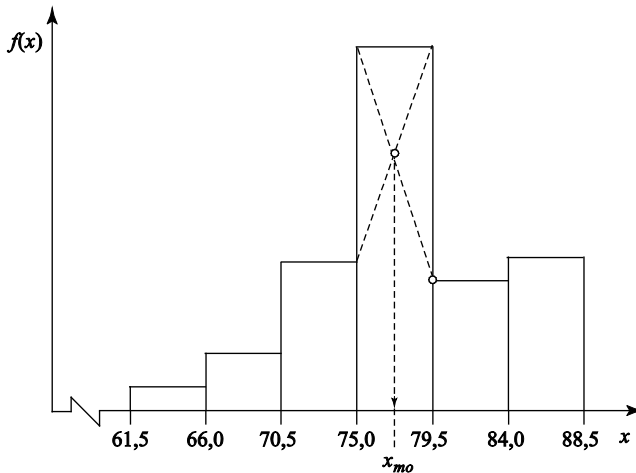
Przykład 3.10

Wyznamy modalną z szeregu rozdzielczego przedstawionego w tabeli 3.4. Największa liczebność występuje w przedziale klasowym o numerze $j = r = 4$, o kresie dolnym $x_{d,4} = 75,00$. Wszystkie przedziały klasowe mają jednakową długość $l = 4,50$. Z tabeli 3.4 odczytujemy liczebności: $f_3 = 6, f_4 = 15, f_5 = 5$. Podstawiając te wartości do wzoru (3.59) otrzymujemy:

$$x_{mo} = 75 + \frac{(15-6) \cdot 4,5}{(15-6) + (15-5)} = 75 + 2,13 = 77,13.$$

Uzyskany wynik przyjmuje się jako wartość najczęstszą w zbiorze realizacji X_N . Zauważmy jednak, że wartość $x_{mo} = 77,13$ w zbiorze tym w ogóle nie występuje. W otoczeniu punktu 77,13 daje się natomiast zauważyć wyraźne „zagęszczenie” realizacji $x_i \in X_N$. Jeśli przyjąć – na przykład – że otoczeniem punktu 77,13 jest przedział $[75,13; 79,13]$, to w przedziale tym występuje 41,7% wszystkich realizacji, podczas gdy w sąsiednich przedziałach o tej samej długości mamy 16,7% i 8,3%.

Na rys. 3.8 pokazano graficzną metodę wyznaczania wartości modalnej w rozważanym szeregu rozdzielczym. ■



Rys. 3.8. Graficzne wyznaczanie modalnej w szeregu rozdzielczym ciągłej zmiennej losowej

3.2.3. Miary rozproszenia

Wewnętrzne zróżnicowanie zbioru realizacji X_N jest skutkiem działania nie kontrolowanych czynników losowych w każdym rzeczywistym doświadczeniu. Zróżnicowanie elementów $x_i \in X_N$ może być większe lub mniejsze. Powstaje więc problem pomiaru natężenia tego zjawiska. Służą do tego celu funkcje pomiarowe określone na zbiorze X_N i o wartościach w zbiorze $\mathbf{R}^+ \cup \{0\}$. Funkcje te nazywane są miarami rozproszenia. Są one konstruowane w ten sposób, by

przyjmowały wartość zero, gdy wszystkie realizacje obserwowanej zmiennej losowej mają tę samą wartość, a więc gdy $\bigwedge_i x_i = x$, a także by przyjmowały tym większe wartości, im większe jest natężenie owych czynników losowych, pozostających poza kontrolą eksperymentatora.

3.2.3.1. Wariancja i odchylenie standardowe

Przypomnijmy (zob. punkt 3.2.2.1), że wartość $x_i \in \mathbf{X}_N$ można rozważać jako sumę:

$$x_i = \bar{x} + e_i,$$

w której e_i jest wartością składnika losowego w i -tej realizacji zmiennej losowej X . W konsekwencji mamy więc:

$$\left. \begin{array}{l} e_1 = x_1 - \bar{x} \\ e_2 = x_2 - \bar{x} \\ \vdots \\ e_i = x_i - \bar{x} \\ \vdots \\ e_N = x_N - \bar{x} \end{array} \right\} \quad (3.60)$$

Mogłoby więc wydawać się, że biorąc średnią arytmetyczną:

$$\bar{e} = \frac{e_1 + e_2 + \dots + e_i + \dots + e_n}{N} = \frac{1}{N} \sum_{i=1}^N e_i \quad (3.61)$$

uzyskamy ocenę udziału składnika losowego w realizacjach $x_i \in \mathbf{X}_N$. Taka propozycja funkcji mierzącej wewnętrzne zróżnicowanie zbioru \mathbf{X}_N nie może być jednak przyjęta, albowiem funkcja (3.61) nie reaguje na jakiegokolwiek zmiany tego zróżnicowania. W każdej sytuacji $\bar{e} = 0$, ponieważ $\sum_{i=1}^N e_i = 0$; zob. wzory (3.37), (3.37a).

W celu ominięcia tej przeszkody, zamiast średniej potęgowej stopnia $r = 1$ (czyli średniej arytmetycznej postaci (3.61)) bierze się średnią potęgową stopnia $r = 2$, czyli średnią postaci:

$$m_e^2 = \left(\frac{1}{N} \sum_{i=1}^N e_i^2 \right)^{\frac{1}{2}}. \quad (3.62)$$

Przypomnijmy, że o tego typu średnich wspominaliśmy w punkcie 3.2.2.3; zob. wzory (3.43)–(3.48). Charakterystykę postaci (3.62) przyjęto nazywać odchyleniem standardowym (albo odchyleniem średnim). Odchylenie to oznaczymy symbolem s_x i zdefiniujemy je następująco:

$$s_x = \sqrt{\frac{\sum_{i=1}^N (x_i - \bar{x})^2}{N}}. \quad (3.63)$$

Kwadrat odchylenia standardowego (s_x^2) nazywany jest wariancją. Mamy mianowicie:

$$s_x^2 = \frac{\sum_{i=1}^N (x_i - \bar{x})^2}{N}. \quad (3.64)$$

Wzory (3.63) i (3.64) mogą być stosowane w odniesieniu do szczegółowego szeregu statystycznego (pierwotnego lub pozycyjnego). W przypadku szeregu rozdzielczego mamy natomiast:

$$s_x^2 = \frac{\sum_{j=1}^k f_j (x_j^\circ - \bar{x})^2}{\sum_{j=1}^k f_j} = \frac{\sum_{j=1}^k f_j (x_j^\circ - \bar{x})^2}{N} \quad (3.65)$$

lub

$$s_x^2 = \sum_{j=1}^k v_j (x_j^\circ - \bar{x})^2. \quad (3.66)$$

Jeśli badana zmienna losowa X jest dyskretna (skokowa), to wartości x_j° są elementami zbioru X° . W przypadku ciągłych zmiennych losowych są to środki przedziałów klasowych.

Biorąc dodatni pierwiastek kwadratowy z wariancji uzyskujemy odchylenie standardowe:

$$s_x = \sqrt{s_x^2} . \quad (3.67)$$

Obliczanie wariancji i odchylenia standardowego można niekiedy uprościć po dokonaniu odpowiednich przekształceń wyrażenia:

$$\sum_{i=1}^N (x_i - \bar{x})^2 \quad (3.68)$$

lub wyrażenia

$$\sum_{j=1}^k f_j (x_j^\circ - \bar{x})^2 . \quad (3.69)$$

Mamy mianowicie:

$$\sum_{i=1}^N (x_i - \bar{x})^2 = \sum_{i=1}^N x_i^2 - 2\bar{x} \sum_{i=1}^N x_i + N\bar{x}^2 = \quad (3.70)$$

$$= \sum_{i=1}^N x_i^2 - 2 \frac{\left(\sum_{i=1}^N x_i \right)^2}{N} + \frac{\left(\sum_{i=1}^N x_i \right)^2}{N} = \sum_{i=1}^N x_i^2 - \frac{1}{N} \left(\sum_{i=1}^N x_i \right)^2 .$$

Po podstawieniu do wzoru (3.70) mamy więc:

$$s_x^2 = \frac{1}{N} \left[\sum_{i=1}^N x_i^2 - \frac{1}{N} \left(\sum_{i=1}^N x_i \right)^2 \right] = \quad (3.71)$$

$$= \frac{1}{N} \sum_{i=1}^N x_i^2 - \left(\frac{1}{N} \sum_{i=1}^N x_i \right)^2 =$$

$$= \overline{x^2} - \bar{x}^2 . \quad (3.72)$$

Wariancja jest więc średnim kwadratem pomniejszonym o kwadrat średniej. Odchylenie standardowe jest – zgodnie ze wzorem (3.67) – dodatnim pierwiast-

kiem kwadratowym z wariancji. Wyrażenie (3.71) dotyczy szczegółowego szeregu statystycznego. W przypadku szeregu rozdzielczego mamy natomiast:

$$\sum_{j=1}^k f_j (x_j^\circ - \bar{x})^2 = \sum_{j=1}^k f_j (x_j^\circ)^2 - \frac{1}{N} \left(\sum_{j=1}^k f_j x_j^\circ \right)^2 \quad (3.73)$$

i w konsekwencji

$$s_x^2 = \frac{1}{N} \left[\sum_{j=1}^k f_j (x_j^\circ)^2 - \frac{1}{N} \left(\sum_{j=1}^k f_j x_j^\circ \right)^2 \right],$$

$$s_x^2 = \frac{1}{N} \sum_{j=1}^k f_j (x_j^\circ)^2 - \left(\frac{1}{N} \sum_{j=1}^k f_j x_j^\circ \right)^2, \quad (3.74)$$

$$s_x^2 = \sum_{j=1}^k v_j (x_j^\circ)^2 - \left(\sum_{j=1}^k v_j x_j^\circ \right)^2. \quad (3.75)$$

Mamy więc – analogicznie jak poprzednio – średni kwadrat:

$$\overline{x^2} = \frac{1}{N} \sum_{j=1}^k f_j (x_j^\circ)^2 = \sum_{j=1}^k v_j (x_j^\circ)^2$$

pomniejszony o kwadrat średniej:

$$\bar{x}^2 = \left(\frac{1}{N} \sum_{j=1}^k f_j x_j^\circ \right)^2 = \left(\sum_{j=1}^k v_j x_j^\circ \right)^2.$$

Przykład 3.11

Obliczymy wariancję i odchylenie standardowe z szeregu rozdzielczego ciągłej zmiennej losowej X , pokazanego w tabelicy 3.4. Zastosujemy kolejno wzory (3.65), (3.74) i (3.75).

W tablicy 3.12 przedstawiono obliczenia niezbędne do wyznaczenia wariancji za pomocą wzoru (3.65). Iloczyny $x_j^\circ f_j$ zestawione w kolumnie 5 są potrzebne do obliczenia średniej arytmetycznej $\bar{x} = 2790/36 = 77,50$. Wartość tę wyznaczyliśmy już w przykładzie 3.5. Obecnie powtarzamy te obliczenia w celu zachowania ciągłości rozumowania. W kolumnach 6–8 pokazano kolejne etapy postępowania zmierzającego do wyznaczenia wartości sumy kwadratów odchyleń (3.69). Mając tę wartość obliczamy wariancję:

Tablica 3.12

Obliczanie wariancji z szeregu rozdzielczego ciągłej zmiennej losowej; wzór (3.65)

j	$(x_{d,j}; x_{g,j}]$	f_j	x_j°	$x_j^\circ f_j$	$x_j^\circ - 77,50$	$(x_j^\circ - 77,50)^2$	$f_j(x_j^\circ - 77,50)^2$
1	2	3	4	5	6	7	8
1	(61,50; 66,00]	1	63,75	63,75	-13,75	189,0625	189,0625
2	(66,00; 70,50]	3	68,25	204,75	-9,25	85,5625	256,6875
3	(70,50; 75,00]	6	72,75	436,50	-4,75	22,5625	135,3750
4	(75,00; 79,50]	15	77,25	1158,75	-0,25	0,0625	0,9375
5	(79,50; 84,00]	5	81,75	408,75	4,25	18,0625	90,3125
6	(84,00; 88,50]	6	86,25	517,50	8,75	76,5625	459,3750
Σ	*	36	*	2790,00	*	*	1131,7500

Źródło: opracowanie własne.

$$s_x^2 = \frac{1131,7500}{36} \approx 31,4375$$

i odchylenie standardowe:

$$s_x = \sqrt{31,4375} \approx 5,61.$$

Obliczenia niezbędne w celu wyznaczenia wariancji za pomocą wzoru (3.74) zawiera tablica 3.13. Mając sumy kolumn 3, 5 i 7 obliczamy wartość wariancji:

$$s_x^2 = \frac{217356,75}{36} - \left(\frac{2790}{36} \right)^2 = 6037,6875 - 6006,2500 \approx 31,4375.$$

Biorąc pierwiastek kwadratowy, otrzymujemy – podobnie jak poprzednio – wartość odchylenia standardowego.

W tablicy 3.14 przedstawiono obliczenia, które należy wykonać w celu wyznaczenia wariancji za pomocą wzoru (3.75). Mając sumy kolumn 5 i 7 obliczamy wartość wariancji:

$$s_x^2 = 6037,6875 - (77,5003)^2 \approx 31,4268,$$

a następnie wartość odchylenia standardowego:

$$s_x = \sqrt{31,4268} \approx 5,61.$$

Tablica 3.13

Obliczanie wariancji z szeregu rozdzielczego ciągłej zmiennej losowej; wzór (3.74)

j	$(x_{d,j}; x_{g,j}]$	f_j	x_j°	$x_j^\circ f_j$	$(x_j^\circ)^2$	$(x_j^\circ)^2 f_j$
1	2	3	4	5	6	7
1	(61,50; 66,00]	1	63,75	63,75	4064,0625	4064,0625
2	(66,00; 70,50]	3	68,25	204,75	4658,0625	13974,1875
3	(70,50; 75,00]	6	72,75	433,50	5292,5625	31755,3750
4	(75,00; 79,50]	15	77,25	1158,75	5967,5625	89513,4375
5	(79,50; 84,00]	5	81,75	408,75	6683,0625	33415,3125
6	(84,00; 88,50]	6	86,25	517,50	7439,0625	44634,3750
Σ	*	36	*	2790,00	*	217356,7500

Źródło: opracowanie własne.

Tablica 3.14

Obliczanie wariancji z szeregu rozdzielczego ciągłej zmiennej losowej; wzór (3.75)

j	$(x_{d,j}; x_{g,j}]$	v_j	x_j°	$x_j^\circ v_j$	$(x_j^\circ)^2$	$(x_j^\circ)^2 v_j$
1	2	3	4	5	6	7
1	(61,50; 66,00]	0,0278	63,75	1,7723	4064,0625	112,9809
2	(66,00; 70,50]	0,0833	68,25	5,6852	4658,0625	388,0166
3	(70,50; 75,00]	0,1667	72,75	12,1274	5292,5625	882,2702
4	(75,00; 79,50]	0,4166	77,25	32,1824	5967,5625	2486,0865
5	(79,50; 84,00]	0,1389	81,75	11,3551	6683,0625	928,2774
6	(84,00; 88,50]	0,1667	86,25	14,3779	7439,0625	1240,0917
Σ	*	1,0000	*	77,5003	*	6037,7233

Źródło: opracowanie własne.

Tablice 3.12, 3.13 i 3.14 pozwalają porównać skalę problemów numerycznych w poszczególnych metodach wyznaczania s_x^2 i s_x . ■

Zarówno wariancja (s_x^2), jak i odchylenie standardowe (s_x) są wielkościami mianowanymi. Odchylenie standardowe wyrażone jest w takich samych jednostkach miary, w jakich wyrażone są pierwotne wyniki obserwacji, należące do zbioru \mathbf{X}_N , natomiast wariancja wyrażona jest w jednostkach miary podniesionych do kwadratu. Ogranicza to możliwości porównywania wewnętrznego zróżnicowania zbiorów \mathbf{X}_N do tych przypadków, w których wariancje i odchylenia standardowe mają takie same miana. W praktyce pole porównań jest jeszcze mniejsze, albowiem w wewnętrznym zróżnicowaniu zbiorów \mathbf{X}_N ujawnia się często efekt skali. Ma to ten skutek, że dwa odchylenia standardowe $s_{x,1}$ i $s_{x,2}$ mogą być porównywane w kategoriach merytorycznych przede wszystkim wówczas, gdy różnica między odpowiednimi średnimi arytmetycznymi \bar{x}_1 i \bar{x}_2 nie jest zbyt duża, a więc gdy zbiory realizacji $\mathbf{X}_{N,1}$ i $\mathbf{X}_{N,2}$ mają zbliżone położenie na osi liczb rzeczywistych. Niedogodności tych można uniknąć wykorzystując do porównań wewnętrznego zróżnicowania zbiorów $\mathbf{X}_{N,1}$ i $\mathbf{X}_{N,2}$ współczynnik zmienności:

$$v_x = \frac{s_x}{\bar{x}}. \quad (3.76)$$

Ponieważ miana \bar{x} i s_x są takie same, zatem v_x jest wielkością niemianowaną, a ponadto miara ta nie jest obciążona skutkami efektu skali.

3.2.3.2. Odchylenie przeciętne

Przypomnijmy, że suma odchyłeń od średniej arytmetycznej zawsze równa jest zero; zob. wzory (3.37), (3.37a). Przy konstruowaniu odchylenia standardowego (s_x) uniknięto konsekwencji tego faktu, definiując s_x jako średnią potęgową stopnia $r = 2$, a nie jako średnią potęgową stopnia $r = 1$, odchyłeń od średniej arytmetycznej \bar{x} . Przy konstruowaniu odchylenia przeciętnego (d_x) zastosowano inną metodę. Odchylenia poszczególnych realizacji x_i ($i = 1, 2, \dots, N$) od średniej arytmetycznej (\bar{x}) zastępuje się ich bezwzględными wartościami. W przypadku szczegółowego szeregu statystycznego mamy więc:

$$d_x = \frac{1}{N} \sum_{i=1}^N |x_i - \bar{x}|. \quad (3.77)$$

Jeśli natomiast mamy do czynienia z szeregiem rozdzielczym, to:

$$d_x = \frac{\sum_{j=1}^k f_j |x_j^\circ - \bar{x}|}{\sum_{j=1}^k f_j} = \frac{1}{N} \sum_{j=1}^k f_j |x_j^\circ - \bar{x}| \quad (3.78)$$

lub

$$d_x = \sum_{j=1}^k v_j |x_j^\circ - \bar{x}|, \quad (3.79)$$

gdzie x_j° jest elementem zbioru wartości (\mathbf{X}°) obserwowanej zmiennej losowej X albo środkiem j -tego przedziału klasowego – w przypadku ciągłych zmiennych losowych.

Przykład 3.12

Obliczymy odchylenie przeciętne z szeregu rozdzielczego ciągłej zmiennej losowej, przedstawionego w tabelicy 3.4. Kolejne etapy obliczeń pokazano w tabelicy 3.15. Sumy kolumn 3 i 7 podstawiamy do wzoru (3.78):

$$d_x = \frac{147,5}{36} \approx 4,10.$$

Tabela 3.15

Obliczanie odchylenia przeciętnego z szeregu rozdzielczego ciągłej zmiennej losowej

j	$(x_{d,j}; x_{g,j}]$	f_j	x_j°	$x_j^\circ - 77,50$	$ x_j^\circ - 77,50 $	$f_j x_j^\circ - 77,50 $
1	2	3	4	5	6	7
1	(61,50; 66,00]	1	63,75	-13,75	13,75	13,75
2	(66,00; 70,50]	3	68,25	-9,25	9,25	27,75
3	(70,50; 75,00]	6	72,75	-4,75	4,75	28,50
4	(75,00; 79,50]	15	77,25	-0,25	0,25	3,75
5	(79,50; 84,00]	5	81,75	4,25	4,25	21,25
6	(84,00; 88,50]	6	86,25	8,75	8,75	52,50
Σ	*	36	*	*	*	147,40

Źródło: opracowanie własne.

Przypomnijmy (zob. przykład 3.11), że odchylenie standardowe obliczone na podstawie tego samego szeregu rozdzielczego wynosi $s_x \approx 5,61$. ■

Odchylenie standardowe i odchylenie przeciętne są dwiema różnymi miarami wewnętrznego zróżnicowania zbioru \mathbf{X}_N .

3.2.3.3. Miary rozproszenia oparte na statystykach pozycyjnych

Najprostszą charakterystyką tej klasy jest rozstęp (r), czyli różnica między największą i najmniejszą wartością w zbiorze realizacji \mathbf{X}_N . Mamy więc:

$$r = x_{\max} - x_{\min} = x_{(N)} - x_{(1)}, \quad (3.80)$$

gdzie $x_{\min} = x_{(1)}$ jest pierwszym, natomiast $x_{\max} = x_{(N)}$ jest ostatnim elementem w pozycyjnym szeregu statystycznym, czyli w zbiorze \mathbf{X}_N uporządkowanym za pomocą relacji „ \leq ”; zob. wzory (3.19), (3.20).

Inną miarą rozproszenia opartą na statystykach pozycyjnych jest tzw. odchylenie ćwiartkowe:

$$q = \frac{x_{0,75} - x_{0,25}}{2}, \quad (3.81)$$

w którym $x_{0,25}$ jest pierwszym, natomiast $x_{0,75}$ jest trzecim kwartylem.

Definiowane są również podobne charakterystyki rozproszenia oparte na innych kwantylach, w szczególności na decylach i centylach.

Przykład 3.13

Wyznamy rozstęp (r), a następnie odchylenie ćwiartkowe (q) na podstawie zbioru realizacji \mathbf{X}_N , przedstawionego w tabelicy 3.3. Posłużymy się w tym celu szeregiem pozycyjnym pokazanym w tabelicy 3.6. Wykorzystując tę tabelicę mamy:

$$r = x_{(36)} - x_{(1)} = 88,21 - 61,64 = 26,57,$$

$$x_{0,25} = \frac{x_{(9)} + x_{(10)}}{2} = \frac{73,70 + 74,20}{2} = 73,95,$$

$$x_{0,75} = \frac{x_{(27)} + x_{(28)}}{2} = \frac{81,98 + 82,49}{2} = 82,24,$$

$$q = \frac{x_{0,75} - x_{0,25}}{2} = \frac{82,24 - 73,95}{2} = 4,145.$$

Zauważmy, że kwartyle $x_{0,25}$ i $x_{0,75}$ można również wyznaczyć z szeregu rozdzielczego przedstawionego w tablicy 3.4. Należy w tym celu posłużyć się wzorami interpolacyjnymi, skonstruowanymi analogicznie jak wzór (3.58). ■

3.2.4. Momenty empiryczne

Pod pojęciem momentu empirycznego będziemy rozumieć charakterystykę postaci:

$$M_{x,r} = \frac{1}{N} \sum_{i=1}^N (x_i - c)^r, \quad (3.82)$$

gdzie r jest liczbą naturalną i oznacza rząd (stopień) momentu, natomiast c jest dowolną liczbą rzeczywistą.

Jeśli $c = 0$, mamy do czynienia z momentem zwykłym:

$$M'_{x,r} = \frac{1}{N} \sum_{i=1}^N x_i^r. \quad (3.83)$$

Jeśli natomiast podstawimy $c = \bar{x}$, to otrzymamy moment centralny:

$$M''_{x,r} = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^r. \quad (3.84)$$

W przypadku gdy zbiór realizacji zmiennej losowej X ma postać szeregu rozdzielczego, wzory (3.83) i (3.84) przybierają postać:

$$M'_{x,r} = \frac{1}{N} \sum_{j=1}^k x_j^\circ f_j, \quad (3.85)$$

$$M''_{x,r} = \frac{1}{N} \sum_{j=1}^k (x_j^\circ - \bar{x})^r f_j, \quad (3.86)$$

przy czym $N = \sum_{j=1}^k f_j$.

Zauważmy, że omówione powyżej miary położenia (punkt 3.2.2) i miary rozproszenia (punkt 3.2.3) mogą być identyfikowane jako momenty. Średnia arytmetyczna (\bar{x}) jest momentem zwykłym rzędu $r = 1$. Wariancja (s_x^2) jest momentem centralnym rzędu $r = 2$. Również odchylenie przeciętne (d_x) jest momentem. Jest to tzw. absolutny moment centralny rzędu pierwszego.

3.3. Wielowymiarowe zmienne losowe

3.3.1. Rozkład empiryczny dwuwymiarowej zmiennej losowej

Podstawową charakterystyką empirycznego rozkładu dwuwymiarowej zmiennej losowej (X, Y) jest dwuwymiarowy szereg rozdzielczy. W przypadku gdy obydwie zmienne składowe są dyskretne, mamy:

$$\{(x_j^\circ, y_p^\circ), f_{jp}\}; \quad j = 1, \dots, k; \quad p = 1, \dots, z \quad (3.87)$$

albo

$$\{(x_j^\circ, y_p^\circ), v_{jp}\}; \quad j = 1, \dots, k; \quad p = 1, \dots, z, \quad (3.88)$$

przy czym $f_{jp} = f(x_j^\circ, y_p^\circ)$ oznacza liczbę obiektów $A_i \in \mathbf{A}$, w których doszło do skojarzenia wartości $x_j^\circ \in \mathbf{X}^\circ$ z wartością $y_p^\circ \in \mathbf{Y}^\circ$, natomiast $v_{jp} = v(x_j^\circ, y_p^\circ)$ jest odpowiednią częstością względną. Wartości v_{jp} wynikają z następującego wzoru:

$$v_{jp} = \frac{f_{jp}}{\sum_{j=1}^k \sum_{p=1}^z f_{jp}} = \frac{f_{jp}}{N}, \quad (3.89)$$

przy czym

$$\sum_{j=1}^k \sum_{p=1}^z v_{jp} = 1. \quad (3.90)$$

Jeśli obydwie zmienne składowe są ciągłe, to wzory (3.87), (3.88) przybierają następującą postać:

$$\{(x_{d,j}, x_{g,j}), (y_{d,p}, y_{g,p}), f_{jp}\}; \quad j = 1, \dots, k; \quad p = 1, \dots, z \quad (3.91)$$

$$\{(x_{d,j}, x_{g,j}), (y_{d,p}, y_{g,p}), v_{jp}\}; \quad j = 1, \dots, k; \quad p = 1, \dots, z. \quad (3.92)$$

Symbol f_{jp} oznacza tu liczbę realizacji (x_i, y_i) zmiennej losowej (X, Y) takich, że $x_{d,j} < x_i \leq x_{g,j}$ i jednocześnie $y_{d,p} < y_i \leq y_{d,p}$, natomiast v_{jp} – analogicznie jak poprzednio – jest odpowiednią częstością względną. Jeśli jedna ze zmiennych składowych jest dyskretna, a druga ciągła, to empiryczny rozkład dwuwymiarowy jest kombinacją zbiorów (3.87) i (3.91) albo (3.88) i (3.92).

Dwuwymiarowe rozkłady empiryczne przedstawia się zwykle w postaci tzw. tablic korelacyjnych (w przypadku zmiennych ciągłych) albo wielozdzielczych tablic liczebności lub częstości względnych (nazywanych też tablicami kontyngencji) w przypadku zmiennych dyskretnych. Przykład takiej tablicy przedstawiono na rys. 3.9. Na rysunku tym pokazano rozkład dwuwymiarowy postaci (3.88). Po zsumowaniu częstości względnych w poszczególnych wierszach tablicy otrzymujemy rozkład brzegowy, równoważny rozkładowi jednowymiarowej zmiennej losowej X . Jest to zbiór:

$$\{(x_j^\circ, v_{j\cdot}); j = 1, 2, \dots, k\}, \quad (3.93)$$

przy czym

$$v_{j\cdot} = \sum_{p=1}^z v_{jp}, \quad (3.94)$$

$$\sum_{j=1}^k v_{j\cdot} = 1. \quad (3.95)$$

Analogicznie postępujemy w odniesieniu do drugiej zmiennej składowej Y . Mamy mianowicie:

$$(y_p^\circ, v_{\cdot p}); p = 1, 2, \dots, z, \quad (3.96)$$

$$v_{\cdot p} = \sum_{j=1}^k v_{jp}, \quad (3.97)$$

$$\sum_{p=1}^z v_{\cdot p} = 1. \quad (3.98)$$

Można również rozważać rozkłady zmiennej losowej X przy ustalonych wartościach y_p° , a także odwrotnie, rozkłady zmiennej losowej Y przy ustalonych wartościach x_j° . Są to rozkłady warunkowe. W rozpatrywanym przypadku (zob. rys. 3.8) mamy z rozkładów warunkowych zmiennej losowej X oraz k rozkładów warunkowych zmiennej Y .

		Y^o						Empiryczny rozkład brzegowy X
		y_1^o	y_2^o	...	y_p^o	...	y_z^o	
X^o	x_1^o	v_{11}	v_{12}	...	v_{1p}	...	v_{1z}	$v_{1.} = \sum_{i=1}^z v_{1ip}$
	x_2^o	v_{21}	v_{22}	...	v_{2p}	...	v_{2z}	$v_{2.} = \sum_{i=1}^z v_{2ip}$
	\vdots	\vdots	\vdots	...	\vdots	...	\vdots	\vdots
	x_j^o	v_{j1}	v_{j2}	...	v_{jp}	...	v_{jz}	$v_{j.} = \sum_{i=1}^z v_{jip}$
	\vdots	\vdots	\vdots	...	\vdots	...	\vdots	\vdots
	x_k^o	v_{k1}	v_{k2}	...	v_{kp}	...	v_{kz}	$v_{k.} = \sum_{i=1}^z v_{kip}$
	Empiryczny rozkład brzegowy Y	$v_{.1} = \sum_{j=1}^k v_{j1}$	$v_{.2} = \sum_{j=1}^k v_{j2}$...	$v_{.p} = \sum_{j=1}^k v_{jp}$...	$v_{.z} = \sum_{j=1}^k v_{jz}$	$\sum_{i=1}^z v_{i.} = \sum_{j=1}^k v_{.j} = 1$

Rys. 3.9. Dwuwymiarowy rozkład empiryczny

Zwróćmy uwagę na rozkłady warunkowe przy zadanych wartościach y_p° ; $p = 1, \dots, z$. Przy wyznaczaniu tych rozkładów należy tak skorygować częstości względne v_{jp} , by ich suma w każdej kolumnie tablicy była równa jedności. Te skorygowane częstości względne (v'_{jp}) wyznacza się ze wzoru:

$$v'_{jp} = \frac{v_{jp}}{v_{\bullet p}}. \quad (3.99)$$

W konsekwencji w każdej ($p = 1, \dots, z$) kolumnie tablicy mamy:

$$\sum_{j=1}^k v'_{jp} = 1. \quad (3.100)$$

Rozkład warunkowy zmiennej losowej X przy ustalonej wartości y_p° można więc napisać następująco:

$$\{(x_j^\circ, v'_{jp}); j = 1, \dots, k\}, \quad (3.101)$$

$$p = 1, \dots, z.$$

Analogiczne rozumowanie można powtórzyć w odniesieniu do poszczególnych wierszy tablicy pokazanej na rys. 3.9, a więc w odniesieniu do warunkowych rozkładów zmiennej losowej Y , przy ustalonych wartościach x_j° ; $j = 1, \dots, k$. W każdym wierszu tablicy wyznaczamy skorygowane częstości względne (v''_{jp}) według wzoru:

$$v''_{jp} = \frac{v_{jp}}{v_{j\bullet}}. \quad (3.102)$$

Wartości te spełniają warunek:

$$\sum_{p=1}^z v''_{jp} = 1. \quad (3.103)$$

Warunkowy rozkład zmiennej losowej Y przy ustalonej wartości x_j° ma więc postać:

$$\{(y_p^\circ, v''_{jp}); p = 1, \dots, z\}, \quad (3.104)$$

$$j = 1, \dots, k.$$

3.3.2. Współzależność dwóch zmiennych losowych

Jednym z podstawowych problemów, jakie należy rozwiązać przy badaniu dwuwymiarowych zmiennych losowych (X, Y) , jest pomiar współzależności między zmiennymi X i Y . Omówimy trzy miary owej współzależności, a mianowicie współczynnik korelacji liniowej r_{xy} – Pearsona, współczynnik korelacji dwuseryjnej $r_{d,xy}$ oraz współczynnik skojarzenia Q_{xy} – Kendalla. Współczynnik r_{xy} stosowany jest wówczas, gdy realizacje obu zmiennych składowych (X i Y) uzyskuje się w rezultacie pomiarów na skalach mocnych, co najmniej przedziałowych. Jeśli jedna z cech mierzona jest na skali co najmniej przedziałowej, a druga na dwupunktowej skali nominalnej lub porządkowej, to właściwą miarą ścisłości związku między zmiennymi losowymi X i Y jest współczynnik korelacji dwuseryjnej $r_{d,xy}$. Jeśli natomiast pomiary obu cech dokonywane są na najślabszych skalach nominalnych lub porządkowych, to do pomiaru współzależności między zmiennymi losowymi X i Y może być wykorzystany współczynnik skojarzenia Q_{xy} .

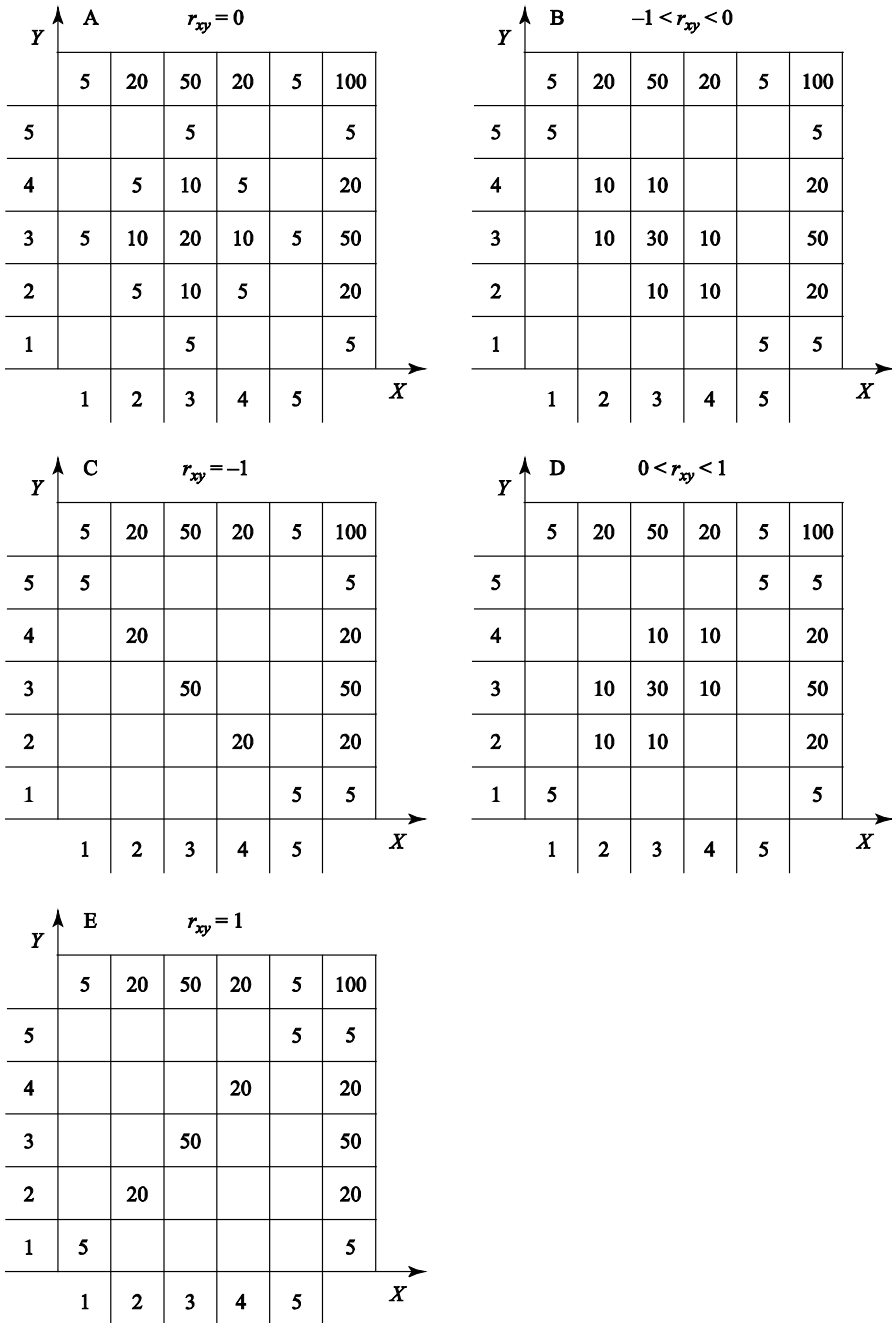
3.3.2.1. Empiryczny współczynnik korelacji liniowej

W badaniach wyczerpujących współczynnik korelacji liniowej r_{xy} definiuje się następująco:

$$r_{xy} = \frac{\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2} \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \bar{y})^2}}. \quad (3.105)$$

Wyrażenie występujące w liczniku tego wzoru jest empiryczną kowariancją (czyli wspólną wariancją) zmiennych X i Y ³. Kowariancja jest miarą współzależności obu zmiennych. Jest to jednak miara niewygodna, albowiem wyraża się w jednostkach mianowanych, a jej przedział zmienności nie jest ograniczony. W celu uniknięcia tych niedogodności dokonuje się zabiegu standaryzacji, dzieląc kowariancję przez iloczyn odchyłeń standardowych zmiennych X i Y . Skonstruowany w ten sposób współczynnik korelacji r_{xy} przybiera wartości z przedziału $[-1; 0]$ albo $[0; 1]$ (zob. uwagi w punkcie 1.5.4).

³ Zauważmy, że kowariancję można identyfikować jako moment. Jest to mieszany moment centralny rzędu drugiego dwóch zmiennych losowych.



Rys. 3.10. Własności empirycznego współczynnika korelacji liniowej

Własności empirycznego współczynnika korelacji liniowej ilustruje rys. 3.10. Przy sporządzaniu tego rysunku przyjęto, że zbiór realizacji zmiennej losowej (X, Y) obejmuje 100 elementów, a rozkłady brzegowe obu zmiennych są jednakowe we wszystkich rozważanych sytuacjach. Przy tych ograniczeniach skonstruowano pięć różnych rozkładów dwuwymiarowej zmiennej losowej (X, Y) , którym odpowiada pięć różnych wartości r_{xy} . Zauważmy, że definicyjny wzór (3.105) można znacznie uprościć. Po podzieleniu licznika i mianownika przez N mamy:

$$r_{xy} = \frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^N (x_i - \bar{x})^2 \cdot \sum_{i=1}^N (y_i - \bar{y})^2}}. \quad (3.106)$$

Po dalszych przekształceniach otrzymujemy:

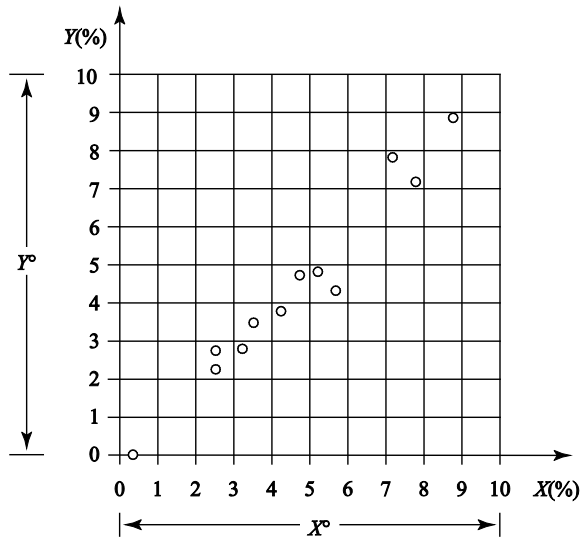
$$r_{xy} = \frac{\sum_{i=1}^N x_i y_i - \frac{1}{N} \sum_{i=1}^N x_i \sum_{i=1}^N y_i}{\sqrt{\left[\sum_{i=1}^N x_i^2 - \frac{1}{N} \left(\sum_{i=1}^N x_i \right)^2 \right] \left[\sum_{i=1}^N y_i^2 - \frac{1}{N} \left(\sum_{i=1}^N y_i \right)^2 \right]}}. \quad (3.107)$$

Kwadrat współczynnika korelacji (r_{xy}^2) nazywany jest współczynnikiem determinacji. Wartość tego współczynnika informuje, jaka część zmienności jednej zmiennej jest zdeterminowana przez zmienność drugiej zmiennej. Wartość współczynnika determinacji można wyrazić w procentach mnożąc r_{xy}^2 przez 100. Różnica $1 - r_{xy}^2$ albo $100(1 - r_{xy}^2)$ nosi nazwę współczynnika indeterminacji.

Przykład 3.14

Porównywano dwie metody oznaczania wilgotności, a mianowicie metodę przemywania acetonem w tyglu Schotta oraz konwencjonalną metodę suszarkową. W tym celu – posługując się obydwoma metodami – zbadano wilgotność $N = 12$ próbek piasku. Na rys. 3.11 pokazano zbiór realizacji obserwowanej zmiennej losowej (X, Y) . Szczegółowe wyniki badań zestawiono w kolumnach 1–3 tablicy 3.16. W dalszych kolumnach tej tablicy przedstawiono obliczenia niezbędne w celu wyznaczenia współczynnika korelacji r_{xy} , według wzoru (3.107). Po podstawieniu wartości otrzymujemy:

$$r_{xy} = \frac{314,1239 - \frac{1}{12} \cdot 55,80 \cdot 52,39}{\sqrt{\left[329,4670 - \frac{1}{12} (55,80)^2\right] \left[300,5121 - \frac{1}{12} (52,39)^2\right]}} = 0,9947.$$



Rys. 3.11. Zbiór wartości i zbiór realizacji dwuwymiarowej zmiennej losowej (X, Y)

Wykorzystując r_{xy} obliczamy wartość współczynnika determinacji i współczynnika indeterminacji. Mamy mianowicie: $r_{xy}^2 = 0,9894$ (98,94%), $1 - r_{xy}^2 = 0,0106$ (1,06%). Tak więc obserwowane zmienne losowe są niemal całkowicie współzależne. Niemal 99% zmienności jednej zmiennej (X lub Y) można wyjaśnić zmiennością drugiej zmiennej (X lub Y). ■

Tablica 3.16

Badanie współzależności między wynikami oznaczania wilgotności piasku dwiema metodami

Nr próbki piasku i	Wilgotność [%]		$x_i y_i$	x_i^2	y_i^2
	Metoda suszarkowa x_i	Metoda przemywania acetonem y_i			
1	2	3	4	5	6
1	7,29	7,62	55,5498	53,1441	58,0644
2	0,08	0,00	0	0,0064	0
3	8,98	8,84	79,3832	80,6404	78,1456
4	3,12	2,77	8,6424	9,7344	7,6729
5	5,25	4,75	24,9375	27,5625	22,5625
6	3,57	3,30	11,7810	12,7449	10,8900
7	7,90	7,40	58,4600	62,4100	54,7600
8	4,17	3,86	16,0962	17,3889	14,8996
9	2,57	2,45	6,2965	6,6049	6,0025
10	5,35	4,71	25,1985	28,6225	22,1841
11	4,84	4,56	22,0704	23,4256	20,7936
12	2,68	2,13	5,7084	7,1824	4,5369
Σ	55,80	52,39	314,1239	329,4670	300,5121

Źródło: badania własne.

3.3.2.2. Empiryczny współczynnik korelacji dwuseryjnej

Niech (X, Y) oznacza dwuwymiarową zmienną losową, taką że:

- X jest zmienną zero-jedynkową⁴ o zbiorze wartości $X^\circ = \{0; 1\}$,
- Y jest zmienną ciągłą w przedziale $[y_d; y_g]$ o zbiorze wartości $Y^\circ = \{y; y_d \leq y \leq y_g\}$.

Zbiór wartości dwuwymiarowej zmiennej losowej (X, Y) jest iloczynem kartezjańskim zbiorów X° i Y° . W takiej sytuacji miarą współzależności między zmiennymi losowymi X i Y jest tzw. współczynnik korelacji dwuseryjnej postaci:

$$r_{d,xy} = \frac{\bar{y}_1 - \bar{y}_0}{s} \sqrt{\frac{N_1 \cdot N_0}{N(N-1)}}, \quad (3.108)$$

⁴ Zero-jedynkową zmienną losową zdefiniowano w punkcie 1.4.1.1; zob. też przykład 1.2.

gdzie:

- \bar{y}_0 – średnia arytmetyczna realizacji zmiennej losowej Y , skojarzonych z realizacjami zmiennej losowej X o wartości 0,
- \bar{y}_1 – średnia arytmetyczna realizacji zmiennej losowej Y , skojarzonych z realizacjami zmiennej losowej X o wartości 1,
- N_0 – liczebność podzbioru zer w zbiorze realizacji X_N ,
- N_1 – liczebność podzbioru jedynek w zbiorze realizacji X_N , przy czym $N_0 + N_1 = N$,
- s – odchylenie standardowe zmiennej losowej Y .

Współczynnik korelacji dwuseryjnej może przyjmować wartości z przedziału $[-1; 0]$ albo z przedziału $[0; 1]$.

Przykład 3.15

W kolumnach 2 i 3 tablicy 3.17 przedstawiono wyniki pewnego eksperymentu technologicznego. Metodą odlewania ciśnieniowego otrzymano $N = 8$ kształtek poliamidowych. Każdy element tego zbioru badano ze względu na:

- barwę (oceniając organoleptycznie, przy zastosowaniu dwupunktowej skali nominalnej),
- twardość (mierzoną metodą Brinella na skali przedziałowej).

Tablica 3.17

Wyniki badania kształtek poliamidowych otrzymanych metodą odlewania ciśnieniowego

i	Barwa x_i	Twardość według Brinella y_i	y_i^2
1	2	3	4
1	1	8,21	67,4041
2	0	9,37	87,7969
3	0	8,90	79,2100
4	1	8,04	64,6416
5	0	8,80	77,4400
6	1	8,43	71,0649
7	1	7,96	63,3616
8	0	9,26	85,7476
Σ	–	68,97	596,6667

Źródło: opracowanie własne.

Uzyskane empirycznie oceny barwy potraktowano jako realizacje zero-jedynkowej zmiennej losowej X , takiej że:

$$X = \begin{cases} 0 & \text{– gdy badany element miał naturalną (żółtawą) barwę tworzywa,} \\ 1 & \text{– w pozostałych przypadkach.} \end{cases}$$

Zastosowana metoda badania twardości pozwala traktować uzyskiwane wyniki jako realizacje ciągłej zmiennej losowej o wartościach w zbiorze \mathbf{R}^+ . Zmienną tę oznaczono symbolem Y .

Do pomiaru korelacji między tak zdefiniowanymi zmiennymi losowymi X i Y można wykorzystać współczynnik korelacji dwuseryjnej postaci (3.108). Obliczone sumy kolumn 3 i 4 tablicy 3.17 podstawiamy do wzoru (3.71):

$$\begin{aligned} s^2 &= \frac{1}{N} \left[\sum_{i=1}^N y_i^2 - \frac{1}{N} \left(\sum_{i=1}^N y_i \right)^2 \right] = \frac{1}{8} \left[596,6667 - \frac{1}{8} (68,97)^2 \right] = \\ &= \frac{1}{8} (596,6667 - 594,6076) = 0,2574. \end{aligned}$$

Biorąc dodatni pierwiastek kwadratowy z wartości wariancji otrzymujemy odchylenie standardowe zmiennej losowej Y :

$$s = \sqrt{s^2} = \sqrt{0,2574} \approx 0,5073 \approx 0,51.$$

Wartości \bar{y}_0 i \bar{y}_1 obliczamy następująco:

$$\bar{y}_0 = \frac{1}{4} (9,37 + 8,90 + 8,80 + 9,26) \approx 9,08,$$

$$\bar{y}_1 = \frac{1}{4} (8,21 + 8,04 + 8,43 + 7,96) \approx 8,16.$$

Z tablicy 3.17 wynika bezpośrednio, że $N_0 = N_1 = 4$. Podstawiając wartości do wzoru (3.108) otrzymujemy:

$$r_{d,xy} = \frac{8,16 - 9,08}{0,51} \sqrt{\frac{4 \cdot 4}{8 \cdot 7}} = -0,9642.$$

Uzyskany wynik świadczy o tym, że w badanym zbiorze kształtek poliamidowych zachodzi bardzo ścisły związek między barwą materiału a jego twardością, przy czym zależność ta ma charakter funkcji malejącej, albowiem $r_{d,xy} < 0$. Dodajmy, że między barwą materiału a jego twardością nie ma oczywiście związku przyczynowo-skutkowego. Barwa materiału jest tu jedynie cechą diagnostyczną, zależną od pewnych własności jego makrostruktury, a te z kolei wywierają wyraźny wpływ na twardość warstwy powierzchniowej. W rozważanym przypadku zbyt jasna barwa może świadczyć o dużej liczbie mikropęcherzyków dwutlenku węgla – używanego w procesie technologicznym – zaokludowanych w powierzchniowej warstwie materiału. Konsekwencją takiego stanu jest obniżona twardość. Zbyt ciemna barwa materiału może świadczyć o destrukcji polimeru, będącej skutkiem obróbki termicznej w procesie technologicznym. Materiał taki może mieć również obniżoną twardość. ■

3.3.2.3. Współczynnik skojarzenia

Założmy, że obydwie składowe dwuwymiarowej zmiennej losowej (X, Y) są zmiennymi losowymi zero-jedynkowymi. Mamy więc $X^\circ = \{0; 1\}$ oraz $Y^\circ = \{0; 1\}$. Zbiór wartości dwuwymiarowej zmiennej losowej (X, Y) jest iloczynem kartezjańskim zbiorów X° i Y° . Tak więc:

$$X^\circ \times Y^\circ = \{(0; 0), (0; 1), (1; 0), (1; 1)\}. \quad (3.109)$$

Zbiór realizacji zmiennej losowej (X, Y) można więc rozciąć na cztery rozłączne podzbiory o liczebnościach $f(0; 0)$, $f(0; 1)$, $f(1; 0)$ i $f(1; 1)$, przy czym:

$$f(0; 0) + f(0; 1) + f(1; 0) + f(1; 1) = N.$$

Współczynnik skojarzenia Q wyznacza się według wzoru:

$$Q_{xy} = \frac{f(0; 0)f(1; 1) - f(0; 1)f(1; 0)}{f(0; 0)f(1; 1) + f(0; 1)f(1; 0)}. \quad (3.110)$$

Współczynnik ten jest miarą unormowaną i przyjmuje wartości z przedziału $[-1; 0]$ albo $[0; 1]$. Własności współczynnika skojarzenia Q ilustruje rys. 3.12. Przy sporządzaniu tego rysunku założono, że zbiór realizacji ma liczebność $N = 100$, a wszystkie rozkłady brzegowe są jednakowe.

A	$Q_{xy} = -1$																
<table border="1" style="border-collapse: collapse; text-align: center;"> <tr> <td rowspan="2" style="padding: 5px;">$X \backslash Y$</td> <td colspan="2" style="padding: 5px;">Y^o</td> <td rowspan="2" style="padding: 5px;">Σ</td> </tr> <tr> <td style="padding: 5px;">0</td> <td style="padding: 5px;">1</td> </tr> <tr> <td rowspan="2" style="padding: 5px;">X^o</td> <td style="padding: 5px;">0</td> <td style="padding: 5px;">50</td> <td style="padding: 5px;">50</td> </tr> <tr> <td style="padding: 5px;">1</td> <td style="padding: 5px;">0</td> <td style="padding: 5px;">50</td> </tr> <tr> <td colspan="2" style="padding: 5px;">Σ</td> <td style="padding: 5px;">50</td> <td style="padding: 5px;">100</td> </tr> </table>	$X \backslash Y$	Y^o		Σ	0	1	X^o	0	50	50	1	0	50	Σ		50	100
$X \backslash Y$		Y^o			Σ												
	0	1															
X^o	0	50	50														
	1	0	50														
Σ		50	100														

B	$Q_{xy} = 0$																
<table border="1" style="border-collapse: collapse; text-align: center;"> <tr> <td rowspan="2" style="padding: 5px;">$X \backslash Y$</td> <td colspan="2" style="padding: 5px;">Y^o</td> <td rowspan="2" style="padding: 5px;">Σ</td> </tr> <tr> <td style="padding: 5px;">0</td> <td style="padding: 5px;">1</td> </tr> <tr> <td rowspan="2" style="padding: 5px;">X^o</td> <td style="padding: 5px;">0</td> <td style="padding: 5px;">25</td> <td style="padding: 5px;">50</td> </tr> <tr> <td style="padding: 5px;">1</td> <td style="padding: 5px;">25</td> <td style="padding: 5px;">50</td> </tr> <tr> <td colspan="2" style="padding: 5px;">Σ</td> <td style="padding: 5px;">50</td> <td style="padding: 5px;">100</td> </tr> </table>	$X \backslash Y$	Y^o		Σ	0	1	X^o	0	25	50	1	25	50	Σ		50	100
$X \backslash Y$		Y^o			Σ												
	0	1															
X^o	0	25	50														
	1	25	50														
Σ		50	100														

C	$Q_{xy} = 1$																
<table border="1" style="border-collapse: collapse; text-align: center;"> <tr> <td rowspan="2" style="padding: 5px;">$X \backslash Y$</td> <td colspan="2" style="padding: 5px;">Y^o</td> <td rowspan="2" style="padding: 5px;">Σ</td> </tr> <tr> <td style="padding: 5px;">0</td> <td style="padding: 5px;">1</td> </tr> <tr> <td rowspan="2" style="padding: 5px;">X^o</td> <td style="padding: 5px;">0</td> <td style="padding: 5px;">50</td> <td style="padding: 5px;">50</td> </tr> <tr> <td style="padding: 5px;">1</td> <td style="padding: 5px;">0</td> <td style="padding: 5px;">50</td> </tr> <tr> <td colspan="2" style="padding: 5px;">Σ</td> <td style="padding: 5px;">50</td> <td style="padding: 5px;">100</td> </tr> </table>	$X \backslash Y$	Y^o		Σ	0	1	X^o	0	50	50	1	0	50	Σ		50	100
$X \backslash Y$		Y^o			Σ												
	0	1															
X^o	0	50	50														
	1	0	50														
Σ		50	100														

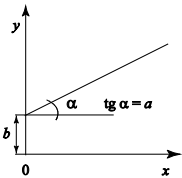
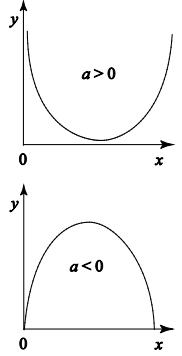
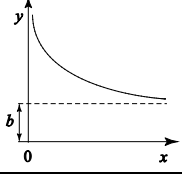
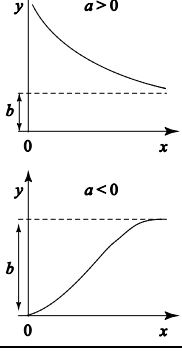
Rys. 3.12. Własności współczynnika skojarzenia Q_{xy}

3.3.3. Regresja dwóch zmiennych

W przypadku gdy realizacje obydwu zmiennych losowych (X i Y) otrzymuje się w rezultacie pomiarów na mocnych skalach (przedziałowych lub ilorazowych), badaniu poddaje się nie tylko ścisłość współzależności między tymi zmiennymi, ale również postać funkcji, za pomocą której można aproksymować zależność między nimi. Jest to analiza regresji. Badaniu można poddać zarówno regresję Y względem X , jak i regresję X względem Y . Jeśli badamy regresję Y względem X , to zmienną X traktujemy jako zmienną objaśniającą, natomiast Y jako zmienną objaśnianą. W takiej sytuacji obserwowaną zmienność Y chcemy wyjaśnić poprzez związek tej zmiennej ze zmienną X . Jeśli odwrócimy role zmiennych, to wpływem zmiennej Y chcemy wyjaśnić obserwowaną zmienność X . Badamy wówczas regresję X względem Y . Zmienna Y jest w takiej sytuacji zmienną objaśniającą, a X zmienną objaśnianą.

Tablica 3.18

Przegląd funkcji najczęściej wykorzystywanych w analizie regresji dwóch zmiennych

Wykres funkcji	Równanie	Siatka funkcyjna linearyzująca wykres	Równania normalne
1	2	3	4
	$y = ax + b$	—	$\sum_{i=1}^N y_i = a \sum_{i=1}^N x_i + bN$ $\sum_{i=1}^N x_i y_i = a \sum_{i=1}^N x_i^2 + b \sum_{i=1}^N x_i$
	$y = ax^2 + bx + c$	—	$\sum_{i=1}^N y_i = a \sum_{i=1}^N x_i^2 + b \sum_{i=1}^N x_i + Nc$ $\sum_{i=1}^N x_i y_i = a \sum_{i=1}^N x_i^3 + b \sum_{i=1}^N x_i^2 + c \sum_{i=1}^N x_i$ $\sum_{i=1}^N x_i^2 y_i = a \sum_{i=1}^N x_i^4 + b \sum_{i=1}^N x_i^3 + c \sum_{i=1}^N x_i^2$
	$y = ax^{-1} + b$	$u = x^{-1}$ $v = y$	$\sum_{i=1}^N y_i = a \sum_{i=1}^N x_i^{-1} + bN$ $\sum_{i=1}^N x_i^{-1} y_i = a \sum_{i=1}^N x_i^{-2} + b \sum_{i=1}^N x_i^{-1}$
	$y = e^{ax} b$	$u = x^{-1}$ $v = \ln y$ lub $v = \lg y$	$\sum_{i=1}^N \ln y_i = a \sum_{i=1}^N x_i^{-2} + N \ln b$ $\sum_{i=1}^N x_i^{-1} \ln y_i = a \sum_{i=1}^N x_i^{-2} + \ln b \sum_{i=1}^N x_i^{-1}$ $\sum_{i=1}^N \lg y_i = 0,4343a \sum_{i=1}^N x_i^{-1} + N \lg b$ $\sum_{i=1}^N x_i^{-1} \lg y_i = 0,4343a \sum_{i=1}^N x_i^{-1} + \lg b \sum_{i=1}^N x_i^{-1}$

cd. tablicy 3.18

1	2	3	4
	$y = bx^a$	$u = \ln x$ lub $u = \lg x$ $v = \ln y$ lub $v = \lg y$	$\sum_{i=1}^N \ln y_i = a \sum_{i=1}^N \ln x_i = N \ln b$ lub $\sum_{i=1}^N \ln x_i \ln y_i = a \sum_{i=1}^N (\ln x_i)^2 + \ln b \sum_{i=1}^N \ln x_i$
	$y = be^{ax}$	$u = x$ $v = \ln y$ lub $v = \lg y$	$\sum_{i=1}^N \ln y_i = a \sum_{i=1}^N x_i + N \ln b$ lub $\sum_{i=1}^N x_i \ln y_i = a \sum_{i=1}^N x_i^2 + \ln b \sum_{i=1}^N x_i$ lub $\sum_{i=1}^N \lg y_i = 0,4343a \sum_{i=1}^N x_i + N \lg b$ lub $\sum_{i=1}^N x_i \lg y_i = 0,4343a \sum_{i=1}^N x_i^2 + \lg b \sum_{i=1}^N x_i$
	$y = \frac{a}{x^{-1} + b}$	$u = x^{-1}$ $v = y^{-1}$	$\sum_{i=1}^N y_i^{-1} = a^{-1} \sum_{i=1}^N x_i^{-1} + N \frac{b}{a}$ lub $\sum_{i=1}^N x_i^{-1} y_i^{-1} = a^{-1} \sum_{i=1}^N x_i^{-2} + \frac{b}{a} \sum_{i=1}^N x_i^{-1}$
	$y = ae^{-x} + b$	$u = e^{-x}$	$\sum_{i=1}^N y_i = a \sum_{i=1}^N e^{-x_i} + Nb$ lub $\sum_{i=1}^N y_i e^{-x_i} = a \sum_{i=1}^N e^{-2x_i} + b \sum_{i=1}^N e^{-x_i}$

Przebieg analizy regresji dwóch zmiennych nie zależy od tego, którą z nich traktujemy jako objaśniającą, a którą jako objaśnianą. Zwróćmy zatem uwagę na przypadek, gdy badaniu podlega zależność ogólnej postaci $Y = f(X)$. Podstawowe funkcje wykorzystywane do aproksymacji szczegółowej postaci tej zależności zestawiono w tabelicy 3.18. O wyborze właściwego modelu powinny decydować przesłanki merytoryczne, a przede wszystkim mechanizm badanego zjawiska lub struktura i właściwości obiektu, który podlega badaniu. Ponieważ jednak najczęściej przesłanki te są niewystarczające, zatem podstawową rolę odgrywa tu wizualna ocena korelogramu, czyli wykresu punktowego przedstawiającego empiryczną zależność między zmiennymi X i Y . Przykładem takiego korelogramu jest wykres pokazany na rys. 3.11. Najczęściej poprzestajemy na badaniu regresji liniowej. Do danych empirycznych dopasowujemy wówczas funkcję liniową postaci:

$$y = ax + b. \quad (3.111)$$

Wykorzystując metodę najmniejszych kwadratów (por. punkt 3.2.2.1) dążymy do ustalenia wartości parametrów a i b na takim poziomie, by:

$$\sum_{i=1}^N (y_i - \hat{y}_i)^2 = \text{minimum}. \quad (3.112)$$

W wyrażeniu tym y_i jest empirycznie uzyskaną realizacją zmiennej losowej Y (zmiennej objaśnianej), natomiast \hat{y}_i wynika z równania regresji:

$$\hat{y}_i = ax_i + b, \quad (3.113)$$

w którym x_i jest empirycznie uzyskaną realizacją zmiennej objaśniającej (X).

Wykorzystując równanie (3.113), warunek (3.112) można zapisać następująco:

$$\sum_{i=1}^N (y_i - (ax_i + b))^2 = \sum_{i=1}^N (y_i - ax_i - b)^2 = \text{minimum}. \quad (3.114)$$

Wyrażenie to potraktujemy jako funkcję a i b :

$$f(a, b) = \sum_{i=1}^N (y_i - ax_i - b)^2. \quad (3.115)$$

Warunek osiągnięcia przez tę funkcję minimum przedstawia się następująco:

$$\frac{\partial f(a, b)}{\partial a} = 0, \quad \frac{\partial f(a, b)}{\partial b} = 0.$$

Różniczkując funkcję (3.115) względem a i b otrzymujemy:

$$\frac{\partial f(a, b)}{\partial a} = -2 \sum_{i=1}^N y_i + 2Nb + 2a \sum_{i=1}^N x_i, \quad (3.116)$$

$$\frac{\partial f(a, b)}{\partial b} = -2 \sum_{i=1}^N x_i y_i + 2b \sum_{i=1}^N x_i + 2a \sum_{i=1}^N x_i^2. \quad (3.117)$$

Po przyrównaniu każdej z tych funkcji do zera i po odpowiednich przekształceniach otrzymujemy następujący układ równań:

$$\left. \begin{aligned} a \sum_{i=1}^N x_i + bN &= \sum_{i=1}^N y_i \\ a \sum_{i=1}^N x_i^2 + b \sum_{i=1}^N x_i &= \sum_{i=1}^N x_i y_i \end{aligned} \right\}. \quad (3.118)$$

Są to tzw. równania normalne (zob. tablica 3.18). Rozwiązując ten układ równań względem a i b otrzymujemy:

$$a = \frac{\sum_{i=1}^N x_i y_i - \frac{\sum_{i=1}^N x_i \sum_{i=1}^N y_i}{N}}{\sum_{i=1}^N x_i^2 - \frac{\left(\sum_{i=1}^N x_i\right)^2}{N}}, \quad (3.119)$$

$$b = \frac{\sum_{i=1}^N y_i}{N} - a \frac{\sum_{i=1}^N \bar{x}_i}{N} = \bar{y} - a\bar{x}. \quad (3.120)$$

Ze wzoru (3.119) wynika bezpośrednio, że współczynnik kierunkowy (a) prostej regresji (3.113) określony jest przez stosunek empirycznej kowariancji do wa-

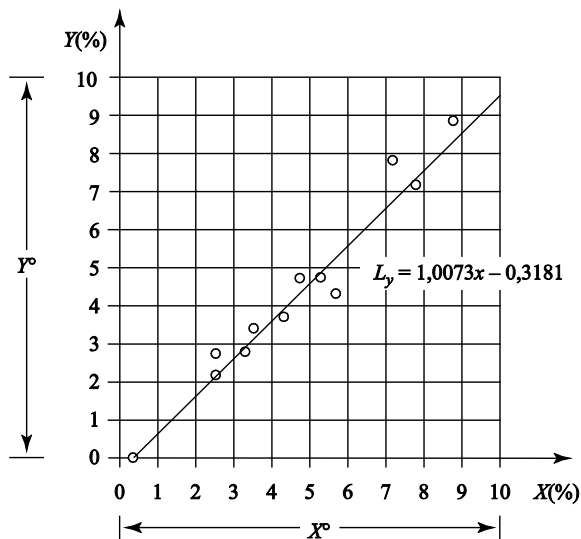
riancji tej zmiennej, która w konkretnym przypadku jest traktowana jako zmienna objaśniająca.

Przykład 3.16

Powróćmy do wyników eksperymentu przedstawionego w przykładzie 3.14. Położenie punktów empirycznych na rys. 3.11 wskazuje na to, że mamy do czynienia z regresją liniową. Wykorzystując dane liczbowe zestawione w tabelicy 3.16 wyznaczmy parametry (a , b) prostej regresji. Na podstawie wzorów (3.119) i (3.120) mamy:

$$a = \frac{314,1239 - \frac{55,80 \cdot 52,39}{12}}{329,4670 - \frac{(55,80)^2}{12}} = \frac{70,5104}{69,9970} = 1,0073[\%],$$

$$b = \frac{52,39}{12} - 1,0073 \frac{55,80}{12} = -0,3181[\%].$$



Rys. 3.13. Dopasowanie prostej regresji do danych empirycznych z tabelicy 3.16

Równanie prostej regresji przedstawia się więc następująco:

$$\hat{y}_i = 1,0073 x_i - 0,3181.$$

Wykres tej prostej pokazano na rys. 3.13.□

3.3.4. Uogólnienie współczynnika korelacji

Ogólną zmienność zmiennej Y w danym układzie doświadczalnym charakteryzuje suma:

$$\sum_{i=1}^N (y_i - \bar{y})^2. \quad (3.121)$$

Zależność $Y = f(X)$ wyjaśnia i usuwa część tej zmienności. Tę mianowicie część, która wynika z faktu, że każda zrealizowana wartość zmiennej Y jest w pewnym stopniu zdeterminowana przez odpowiadającą jej realizację zmiennej losowej X . Nie wyjaśniona pozostaje ta część ogólnej zmienności zmiennej Y , która jest charakteryzowana przez sumę:

$$\sum_{i=1}^N (y_i - \hat{y}_i)^2. \quad (3.122)$$

Jest to tzw. zmienność resztowa i jest ona oczywiście tym większa, im gorzej dopasowana jest linia regresji do „chmury” punktów empirycznych na korelogramie. Zauważmy, że gdyby wszystkie punkty owej „chmury” leżały dokładnie na linii regresji, to:

$$\bigwedge_i y_i = \hat{y}_i, \quad (3.123)$$

a w konsekwencji suma (3.122) przyjęłaby wartość zero. Można zatem przeprowadzić następujące rozumowanie:

skoro $\sum_{i=1}^N (y_i - \bar{y})^2$ stanowi 100% zmienności zmiennej losowej Y ,

to $\sum_{i=1}^N (y_i - \hat{y}_i)^2$ stanowi $D\%$ zmienności zmiennej losowej Y .

W konsekwencji

$$D [\%] = \frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{\sum_{i=1}^N (y_i - \bar{y})^2} 100. \quad (3.124)$$

Możemy zatem zmierzyć (na przykład w procentach), jaka część ogólnej zmienności zmiennej objaśnianej pozostaje nie wyjaśniona, po uwzględnieniu faktu, że pewna część tej zmienności wynika z zależności Y od X . Zdefiniowana w ten sposób unormowana miara braku determinacji w układzie dwóch zmiennych losowych (X, Y) odpowiada pojęciowo współczynnikowi indeterminacji ($1 - r_{xy}^2$), nazywanemu też współczynnikiem zbieżności, zdefiniowanemu w punkcie 3.3.2.1.

Możemy więc napisać:

$$1 - R_{xy}^2 = \frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{\sum_{i=1}^N (y_i - \bar{y})^2}. \quad (3.125)$$

Po niezbędnych przekształceniach mamy:

$$R_{xy}^2 = 1 - \frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{\sum_{i=1}^N (y_i - \bar{y})^2} \quad (3.126)$$

i w konsekwencji

$$R_{xy} = \sqrt{R_{xy}^2}. \quad (3.127)$$

Jest to uogólniony współczynnik korelacji. Nazywany jest on najczęściej współczynnikiem korelacji wielorakiej. Znajduje to uzasadnienie w fakcie, że współczynnik ten może być wykorzystany jako unormowana miara ścisłości związku korelacyjnego wielu zmiennych. Może on być także wykorzystany jako współczynnik korelacji krzywoliniowej dwóch zmiennych, a więc wówczas, gdy regresja nie ma charakteru liniowego. W przypadku regresji liniowej uogólniony

współczynnik korelacji R_{xy} jest równoważny współczynnikowi korelacji liniowej r_{xy} danemu wzorami (3.105)–(3.107). Zauważmy jednak, że o ile współczynnik korelacji liniowej r_{xy} , w zależności od znaku kowariancji, może przyjmować wartości z przedziału $[-1; 0]$ albo z przedziału $[1; 0]$, o tyle współczynnik R_{xy} przyjmuje wartości z przedziału $[0; 1]$, albowiem jest on zdefiniowany jako dodatni pierwiastek kwadratowy z odpowiedniego współczynnika determinacji; zob. wzór (3.127). W przypadku gdy współczynnik regresji a w równaniu (3.113) jest ujemny ($a < 0$), współczynnikowi R_{xy} przypisuje się znak minus. W celu zilustrowania związku między r_{xy} i R_{xy} posłużymy się przykładem numerycznym.

Przykład 3.17

Odwołajmy się powtórnie do danych empirycznych przedstawionych w przykładzie 3.14. Na podstawie równania regresji:

$$\hat{y}_i = 1,0073 x_i - 0,3181$$

wyznaczonego w przykładzie 3.16 możemy obliczyć wartość resztowej sumy kwadratów (3.122). Obliczenia przedstawiono w tabelicy 3.19. Ogólną sumę kwadratów obliczymy na podstawie danych zaczerpniętych z tabelicy 3.16. Mamy mianowicie:

$$\sum_{i=1}^N (y_i - \bar{y})^2 = 300,5121 - \frac{(52,39)^2}{12} = 71,7861.$$

Na podstawie wzorów (3.126) i (3.127) otrzymujemy:

$$R_{xy}^2 = 1 - \frac{0,7516}{71,7861} = 0,9895,$$

$$R_{xy} = \sqrt{0,9894} = 0,9948.$$

W przykładzie 3.14 otrzymaliśmy $r_{xy} = 0,9947$. Różnica między uzyskanymi wynikami ($R_{xy} - r_{xy} = 0,0001$) nie ma żadnego praktycznego znaczenia i wynika z zaokrągleń w operacjach numerycznych. ■

Współczynnik korelacji R_{xy} może być stosowany przy dowolnej postaci funkcji regresji, podczas gdy współczynnik korelacji liniowej r_{xy} powinno się oczywiście stosować tylko wówczas, gdy spełnione jest założenie o liniowości funkcji regresji. Posługując się wzorami (3.125)–(3.127) można więc obliczyć

wartość współczynnika korelacji, determinacji i indeterminacji niezależnie od tego, która spośród funkcji przedstawionych w tabelicy 3.18 zostanie wykorzystana do aproksymacji zależności między zmiennymi X i Y .

Tablica 3.19

Obliczanie resztowej sumy kwadratów

i	y_i	\hat{y}_i	$y_i - \hat{y}_i$	$(y_i - \hat{y}_i)^2$
1	2	3	4	5
1	7,62	7,03	0,59	0,3481
2	0,00	-0,24	0,24	0,0576
3	8,84	8,73	0,11	0,0121
4	2,77	2,82	-0,05	0,0025
5	4,75	4,97	-0,22	0,0484
6	3,30	3,28	0,02	0,0004
7	7,40	7,64	-0,24	0,0576
8	3,86	3,88	-0,02	0,0004
9	2,45	2,27	0,18	0,0324
10	4,71	5,07	-0,36	0,1296
11	4,56	4,56	0,00	0,0000
12	2,13	2,38	-0,25	0,0625
Σ	52,39	52,39	0,00	0,7516

Źródło: obliczenia własne.

3.3.5. Regresja wielowymiarowa oraz korelacja wieloraka i cząstkowa

Jeśli model regresji z jedną zmienną objaśniającą (zob. tablica 3.18) nie generuje dostatecznie wysokiej wartości współczynnika determinacji R_{xy}^2 (zob. wzór (3.126)), to w celu zmniejszenia zmienności resztowej wprowadza się do modelu dodatkowe zmienne objaśniające. Przechodzimy więc do badania kolejnych modeli ogólnej postaci $Y = f(X_1, X_2, \dots, X_k)$. Jest to model regresji wielowymiarowej. Siłę związku między wektorem zmiennych objaśniających (X_1, X_2, \dots, X_k) a zmienną objaśnianą (Y) mierzy wówczas współczynnik korelacji wielorakiej, wyznaczany według wzoru (3.127). Powiększanie wektora zmiennych objaśniających o kolejne zmienne ma oczywiście sens tylko wówczas, gdy zwiększa to wartość współczynnika determinacji. Jeśli nie osiągamy tego efektu, to wprowadzamy

dzanie do modelu dodatkowych zmiennych objaśniających jest nieracjonalne, albowiem zwiększa ilość niezbędnych obliczeń, nie przyczyniając się jednocześnie do pełniejszego wyjaśnienia zmienności właściwości obserwowanych obiektów lub powtórzeń zjawiska.

W procesie badania regresji wielowymiarowej i korelacji wielorakiej podstawową rolę odgrywają modele liniowe postaci:

$$y = a_0 + a_1x_1 + a_2x_2 + \dots + a_kx_k. \quad (3.128)$$

W przypadku dwóch zmiennych objaśniających (X_1, X_2) równanie płaszczyzny regresji przedstawia się więc następująco:

$$y = a_0 + a_1x_1 + a_2x_2. \quad (3.129)$$

Stosując metodę najmniejszych kwadratów uzyskujemy następujący układ równań normalnych:

$$\left. \begin{aligned} a_0N + a_1 \sum_{i=1}^N x_{1i} + a_2 \sum_{i=1}^N x_{2i} &= \sum_{i=1}^N y_i \\ a_0 \sum_{i=1}^N x_{1i} + a_1 \sum_{i=1}^N x_{1i}^2 + a_2 \sum_{i=1}^N x_{1i}x_{2i} &= \sum_{i=1}^N x_{1i}y_i \\ a_0 \sum_{i=1}^N x_{2i} + a_1 \sum_{i=1}^N x_{1i}x_{2i} + a_2 \sum_{i=1}^N x_{2i}^2 &= \sum_{i=1}^N x_{2i}y_i \end{aligned} \right\}. \quad (3.130)$$

W równaniach tych y_i oznacza i -tą realizację zmiennej objaśnianej Y , natomiast x_{1i} oraz x_{2i} są odpowiednimi realizacjami zmiennych objaśniających X_1 i X_2 . Rozwiązując układ równań (3.130) względem a_0, a_1 i a_2 otrzymujemy parametry równania (3.129).

Istotne uproszczenie obliczeń związanych z wyznaczeniem wartości tych parametrów można osiągnąć stosując rachunek macierzowy. Zauważmy, że w rezultacie doświadczenia uzyskujemy macierz realizacji trójwymiarowej zmiennej losowej (Y, X_1, X_2):

$$\begin{bmatrix} y_1 & x_{11} & x_{12} \\ y_2 & x_{12} & x_{22} \\ \vdots & \vdots & \vdots \\ y_i & x_{1i} & x_{2i} \\ \vdots & \vdots & \vdots \\ y_N & x_{1N} & x_{2N} \end{bmatrix}. \quad (3.131)$$

Każdą wartość y_i traktujemy tu jako liniową funkcję X_1 i X_2 postaci (3.129). Zaczodzi więc następująca zależność:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_i \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} a_0 + a_1 x_{11} + a_2 x_{12} \\ a_0 + a_1 x_{12} + a_2 x_{22} \\ \vdots \\ a_0 + a_1 x_{1i} + a_2 x_{2i} \\ \vdots \\ a_0 + a_1 x_{1N} + a_2 x_{2N} \end{bmatrix}. \quad (3.132)$$

Prawą stronę tej równości można przedstawić w formie iloczynu zmodyfikowanej macierzy realizacji zmiennych objaśniających i wektora kolumnowego parametrów równania (3.129). Mamy mianowicie:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_i \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{12} & x_{22} \\ \vdots & \vdots & \vdots \\ 1 & x_{1i} & x_{2i} \\ \vdots & \vdots & \vdots \\ 1 & x_{1N} & x_{2N} \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix}. \quad (3.133)$$

Wynika stąd bezpośrednio, że:

$$\mathbf{a} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}, \quad (3.134)$$

gdzie \mathbf{a} jest wektorem parametrów, \mathbf{Y} oznacza wektor realizacji zmiennej objaśnianej Y , natomiast \mathbf{X} jest zmodyfikowaną macierzą realizacji zmiennych objaśniających. Przedstawione powyżej rozważania zilustrujemy na przykładzie numerycznym.

Przykład 3.18

Dwunastu studentów ($N = 12$) zdawało egzamin z pewnego przedmiotu. Poziom wiedzy studentów oceniano na skali 100-punktowej. Oceny uzyskane przez poszczególnych studentów (y_i) zestawione są w kolumnie 2 tablicy 3.20. Oceny te potraktujemy jako realizacje zmiennej losowej Y . Będzie to zmienna objaśniana. Zmiennymi objaśniającymi będą:

X_1 – iloraz inteligencji,

X_2 – liczba godzin poświęcanych tygodniowo na przygotowanie się do egzaminu.

Realizacje tych zmiennych losowych (x_{1i}, x_{2i}) zestawiono w kolumnach 3 i 4 tablicy 3.20.

Tablica 3.20

Dane liczbowe do przykładu 3.18

i	Wynik egzaminu y_i	Iloraz inteligencji x_{1i}	Liczba godzin nauki x_{2i}
1	2	3	4
1	83	112	9
2	77	115	6
3	95	129	14
4	49	103	4
5	63	117	8
6	80	115	12
7	91	124	10
8	79	113	9
9	36	106	5
10	58	114	7
11	93	136	8
12	84	127	3

Źródło: J.E. Freund, *Podstawy nowoczesnej statystyki*, PWE, Warszawa 1968, s. 336 (podane w tablicy dane liczbowe są treścią zadania przeznaczonego do rozwiązania przez Czytelnika).

Rozważmy kolejno trzy zależności, a mianowicie:

– zależność wyniku egzaminu od ilorazu inteligencji

$$Y_{(1)} = a_{01} + a_{11}X_1, \quad (\text{i})$$

– zależność wyniku egzaminu od liczby godzin poświęconych na przygotowanie się do egzaminu

$$Y_{(2)} = a_{02} + a_{12}X_2, \quad (\text{ii})$$

– zależność wyniku egzaminu od obu wymienionych powyżej zmiennych objaśniających traktowanych łącznie

$$Y = a_0 + a_1X_1 + a_2X_2. \quad (\text{iii})$$

Załóżmy, że poszukujemy takiego modelu regresji, który pozwoli wyjaśnić co najmniej 70% zaobserwowanej zmienności zmiennej losowej Y . Najmniejsza akceptowana wartość współczynnika determinacji wynosi więc 0,70.

Ad (i)

Na podstawie danych przedstawionych w tabelicy 3.20 mamy:

$$\begin{aligned}\sum_{i=1}^{12} y_i &= 888, & \sum_{i=1}^{12} y_i^2 &= 69500, \\ \sum_{i=1}^{12} x_{1i} &= 1411, & \sum_{i=1}^{12} x_{1i}^2 &= 166935, \\ \sum_{i=1}^{12} x_{1i} y_i &= 105979.\end{aligned}$$

Wykorzystując te sumy obliczamy wartości parametrów liniowego równania regresji (i):

$$Y_{(1)} = -105,5497 + 1,5270 X_1$$

oraz współczynnika korelacji liniowej ($r_{yx_1} = 0,7943$), współczynnika determinacji ($r_{yx_1}^2 = 0,6309$) i współczynnika indeterminacji ($1 - r_{yx_1}^2 = 0,3691$). Okazuje się więc, że zróżnicowanie ilorazu inteligencji wyjaśnia nieco ponad 63% ogólnej zmienności wyników egzaminu.

Ad (ii)

Postępując analogicznie jak poprzednio, wyznaczamy wartości następujących sum:

$$\sum_{i=1}^{12} x_{2i} = 95, \quad \sum_{i=1}^{12} x_{2i}^2 = 865, \quad \sum_{i=1}^{12} x_{2i} y_i = 7402.$$

W konsekwencji mamy:

$$Y_{(2)} = 47,9185 + 3,2945 X_2$$

a także

$$r_{yx_2} = 0,5688, \quad r_{yx_2}^2 = 0,3235, \quad 1 - r_{yx_2}^2 = 0,6765.$$

Zatem tylko 32,35% ogólnej zmienności zmiennej losowej Y można wyjaśnić zróżnicowaniem czasu, jaki poświęcili poszczególni studenci na przygotowanie się do egzaminu.

Ad (iii)

Z przedstawionych powyżej obliczeń wynika, że jeśli zmienne objaśniające X_1 i X_2 wykorzystywane są w osobnych modelach regresji, to żadna z nich nie zapewnia odpowiednio wysokiej wartości współczynnika determinacji. Obecnie potraktujemy te zmienne łącznie, w jednym modelu regresji. Wykorzystując dane empiryczne przedstawione w tabelicy 3.20 możemy napisać:

$$Y = \begin{bmatrix} 83 \\ 77 \\ 95 \\ 49 \\ 63 \\ 80 \\ 91 \\ 79 \\ 36 \\ 58 \\ 93 \\ 84 \end{bmatrix} \quad X = \begin{bmatrix} 1 & 112 & 9 \\ 1 & 115 & 6 \\ 1 & 129 & 14 \\ 1 & 103 & 4 \\ 1 & 117 & 8 \\ 1 & 115 & 12 \\ 1 & 124 & 10 \\ 1 & 113 & 9 \\ 1 & 106 & 5 \\ 1 & 114 & 7 \\ 1 & 136 & 8 \\ 1 & 127 & 3 \end{bmatrix} .$$

Poszukiwany wektor parametrów równania (iii) ma postać:

$$a = \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} .$$

Na podstawie wzoru (3.134) mamy:

$$a = (X^T X)^{-1} X^T Y = \begin{bmatrix} -95,085 \\ 1,306 \\ 1,957 \end{bmatrix} .$$

Równanie (iii) przedstawia się więc następująco:

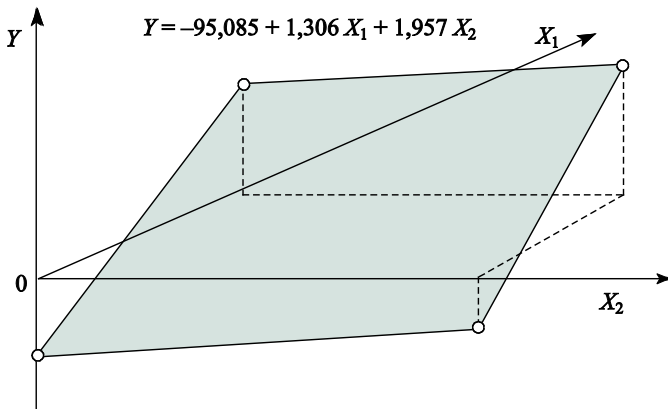
$$Y = -95,085 + 1,306 X_1 + 1,957 X_2.$$

Wykres tej płaszczyzny regresji pokazano na rys. 3.14. Zapisując otrzymane równanie w postaci:

$$\hat{y}_i = -95,085 + 1,306 x_{1i} + 1,957 x_{2i}$$

można wyznaczyć wartość resztowej sumy kwadratów (3.122). Obliczenia przebiegają analogicznie jak w tabelicy 3.19 (zob. przykład 3.17). Po wykonaniu tych obliczeń otrzymujemy:

$$\sum_{i=1}^{12} (y_i - \hat{y}_i)^2 = 1018,781.$$



Rys. 3.14. Płaszczyzna regresji

Mając wartość tej sumy można wyznaczyć współczynnik determinacji i indeterminacji, a także współczynnik korelacji wielorakiej. Mamy mianowicie:

$$1 - R^2_{y \cdot x_1 x_2} = \frac{\sum_{i=1}^{12} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{121} (y_i - \bar{y})^2} = 1 - \frac{1018,781}{378} = 0,269.$$

W konsekwencji

$$R^2_{y.x_1x_2} = 0,731, \quad R_{y.x_1x_2} = 0,855.$$

Okazuje się więc, że po wprowadzeniu do modelu dwóch zmiennych objaśniających potrafimy wyjaśnić 73,1% ogólnej obserwowanej zmienności zmiennej Y . \square

Współczynnik korelacji wielorakiej $R_{y.x_1x_2\dots x_k}$ jest syntetyczną miarą współzależności między zmienną objaśnianą Y a zmiennymi objaśniającymi X_1, X_2, \dots, X_k , traktowanymi łącznie. Miarami współzależności między zmienną Y a poszczególnymi zmiennymi objaśniającymi X_1, X_2, \dots, X_k , traktowanymi z osobna, są współczynniki korelacji cząstkowej. Mierzą one siłę związku między zmienną Y a pojedynczą zmienną objaśniającą, przy wyeliminowaniu wpływu wszystkich pozostałych zmiennych objaśniających.

W celu wyjaśnienia tych problemów zwróćmy ponownie uwagę na trójwymiarową zmienną losową (Y, X_1, X_2) . W przedstawionych poniżej rozważaniach będziemy posługiwać się uproszczonym zapisem subskryptów. Poszczególne zmienne będziemy oznaczać indeksami 0, 1 oraz 2. Tak więc na przykład symbol r_{01} będzie oznaczał współczynnik korelacji między zmienną Y (indeks 0) i zmienną X_1 (indeks 1), natomiast r_{12} będzie współczynnikiem korelacji między X_1 i X_2 .

W rozważanym przypadku trójwymiarowej zmiennej losowej mamy trzy współczynniki korelacji cząstkowej, a mianowicie:

$$r_{01.2} = \frac{r_{01} - r_{02} r_{12}}{\sqrt{(1 - r_{02}^2)(1 - r_{12}^2)}}, \quad (3.135)$$

$$r_{02.1} = \frac{r_{02} - r_{01} r_{12}}{\sqrt{(1 - r_{01}^2)(1 - r_{12}^2)}}, \quad (3.136)$$

$$r_{12.0} = \frac{r_{12} - r_{01} r_{02}}{\sqrt{(1 - r_{01}^2)(1 - r_{02}^2)}}, \quad (3.137)$$

gdzie:

- $r_{01.2}$ – współczynnik korelacji między Y i X_1 , po wyeliminowaniu wpływu zmiennej X_2 ,
- $r_{02.1}$ – współczynnik korelacji między Y i X_2 , po wyeliminowaniu wpływu zmiennej X_1 ,
- $r_{12.0}$ – współczynnik korelacji między X_1 i X_2 , po wyeliminowaniu wpływu zmiennej Y .

Ze wzorów tych wynika bezpośrednio, że w celu wyznaczenia wartości współczynników korelacji cząstkowej niezbędna jest znajomość macierzy współczynników korelacji. Macierz ta ma postać:

$$\mathbf{R} = \begin{bmatrix} r_{00} & r_{01} & r_{02} \\ r_{10} & r_{11} & r_{12} \\ r_{20} & r_{21} & r_{22} \end{bmatrix} = \begin{bmatrix} 1 & r_{01} & r_{02} \\ r_{10} & 1 & r_{12} \\ r_{20} & r_{21} & 1 \end{bmatrix}. \quad (3.138)$$

Ponieważ współczynnik korelacji liniowej jest miarą symetryczną w tym sensie, że $r_{ij} = r_{ji}$, zatem $r_{10} = r_{01}$, $r_{20} = r_{02}$ i $r_{21} = r_{12}$. Na głównej przekątnej macierzy mamy $r_{00} = r_{11} = r_{22} = 1$. Macierz \mathbf{R} można wykorzystać do wyznaczania wartości współczynników korelacji cząstkowej metodami rachunku macierzowego.

W przypadku trójwymiarowej zmiennej losowej mamy:

$$r_{ij.k} = \frac{-\mathbf{R}_{ij}}{\sqrt{\mathbf{R}_{ii}\mathbf{R}_{jj}}}, \quad (1.139)$$

gdzie \mathbf{R}_{ij} oraz \mathbf{R}_{ii} i \mathbf{R}_{jj} są dopełnieniami algebraicznymi macierzy \mathbf{R} , przy czym indeksy i, j oraz k mogą przyjmować wartości 0, 1, 2. Dopełnienie algebraiczne \mathbf{R}_{ij} otrzymujemy skreślając i -ty wiersz oraz j -tą kolumnę w macierzy \mathbf{R} . Dopełnienia \mathbf{R}_{ii} i \mathbf{R}_{jj} uzyskujemy natomiast przez skreślenie i -tego wiersza oraz i -tej kolumny albo j -tego wiersza oraz j -tej kolumny w macierzy \mathbf{R} .

Zastosowanie wzoru (3.139) prowadzi oczywiście do takiego samego rezultatu, jaki uzyskujemy stosując jeden z podanych powyżej wzorów (3.135)–(3.137).

Zwróćmy uwagę na współczynnik korelacji cząstkowej $r_{01.2}$. Po wykonaniu odpowiednich działań otrzymujemy:

$$r_{01.2} = \frac{-\begin{vmatrix} r_{10} & r_{12} \\ r_{20} & 1 \end{vmatrix}}{\sqrt{\begin{vmatrix} r_{11} & r_{12} \\ r_{21} & r_{22} \end{vmatrix} \begin{vmatrix} r_{00} & r_{02} \\ r_{20} & r_{22} \end{vmatrix}}} = \frac{r_{10} - r_{20} r_{12}}{\sqrt{(1 - r_{12}^2)(1 - r_{02}^2)}}.$$

Przykład 3.18 (kontynuacja)

Macierz współczynników korelacji \mathbf{R} przedstawia się następująco:

$$\mathbf{R} = \begin{bmatrix} 1 & 0,7943 & 0,5688 \\ 0,7943 & 1 & 0,3398 \\ 0,5688 & 0,3398 & 1 \end{bmatrix}.$$

Współczynniki korelacji $r_{01} = r_{yx_1} = 0,7943$ oraz $r_{02} = r_{yx_2} = 0,5688$ wyznaczyliśmy w poprzedniej części przykładu. Obecnie, w celu określenia macierzy \mathbf{R} , należy dodatkowo obliczyć wartość współczynnika $r_{12} = r_{x_1x_2}$. Na podstawie danych zestawionych w tablicy 3.20 otrzymujemy $r_{12} = 0,3398$.

Na podstawie wzorów (3.135) i (3.136) mamy:

$$r_{01.2} = \frac{0,7943 - 0,5688 \cdot 0,3398}{\sqrt{(1 - 0,5688^2)(1 - 0,3398^2)}} = 0,7770,$$

$$r_{02.1} = \frac{0,5688 - 0,7943 \cdot 0,3398}{\sqrt{(1 - 0,7943^2)(1 - 0,3398^2)}} = 0,6055.$$

Dla kompletności rozważań obliczymy również wartość współczynnika korelacji cząstkowej $r_{12.0} = r_{x_1x_2.y}$. Wykorzystując wzór (3.137) otrzymujemy:

$$r_{12.0} = \frac{0,3398 - 0,7943 \cdot 0,5688}{\sqrt{(1 - 0,7943^2)(1 - 0,5688^2)}} = -0,2241.$$

Na uwagę zasługuje fakt, że $r_{01} = 0,7943 > r_{01.2} = 0,7770$. Tak więc jeśli rozważany układ doświadczalny uwolnić od wpływu zmiennej objaśniającej X_2 , to wpływ ogólnej wydolności intelektualnej (mierzonej ilorazem inteligencji) na wynik egzaminu jest nieco mniejszy, niż można było pierwotnie przypuszczać. Zauważmy też, że $r_{02} = 0,5688 < r_{02.1} = 0,6055$. Oznacza to, że jeśli wyłączymy z rozważań zmienną objaśniającą X_1 (iloraz inteligencji), to rośnie wpływ liczby godzin poświęconych na przygotowanie się do egzaminu (zmienna X_2) na wynik tego egzaminu (Y). Uzyskana wartość współczynnika $r_{12.0}$ wskazuje na to, że istnieje niewielka ujemna korelacja między X_1 i X_2 . Oznacza to, że studenci o mniejszej wydolności intelektualnej wykazują pewną skłonność do poświęcania większej ilości czasu na naukę. \square

Mając macierz współczynników korelacji \mathbf{R} można obliczyć wartość współczynnika korelacji wielorakiej $R_{y.x_1x_2\dots x_k}$ bez wyznaczania odpowiedniego równania regresji i określania zmienności resztowej. W przypadku trójwymiarowej zmiennej losowej (Y, X_1, X_2) współczynnik korelacji wielorakiej $R_{y.x_1x_2} = R_{0.12}$ wynika z następującego wzoru:

$$R_{0.12} = \sqrt{\frac{r_{01}^2 + r_{02}^2 - 2r_{01}r_{02}r_{12}}{1 - r_{12}^2}}. \quad (3.140)$$

Przykład 3.18 (zakończenie)

Po podstawieniu wyznaczonych powyżej elementów macierzy \mathbf{R} do wzoru (3.140) otrzymujemy:

$$R_{0.12} = \sqrt{\frac{0,7943^2 + 0,5688^2 - 2 \cdot 0,7943 \cdot 0,5688 \cdot 0,3398}{1 - 0,3398^2}} = 0,856.$$

Obliczona wartość jest większa o 0,001 od wartości obliczonej poprzednio na podstawie wzorów (3.125)–(3.127). Różnica ta wynika z zaokrągleń w operacjach numerycznych i nie ma żadnego praktycznego znaczenia. ■

Przy opracowywaniu dużych zbiorów realizacji wielowymiarowych zmiennych losowych powszechnie stosuje się obecnie technikę komputerową. Opracowane do tego celu programy wykorzystują przede wszystkim procedury obliczeniowe oparte na rachunku macierzowym. W takiej sytuacji analiza materiału empirycznego rozpoczyna się od macierzy obserwacji postaci (3.131). Dalsza analiza biegnie dwiema równoległymi drogami. Jeden ciąg obliczeń ma na celu wyznaczenie równania płaszczyzny regresji. Drugi natomiast zmierza do obliczenia odpowiednich miar ścisłości związku korelacyjnego, a przede wszystkim współczynnika korelacji wielorakiej oraz współczynników korelacji cząstkowej. Nie są to drogi niezależne, a szczegółowy przebieg obliczeń zależy od klasy wybranego modelu regresji. W przypadku modeli liniowych postaci (3.128), które dominują w omawianej tu analizie, podstawowe znaczenie ma wyznaczenie macierzy współczynników korelacji \mathbf{R} (zob. wzór (3.138)). Mając tę macierz można wyznaczyć zarówno współczynnik korelacji wielorakiej (wzór (3.140)) i współczynniki korelacji cząstkowej (wzór (3.139)), jak i parametry równania płaszczyzny regresji.

Rozdział 4

ESTYMACJA PARAMETRÓW ZMIENNYCH LOSOWYCH

4.1. Wprowadzenie

W tym rozdziale, a także w rozdziale 5, będziemy zakładać, że zbiorowość generalna A poddawana jest badaniom niewyczerpującym (częściowym), a jedynym empirycznie dostępnym źródłem informacji o tej zbiorowości jest próba \underline{A} . W głównym nurcie rozważań będziemy przyjmować, że \underline{A} jest prostą próbą losową (zob. rozdział 2), a wyjątkowo będziemy również odwoływać się do próby uzyskiwanej na drodze losowej bez zwracania. Rozważania ograniczymy – analogicznie jak dotychczas – do jednowymiarowych i dwuwymiarowych zmiennych losowych. Przypomnijmy (zob. rozdział 2), że w przypadku jednowymiarowej zmiennej losowej X , w rezultacie przeprowadzonych badań otrzymuje się zbiór realizacji postaci (2.25):

$$\mathbf{X}_n = \{x_i; i = 1, 2, \dots, n\}.$$

W przypadku dwuwymiarowej zmiennej losowej (X_1, X_2) zbiór realizacji uzyskiwany w rezultacie doświadczenia ma postać (2.27):

$$(\mathbf{X}_1, \mathbf{X}_2)_n = \{(X_{1i}, X_{2i}); i = 1, 2, \dots, n\}.$$

W literaturze przedmiotu, zwłaszcza w pozycjach nie zorientowanych na zastosowania metod statystycznych w naukach empirycznych, zbiory realizacji \mathbf{X}_n , $(\mathbf{X}_1, \mathbf{X}_2)_n$ nazywane są próbkami. Będziemy również używać tego określenia, pamiętając jednak, że jest to uproszczenie, albowiem należy rozróżnić próbkę (\underline{A}) i jej liczbowy obraz, jakim jest odpowiedni zbiór realizacji. Rozróżnienie to jest istotne, ponieważ jedna próbka może mieć wiele liczbowych obrazów, w zależności od cech wybranych do badania, a także zastosowanych metod badania i funkcji pomiarowych. Charakterystyki obserwowanych zmiennych losowych X i (X_1, X_2) , wyznaczane na podstawie zbiorów realizacji \mathbf{X}_n i $(\mathbf{X}_1, \mathbf{X}_2)_n$, określane są jako charakterystyki z próby albo statystyki. Charakterystyki te są zmiennymi

losowymi, a wartości liczbowe uzyskiwane w rezultacie badania konkretnych próbek \underline{A} są realizacjami tych zmiennych losowych. Wynika to z faktu, że każda zbiorowość generalna A , poddawana badaniom niewyczerpującym generuje pewną przestrzeń prób \underline{A}° , a wylosowana próba \underline{A} jest tylko jednym elementem tej przestrzeni. Problemy te – mające podstawowe znaczenie w procesie szacowania (estymacji) parametrów zmiennych losowych w zbiorowości generalnej – wyjaśnimy na przykładzie.

Przykład 4.1

Niech zbiór

$$A = \{A_1, A_2, A_3, A_4, A_5\}$$

będzie obserwowaną zbiorowością generalną. Założmy, że – w celu zbadania własności tej zbiorowości ze względu na pewną cechę B – postanowiono pobrać próbkę o licznosci $n = 2$. Ile różnych prób o tej licznosci można wylosować ze zbiorowości A ? W przypadku losowania ze zwracaniem każda taka próba jest dwuelementową wariacją z powtórzeniami. Wariacji takich jest $N^n = 5^2 = 25$.

Przestrzeń prób \underline{A}° jest więc zbiorem 25-elementowym. Mamy mianowicie:

$$\begin{aligned} \underline{A}^\circ = \{ & (A_1, A_1), (A_1, A_2), (A_1, A_3), (A_1, A_4), (A_1, A_5), (A_2, A_1), (A_2, A_2), \\ & (A_2, A_3), (A_2, A_4), (A_2, A_5), (A_3, A_1), (A_3, A_2), (A_3, A_3), (A_3, A_4), \\ & (A_3, A_5), (A_4, A_1), (A_4, A_2), (A_4, A_3), (A_4, A_4), (A_4, A_5), (A_5, A_1), \\ & (A_5, A_2), (A_5, A_3), (A_5, A_4), (A_5, A_5) \}. \end{aligned}$$

Dokonując losowania próby \underline{A} o licznosci $n = 2$ realizujemy jedną spośród tych 25 możliwości. Na podstawie badania wylosowanej próby chcemy oszacować interesujący nas parametr obserwowanej zmiennej losowej w zbiorowości generalnej. W celu wyjaśnienia tego procesu założmy, że badanej zbiorowości A odpowiada hipotetyczny zbiór realizacji:

$$\begin{aligned} \tilde{X}_N = \{x_1 = h[B(A_1)] = 1, \quad x_2 = h[B(A_2)] = 2, \quad x_3 = h[B(A_3)] = 3, \\ x_4 = h[B(A_4)] = 4, \quad x_5 = h[B(A_5)] = 5\} = \{1, 2, 3, 4, 5\}. \end{aligned}$$

Zakładamy więc, że obserwowana zmienna losowa X ma w zbiorowości generalnej A rozkład prostokątny o wartości przeciętnej $\bar{x}_N = 3$. Rozkład ten pokazano na rys. 4.1A. Charakterystyką z próby, którą wykorzystamy do szacowania wartości przeciętnej \bar{X}_A , jest średnia arytmetyczna z próby \bar{X}_N . W rozważanym przypadku jest to zmienna losowa o następującym zbiorze wartości:

$$\{1; 1,5; 2; 2,5; 3; 3,5; 4; 4,5; 5\}.$$

Liczbowym obrazem przestrzeni próbek jest zbiór wszystkich możliwych realizacji (\bar{x}_n) zmiennej losowej \bar{X}_n . Wykorzystując wzór (3.8) zbiór ten można przedstawić następująco:

$$\{(1; 1), (1,5; 2), (2; 3), (2,5; 4), (3; 5), (3,5; 4), (4; 3), (4,5; 2), (5; 1)\}.$$

Opierając się na tym zbiorze skonstruowano funkcję prawdopodobieństwa zmiennej losowej \bar{X}_n . Wykres tej funkcji pokazano na rys. 4.1B. Z rysunku tego wynika bezpośrednio, że:

$$P(|\bar{X}_n - \bar{X}_N| \leq 0,5) = 13/25 = 0,52,$$

$$P(|\bar{X}_n - \bar{X}_N| \leq 1) = 19/25 = 0,76,$$

$$P(|\bar{X}_n - \bar{X}_N| \leq 1,5) = 23/25 = 0,92,$$

$$P(|\bar{X}_n - \bar{X}_N| \leq 2) = 25/25 = 1.$$

W przypadku losowania zależnego (bez zwracania) próbka \underline{A} jest – w rozważanym przypadku – dwuelementową wariacją bez powtórzeń. Wariacji takich mamy $N!/(N-n)! = 5!/3! = 20$. Przestrzeń próbek przedstawia się więc następująco:

$$\begin{aligned} \underline{A}^\circ = \{ & (A_1, A_2), (A_1, A_3), (A_1, A_4), (A_1, A_5), (A_2, A_1), (A_2, A_3), (A_2, A_4), \\ & (A_2, A_5), (A_3, A_1), (A_3, A_2), (A_3, A_4), (A_3, A_5), (A_4, A_1), (A_4, A_2), \\ & (A_4, A_3), (A_4, A_5), (A_5, A_1), (A_5, A_2), (A_5, A_3), (A_5, A_4) \}. \end{aligned}$$

Średnia arytmetyczna z próby \bar{X}_n jest w tym przypadku zmienną losową o następującym zbiorze wartości:

$$\{1,5; 2; 2,5; 3; 3,5; 4; 4,5\}.$$

Zbiór wszystkich możliwych realizacji tej zmiennej losowej przedstawimy w postaci (3.8). Mamy więc:

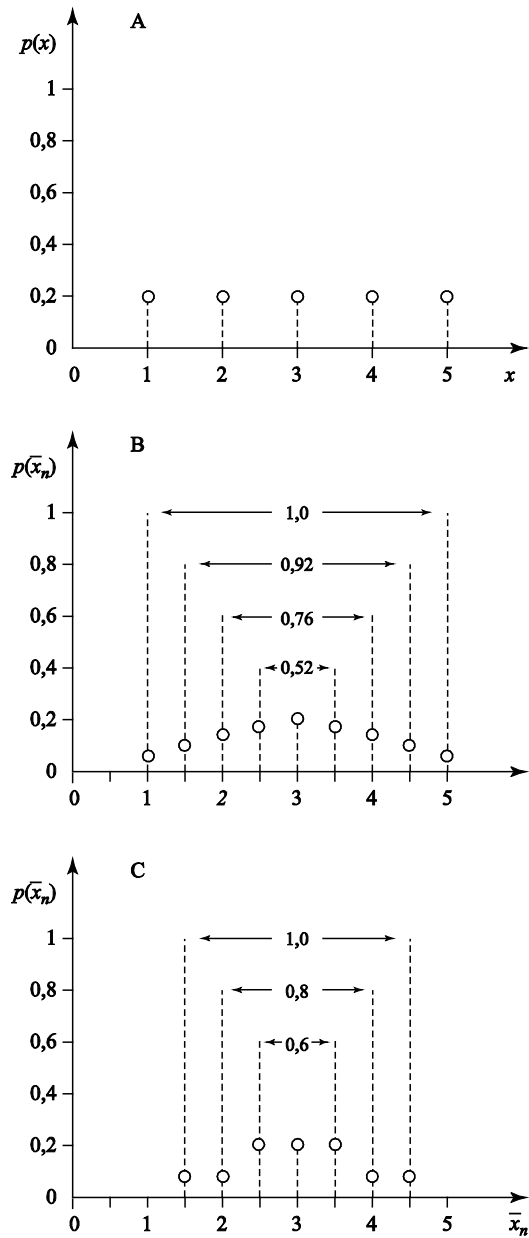
$$\{(1,5; 2), (2; 2), (2,5; 4), (3; 4), (3,5; 4), (4; 2), (4,5; 2)\}.$$

Funkcję prawdopodobieństwa skonstruowaną na podstawie tego zbioru pokazano na rys. 4.1C. Z rysunku tego wynika, że:

$$P(|\bar{X}_n - \bar{X}_N| \leq 0,5) = 12/20 = 0,60,$$

$$P(|\bar{X}_n - \bar{X}_N| \leq 1) = 16/20 = 0,80,$$

$$P(|\bar{X}_n - \bar{X}_N| \leq 1,5) = 20/20 = 1.$$



Rys. 4.1. Empiryczny rozkład średniej arytmetycznej z próby

Okazuje się więc, że w przypadku losowania bez zwracania prawdopodobieństwo dużego błędu szacowania jest mniejsze niż w przypadku losowania ze zwracaniem. ■

Przedstawiony przykład pozwala wyróżnić trzy ważne pojęcia teorii estymacji, a mianowicie:

- szacowany parametr,
- estymator,
- ocenę parametru (albo oszacowanie).

Szacowanym parametrem jest tu wartość przeciętna (\bar{X}_A) obserwowanej zmiennej losowej (X) w zbiorowości generalnej (A). Estymatorem jest średnia arytmetyczna (\bar{X}_n), rozumiana jako funkcja określona na zbiorze realizacji X_n . Oceną (oszacowaniem) jest natomiast wartość średniej arytmetycznej z próby (\bar{x}_n), obliczona na podstawie uzyskanego zbioru realizacji. Przytoczony przykład ilustruje tylko jeden typ estymacji, a mianowicie *punktową estymację parametryczną*. Jest to estymacja „parametryczna”, albowiem jej celem jest oszacowanie nieznannej wartości parametru zmiennej losowej w zbiorowości generalnej. Określenie „punktowa” oznacza natomiast, że efektem postępowania estymacyjnego jest jedna liczba, a więc jeden punkt na osi liczb rzeczywistych. W przypadku *przedziałowej estymacji parametrycznej* produktem szacowania jest pewien przedział na osi liczb rzeczywistych (tzw. przedział ufności), który z określonym prawdopodobieństwem pokrywa nieznaną wartość parametru. Wyróżnia się także *estymację nieparametryczną*, której celem jest ocena postaci funkcyjnej dystrybuanty obserwowanej zmiennej losowej w zbiorowości generalnej.

4.2. Estymatory i ich podstawowe własności

Szacowany parametr oznaczmy symbolem Q . Może to być – na przykład – parametr p w rozkładzie dwumianowym, wartość oczekiwana (μ) – albo wariancja (σ^2) – normalnej zmiennej losowej, parametr λ w rozkładzie wykładniczym itd. (zob. rozdział 1). Estymator parametru Q oznaczmy symbolem \hat{Q}_n . Symbolem tym będziemy również oznaczać ocenę (oszacowanie) parametru Q . Estymatorem \hat{Q}_n może być każda charakterystyka z próby, której rozkład zależy od parametru Q . Nie wszystkie statystyki spełniające ten warunek mają odpo-

wiednie własności operacyjne. Od dobrego estymatora wymaga się – przede wszystkim – by był on nieobciążony, zgodny i efektywny¹.

Estymator \hat{Q}_n jest nieobciążony, jeśli:

$$E(\hat{Q}_n) = Q. \quad (4.1)$$

Estymatorem nieobciążonym jest więc taka charakterystyka z próby (statystyka), której wartość oczekiwana jest równa szacowanemu parametrowi. Estymatory, które nie mają własności (4.1), są estymatorami obciążonymi. Estymatory takie generują obciążone oszacowania. Wielkość obciążenia wynika z różnicy:

$$\delta_n = E(\hat{Q}_n) - Q. \quad (4.2)$$

Jeśli estymator obciążony ma własność polegającą na tym, że:

$$\lim_{n \rightarrow \infty} \delta_n = 0, \quad (4.3)$$

to mamy do czynienia z estymatorem asymptotycznie nieobciążonym. Estymator \hat{Q}_n jest zgodny, jeśli:

$$\lim_{n \rightarrow \infty} P(|\hat{Q}_n - Q| < \varepsilon) = 1, \quad (4.4)$$

gdzie ε jest dowolnie małą, dodatnią liczbą rzeczywistą. Estymatorem zgodnym jest więc taka charakterystyka z próby, która podlega prawu wielkich liczb. Oznacza to, że zwiększając licznosc próbki (n) można zmniejszyć błąd szacowania (oceny).

Miarą efektywności estymatora jest jego wariancja $D^2(\hat{Q}_n)$. Estymatorem najefektywniejszym jest estymator o najmniejszej wariancji. Do oceny efektywności estymatora wykorzystuje się następującą charakterystykę:

$$e(\hat{Q}_n) = \frac{D^2(\hat{Q}_n^*)}{D^2(\hat{Q}_n)}, \quad (4.5)$$

gdzie \hat{Q}_n jest estymatorem ocenianym, natomiast \hat{Q}_n^* oznacza estymator najefektywniejszy.

¹ Rozważa się również inne własności, wśród których szczególnie istotną rolę odgrywa dostateczność estymatora. Zob. J. Greń, *Statystyka matematyczna – podręcznik programowy*, PWN, Warszawa 1987.

Estymator \hat{Q}_n posiadający własność polegającą na tym, że:

$$\lim_{n \rightarrow \infty} e(\hat{Q}_n) = 1 \quad (4.6)$$

jest estymatorem asymptotycznie najefektywniejszym. Jeśli estymator jest nie-obciążony, to wysoka efektywność zapewnia małe prawdopodobieństwo dużego błędu oszacowania (oceny).

4.3. Metoda największej wiarygodności

Wśród metod uzyskiwania estymatorów podstawowe znaczenie mają:

- metoda momentów,
- metoda najmniejszych kwadratów,
- metoda największej wiarygodności.

Metodę momentów zastosowaliśmy intuicyjnie w przykładzie 4.1, a w sposób bardziej usystematyzowany wykorzystano tę metodę w punkcie 4.7 przy konstruowaniu estymatora współczynnika korelacji liniowej. Postępowanie charakterystyczne dla metody najmniejszych kwadratów zastosowaliśmy w rozdziale 3 przy wyprowadzaniu wzoru na średnią arytmetyczną. Problemów tych nie będziemy szerzej omawiać. Przedstawimy natomiast dokładniej metodę największej wiarygodności, która ma największe znaczenie praktyczne.

Funkcja wiarygodności ma postać:

$$L(\mathbf{X}_n; Q) = \prod_{i=1}^n f(x_i; Q), \quad (4.7)$$

gdzie $f(x_i; Q)$ oznacza prawdopodobieństwo uzyskania wartości $x_i \in \mathbf{X}_n$ przy pewnej nieznannej wartości parametru Q albo odpowiednią gęstość prawdopodobieństwa (w przypadku ciągłych zmiennych losowych).

Funkcja (4.7) wyraża więc prawdopodobieństwo uzyskania zbioru realizacji \mathbf{X}_n , przy pewnej wartości parametru Q . Zauważmy, że u podstaw funkcji wiarygodności postaci (4.7) leży założenie, że elementy zbiorowości generalnej wybierane są do próby na drodze losowania niezależnego. Tylko wówczas bowiem prawdopodobieństwo uzyskania zbioru realizacji $\mathbf{X}_n = \{x_1, x_2, \dots, x_i, \dots, x_n\}$ jest równe iloczynowi prawdopodobieństw uzyskania poszczególnych realizacji $x_i \in \mathbf{X}_n$. Idea metody największej wiarygodności polega na poszukiwaniu takiej wartości Q , przy której funkcja $L(\mathbf{X}_n; Q)$ osiąga maksimum – stąd nazwa metody.

Rezultatem postępowania jest funkcja:

$$\hat{Q}_n = h(\mathbf{X}_n), \quad (4.8)$$

będąca poszukiwanym estymatorem parametru Q .

Estymacja parametru p w rozkładzie dwupunktowym zero-jedynkowej zmiennej losowej

Niech X oznacza zero-jedynkową zmienną losową. Jedynym parametrem rozkładu takiej zmiennej jest $p = P(X = 1)$ (zob. rozdział 1). Elementami zbioru realizacji \mathbf{X}_n są – *ex definitione* – zera i jedynki. Funkcję wiarygodności zapiszemy następująco:

$$L = \prod_{i=1}^n p^{x_i} (1-p)^{1-x_i}. \quad (4.9)$$

Jeśli $x_i = 0$, to i -ty czynnik iloczynu równy jest $1 - p$. Jeśli natomiast $x_i = 1$, to czynnik ten przyjmuje wartość p . Po wykonaniu mnożenia mamy:

$$L = p^{\sum_{i=1}^n x_i} (1-p)^{n - \sum_{i=1}^n x_i}.$$

Logarytmując to wyrażenie otrzymujemy:

$$\ln L = \sum_{i=1}^n x_i \ln p + (n - \sum_{i=1}^n x_i) \ln(1-p).$$

Wyznaczamy pierwszą pochodną funkcji $\ln L$ względem p :

$$\frac{d \ln L}{dp} = \frac{1}{p} \sum_{i=1}^n x_i - \frac{1}{1-p} \left(n - \sum_{i=1}^n x_i \right).$$

Po przyrównaniu tej pochodnej do zera i rozwiązaniu otrzymanego równania względem p mamy:

$$p = \frac{1}{n} \sum_{i=1}^n x_i.$$

Poszukiwana funkcja (4.8) ma więc postać:

$$\hat{p}_n = \frac{1}{n} \sum_{i=1}^n x_i. \quad (4.10)$$

Okazuje się więc, że estymatorem parametru p jest średnia arytmetyczna z próby. Jest to również estymator parametru p w rozkładzie dwumianowym, albowiem zmienna losowa $Y \sim B(n, p)$ jest sumą n niezależnych zmiennych $X \sim D(p)$. Można wykazać, że \hat{p}_n jest estymatorem nieobciążonym, zgodnym i najefektywniejszym.

Estymacja parametru λ w rozkładzie Poissona

Niech X oznacza zmienną losową o rozkładzie Poissona, o nieznanym parametrze λ . Funkcja wiarygodności ma postać:

$$L = \prod_{i=1}^n \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}. \quad (4.11)$$

Postępując analogicznie jak poprzednio mamy:

$$L = \frac{\lambda^{\sum_{i=1}^n x_i} e^{-n\lambda}}{\prod_{i=1}^n x_i!},$$

$$\ln L = \sum_{i=1}^n x_i \ln \lambda - n\lambda - \sum_{i=1}^n \ln x_i!,$$

$$\frac{d \ln L}{d\lambda} = -n + \frac{1}{\lambda} \sum_{i=1}^n x_i,$$

$$\lambda = \frac{1}{n} \sum_{i=1}^n x_i.$$

Poszukiwany estymator parametru λ ma więc postać:

$$\hat{\lambda}_n = \frac{1}{n} \sum_{i=1}^n x_i. \quad (4.12)$$

Jest to estymator nieobciążony, zgodny i najefektywniejszy.

Estymacja parametru μ w rozkładzie normalnym

Niech X oznacza normalną zmienną losową o nieznanym parametrze μ . Funkcja wiarygodności ma postać:

$$\begin{aligned} L &= \prod_{i=1}^n \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x_i - \mu}{\sigma}\right)^2\right] = \\ &= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2\right]. \end{aligned} \quad (4.13)$$

Po zlogarytmowaniu mamy:

$$\ln l = n \ln \frac{1}{\sigma} + n \ln \frac{1}{\sqrt{2\pi}} - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2. \quad (4.14)$$

Wyznaczamy pochodną tej funkcji względem μ i przyrównujemy ją do zera:

$$\frac{d \ln L}{d\mu} = \frac{1}{\sigma^2} \sum_{i=1}^n x_i - \frac{1}{\sigma^2} n\mu = \frac{1}{\sigma^2} \left(\sum_{i=1}^n x_i - n\mu \right),$$

$$\frac{1}{\sigma^2} \left(\sum_{i=1}^n x_i - n\mu \right) = 0,$$

$$\sum_{i=1}^n x_i - n\mu = 0,$$

$$\mu = \frac{1}{n} \sum_{i=1}^n x_i.$$

Poszukiwany estymator parametru μ ma więc postać:

$$\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^n x_i. \quad (4.15)$$

Jest to estymator nieobciążony, zgodny i najefektywniejszy.

Estymacja parametru σ^2 w rozkładzie normalnym

Funkcja wiarygodności ma postać (4.13). Wyznaczamy pochodną funkcji (4.14) względem σ^2 i porównujemy ją do zera:

$$\frac{d \ln L}{d\sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (x_i - \mu)^2,$$

$$-\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (x_i - \mu)^2 = 0,$$

$$\sigma^2 n - \sum_{i=1}^n (x_i - \mu)^2 = 0,$$

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2.$$

W konsekwencji mamy:

$$\hat{\sigma}_n^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2. \quad (4.16)$$

Estymacja parametru λ w rozkładzie wykładniczym

Niech X oznacza zmienną losową o wykładniczym rozkładzie prawdopodobieństwa z nieznanym parametrem λ . Funkcja wiarygodności ma postać:

$$L = \prod_{i=1}^n \lambda e^{-\lambda x_i}. \quad (4.17)$$

Mamy więc:

$$L = \lambda^n e^{-\lambda \sum_{i=1}^n x_i},$$

$$\ln L = n \ln \lambda - \lambda \sum_{i=1}^n x_i.$$

Wyznaczamy pochodną tej funkcji względem λ , a następnie przyrównujemy ją do zera:

$$\frac{d \ln L}{d \lambda} = \frac{n}{\lambda} - \sum_{i=1}^n x_i,$$

$$\frac{n}{\lambda} - \sum_{i=1}^n x_i = 0.$$

Rozwiązujemy otrzymane równanie względem λ :

$$\lambda = \frac{n}{\sum_{i=1}^n x_i}.$$

Poszukiwany estymator parametru λ ma więc postać:

$$\hat{\lambda}_n = \frac{n}{\sum_{i=1}^n x_i}. \quad (4.18)$$

Jak łatwo zauważyć, otrzymany estymator jest odwrotnością średniej arytmetycznej z próby.

4.4. Punktowa estymacja parametrów jednowymiarowej zmiennej losowej

Podstawowymi charakterystykami jednowymiarowej zmiennej losowej są:

- wartość oczekiwana, która charakteryzuje przeciętny poziom badanej cechy w zbiorowości generalnej,
- wariancja i odchylenie standardowe, które charakteryzują zróżnicowanie zbiorowości generalnej ze względu na badaną cechę.

Estymacja tych parametrów jest zwykle jednym z podstawowych celów każdego doświadczenia.

4.4.1. Szacowanie wartości oczekiwanej

4.4.1.1. Średnia arytmetyczna z próby

Średnia arytmetyczna z próby (\bar{X}_n) jest podstawowym estymatorem wartości oczekiwanej $E(X)$, niezależnie od tego, jaki rozkład ma obserwowana zmienna losowa X . Jeśli zbiór realizacji X_n ma postać szczegółowego szeregu statystycznego, to wartość średniej arytmetycznej z próby wyznacza się według wzoru:

$$\bar{x}_n = \frac{1}{n} \sum_{i=1}^n x_i, \quad (4.19)$$

w którym x_i oznacza i -tą realizację w zbiorze X_n , natomiast n jest licznością próby.

W przypadku gdy podstawę obliczeń stanowi zbiór realizacji X_n transformowany do postaci szeregu rozdzielczego, wartości \bar{x}_n wyznacza się według wzorów analogicznych do (3.31) i (3.33). Mamy mianowicie:

$$\bar{x}_n = \frac{\sum_{j=1}^k f_j \cdot x_j^\circ}{\sum_{j=1}^k f_j} = \frac{\sum_{j=1}^k f_j \cdot x_j^\circ}{n}, \quad (4.20)$$

$$\bar{x}_n = \sum_{j=1}^k v_j \cdot x_j^\circ, \quad (4.21)$$

przy czym znaczenie symboli w tych wzorach jest takie samo jak w rozdziale 3.

Przypomnijmy, że w przypadku dyskretnych zmiennych losowych x_j° jest elementem zbioru X° , natomiast w przypadku ciągłych zmiennych losowych jest to środek przedziału klasowego $(x_{d,j}; x_{g,j}]$. Symbol f_j oznacza liczbę realizacji $x_i \in X_n$ o wartościach równych x_j° albo liczbę realizacji $x_i \in X_n$, spełniających podwójną nierówność $x_{d,j} < x_i \leq x_{g,j}$, natomiast $v_j = f_j/n$ jest odpowiednią częstością względną.

Formalne własności średniej arytmetycznej z próby są takie same jak w badaniach wyczerpujących; zob. wzory (3.36)–(3.39) oraz (3.36a)–(3.39a).

Średnia arytmetyczna z próby jest estymatorem nieobciążonym, zgodnym i najefektywniejszym. Zauważmy, że we wszystkich przypadkach rozważanych w punkcie 4.3 uzyskaliśmy średnią arytmetyczną z próby, jako estymator odpowiedniej wartości oczekiwanej. Przypomnijmy (zob. rozdział 1), że wartościami oczekiwanymi zmiennych losowych o rozkładzie zero-jedynkowym, Poissona i normalnym są odpowiednio parametry p , λ i μ . W przypadku zmiennej losowej o rozkładzie wykładniczym zachodzi zależność $E(X) = 1/\lambda$. Ponieważ ze wzoru (4.18) wynika, że $\bar{x}_n = 1/\hat{\lambda}_n$, zatem również w tym przypadku uzyskaliśmy średnią arytmetyczną z próby jako estymator wartości oczekiwanej.

4.4.1.2. Mediana z próby

Do estymacji wartości oczekiwanej wykorzystywane są również statystyki pozycyjne z próby. Wśród estymatorów tej klasy podstawowe znaczenie ma mediana z próby $X_{me,n}$. Wartości tej charakterystyki wyznacza się analogicznie jak w przypadku badań wyczerpujących (zob. rozdział 3). Mamy więc:

$$x_{me,n} = x_{\left(\frac{n+1}{2}\right)}. \quad (4.22)$$

$$x_{me,n} = \frac{1}{2} [x_{(n/2)} + x_{(n/2+1)}]. \quad (4.23)$$

$$x_{me,n} = x_{d,r} + \frac{\left(\frac{n}{2} - \sum_{j=1}^{r-1} f_j\right)l}{f_r}. \quad (4.24)$$

Wzory (4.22) i (4.23) znajdują zastosowanie wówczas, gdy zbiór realizacji X_n dany jest w postaci szczegółowego szeregu statystycznego, natomiast wzór (4.24) stosuje się w przypadku szeregu rozdzielczego. Technikę wyznaczania

mediany opisano w rozdziale 3 (zob. punkt 3.2.2.5). Mediana z próby jest znacznie gorszym estymatorem wartości oczekiwanej niż średnia arytmetyczna z próby. Zastosowanie mediany z próby (jako estymatora wartości oczekiwanej) ograniczone jest do ciągłych zmiennych losowych o rozkładach wykazujących co najwyżej umiarkowaną asymetrię. W przypadku takich rozkładów $X_{me,n}$ jest estymatorem asymptotycznie nieobciążonym. Efektywność tego estymatora wynosi około 64% efektywności średniej arytmetycznej z próby², a ponadto $X_{me,n}$ nie jest estymatorem asymptotycznie najefektywniejszym. Mediana z próby nie jest też estymatorem zgodnym.

4.4.2. Szacowanie wariancji i odchylenia standardowego

4.4.2.1. Wariancja z próby

Wyróżnia się trzy charakterystyki określane jako wariancja z próby. Statystyki te oznaczymy symbolami \underline{s}^2 , \underline{S}^2 i S^2 . Jeśli zbiór realizacji \mathbf{X}_n ma postać szczegółowego szeregu statystycznego, to realizacje zmiennej losowej \underline{S}^2 wynikają ze wzoru:

$$\underline{s}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2. \quad (4.25)$$

W przypadku zmiennej losowej \underline{S}^2 mamy:

$$\underline{s}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}_n)^2 = \quad (4.26a)$$

$$= \frac{1}{n} \left[\sum_{i=1}^n x_i^2 - \frac{1}{n} \sum_{i=1}^n x_i \right]^2 = \quad (4.26b)$$

$$= \frac{1}{n} \sum_{i=1}^n x_i^2 - \left(\frac{1}{n} \sum_{i=1}^n x_i \right)^2. \quad (4.26c)$$

² Dokładna wartość $e(x_{me,n}) = 2/\pi$, zob. J. Greń, *op. cit.*, s. 213.

Realizacje zmiennej losowej S^2 obliczamy według wzorów:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x}_n)^2 = \quad (4.27a)$$

$$= \frac{1}{n-1} \left[\sum_{i=1}^n x_i^2 - \frac{1}{n} \left(\sum_{i=1}^n x_i \right)^2 \right] = \quad (4.27b)$$

$$= \frac{1}{n-1} \sum_{i=1}^n x_i^2 - \frac{1}{n(n-1)} \left(\sum_{i=1}^n x_i \right)^2. \quad (4.27c)$$

Każdy z tych wzorów można oczywiście transformować do postaci znajdującej zastosowanie w przypadku szeregu rozdzielczego. Rezultatem tych przekształceń są wzory analogiczne do tych, które przedstawiono w punkcie 3.2.3. Mamy mianowicie:

$$\underline{s}^2 = \frac{\sum_{j=1}^k f_j (x_j^\circ - \mu)^2}{\sum_{j=1}^k f_j} = \frac{1}{n} \sum_{j=1}^k f_j (x_j^\circ - \mu)^2. \quad (4.28)$$

$$\underline{s}^2 = \sum_{j=1}^k v_j (x_j^\circ - \mu)^2. \quad (4.29)$$

$$\underline{s}^2 = \frac{\sum_{j=1}^k f_j (x_j^\circ - \bar{x}_n)^2}{\sum_{j=1}^k f_j} = \frac{1}{n} \sum_{j=1}^k f_j (x_j^\circ - \bar{x}_n)^2 = \quad (4.30a)$$

$$\begin{aligned}
& \frac{\sum_{j=1}^k f_j (x_j^\circ)^2 - \frac{\left(\sum_{j=1}^k f_j x_j^\circ\right)^2}{\sum_{j=1}^k f_j}}{\sum_{j=1}^k f_j} = \\
& = \frac{1}{n} \left[\sum_{j=1}^k f_j (x_j^\circ)^2 - \frac{1}{n} \left(\sum_{j=1}^k f_j x_j^\circ \right)^2 \right] = \tag{4.30b}
\end{aligned}$$

$$= \frac{1}{n} \sum_{j=1}^k f_j (x_j^\circ)^2 - \left(\frac{1}{n} \sum_{j=1}^k f_j x_j^\circ \right)^2. \tag{4.30c}$$

$$\underline{s}^2 = \sum_{j=1}^k v_j (x_j^\circ - \bar{x}_n)^2 = \tag{4.31a}$$

$$= \sum_{j=1}^k v_j (x_j^\circ)^2 - \left(\sum_{j=1}^k v_j x_j^\circ \right)^2. \tag{4.31b}$$

$$s^2 = \frac{\sum_{j=1}^k f_j (x_j^\circ - \bar{x}_n)^2}{\sum_{j=1}^k f_j - 1} = \frac{1}{n-1} \sum_{j=1}^k f_j (x_j^\circ - \bar{x}_n)^2 = \tag{4.32a}$$

$$\begin{aligned}
& \frac{\sum_{j=1}^k f_j (x_j^\circ)^2 - \frac{\left(\sum_{j=1}^k f_j x_j^\circ\right)^2}{\sum_{j=1}^k f_j}}{\sum_{j=1}^k f_j - 1} =
\end{aligned}$$

$$= \frac{1}{n-1} \left[\sum_{j=1}^k f_j (x_j^\circ)^2 - \frac{1}{n} \left(\sum_{j=1}^k f_j x_j^\circ \right)^2 \right] = \quad (4.32b)$$

$$= \frac{1}{n-1} \sum_{j=1}^k f_j (x_j^\circ)^2 - \frac{1}{n(n-1)} \left(\sum_{j=1}^k f_j x_j^\circ \right)^2. \quad (4.32c)$$

Zauważmy, że charakterystykę (4.25) otrzymaliśmy w punkcie 4.3, wyznaczając estymator parametru σ^2 zmiennej losowej $X \sim N(\mu, \sigma)$; zob. wzór (4.16). Statystyka $\underline{\underline{S}}^2$ jest nieobciążonym, zgodnym i najefektywniejszym estymatorem wariancji. Zastosowanie tego estymatora jest jednak ograniczone do tych – raczej nieczęstych – przypadków, gdy znana jest wartość oczekiwana obserwowanej zmiennej losowej. Zastępując wartość oczekiwaną jej oszacowaniem (\bar{x}_n), użyjemy estymatory \underline{S}^2 i S^2 . Statystyka \underline{S}^2 jest obciążonym estymatorem wariancji w zbiorowości generalnej. Wielkość tego obciążenia wynika ze wzoru:

$$\delta_n = E(\underline{S}^2) - \sigma^2 = \frac{n-1}{n} \sigma^2 - \sigma^2 = -\frac{\sigma^2}{n} < 0. \quad (4.33)$$

Ponieważ

$$\lim_{n \rightarrow \infty} -\frac{\sigma^2}{n} = 0, \quad (4.34)$$

zatem \underline{S}^2 jest estymatorem asymptotycznie nieobciążonym. Oznacza to, że systematyczny błąd oceny (oszacowania) parametru maleje ze wzrostem liczności próbki (n). W przypadku małych próbek nieobciążonym, zgodnym i najefektywniejszym estymatorem wariancji jest statystyka S^2 , o wartościach wyznaczonych ze wzorów (4.27) i (4.32). We wzorach tych sumę kwadratów odchyłeń od średniej arytmetycznej dzielimy przez *liczbę stopni swobody*, czyli przez liczbę niezależnych obserwacji; a ponieważ jeden stopień swobody tracimy wyznaczając średnią arytmetyczną ze zbioru \mathbf{X}_n , więc przy obliczaniu wartości s^2 mamy $n - 1$ niezależnych obserwacji. Zauważmy, że przy obliczaniu wartości $\underline{\underline{s}}^2$ liczba stopni swobody jest równa liczbie obserwacji (n), albowiem zastosowanie wzoru (4.25) nie wymaga wcześniejszego obliczania wartości średniej arytmetycznej \bar{x}_n .

4.4.2.2. Odchylenie standardowe z próby

Odchylenie standardowe z próby jest dodatnim pierwiastkiem kwadratowym z wariancji. Istnieją więc trzy charakterystyki określane wspólnym mianem odchylenia standardowego z próby, a mianowicie: $\underline{\underline{S}} = \sqrt{\underline{\underline{S}}^2}$, $\underline{S} = \sqrt{\underline{S}^2}$ i $S = \sqrt{S^2}$. Wartości tych charakterystyk wyznacza się biorąc dodatnie pierwiastki kwadratowe z wartości $\underline{\underline{S}}^2$, \underline{S}^2 , S^2 , obliczonych według wzorów (4.25)–(4.32). Mimo że statystyki $\underline{\underline{S}}^2$ i S^2 są nieobciążonymi estymatorami wariancji, to jednak zarówno $\underline{\underline{S}}$ jak i S są tylko asymptotycznie nieobciążonymi estymatorami odchylenia standardowego w zbiorowości generalnej. W przypadku zmiennej losowej $X \sim N(\mu, \sigma)$ obciążenie to można usunąć. Nieobciążonymi estymatorami odchylenia standardowego (σ) są statystyki $\underline{\underline{S}}^*$ i S^* , których wartości wyznacza się ze wzorów:

$$\underline{\underline{S}}^* = \underline{\underline{S}} \cdot c_r, \quad (4.35)$$

$$S^* = S \cdot c_r, \quad (4.36)$$

w których

$$c_r = \frac{\Gamma\left(\frac{r}{2}\right)\sqrt{r}}{\Gamma\left(\frac{r+1}{2}\right)\sqrt{2}}, \quad (4.37)$$

gdzie r oznacza liczbę stopni. We wzorze (4.35) mamy $r = n$. Jeśli natomiast stosujemy wzór (4.36), to $r = n - 1$. Wartości funkcji gamma wyznacza się następująco:

$$\Gamma\left(\frac{r}{2}\right) = \begin{cases} \left(\frac{r}{2}-1\right)! & \text{gdy } r \text{ jest liczbą parzystą, } n \geq 2 \\ \left(\frac{r-2}{2}\right)\left(\frac{r-4}{2}\right)\dots\left(\frac{1}{2}\right)\sqrt{\pi} & \text{gdy } r \text{ jest liczbą nieparzystą, } n \geq 3 \end{cases}$$

$$\Gamma\left(\frac{r+1}{2}\right) = \begin{cases} \left(\frac{r-1}{2}\right)\left(\frac{r-3}{2}\right)\dots\left(\frac{1}{2}\right)\sqrt{\pi} & \text{gdy } r \text{ jest liczbą parzystą, } n \geq 2 \\ \left(\frac{r+1}{2}-1\right)! & \text{gdy } r \text{ jest liczbą nieparzystą, } n \geq 1 \end{cases}$$

Przykład 4.2

Załóżmy, że $r = 5$. Chcemy obliczyć wartość $c_r = c_5$. Po podstawieniu $r = 5$ do wzoru (4.37) mamy:

$$c_5 = \frac{\Gamma\left(\frac{5}{2}\right)\sqrt{5}}{\Gamma\left(\frac{5+1}{2}\right)\sqrt{2}} = \frac{\frac{3}{2} \cdot \frac{1}{2} \sqrt{\pi} \sqrt{5}}{\left(\frac{6}{2}-1\right)!\sqrt{2}} = \frac{0,75\sqrt{\pi}\sqrt{5}}{2\sqrt{2}} \approx \frac{2,9725}{2,8284} \approx 1,0509.$$

Podstawiając tę wartość do wzoru (4.35) albo (4.36), uzyskujemy nieobciążoną ocenę odchylenia standardowego w zbiorowości generalnej. ■

4.4.2.3. Szacowanie odchylenia standardowego normalnej zmiennej losowej na podstawie rozstępu z próby

Rozstęp z próby (R_n) jest zmienną losową, której realizacje (r_n) wyznacza się analogicznie jak rozstęp w badaniach wyczerpujących (zob. punkt 3.2.3.3). Mamy więc:

$$r_n = x_{\max} - x_{\min} = x_{(n)} - x_{(1)}, \quad (4.38)$$

gdzie $x_{\min} = x_{(1)}$ oznacza pierwszą, natomiast $x_{\max} = x_{(n)}$ ostatnią realizację w zbiorze \mathbf{X}_n , uporządkowanym za pomocą relacji „ \leq ”. Jeśli obserwowana zmienna losowa ma rozkład normalny z odchyleniem standardowym σ , to statystyka R_n/σ ma następujące parametry:

$$E\left(\frac{R_n}{\sigma}\right) = d_n, \quad (4.39)$$

$$D\left(\frac{R_n}{\sigma}\right) = k_n, \quad (4.40)$$

W zamieszczonej w aneksie tablicy V zestawiono wartości d_n , k_n i $1/d_n$ dla $n = 2, (1), 10$.

Zależność (4.39) można wykorzystać do skonstruowania estymatora odchylenia standardowego σ . Przekształcając to wyrażenie otrzymujemy:

$$d_n = \frac{E(R_n)}{\sigma},$$

$$\sigma = \frac{E(R_n)}{d_n}.$$

Poszukiwany estymator parametru σ ma więc postać:

$$\hat{\sigma}_n = \frac{\bar{r}_n}{d_n}, \quad (4.41)$$

gdzie \bar{r}_n jest średnią arytmetyczną rozstępu w próbce o liczności n . Na podkreślenie zasługuje fakt, że względna efektywność estymatora (4.41) maleje ze wzrostem liczebności próby. Dlatego też, jeśli $n \leq 10$, to rozstęp (r_n) uzyskany w rezultacie badania jednej próbki można przyjąć jako \bar{r}_n we wzorze (4.41).

W przypadku próbek o większych liczebnościach postępuje się następująco:

- zbiór realizacji o liczności $m > 10$ rozcina się na k podzbiorów o jednakowej liczności $n \leq 10$; tak więc $m = kn$;
- w każdym podzbiornie wyznacza się rozstęp według wzoru (4.38), a następnie oblicza się średni rozstęp według wzoru:

$$\bar{r}_n = \frac{1}{k} \sum_{h=1}^k r_{n,h}, \quad (4.42)$$

w którym $r_{n,h}$ oznacza rozstęp w h -tym podzbiornie o liczności n ;

- uzyskaną wartość \bar{r}_n podstawia się do wzoru (4.41).

Przykład 4.3

Wyniki badania doraźnej wytrzymałości na rozciąganie zestawione w tablicy 3.3 potraktujemy obecnie jako próbkę (o liczności $n = 36$) z nieskończonej liczebnej zbiorowości generalnej. Kształt histogramu pokazanego na rys. 3.2 pozwala przyjąć założenie, że mamy do czynienia ze zmienną losową $X \sim N(\mu, \sigma)$, przy

czym μ i σ nie są znane. Oszacujemy odchylenie standardowe (σ) tej zmiennej losowej, stosując kolejno odchylenie standardowe z próby (s) oraz estymator oparty na rozstępie (4.41). Wykorzystując obliczenia przedstawione w tablicy 3.12 wyznaczamy wartość statystyki S^2 według wzoru (4.32a):

$$s^2 = \frac{\sum_{j=1}^k f_j (x_j^\circ - \bar{x}_n)^2}{\sum_{j=1}^k f_j - 1} \approx \frac{1131,75}{35} \approx 32,3357.$$

W konsekwencji mamy:

$$s \approx \sqrt{32,3357} \approx 5,69.$$

Obecnie dokonamy oszacowania wartości σ za pomocą estymatora wyrażonego wzorem (4.41). Liczność próby wynosi $m = 36$. Przyjmujemy $k = 6$. W konsekwencji $n = 6$. Rozcinamy zbiór realizacji przedstawiony w tablicy 3.3 na 6 podzbiorów i w każdym z nich wyznaczamy rozstęp. Mamy więc:

$$r_{6,1} = 87,28 - 61,64 = 25,64$$

$$r_{6,2} = 86,43 - 72,34 = 14,09$$

$$r_{6,3} = 77,43 - 68,12 = 9,31$$

$$r_{6,4} = 88,21 - 74,20 = 14,01$$

$$r_{6,5} = 84,35 - 67,88 = 16,47$$

$$r_{6,6} = 87,50 - 75,57 = 11,93$$

$$\sum_{h=1}^6 r_{6,h} = 91,45,$$

$$\bar{r}_6 = \frac{91,45}{6} \approx 15,24,$$

$$\hat{\sigma}_n = \bar{r}_6 \cdot \frac{1}{d_n} = 15,24 \cdot 0,395 \approx 6,02$$

Wartość $d_6 = 0,395$ odczytano z tablicy V zamieszczonej w aneksie. Niewielka rozbieżność między uzyskanymi ocenami parametru σ nie ma praktycznego znaczenia. ■

4.5. Rozkłady podstawowych charakterystyk z próby

4.5.1. Rozkład średniej arytmetycznej z próby

Jeśli obserwowana zmienna losowa X ma rozkład normalny o parametrach $E(X) = \mu$ i $D(X) = \sigma$, to średnia arytmetyczna z próby \bar{X}_n – o realizacjach uzyskiwanych według wzorów (4.19), (4.20), (4.21) – ma również rozkład normalny o parametrach:

$$E(\bar{X}_n) = \mu(\bar{x}) = \mu, \quad (4.43)$$

$$D(\bar{X}_n) = \sigma(\bar{x}) = \frac{\sigma}{\sqrt{n}}. \quad (4.44)$$

W konsekwencji wartości u obliczane według wzoru:

$$u = \frac{\bar{x}_n - \mu(\bar{x})}{s(\bar{x})} = \frac{x_n - \mu}{\sigma} \sqrt{n} \quad (4.45)$$

są realizacjami zmiennej losowej $U \sim N(0; 1)$.

Relacje między rozkładem zmiennej losowej X i rozkładem średniej arytmetycznej \bar{X}_n ilustruje rys. 4.2. Przy sporządzaniu tego rysunku przyjęto, że $n = 4$.

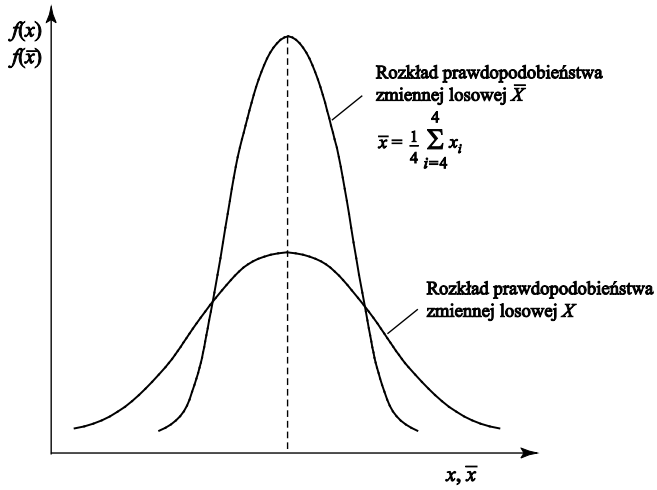
W konsekwencji odchylenie standardowe średniej arytmetycznej jest $\sqrt{4} = 2$ razy mniejsze niż odchylenie standardowe zmiennej losowej X .

Parametr $\sigma(\bar{x})$ nazywany jest błędem średnim średniej arytmetycznej. W przypadku gdy wartość σ nie jest znana, parametr $\sigma(\bar{x})$ należy oszacować z próby. Jeśli wykorzystamy do tego celu statystykę \underline{S} , to oszacowanie parametru $\sigma(\bar{x})$ uzyskuje się ze wzoru:

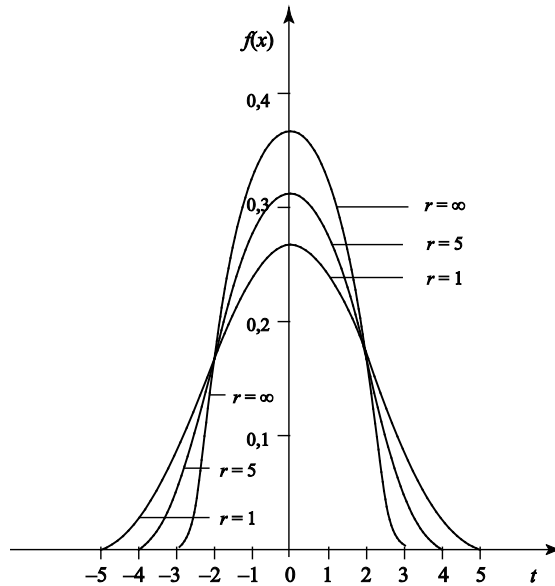
$$s(\bar{x}) = \frac{s}{\sqrt{n-1}} = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x}_n)^2}{n(n-1)}}. \quad (4.46)$$

Jeśli natomiast zastosujemy statystykę S , to:

$$s(\bar{x}) = \frac{s}{\sqrt{n}} = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x}_n)^2}{n(n-1)}}. \quad (4.47)$$



Rys. 4.2. Porównanie rozkładu obserwowanej zmiennej losowej (X) z rozkładem średniej arytmetycznej z próby (\bar{X})



Rys. 4.3. Rozkład Studenta

Mając $s(\bar{x})$ możemy dokonać transformacji (tzw. studentyzacji) charakterystyki (4.45). Mamy mianowicie:

$$t = \frac{\bar{x}_n - \mu}{s(\bar{x})} = \frac{\bar{x}_n - \mu}{\underline{s}} \sqrt{n-1} = \frac{\bar{x}_n - \mu}{s} \sqrt{n}. \quad (4.48)$$

Okazało się, że tak stransformowana charakterystyka (4.45) nie ma rozkładu $N(0; 1)$. Zmienna losowa t o realizacjach wynikających ze wzoru (4.48) ma rozkład Studenta³ o $r = n - 1$ stopniach swobody, przy czym:

$$E(t) = 0, \quad (4.49)$$

$$D(t) = \frac{n-1}{n-3}; \quad n > 3. \quad (4.50)$$

Szczegółowa postać funkcji prawdopodobieństwa zmiennej losowej t uzależniona jest od liczby stopni swobody (r) przy szacowaniu wariancji i wyraża się wzorem:

$$f_r(t) = \frac{1}{\sqrt{\pi r}} \frac{\Gamma\left(\frac{r+1}{2}\right)}{\Gamma\left(\frac{r}{2}\right)} \left(1 + \frac{t^2}{r}\right)^{-\frac{r+1}{2}}, \quad (4.51)$$

gdzie $r = n - 1$.

Własności tej funkcji ilustruje rys. 4.3. Wykres funkcji $f_r(t)$ jest symetryczny względem $t = 0$ i tym bardziej „spłaszczony”, im mniejsza jest liczba stopni swobody $r = n - 1$. Oznacza to, że im mniej niezależnych obserwacji stanowi podstawę do obliczenia \bar{x}_n , tym bardziej średnia ta może odbiegać od wartości oczekiwanej $E(\bar{X}_n) = \mu(\bar{x}_n) = \mu$. Ze wzrostem liczby stopni swobody rozkład Studenta zmierza do rozkładu $N(0; 1)$. Zbieżność ta jest bardzo szybka i już przy $n \approx 30$ różnice między dystrybuantą zmiennej losowej Studenta i dystrybuantą zmiennej losowej $U \sim N(0; 1)$ są tak małe, iż nie mają praktycznego znaczenia.

³ Jest to pseudonim naukowy angielskiego statystyka Williama Gosseta (1876–1937). Zarówno zmienna losowa Studenta, jak i jej realizacja oznaczane są symbolem t . Nie jest tu przestrzegana reguła, według której zmienne losowe należy oznaczać dużymi literami, a ich realizacje odpowiednimi małymi literami.

Dlatego też $n = 30$ stanowi umowną granicę między małą i dużą próbą. Dla $n \geq 30$ rozkład średniej arytmetycznej z próby (\bar{X}_n) może być aproksymowany przez rozkład normalny, niezależnie od tego, czy wartość σ jest znana, czy też musi ona być szacowana z próby. Rozkład normalny jest też rozkładem granicznym dla rozkładów średnich arytmetycznych w tych przypadkach, gdy rozkład obserwowanej zmiennej losowej X nie jest rozkładem normalnym. Tę własność średniej arytmetycznej z próby ilustruje przykład 4.1.

Rozkład zmiennej losowej \bar{X}_n pozostaje normalny także wówczas, gdy próbka powstaje na drodze losowania zależnego (bez zwracania). W takiej sytuacji średnia arytmetyczna z próby ma wartość oczekiwaną $E(\bar{X}_n) = \mu(\bar{x}_n) = \mu$ – analogicznie jak w przypadku losowania ze zwracaniem – natomiast wariancja tej zmiennej losowej ma postać:

$$D^2(\bar{X}_n) = \sigma^2(\bar{x}) = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}. \quad (4.52)$$

Jak łatwo zauważyć, wartość $\sigma^2(\bar{x})$ jest tym mniejsza, im większą funkcję zbiorowości generalnej stanowi próba. W przypadku gdy $n = N$ – a więc w przypadku badań wyczerpujących – mamy $\sigma^2(\bar{x}) = 0$, a \bar{X}_n przestaje być zmienną losową, przestrzeń prób bowiem jest jednoelementowa.

4.5.2. Rozkład wariancji i odchylenia standardowego z próby

Jeśli obserwowana zmienna losowa X ma rozkład normalny o wariancji σ^2 , to wariancja z próby ma rozkład asymptotycznie normalny, przy czym:

$$E(\underline{S}^2) = \sigma^2, \quad D(\underline{S}^2) = \sigma^2 \sqrt{\frac{2}{n}}, \quad (4.53)$$

$$E(\underline{S}^2) = \frac{n-1}{n} \sigma^2, \quad D(\underline{S}^2) = \frac{\sigma^2}{n} \sqrt{2(n-1)}, \quad (4.54)$$

$$E(S^2) = \sigma^2, \quad D(S^2) = \sigma^2 \sqrt{\frac{2}{n-1}}. \quad (4.55)$$

Jeśli $X \sim N(\mu, \sigma)$, to również odchylenie standardowe z próby $S = \sqrt{S^2}$ ma rozkład asymptotycznie normalny o parametrach:

$$D(S) = \sigma, \quad D(S) = \sigma \frac{1}{\sqrt{2n}}. \quad (4.56)$$

W praktyce wnioskowania statystycznego podstawową rolę odgrywają małe próbki, o licznosci $n \leq 30$. Dlatego też przytoczone powyżej wzory mają ograniczone zastosowanie. Mogą być one stosowane w przypadku próbek o dużej licznosci. W przypadku małych próbek wykorzystuje się fakt, że statystyka:

$$\frac{S^2}{\sigma^2} = \frac{(n-1)S^2}{\sigma^2} \quad (4.57)$$

ma rozkład chi-kwadrat (χ^2) o $r = n - 1$ stopniach swobody.

Zmienna losowa χ_r^2 jest sumą kwadratów n niezależnych zmiennych losowych $U \sim N(0; 1)$. Ten definicyjny warunek jest spełniony w rozważanym przypadku, albowiem każdą wartość statystyki (4.57) można przedstawić następująco:

$$\frac{(n-1)s^2}{\sigma^2} = \frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \bar{x}_n)^2 = \sum_{i=1}^n \left(\frac{x_i - \bar{x}_n}{\sigma} \right)^2 = \sum_{i=1}^n u_i^2,$$

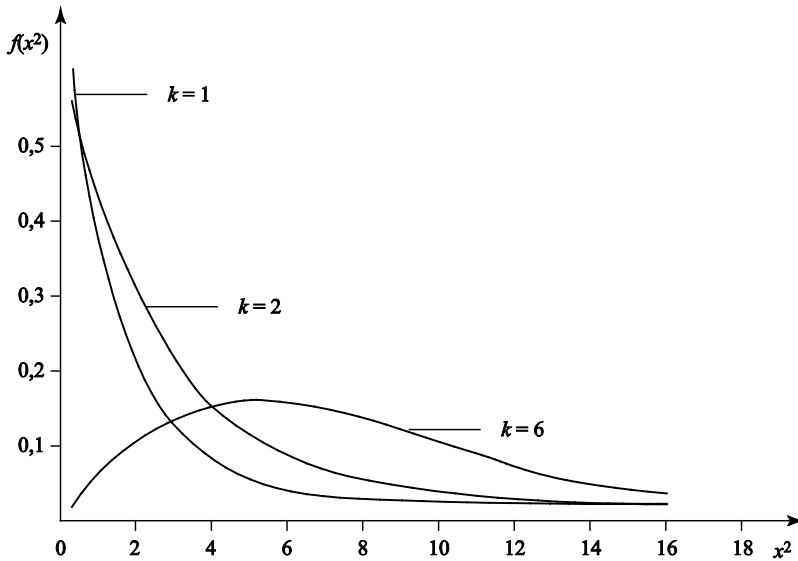
gdzie u_i jest wartością zmiennej losowej $U \sim N(0; 1)$. Funkcja gęstości prawdopodobieństwa rozkładu chi-kwadrat ma postać:

$$f_r(\chi^2) = \frac{(\chi^2)^{\frac{r}{2}-1}}{2^{\frac{r}{2}} \Gamma\left(\frac{r}{2}\right)} \exp\left(-\frac{r}{2} \chi^2\right), \quad (4.58)$$

przy czym χ^2 oznacza tu – na mocy definicji zmiennej losowej chi-kwadrat – dodatnią liczbę rzeczywistą. Wartość oczekiwana i wariancja zmiennej losowej chi-kwadrat o r stopniach swobody przedstawiają się następująco:

$$E(\chi_r^2) = r = n - 1, \quad (4.59)$$

$$D^2(\chi_r^2) = 2r = 2(n - 1). \quad (4.60)$$



Rys. 4.4. Rozkład chi-kwadrat

Na rys. 4.4 pokazano wykresy funkcji gęstości (4.58) dla trzech wybranych wartości r . Ze wzrostem liczby stopni swobody wykres funkcji (4.58) symetryzuje się. Przy $n > 30$ zmienna losowa $\sqrt{2\chi_r^2}$ ma rozkład zbliżony do rozkładu normalnego $N(\sqrt{2r-1}; 1)$. Tak więc, jeśli $n > 30$, to zmienna losowa $\sqrt{2\chi_r^2} - \sqrt{2r-1}$ ma rozkład normalny $N(0; 1)$.

Przedstawione zależności funkcjonują wówczas, gdy obserwowana zmienna losowa X ma rozkład normalny. Można jednak wykazać⁴, że niezależnie od rozkładu zmiennej losowej X , statystyka S^2 ma parametry:

$$E(S^2) = \sigma^2, \quad D(S^2) = \sqrt{\frac{1}{n} \left(\mu_4 - \frac{n-3}{n-1} \sigma^4 \right)}, \quad (4.61)$$

gdzie μ_4 jest czwartym momentem centralnym; por. wzór (3.84).

⁴ Zob. J. Greń, *op. cit.*, s. 155 i następn.

4.6. Przedziałowa estymacja parametrów jednowymiarowej zmiennej losowej

4.6.1. Ogólne zasady konstruowania przedziałów ufności

Estymacja przedziałowa możliwa jest wówczas, gdy znany jest rozkład estymatora punktowego (\hat{Q}_n) parametru Q . Wynikiem postępowania estymacyjnego jest przedział liczbowy:

$$(h_d(Q); h_g(Q))\gamma, \quad (4.62)$$

który z określonym prawdopodobieństwem γ pokrywa nieznaną wartość parametru Q . Mamy więc:

$$P(h_d(Q) < Q < h_g(Q)) = \gamma, \quad (4.63)$$

przy czym wartość γ może być dowolnie bliska jedności, ale różna od jedności. Należy podkreślić, że parametr Q jest tu traktowany jako wartość nielosowa, natomiast końce przedziału (4.62) są realizacjami zmiennych losowych $H_d(Q)$ i $H_g(Q)$, o rozkładach zależnych od Q ⁵. Przedział postaci (4.62) nazywany jest przedziałem ufności, natomiast prawdopodobieństwo γ określane jest jako współczynnik ufności. Estymacja przedziałowa ma tę przewagę nad estymacją punktową, że każdej ocenie przyporządkowane jest prawdopodobieństwo prawdziwości tej oceny. W estymacji punktowej jest to niemożliwe. Co więcej, jeśli statystyka \hat{Q}_n jest ciągłą zmienną losową, to $P(\hat{Q} = Q) = 0$. Praktyczną wartość mają przedziały o możliwie najmniejszej długości $h_g(Q) - h_d(Q)$ i o możliwie największym współczynniku ufności γ . Jednoczesne spełnienie tych dwóch wymagań jest możliwe tylko na drodze zwiększenia liczebności próby (n). Jest to równoznaczne ze zwiększeniem ilości informacji o badanym obiekcie lub zjawisku.

4.6.2. Przedział ufności dla wartości oczekiwanej

Zwróćmy uwagę na przypadek, gdy obserwowana zmienna losowa X ma rozkład normalny o znanym odchyleniu standardowym σ . W takiej sytuacji średnia arytmetyczna z próby (\bar{X}_n) ma rozkład $N(\mu(\bar{x}), \sigma(\bar{x})) = N(\mu, \sigma/\sqrt{n})$. Możemy więc napisać:

⁵ Takie ujęcie problemu przedstawił Jerzy Sława-Neyman (1894–1980), polski matematyk i statystyk, pracujący w USA.

$$u = \frac{x_n - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\bar{x}_n - \mu}{\sigma} \sqrt{n}, \quad (4.64)$$

gdzie u jest wartością zmiennej losowej $U \sim N(0; 1)$. Ponieważ:

$$P(|u| < u_{\alpha/2}) = 1 - \alpha = \gamma, \quad (4.65)$$

zatem

$$P\left(\left|\frac{x_n - \mu}{\sigma} \sqrt{n}\right| < u_{\alpha/2}\right) = 1 - \alpha = \gamma,$$

$$P\left(-u_{\alpha/2} < \frac{\bar{x}_n - \mu}{\sigma} \sqrt{n} < u_{\alpha/2}\right) = 1 - \alpha = \gamma,$$

$$P\left(\bar{x}_n - u_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu < \bar{x}_n + u_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha = \gamma.$$

W konsekwencji mamy:

$$h_d(\mu) = \bar{x}_n - u_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \quad (4.66)$$

$$h_g(\mu) = \bar{x}_n + u_{\alpha/2} \frac{\sigma}{\sqrt{n}}. \quad (4.67)$$

Przykład 4.4

Załóżmy, że procentowe stężenie tłuszczu w pewnym produkcie spożywczym jest zmienną losową $X \sim N(\mu, 0,2)$. W celu oszacowania wartości μ wykonano $n = 25$ niezależnych oznaczeń i uzyskano średnią arytmetyczną $\bar{x}_n = 31,50$. Należy skonstruować przedział ufności, przyjmując $\gamma = 0,95$. Poszukujemy więc przedziału liczbowego, który przeciętnie w 95 przypadkach na 100 powtórzeń doświadczenia będzie pokrywał nieznaną wartość μ . Ponieważ $\gamma = 1 - \alpha = 0,95$, więc $\alpha/2 = 0,025$. Przyjmując $\varepsilon = 0,025$ odczytujemy w tablicy II (zob. aneks) wartość $u_{\alpha/2} = u_{0,025} = 1,96$. Przyjmujemy przy tym $r = \infty$, albowiem odchylenie

standardowe (σ) jest znane. Po podstawieniu wartości do wzorów (4.66), (4.67) mamy:

$$h_d(\mu) = 31,5 - 1,96 \cdot \frac{0,2}{\sqrt{25}} = 31,42,$$

$$h_g(\mu) = 31,5 + 1,96 \cdot \frac{0,2}{\sqrt{25}} = 31,58.$$

Zatem przedział (31,42; 31,58) z prawdopodobieństwem $\gamma = 0,95$ pokrywa nieznaną średnią zawartość tłuszczu w badanym produkcie. ■

Jeśli nie jest znane odchylenie standardowe (σ) obserwowanej zmiennej losowej $X \sim N(\mu, \sigma)$, to przedział ufności dla wartości oczekiwanej μ konstruować należy w oparciu o rozkład Studenta. W przypadku gdy do szacowania σ wykorzystujemy statystykę S , mamy:

$$t_r = \frac{\bar{x}_n - \mu}{\frac{s}{\sqrt{n}}} = \frac{\bar{x}_n - \mu}{s} \sqrt{n}, \quad (4.68)$$

$$P(|t_r| < t_{r;\alpha/2}) = 1 - \alpha = \gamma, \quad (4.69)$$

gdzie t_r jest realizacją zmiennej losowej t -Studenta o $r = n - 1$ stopniach swobody. Możemy więc napisać:

$$P\left(\left|\frac{\bar{x}_n - \mu}{s} \sqrt{n}\right| < t_{r;\alpha/2}\right) = 1 - \alpha = \gamma,$$

$$P\left(\bar{x}_n - t_{r;\alpha/2} \frac{s}{\sqrt{n}} < \mu < \bar{x}_n + t_{r;\alpha/2} \frac{s}{\sqrt{n}}\right) = 1 - \alpha = \gamma$$

i w konsekwencji

$$h_d(\mu) = \bar{x}_n - t_{r;\alpha/2} \frac{s}{\sqrt{n}}, \quad (4.70)$$

$$h_g(\mu) = \bar{x}_n + t_{r;\alpha/2} \frac{s}{\sqrt{n}}. \quad (4.71)$$

W przypadku gdy estymatorem σ jest \underline{s} , wzory te przybierają postać:

$$h_d(\mu) = \bar{x}_n - t_{r;\alpha/2} \frac{s}{\sqrt{n-1}}, \quad (4.72)$$

$$h_g(\mu) = \bar{x}_n + t_{r;\alpha/2} \frac{s}{\sqrt{n-1}}, \quad (4.73)$$

albowiem mamy wówczas:

$$t_r = \frac{\bar{x}_n - \mu}{\frac{\underline{s}}{\sqrt{n-1}}} = \frac{\bar{x}_n - \mu}{\underline{s}} \sqrt{n-1}. \quad (4.74)$$

Różnice między wzorami (4.68) i (4.74) wynikają z faktu, że oszacowanie błędu średniego średniej arytmetycznej można uzyskać za pomocą wzoru (4.46) albo (4.47).

4.6.3. Przedział ufności dla różnicy dwóch wartości oczekiwanych

Jeśli $X_1 \sim N(\mu_1, \sigma_1)$ i $X_2 \sim N(\mu_2, \sigma_2)$, to zmienna losowa $(X_1 - X_2)$ ma również rozkład normalny o wartości oczekiwanej $\mu_1 - \mu_2$ i odchyleniu standardowym $\sqrt{\sigma_1^2 + \sigma_2^2}$. W konsekwencji mamy:

$$(\bar{X}_1 - \bar{X}_2) \sim N \left(\mu_1 - \mu_2; \sqrt{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \right)} \right), \quad (4.75)$$

$$u = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}, \quad (4.76)$$

gdzie u jest wartością zmiennej losowej $U \sim N(0; 1)$. Wykorzystując wzór (4.65) można więc napisać:

$$P \left(\left| \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \right| < u_{\alpha/2} \right) = 1 - \alpha = \gamma ,$$

$$P \left(-u_{\alpha/2} < \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} < u_{\alpha/2} \right) = 1 - \alpha = \gamma ,$$

$$P \left[(\bar{x}_1 - \bar{x}_2) - \mu_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + u_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right] = 1 - \alpha = \gamma .$$

Tak więc:

$$h_d(\mu_1 - \mu_2) = (\bar{x}_1 - \bar{x}_2) - u_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} , \quad (4.77)$$

$$h_g(\mu_1 - \mu_2) = (\bar{x}_1 - \bar{x}_2) + u_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} . \quad (4.78)$$

Jeśli wariancje σ_1^2 i σ_2^2 szacowane są z próby, to przedział ufności dla różnicy $\mu_1 - \mu_2$ konstruowany jest w oparciu o rozkład Studenta.

Zwróćmy uwagę na przypadek, gdy wariancje σ_1^2 i σ_2^2 szacowane są za pomocą estymatora S^2 . Mamy wówczas:

$$t_r = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{n_1 + n_2}{n_1 n_2}}} , \quad (4.79)$$

gdzie $r = (n_1 - 1) + (n_2 - 1) = n_1 + n_2 - 2$, natomiast

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 2)}}. \quad (4.80)$$

Jest to tzw. uogólnione oszacowanie odchylenia standardowego $\sqrt{\sigma_1^2 + \sigma_2^2}$. Na podstawie (4.69) możemy napisać:

$$P \left(\left| \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{n_1 + n_2}{n_1 n_2}}} \right| < t_{r; \alpha/2} \right) = 1 - \alpha = \gamma,$$

$$P \left(-t_{r; \alpha/2} < \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{n_1 + n_2}{n_1 n_2}}} < t_{r; \alpha/2} \right) = 1 - \alpha = \gamma,$$

$$P \left[(\bar{x}_1 - \bar{x}_2) - t_{r; \alpha/2} s_p \sqrt{\frac{n_1 + n_2}{n_1 n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + t_{r; \alpha/2} s_p \sqrt{\frac{n_1 + n_2}{n_1 n_2}} \right] = 1 - \alpha = \gamma.$$

Mamy więc:

$$h_d(\mu_1 - \mu_2) = (\bar{x}_1 - \bar{x}_2) - t_{r; \alpha/2} \cdot s_p \sqrt{\frac{n_1 + n_2}{n_1 n_2}}, \quad (4.81)$$

$$h_g(\mu_1 - \mu_2) = (\bar{x}_1 - \bar{x}_2) + t_{r; \alpha/2} \cdot s_p \sqrt{\frac{n_1 + n_2}{n_1 n_2}}. \quad (4.82)$$

Dodajmy, że analogicznie można skonstruować przedział ufności dla sumy wartości oczekiwanych dwóch normalnych zmiennych losowych.

4.6.4. Przedziały ufności dla wariancji i odchylenia standardowego

Przypomnijmy (zob. punkt 4.5.2), że jeśli $X \sim N(\mu, \sigma)$, to statystyka $(n-1) \cdot s^2/\sigma^2$ ma rozkład chi-kwadrat o $r = n - 1$ stopniach swobody. Możemy więc napisać:

$$\chi_r^2 = \frac{(n-1)s^2}{\sigma^2}, \quad (4.83)$$

$$P(\chi_{r;1-\alpha/2}^2 < \chi_r^2 < \chi_{r;\alpha/2}^2) = 1 - \alpha = \gamma. \quad (4.84)$$

Wynika stąd, że:

$$P\left(\chi_{r;1-\alpha/2}^2 < \frac{(n-1)s^2}{\sigma^2} < \chi_{r;\alpha/2}^2\right) = 1 - \alpha = \gamma,$$

$$P\left(\frac{(n-1)s^2}{\chi_{r;\alpha/2}^2} < \sigma^2 < \frac{(n-1)s^2}{\chi_{r;1-\alpha/2}^2}\right) = 1 - \alpha = \gamma.$$

Mamy zatem:

$$h_d(\sigma^2) = \frac{(n-1)s^2}{\chi_{r;\alpha/2}^2}, \quad (4.85)$$

$$h_g(\sigma^2) = \frac{(n-1)s^2}{\chi_{r;1-\alpha/2}^2}. \quad (4.86)$$

Przykład 4.5

Wykonano $n = 5$ niezależnych oznaczeń zawartości żelaza w roztworze chlorku żelazowego i otrzymano następujące wyniki (w mg): 11,2; 10,8; 10,9; 10,7; 11,0. Przyjmując, że rozkład obserwowanej zmiennej losowej jest zbliżony do normalnego, wyznaczmy przedział ufności dla wariancji, przy współczynniku ufności $\gamma = 0,95$.

Punktową ocenę σ^2 wyznaczmy według wzoru (4.27b). Niezbędne obliczenia przedstawiają się następująco:

i	1	2	3	4	5	$\sum_{i=1}^5$
x_i	11,2	10,8	10,9	10,7	11,0	54,6
x_i^2	125,44	116,64	118,81	114,49	121,00	596,38

Mamy więc:

$$s^2 = \frac{1}{n-1} \left[\sum_{i=1}^n x_i^2 - \frac{1}{n} \left(\sum_{i=1}^n x_i \right)^2 \right] = \frac{1}{4} [596,38 - \frac{1}{5} (54,6)^2] = 0,0370.$$

Wartość kwantyli $\chi_{r;1-\alpha/2}^2 = \chi_{4;0,975}^2$ i $\chi_{r;\alpha/2}^2 = \chi_{4;0,025}^2$ odczytujemy z tablicy III zamieszczonej w aneksie, przyjmując kolejno $\varepsilon = 0,975$ i $\varepsilon = 0,025$. Mamy mianowicie: $\chi_{4;0,975}^2 = 0,484$; $\chi_{4;0,025}^2 = 11,143$. Po podstawieniu wartości do wzorów (4.85) i (4.86) otrzymujemy:

$$h_d(\sigma^2) = 4 \cdot 0,0370/11,143 = 0,0133,$$

$$h_g(\sigma^2) = 4 \cdot 0,0370/0,484 = 0,3058.$$

Biorąc pierwiastki kwadratowe z tych wartości otrzymujemy przybliżony przedział ufności dla odchylenia standardowego (σ):

$$h_d(\sigma) \approx \sqrt{0,0133} \approx 0,1153,$$

$$h_g(\sigma) \approx \sqrt{0,3058} \approx 0,5530.$$

Zauważmy, że w rozważanym przypadku wariancja i odchylenie standardowe są miarami precyzji zastosowanej metody badania. Można bowiem przyjąć, że stężenie roztworu – wskutek dyfuzji – jest praktycznie jednakowe we wszystkich punktach. ■

W przypadku gdy próba jest liczna, przedział ufności dla wariancji (σ^2) i odchylenia standardowego (σ) można wyznaczyć w oparciu o rozkład normalny; zob. wzory (4.53)–(4.56). Zwróćmy uwagę na odchylenie standardowe (σ), szacowane za pomocą statystyki S . Wykorzystując wzór (4.56) możemy napisać:

$$u = \frac{s-\sigma}{\frac{\sigma}{\sqrt{2n}}} = \frac{s-\sigma}{\sigma} \sqrt{2n}, \quad (4.87)$$

gdzie u jest wartością zmiennej losowej $U \sim N(0; 1)$. Na podstawie wzoru (4.65) mamy:

$$P\left(\left|\frac{s-\sigma}{\sigma}\sqrt{2n}\right| < u_{\alpha/2}\right) = 1-\alpha = \gamma,$$

$$P\left(-u_{\alpha/2} < \frac{s-\sigma}{\sigma}\sqrt{2n} < u_{\alpha/2}\right) = 1-\alpha = \gamma,$$

$$P\left(\frac{s}{1+\frac{u_{\alpha/2}}{\sqrt{2n}}} < \sigma < \frac{s}{1-\frac{u_{\alpha/2}}{\sqrt{2n}}}\right) = 1-\alpha = \gamma$$

i w konsekwencji

$$h_d(\sigma) = \frac{s}{1+\frac{u_{\alpha/2}}{\sqrt{2n}}}, \quad (4.88)$$

$$h_g(\sigma) = \frac{s}{1-\frac{u_{\alpha/2}}{\sqrt{2n}}}. \quad (4.89)$$

Zauważmy, że powyższe przekształcenie jest możliwe wówczas, gdy $u_{\alpha/2}/\sqrt{2n} < 1$, a więc gdy $n > u_{\alpha/2}^2/2$. W przypadku licznych próbek warunek ten jest zawsze spełniony, albowiem nawet przy ekstremalnie wysokich wartościach γ , $u_{\alpha/2}$ nie przekracza wartości 4.

4.6.5. Przedział ufności dla stosunku dwóch wariancji

Jeśli $X_1 \sim N(\mu_1, \sigma_1)$ i $X_2 \sim N(\mu_2, \sigma_2)$, to statystyka:

$$\frac{\chi_{r_1}^2}{r_1} = \frac{(n_1-1)S_1^2}{(n_1-1)\sigma_1^2} = \frac{S_1^2}{\sigma_1^2} = \frac{S_1^2}{S_2^2} \cdot \frac{\sigma_2^2}{\sigma_1^2}$$

$$\frac{\chi_{r_2}^2}{r_2} = \frac{(n_2-1)S_2^2}{(n_2-1)\sigma_2^2} = \frac{S_2^2}{\sigma_2^2}$$

$$(4.90)$$

ma rozkład F -Snedecora o $r_1 = n_1 - 1$ i $r_2 = n_2 - 1$ stopniach swobody. Rozkład ten wykorzystamy do skonstruowania przedziału ufności dla stosunku dwóch wariancji. Mamy mianowicie:

$$F_{r_1, r_2} = \frac{\frac{s_1^2}{\sigma_1^2}}{\frac{s_2^2}{\sigma_2^2}}, \quad (4.91)$$

$$P(F_{r_1, r_2; 1-\alpha/2} < F_{r_1, r_2} < F_{r_1, r_2; \alpha/2}) = 1 - \alpha = \gamma. \quad (4.92)$$

Możemy więc napisać:

$$P \left(F_{r_1, r_2; 1-\alpha/2} < \frac{\frac{s_1^2}{\sigma_1^2}}{\frac{s_2^2}{\sigma_2^2}} < F_{r_1, r_2; \alpha/2} \right) = 1 - \alpha = \gamma,$$

$$P \left(\frac{\frac{s_1^2}{\sigma_1^2}}{\frac{s_2^2}{\sigma_2^2}} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{\frac{s_1^2}{\sigma_1^2}}{\frac{s_2^2}{\sigma_2^2}} \right) = 1 - \alpha = \gamma.$$

W konsekwencji

$$h_d \left(\frac{\sigma_1^2}{\sigma_2^2} \right) = \frac{\frac{s_1^2}{\sigma_1^2}}{F_{r_1, r_2; \alpha/2}}, \quad (4.93)$$

$$h_g \left(\frac{\sigma_1^2}{\sigma_2^2} \right) = \frac{\frac{s_1^2}{\sigma_1^2}}{F_{r_1, r_2; 1-\alpha/2}}, \quad (4.94)$$

gdzie $F_{r_1, r_2; \alpha/2}$ jest taką wartością zmiennej losowej F_{r_1, r_2} , że:

$$P(F_{r_1, r_2} > F_{r_1, r_2; \alpha/2}) = \frac{\alpha}{2},$$

natomiast $F_{r_1, r_2; 1-\alpha/2}$ spełnia warunek:

$$P(F_{r_1, r_2} < F_{r_1, r_2; 1-\alpha/2}) = \frac{\alpha}{2},$$

przy czym:

$$F_{r_1, r_2; 1-\alpha/2} < F_{r_1, r_2; \alpha/2}.$$

Tablice kwantyli rozkładu F -Snedecora (zob. tablice IV/1 i IV/2 zamieszczone w aneksie) podają tylko wartości $F_{r_1, r_2; \alpha/2}$. Wartości $F_{r_1, r_2; 1-\alpha/2}$ potrzebne do wyznaczenia wartości $h_g(\sigma_1^2 / \sigma_2^2)$ oblicza się ze wzoru:

$$F_{r_1, r_2; 1-\alpha/2} = \frac{1}{F_{r_2, r_1; \alpha/2}}. \quad (4.95)$$

4.6.6. Przedział ufności dla wskaźnika struktury

Parametr p w dwupunktowym rozkładzie zero-jedynkowej zmiennej losowej może być interpretowany jako tzw. wskaźnik struktury, czyli jako unormowana miara udziału w zbiorowości generalnej tych elementów (obiektów lub powtórzeń zjawiska), w których zrealizował się pewien wyróżniony stan obserwowanej cechy.

Estymatorem parametru p jest średnia arytmetyczna z próby (zob. punkt 4.3), która w rozważanym przypadku jest częstością względną realizacji o wartości 1, w n -elementowej próbie losowej. Mamy więc:

$$\hat{p} = \frac{Z}{n}, \quad (4.96)$$

gdzie:

$$Z = \sum_{i=1}^n X_i, \quad (4.97)$$

przy czym X_i ($i = 1, 2, \dots, n$) są zero-jedynkowymi zmiennymi losowymi, których realizacje (x_i) uzyskuje się w rezultacie badania kolejnych elementów próby.

Statystyka \hat{p} ma takie same własności jak zmienna losowa W , omówiona w rozdziale 1. Wynika to bezpośrednio z porównania wzorów (4.96) i (4.97) ze wzorem (1.67). W konsekwencji mamy:

$$E(\hat{p}) = p, \quad (4.98)$$

$$D^2(\hat{p}) = \frac{p(1-p)}{n} = \frac{pq}{n}. \quad (4.99)$$

Z ogólnych własności średniej arytmetycznej z próby wynika, że jeśli próba jest duża, to statystyka p ma asymptotyczny rozkład normalny. Można więc napisać:

$$u = \frac{\frac{z}{n} - p}{\sqrt{\frac{\frac{z}{n}\left(1 - \frac{z}{n}\right)}{n}}} = \frac{w - p}{\sqrt{\frac{w(1-w)}{n}}}, \quad (4.100)$$

gdzie $w = z/n$ jest wartością statystyki \hat{p} uzyskaną w rezultacie badania próby, natomiast u oznacza realizację standaryzowanej zmiennej losowej $U \sim N(0; 1)$.

Wykorzystując wzór (4.65) mamy więc:

$$P\left(\left|\frac{w - p}{\sqrt{\frac{w(1-w)}{n}}}\right| < u_{\alpha/2}\right) = 1 - \alpha = \gamma,$$

$$P\left(-u_{\alpha/2} < \frac{w - p}{\sqrt{\frac{w(1-w)}{n}}} < u_{\alpha/2}\right) = 1 - \alpha = \gamma,$$

$$P\left(w - u_{\alpha/2}\sqrt{\frac{w(1-w)}{n}} < p < w + u_{\alpha/2}\sqrt{\frac{w(1-w)}{n}}\right) = 1 - \alpha = \gamma.$$

W konsekwencji mamy:

$$h_d(p) = w - u_{\alpha/2} \sqrt{\frac{w(1-w)}{n}}, \quad (4.101)$$

$$h_g(p) = w + u_{\alpha/2} \sqrt{\frac{w(1-w)}{n}}, \quad (4.102)$$

gdzie $1 - \alpha = \gamma$ jest współczynnikiem ufności.

4.6.7. Przedział ufności dla różnicy między dwoma wskaźnikami struktury

Załóżmy, że z dwóch zbiorowości generalnych o nieznanymi parametrach p_1 i p_2 pobrano próbki o licznosciach n_1 i n_2 oraz że w rezultacie doświadczenia uzyskano wartości z_1 i z_2 według wzoru (4.97). Jeśli obydwie próbki są dostatecznie liczne, to po podstawieniu $w_1 = z_1/n_1$ oraz $w_2 = z_2/n_2$ możemy napisać:

$$u = \frac{(w_1 - w_2) - (p_1 - p_2)}{\sqrt{\frac{w_1(1-w_1)}{n_1} + \frac{w_2(1-w_2)}{n_2}}}, \quad (4.103)$$

gdzie u jest wartością standaryzowanej zmiennej losowej $U \sim N(0; 1)$.

Postępując analogicznie jak w punkcie 4.6.5 mamy:

$$P \left(\left| \frac{(w_1 - w_2) - (p_1 - p_2)}{\sqrt{\frac{w_1(1-w_1)}{n_1} + \frac{w_2(1-w_2)}{n_2}}} \right| < u_{\alpha/2} \right) = 1 - \alpha = \gamma,$$

$$P \left(-u_{\alpha/2} < \frac{(w_1 - w_2) - (p_1 - p_2)}{\sqrt{\frac{w_1(1-w_1)}{n_1} + \frac{w_2(1-w_2)}{n_2}}} < u_{\alpha/2} \right) = 1 - \alpha = \gamma.$$

Poszukiwane końce przedziału ufności wynikają ze wzorów:

$$h_d(p_1 - p_2) = (w_1 - w_2) - u_{\alpha/2} \sqrt{\frac{w_1(1-w_1)}{n_1} + \frac{w_2(1-w_2)}{n_2}}, \quad (4.104)$$

$$h_g(p_1 - p_2) = (w_1 - w_2) + u_{\alpha/2} \sqrt{\frac{w_1(1-w_1)}{n_1} + \frac{w_2(1-w_2)}{n_2}}. \quad (4.105)$$

Przedział liczbowy o tych końcach z prawdopodobieństwem $\gamma = 1 - \alpha$ pokrywa nieznaną wartość różnicy $p_1 - p_2$.

4.6.8. Liczebność próby

Przedstawione powyżej zasady konstruowania przedziałów ufności pozwalają wyznaczyć liczebność próbki niezbędną dla uzyskania przedziału o zadanej z góry długości.

Zwróćmy uwagę na przedział ufności dla wartości oczekiwanej normalnej zmiennej losowej o końcach $h_d(\mu)$ i $h_g(\mu)$, określonych wzorami (4.66) i (4.67). Niech Δx oznacza największą dopuszczalną długość przedziału ufności. Mamy więc:

$$[h_g(\mu) - h_d(\mu)] \leq \Delta x,$$

$$\left(\bar{x}_n + u_{\alpha/2} \frac{\sigma}{\sqrt{n}} - \bar{x}_n + u_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right) \leq \Delta x,$$

$$\frac{2u_{\alpha/2}\sigma}{\sqrt{n}} \leq \Delta x,$$

a stąd

$$n \geq \left(\frac{2u_{\alpha/2}\sigma}{\Delta x} \right)^2. \quad (4.106)$$

Przykład 4.6

W przykładzie 4.4 wyznaczyliśmy przedział ufności dla μ o długości $h_g(\mu) - h_d(\mu) = 31,58 - 31,42 = 0,16$. Przedział ten wyznaczyliśmy w oparciu o $n = 25$ niezależnych realizacji zmiennej losowej $X \sim N(\mu, 0,2)$, przyjmując

współczynnik ufności $\gamma = 0,95$. Obliczymy obecnie licznosc próbki niezbędną dla uzyskania przedziału ufności o długości nie przekraczającej $\Delta x = 0,08$. Podstawiając wartości liczbowe do wzoru (4.94) otrzymujemy:

$$n \geq \left(\frac{2 \cdot 1,96 \cdot 0,2}{0,08} \right)^2 = 96,04.$$

Najmniejszą liczbą naturalną spełniającą nierówność $n \geq 96,04$ jest $n = 97$. Czytelnik zechce sprawdzić, że przy tej liczności próbki uzyskujemy przedział ufności o długości $0,0796 < 0,08$. ■

4.7. Punktowa i przedziałowa estymacja współczynnika korelacji liniowej

Przypomnijmy (zob. punkt 1.5.5), że współczynnik korelacji liniowej ρ_{xy} ma postać:

$$\rho_{xy} = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X) \text{var}(Y)}}.$$

Wariancje są centralnymi momentami drugiego stopnia, natomiast kowariancja jest centralnym momentem mieszanym, również drugiego stopnia. Pozwala to zastosować metodę momentów do konstrukcji estymatora współczynnika korelacji liniowej ρ_{xy} . Należy w tym celu zastąpić wymienione powyżej momenty odpowiednimi empirycznymi momentami z próby, które definiuje się analogicznie jak momenty empiryczne w badaniach wyczerpujących (zob. punkt 3.2.4).

Współczynnik korelacji z próby – będący estymatorem współczynnika korelacji ρ_{xy} – ma więc postać:

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n} \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n}}} = \quad (4.107a)$$

$$= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} = \quad (4.107b)$$

$$= \frac{\sum_{i=1}^n x_i y_i - \frac{1}{n} \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{\sqrt{\left[\sum_{i=1}^n x_i^2 - \frac{1}{n} \left(\sum_{i=1}^n x_i \right)^2 \right] \left[\sum_{i=1}^n y_i^2 - \frac{1}{n} \left(\sum_{i=1}^n y_i \right)^2 \right]}} \quad (4.107c)$$

Otrzymaliśmy więc wzory analogiczne jak w rozdziale 3 w przypadku badań wyczerpujących. Jeśli próba jest liczna i obejmuje kilkaset elementów badanych ze względu na dwie cechy, to współczynnik korelacji z próby (r_{xy}) ma asymptotyczny rozkład normalny o wartości oczekiwanej $E(r_{xy}) = \rho_{xy}$ i odchyleniu standardowym $D(r_{xy}) = (1 - \rho_{xy}^2) / \sqrt{n}$. Mamy więc:

$$u = \frac{r_{xy} - \rho_{xy}}{\frac{1 - \rho_{xy}^2}{\sqrt{n}}} = \frac{r_{xy} - \rho_{xy}}{1 - \rho_{xy}^2} \sqrt{n}, \quad (4.108)$$

gdzie u jest wartością zmiennej losowej U o standardowym rozkładzie normalnym $N(0; 1)$. Wykorzystując wzór (4.65) możemy zapisać:

$$P \left(\left| \frac{r_{xy} - \rho_{xy}}{1 - \rho_{xy}^2} \sqrt{n} \right| < u_{\alpha/2} \right) = 1 - \alpha = \gamma, \quad (4.108a)$$

$$P \left(-u_{\alpha/2} < \frac{r_{xy} - \rho_{xy}}{1 - \rho_{xy}^2} \sqrt{n} < u_{\alpha/2} \right) = 1 - \alpha = \gamma, \quad (4.108b)$$

$$P \left(r_{xy} - u_{\alpha/2} \frac{1 - r_{xy}^2}{\sqrt{n}} < \rho_{xy} < r_{xy} + u_{\alpha/2} \frac{1 - r_{xy}^2}{\sqrt{n}} \right) = 1 - \alpha = \gamma.$$

W konsekwencji mamy:

$$h_d(\rho_{xy}) = r_{xy} - u_{\alpha/2} \frac{1 - r_{xy}^2}{\sqrt{n}}, \quad (4.109)$$

$$h_g(\rho_{xy}) = r_{xy} + u_{\alpha/2} \frac{1 - r_{xy}^2}{\sqrt{n}}. \quad (4.110)$$

W przypadku małych próbek przedział ufności dla współczynnika korelacji (ρ_{xy}) należy wyznaczać w oparciu o dokładny rozkład współczynnika korelacji z próby (r_{xy}). Kwantyle tego rozkładu są stabicowane (zob. R. Zieliński, *Tablice statystyczne*, PWN, Warszawa 1972, nomogramy na s. 183 i 184).

Rozdział 5

WERYFIKACJA HIPOTEZ STATYSTYCZNYCH

5.1. Wprowadzenie

Badania niewyczerpujące (częściowe) nie dają podstaw do formułowania stanowczych stwierdzeń w odniesieniu do obserwowanych zmiennych losowych. Wyniki takich badań pozwalają natomiast formułować i weryfikować (sprawdzać) różnego rodzaju przypuszczenia dotyczące zarówno klasy rozkładu, jak i jego parametrów. Przypuszczenia te nazywane są hipotezami statystycznymi¹. Hipotezy dotyczące parametrów noszą nazwę hipotez parametrycznych, natomiast hipotezy dotyczące klasy rozkładu – a więc postaci funkcyjnej dystrybuanty, ale bez odwoływania się do liczbowych wartości parametrów – nazywamy hipotezami nieparametrycznymi. Procedury służące do sprawdzania (weryfikacji) owych hipotez noszą nazwę testów statystycznych. W celu wyjaśnienia ogólnych zasad funkcjonowania tego typu procedur rozważymy problem weryfikacji hipotez dotyczących różnicy między nieznaną wartością parametru Q zmiennej losowej X , a ustaloną wartością Q_0 . Zadania takie rozwiązuje się za pomocą tzw. parametrycznych testów istotności.

Postępowanie weryfikacyjne rozpoczyna się od sformułowania hipotez. Formułuje się jedną hipotezę zerową i jedną lub kilka hipotez alternatywnych. Hipotezę zerową formułuje się w taki sposób, by orzekała brak różnic (stąd jej nazwa) lub ogólniej – brak wpływu badanego czynnika doświadczalnego na obserwowany parametr Q . Hipoteza zerowa stanowi więc najczęściej zaprzeczenie

¹ Weryfikacji hipotez statystycznych nie należy utożsamiać z treścią następującej anegdoty. Indianie zwrócili się do szamana z pytaniem, jaka będzie w tym roku zima. Szaman odpowiedział – bardzo ostra. Indianie rozpoczęli zbieranie drewna na opał. Zima była łagodna. W następnym roku Indianie ponownie zapytali szamana, który – niestety – znowu się pomylił. Szaman bojąc się trzeciej pomyłki zwrócił się o pomoc do stacji meteorologicznej z zapytaniem, jakie są prognozy na tegoroczną zimę. Meteorolodzy odpowiedzieli – skoro Indianie zbierają drewno już drugi rok, oznacza to, że zima będzie ostra.

hipotezy merytorycznej. W rozważanym przypadku hipoteza zerowa ma najczęściej postać:

$$H_0: Q = Q_0 \quad \text{lub} \quad H_0: Q - Q_0 = 0. \quad (5.1)$$

Zapisana w ten sposób hipoteza zerowa jest hipotezą prostą, albowiem specyfikuje ona jednoznacznie szczegółową postać funkcyjną dystrybuanty zmiennej losowej X . Wysuwamy mianowicie przypuszczenie, że parametr Q ma wartość Q_0 .

Hipoteza zerowa może być również hipotezą złożoną postaci:

$$H_0: Q \leq Q_0 \quad \text{lub} \quad H_0: Q - Q_0 \leq 0 \quad (5.2)$$

albo

$$H_0: Q \geq Q_0 \quad \text{lub} \quad H_0: Q - Q_0 \geq 0. \quad (5.3)$$

Wartość Q_0 jest tu interpretowana jako kres górny (hipoteza (5.2)) albo jako kres dolny (hipoteza (5.3)) pewnego zbioru dopuszczalnych wartości parametru Q .

Hipotezy alternatywne formułowane są następująco:

$$H_1: Q \neq Q_0 \quad \text{lub} \quad H_1: Q - Q_0 \neq 0, \quad (5.4)$$

$$H_1: Q > Q_0 \quad \text{lub} \quad H_1: Q - Q_0 > 0, \quad (5.5)$$

$$H_1: Q < Q_0 \quad \text{lub} \quad H_1: Q - Q_0 < 0. \quad (5.6)$$

Są to hipotezy złożone. Żadna z nich nie specyfikuje jednoznacznie szczegółowej postaci funkcyjnej dystrybuanty obserwowanej zmiennej losowej X . Każda z tych hipotez może być alternatywą hipotezy zerowej (5.1). Alternatywą hipotezy zerowej (5.2) jest hipoteza (5.5). Hipoteza (5.6) jest natomiast alternatywą hipotezy zerowej (5.3).

Przy tak sformułowanych hipotezach zerowych i alternatywnych test – czyli procedura weryfikacyjna – funkcjonuje według podanego poniżej schematu:

– Ze zbiorowości generalnej A , w której określona jest zmienna losowa X , pobiera się n -elementową próbkę losową \underline{A} .

– Na podstawie zbioru realizacji X_n oblicza się wartość $D(X_n)$ statystyki D , będącej sprawdzianem hipotezy zerowej. Statystykę D wybiera się stosownie do treści weryfikowanej hipotezy.

– Na podstawie rozkładu prawdopodobieństwa statystyki D wyznacza się obszar (przedział) krytyczny D_k , taki że:

$$P[D(X_n) \in D_k \mid Q = Q_0] = \alpha, \quad (5.7)$$

gdzie α jest dowolnie małą, ale różną od zera, wartością prawdopodobieństwa. Zauważmy, że w przypadku hipotez (5.2) i (5.3) mamy:

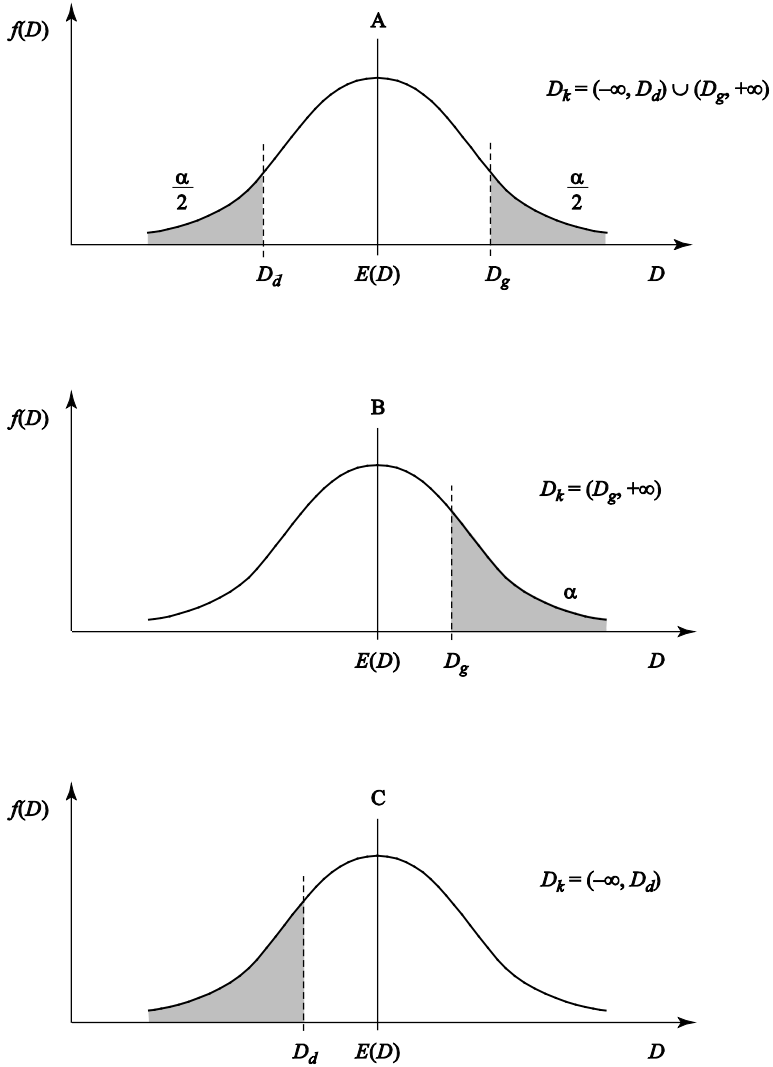
$$P[D(X_n) \in D_k \mid Q \leq Q_0] \leq \alpha, \quad (5.8)$$

$$P[D(X_n) \in D_k \mid Q \geq Q_0] \leq \alpha. \quad (5.9)$$

– Jeśli uzyskana z doświadczenia wartość $D(X_n)$ spełnia warunek:

$$D(X_n) \in \mathbf{D}_k, \quad (5.10)$$

to sprawdzaną hipotezę zerową odrzuca się na korzyść hipotezy alternatywnej.



Rys. 5.1. Konstruowanie obszarów krytycznych

Decyzja taka jest realizacją poglądu opartego na doświadczeniu, że zdarzenia mało prawdopodobne występują rzadko. Jeśli więc wystąpiło zdarzenie, które – przy założeniu prawdziwości H_0 – praktycznie wystąpić nie powinno, to kwestionujemy słuszność przyjętego założenia, odrzucając H_0 .

– Jeśli natomiast:

$$D(X_n) \notin D_k, \quad (5.11)$$

to nie ma podstaw do odrzucenia weryfikowanej hipotezy zerowej.

Funkcjonowanie rozważanej procedury ilustruje rys. 5.1. Przy sporządzaniu tego rysunku założono, że statystyka D ma rozkład normalny o wartości oczekiwanej $E(D) = Q_0$. Na rysunku zaznaczono wartości krytyczne D_d i D_g . Są to kreśły odpowiednich obszarów krytycznych. W praktyce operuje się przede wszystkim tymi właśnie wartościami bez specyfikowania obszarów krytycznych D_k . Jeśli hipoteza zerowa ma postać (5.1) i jest weryfikowana wobec hipotezy alternatywnej (5.4), to test jest dwustronny. Obszar krytyczny D_k składa się z dwóch rozłącznych części, co jest pokazane na rys. 5.1A. Taki podział obszaru krytycznego jest niezbędny, albowiem zarówno zbyt małe, jak i zbyt duże wartości odchyleń $D(X_n) - E(D)$ świadczą przeciw hipotezie zerowej. Jeśli hipoteza (5.5) jest alternatywą wobec hipotezy zerowej (5.2) albo (5.1), to test jest jednostronny (prawostronny); zob. rys. 5.1B. W tym przypadku tylko zbyt duże wartości odchyleń $D(X_n) - E(D)$ przemawiają na niekorzyść hipotezy zerowej (5.5). Jeśli hipoteza zerowa (5.3) albo (5.1) jest weryfikowana wobec hipotezy alternatywnej (5.6), to mamy do czynienia z testem lewostronnym; zob. rys. 5.1C. W takiej sytuacji tylko zbyt małe wartości $D(X_n) - E(D)$ świadczą przeciw hipotezie zerowej. Zauważmy, że za pomocą przedstawionej procedury można albo odrzucić H_0 , albo orzec, że nie ma podstaw do odrzucenia tej hipotezy. Procedury te nie służą natomiast do przyjmowania hipotez zerowych.

W bardziej ogólnych rozważaniach – leżących u podstaw teorii weryfikacji hipotez statystycznych – bierze się pod uwagę zarówno odrzucanie, jak i przyjmowanie hipotez zerowych. Rozróżnia się w związku z tym dwa rodzaje błędów, jakie można popełnić przy weryfikacji każdej hipotezy zerowej. Istotą tych błędów ilustruje rys. 5.2. Błąd I rodzaju polega na odrzuceniu prawdziwej hipotezy zerowej. Prawdopodobieństwo popełnienia tego błędu oznaczane jest symbolem α (zob. wzory (5.7)–(5.9)) i nosi nazwę poziomu istotności testu. Błąd II rodzaju polega na przyjęciu hipotezy zerowej, która w istocie jest fałszywa. Prawdopodobieństwo popełnienia tego błędu oznaczane jest symbolem β .

W standardowych testach istotności tylko wartość α zadawana jest *a priori*, natomiast β jest wartością wynikową, zależną od liczności próbki (n) i rzeczywistej wartości parametru Q . Ponieważ owa rzeczywista wartość parametru Q jest – z założenia – nieznaną, zatem i wartość β pozostaje nieznaną.

		Hipoteza zerowa	
		Prawdziwa	Falszywa
Decyzja	Przyjąć	Decyzja prawidłowa	Błąd II rodzaju
		Odrzucić	Błąd I rodzaju

Rys. 5.2. Błędy popełniane przy weryfikacji hipotez statystycznych

Każdy standardowy test istotności można jednak stransformować do postaci, w której również wartość β może być zadawana *a priori*. Wymaga to odpowiedniego przekształcenia hipotezy alternatywnej i określenia minimalnej liczności próbki. Zwróćmy uwagę na test dwustronny, w którym weryfikacji podlega hipoteza zerowa (5.1) wobec hipotezy alternatywnej (5.4). Hipotezę (5.4) należy zastąpić dwiema hipotezami alternatywnymi postaci:

$$H_1: Q = Q_0 + \Delta Q = Q_1, \quad (5.12)$$

$$H_{-1}: Q = Q_0 - \Delta Q = Q_{-1}, \quad (5.13)$$

gdzie $\Delta Q > 0$ i w konsekwencji $Q_{-1} < Q_0 < Q_1$. Mamy więc:

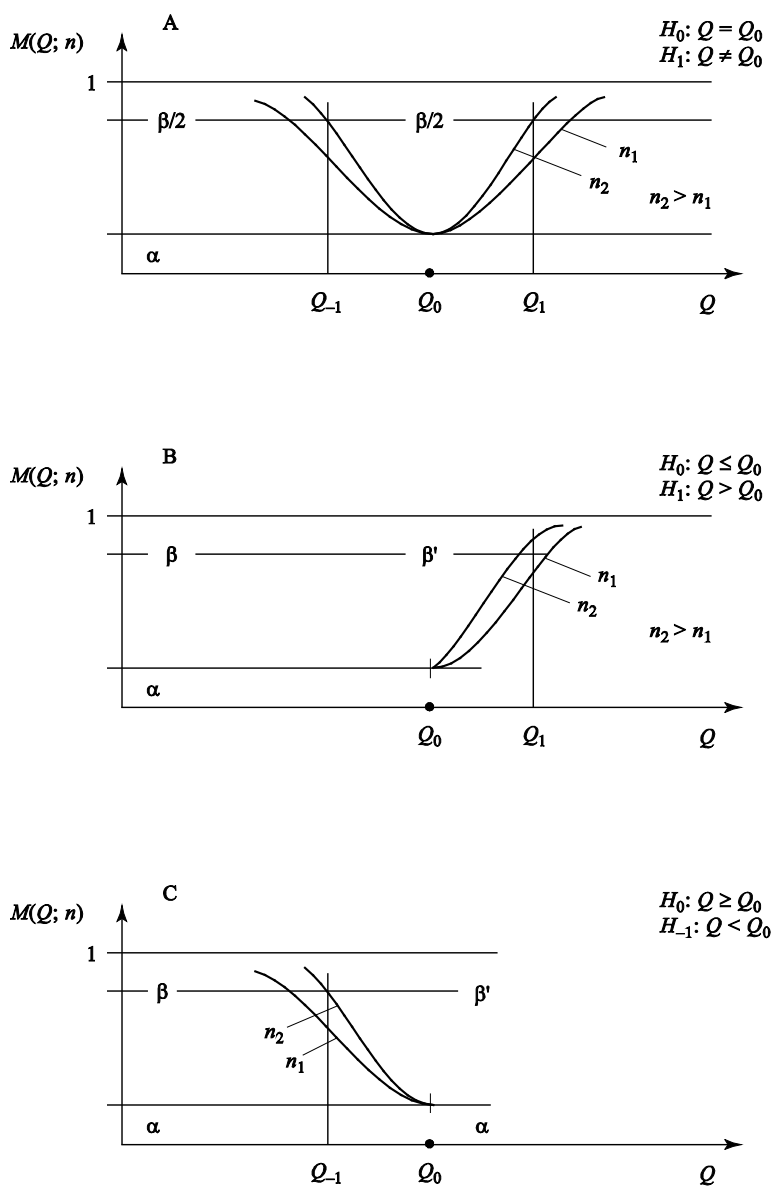
$$\beta' = P[D(\mathbf{X}_n) \notin \mathbf{D}_k | Q = Q_{-1}] + P[D(\mathbf{X}_n) \notin \mathbf{D}_k | Q = Q_1]. \quad (5.14)$$

Jeśli przyjmując dodatkowo, że $Q_0 - Q_{-1} = Q_1 - Q_0$, to:

$$P[D(\mathbf{X}_n) \notin \mathbf{D}_k | Q = Q_{-1}] = \frac{\beta'}{2}, \quad (5.15)$$

$$P[D(\mathbf{X}_n) \notin \mathbf{D}_k | Q = Q_1] = \frac{\beta'}{2}. \quad (5.16)$$

Jeśli wartość ΔQ jest ustalona, to β' jest malejącą funkcją liczności próbki (n). Można więc znaleźć taką licznosc próbki (n'), przy której $\beta' \leq \beta$, gdzie β jest przyjętym *a priori* prawdopodobieństwem błędu II rodzaju. Przy poszukiwaniu wartości n' korzysta się z własności funkcji mocy testu (zob. rys. 5.3).



Rys. 5.3. Krzywe mocy testów statystycznych

Funkcja mocy jest podstawową charakterystyką operacyjną testu i wyraża zależność między prawdopodobieństwem odrzucenia hipotezy zerowej a wartością parametru Q , przy ustalonej liczności próbki (n). Mamy więc:

$$M(Q; n) = P[D(X_n) \in D_k | Q]. \quad (5.17)$$

Na rys. 5.3A pokazano dwa wykresy tej funkcji dla testu dwustronnego, przy $n = n_1$ i $n = n_2$; $n_2 > n_1$. Na rysunku tym zaznaczono wartości Q_{-1} i Q_1 . Przy $n = n_2$ mamy $1 - M(Q_{-1}; n_2) = 1 - M(Q_1; n_2) = \beta'/2 < \beta/2$. Te same problemy występują oczywiście również w przypadku testów jednostronnych. Pokazano to na rys. 5.3 B i C.

5.2. Weryfikacja hipotez dotyczących wartości oczekiwanej

5.2.1. Weryfikacja istotności różnicy między wartością oczekiwaną zmiennej losowej a ustaloną wartością

Niech X oznacza zmienną losową o normalnym rozkładzie prawdopodobieństwa z nieznaną wartością oczekiwaną μ . Drugi parametr rozkładu – czyli odchylenie standardowe σ – może być znany lub może być szacowany z próby. Weryfikacji podlegają hipotezy zerowe postaci (5.1)–(5.3) wobec hipotez alternatywnych postaci (5.4)–(5.6). Mamy więc:

$$H_0: \mu = \mu_0, \quad (5.18)$$

$$H_0: \mu \leq \mu_0, \quad (5.19)$$

$$H_0: \mu \geq \mu_0 \quad (5.20)$$

oraz

$$H_1: \mu \neq \mu_0, \quad (5.21)$$

$$H_1: \mu > \mu_0, \quad (5.22)$$

$$H_1: \mu < \mu_0, \quad (5.23)$$

gdzie μ_0 oznacza pewną ustaloną wartość, wynikającą najczęściej z przesłanek pozastatystycznych. Hipotezy te formułujemy w celu rozstrzygnięcia, czy stwierdzona empirycznie relacja między średnią arytmetyczną z próby (\bar{x}_n) a warto-

ścią μ_0 wynika z działania czynników systematycznych, czy też została ona ukształtowana tylko w rezultacie działania czynników losowych.

5.2.1.1. Test u

Przypomnijmy (zob. rozdział 4), że jeśli odchylenie standardowe jest znane, to:

$$u = \frac{\bar{x}_n - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\bar{x}_n - \mu}{\sigma} \sqrt{n}$$

jest realizacją standaryzowanej zmiennej losowej U o rozkładzie $N(0; 1)$.

Każda ze sformułowanych powyżej hipotez zerowych zawiera przypuszczenie, że $\mu = \mu_0$. Tak więc jeśli hipoteza zerowa jest prawdziwa, to również charakterystyka:

$$U_0 = \frac{\bar{X}_n - \mu_0}{\sigma} \sqrt{n} \quad (5.24)$$

o wartościach uzyskiwanych według wzoru:

$$u_0 = \frac{\bar{x}_n - \mu_0}{\sigma} \sqrt{n} \quad (5.25)$$

ma standardowy rozkład normalny $N(0; 1)$.

Charakterystyka (5.24) może być wykorzystana jako sprawdzian (D) hipotez zerowych (5.18)–(5.20). Wartości tego sprawdzianu, które w punkcie 5.1 oznaczyliśmy symbolem $D(\bar{X}_n)$, oblicza się ze wzoru (5.25). Jeśli test jest dwustronny, to obszar krytyczny ma postać:

$$D_k = (-\infty; -u_{\alpha/2}] \cup [u_{\alpha/2}; \infty), \quad (5.26)$$

przy czym wartości $-u_{\alpha/2} < 0$ i $u_{\alpha/2} > 0$ wyznaczają się w taki sposób, że:

$$P(U \leq -u_{\alpha/2}) = \frac{\alpha}{2}, \quad (5.27)$$

$$P(U \geq u_{\alpha/2}) = \frac{\alpha}{2}. \quad (5.28)$$

W przypadku testów jednostronnych mamy natomiast:

$$D_k = (-\infty; -u_\alpha] \quad (5.29)$$

albo

$$D_k = [u_\alpha; \infty), \quad (5.30)$$

przy czym

$$P(U \leq -u_\alpha) = \alpha, \quad (5.31)$$

$$P(U \geq u_\alpha) = \alpha. \quad (5.32)$$

W przypadku testu dwustronnego reguły decyzyjne są następujące:

– jeśli wartość u_0 obliczona według wzoru (5.25) spełnia nierówność:

$$|u_0| \geq u_{\alpha/2}, \quad (5.33)$$

to należy odrzucić H_0 na korzyść H_1 ,

– jeśli natomiast:

$$|u_0| < u_{\alpha/2}, \quad (5.34)$$

to nie ma podstaw do odrzucenia H_0 .

W przypadku testów jednostronnych (lewostronnych i prawostronnych) reguły decyzyjne przedstawiają się następująco:

– jeśli wartość u_0 obliczona według wzoru (5.25) spełnia nierówność:

$$|u_0| \geq u_\alpha, \quad (5.35)$$

to należy odrzucić H_0 na korzyść H_1 ,

– jeśli natomiast:

$$|u_0| < u_\alpha, \quad (5.36)$$

to nie ma podstaw do odrzucenia H_0 .

Wartości krytyczne u_α albo $u_{\alpha/2}$ należy odczytać z tablicy II, zamieszczonej w aneksie, dla $r = \infty$. Wartości te można również wyznaczyć na podstawie zamieszczonej w aneksie tablicy I. Należy mianowicie znaleźć taką wartość u_α albo $u_{\alpha/2}$, dla której $1 - \Phi(u_\alpha) = \alpha$, $1 - \Phi(u_{\alpha/2}) = \alpha/2$.

Zarówno w teście dwustronnym, jak i jednostronnym prawdopodobieństwo odrzucenia prawdziwej hipotezy zerowej wynosi α .

Przykład 5.1

Nominalna długość pewnego typu sworznia wynosi $L_0 = 100$ mm. Na podstawie długotrwałych obserwacji procesu technologicznego tych sworzni ustalono, że ich długość (L) jest zmienną losową o normalnym rozkładzie prawdopodobieństwa i odchyleniu standardowym $\sigma = 1,8$ mm. W celu sprawdzenia prawidłowości przebiegu procesu technologicznego pobrano z bieżącej produkcji próbkę losową o liczności $n = 25$ sztuk. W wyniku badania tej próbki uzyskano $\bar{l} = 100,4$ mm. Czy można utrzymać hipotezę $H_0: E(L) = 100$ mm przy alternatywie $H_1: E(L) \neq 100$ mm?

Obliczamy:

$$u_0 = \frac{100,4 - 100}{1,8} \sqrt{25} = 1,11.$$

Poziom istotności testu α ustalamy na 0,05. Stąd $\alpha/2 = 0,025$, a $u_{0,025} = 1,96$ (tablica II). Obliczona wartość $u_0 = 1,11$ jest więc mniejsza od wartości krytycznej $u_{0,025} = 1,96$. Wynika stąd, że nie ma podstaw do odrzucenia sformułowanej hipotezy zerowej. Oznacza to, że zaobserwowane odchylenie średniej długości sworznia $\bar{l} = 100,4$ od wartości nominalnej $L_0 = 100$ jest zbyt małe, by można twierdzić, że przeciętna długość sworznia odbiega od wartości nominalnej. ■

W przedstawionej powyżej standardowej wersji testu u pod kontrolą pozostaje tylko błąd I rodzaju. Takie rozwiązanie nie ogranicza własności aplikacyjnych testu, jeśli hipoteza zerowa stanowi zaprzeczenie weryfikowanej hipotezy merytorycznej. W takich przypadkach prawdopodobieństwo α może być bowiem interpretowane jako prawdopodobieństwo przyjęcia fałszywej hipotezy merytorycznej. Istnieje jednak szeroka klasa doświadczeń, w których pożądana jest kontrola zarówno błędu I jak i II rodzaju. Rozważmy obecnie ten problem w odniesieniu do testu prawostronnego, w którym weryfikacji podlega hipoteza zerowa $H_0: \mu = \mu_0$ wobec hipotezy alternatywnej $H_1: \mu > \mu_0$.

Zauważmy, że funkcję testową (5.25) można stransformować do postaci:

$$\bar{x}_n = \mu_0 + u \frac{\sigma}{\sqrt{n}}.$$

Ustalając wartość u na poziomie u_α (zob. wzór (5.30)) otrzymujemy krytyczną wartość średniej arytmetycznej z próby:

$$\bar{x}_{n,g} = \mu_0 + u_\alpha \frac{\sigma}{\sqrt{n}} \quad (5.37)$$

taką że

$$P(\bar{x}_n \geq \bar{x}_{n,g} \mid \mu = \mu_0) = \alpha \quad (5.38)$$

i w konsekwencji

$$P(\bar{x}_n < \bar{x}_{n,g} \mid \mu = \mu_0) = 1 - \alpha. \quad (5.39)$$

W takiej sytuacji sprawdzianem D hipotezy H_0 jest \bar{X}_n , a nie statystyka U o wartościach obliczonych według wzoru (5.25).

W zmodyfikowanej wersji testu – stosowanej szczególnie często w statystycznej kontroli jakości w toku produkcji – stosowane są następujące reguły decyzyjne:

– jeśli \bar{x}_n spełnia nierówność:

$$\bar{x}_n \geq \bar{x}_{n,g}, \quad (5.40)$$

to odrzuca się hipotezę zerową na korzyść hipotezy alternatywnej,

– jeśli natomiast:

$$\bar{x}_n < \bar{x}_{n,g}, \quad (5.41)$$

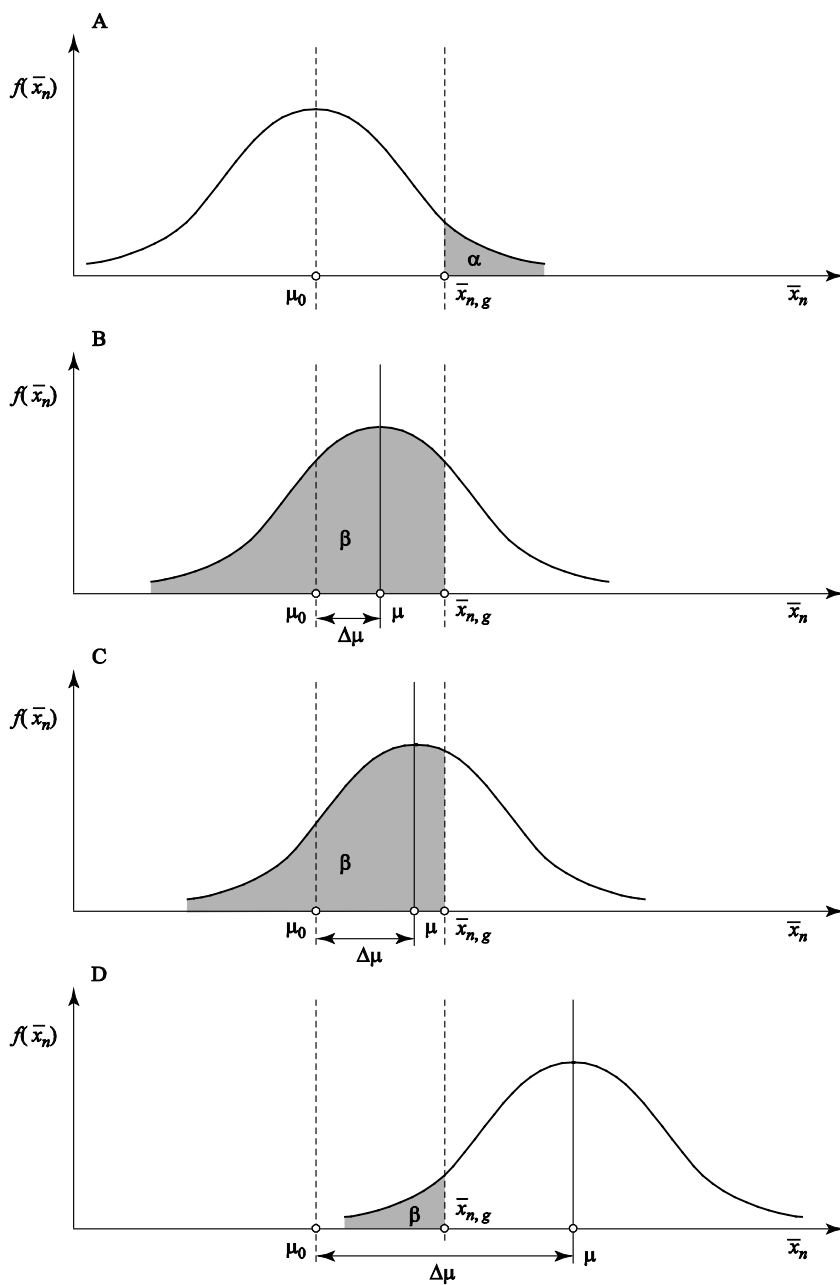
to nie ma podstaw do odrzucenia hipotezy zerowej.

Funkcjonowanie tych reguł ilustruje rys. 5.4, na którym pokazano cztery wykresy funkcji gęstości prawdopodobieństwa zmiennej losowej \bar{X}_n o ustalonym odchyleniu standardowym σ/\sqrt{n} i o różnych wartościach oczekiwanych μ . Rys. 5.4A ilustruje sytuację, gdy $\mu = \mu_0$, a więc gdy prawdziwa jest hipoteza zerowa. Zakreskowane pole odpowiada prawdopodobieństwu popełnienia błędu I rodzaju. Kolejne części (B–D) rys. 5.4 odpowiadają sytuacjom, gdy prawdziwa jest hipoteza alternatywna. W tych częściach rysunku zakreskowane pola odpowiadają więc prawdopodobieństwom popełnienia błędu II rodzaju, jeśli bowiem $\Delta\mu > 0$, to:

$$P(\bar{x}_n < \bar{x}_{n,g} \mid \mu = \mu_0 + \Delta\mu) = \beta. \quad (5.42)$$

Jak łatwo zauważyć, prawdopodobieństwo to maleje ze wzrostem $\Delta\mu$ i osiąga poziom możliwy do zaakceptowania (β') dopiero przy dostatecznie dużej wartości $\Delta\mu'$ (rys. 5.4D). Jeśli wartość $\Delta\mu'$ jest określona *explicite* (na podstawie przesłanek pozastatystycznych), to hipotezę alternatywną można zapisać następująco:

$$H_1: \mu = \mu_0 + \Delta\mu'. \quad (5.43)$$



Rys. 5.4. Prawdopodobieństwo błędu II rodzaju

W konsekwencji mamy:

$$P(\bar{x}_n < \bar{x}_{n,g} \mid \mu = \mu_0 + \Delta\mu') = \beta', \quad (5.44)$$

$$P(\bar{x}_n \geq \bar{x}_{n,g} \mid \mu = \mu_0 + \Delta\mu') = 1 - \beta'. \quad (5.45)$$

Traktując łącznie prawdopodobieństwo (5.38) i (5.39) oraz (5.44) i (5.45) stwierdzamy, że:

– jeśli prawdziwa jest hipoteza zerowa $H_0: \mu = \mu_0$, to zdarzenie losowe $\bar{x}_n < \bar{x}_{n,g}$ realizuje się z prawdopodobieństwem $1 - \alpha$, natomiast zdarzenie losowe $\bar{x}_n \geq \bar{x}_{n,g}$ realizuje się z prawdopodobieństwem α ,

– jeśli prawdziwa jest hipoteza alternatywna $H_0: \mu_0 + \Delta\mu'$, to zdarzenie losowe $\bar{x}_n < \bar{x}_{n,g}$ realizuje się z prawdopodobieństwem β' , natomiast zdarzenie losowe $\bar{x}_n \geq \bar{x}_{n,g}$ realizuje się z prawdopodobieństwem $1 - \beta'$.

Pozwala to sformułować następujące reguły decyzyjne:

– jeśli spełniona jest nierówność $\bar{x}_n \geq \bar{x}_{n,g}$, to należy przyjąć hipotezę alternatywną $H_1: \mu = \mu_0 + \Delta\mu'$,

– jeśli natomiast zachodzi nierówność $\bar{x}_n < \bar{x}_{n,g}$, to należy przyjąć hipotezę zerową $H_0: \mu = \mu_0$.

Dla każdej liczności próbki (n) wielkości β i $\Delta\mu$ powiązane są poprzez funkcję mocy testu (zob. rys. 5.3). Dlatego też jeśli wartości $\Delta\mu'$ i β' są ustalone, to pojawia się problem wyznaczenia takiej minimalnej liczności próbki n' , przy której są spełnione warunki (5.44) i (5.45). Tę minimalną licznosc próbki można wyznaczyć rozwiązując względem n' następujący układ równań:

$$\left. \begin{aligned} \bar{x}_{n,g} &= \mu_0 + u_\alpha \frac{\sigma}{\sqrt{n'}} \\ \bar{x}_{n,g} &= (\mu_0 + \Delta\mu') - u_\beta \frac{\sigma}{\sqrt{n'}} \end{aligned} \right\}, \quad (5.46)$$

gdzie u'_β jest taką wartością zmienną losową U , że:

$$P(U > u'_\beta) = \beta'. \quad (5.47)$$

Mamy więc

$$u_\alpha \frac{\sigma}{\sqrt{n'}} = \Delta\mu' - u'_\beta \frac{\sigma}{\sqrt{n'}},$$

$$\Delta\mu' \sqrt{n'} = \sigma(u_\alpha + u'_\beta)$$

i w konsekwencji

$$n' = \frac{\sigma^2 (u_\alpha + u'_\beta)^2}{(\Delta\mu')^2}. \quad (5.48)$$

Przedstawione powyżej rozumowanie można powtórzyć – bez żadnych istotnych zmian – zarówno w przypadku testu lewostronnego, jak i dwustronnego.

5.2.1.2. Test t

Jeśli wariancja (σ^2) obserwowanej zmiennej losowej $X \sim N(\mu, \sigma)$ nie jest znana, to do weryfikacji hipotez zerowych (5.18)–(5.20) wykorzystuje się test t -Studenta. W przypadku gdy wariancję σ^2 szacuje się za pomocą statystyki \underline{s}^2 , sprawdzian hipotezy zerowej ma postać:

$$t_{r,0} = \frac{\bar{x}_n - \mu_0}{\underline{s}} \sqrt{n-1}. \quad (5.49)$$

Jeśli natomiast wariancję σ^2 szacuje się za pomocą statystyki S^2 , to wówczas:

$$t_{r,0} = \frac{\bar{x}_n - \mu_0}{s} \sqrt{n}. \quad (5.50)$$

Konstrukcja obu tych funkcji testowych oparta jest na fakcie, że zmienna losowa:

$$t_r = \frac{\bar{X}_n - \mu}{\underline{S}} \sqrt{n-1} = \frac{\bar{X}_n - \mu}{S} \sqrt{n}$$

o wartościach

$$t_r = \frac{\bar{x}_n - \mu}{\underline{s}} \sqrt{n-1} = \frac{\bar{x}_n - \mu}{s} \sqrt{n}$$

ma rozkład Studenta o $r = n - 1$ stopniach swobody (zob. rozdział 4). Tak więc jeśli $\mu = \mu_0$, to wartości $t_{r,0}$ obliczone według wzorów (5.49) i (5.50) są realizacjami zmiennej losowej Studenta o $r = n - 1$ stopniach swobody.

W przypadku testu dwustronnego obszar krytyczny ma postać:

$$D_k = (-\infty; -t_{r,\alpha/2}] \cup [t_{r,\alpha/2}; \infty), \quad (5.51)$$

przy czym wartości $-t_{r,\alpha/2} < 0$ i $t_{r,\alpha/2} > 0$ wyznacza się w taki sposób, że:

$$P(t_r \leq -t_{r,\alpha/2}) = \frac{\alpha}{2}, \quad (5.52)$$

$$P(t_r \geq t_{r,\alpha/2}) = \frac{\alpha}{2}, \quad (5.53)$$

gdzie α jest poziomem istotności testu.

Ponieważ $|-t_{r,\alpha/2}| = |t_{r,\alpha/2}|$, zatem w przypadku testu dwustronnego mamy następujące reguły decyzyjne:

– jeśli wartość $t_{r,0}$ obliczona według wzoru (5.49) albo (5.50) spełnia nierówność:

$$|t_{r,0}| \geq t_{r,\alpha/2}, \quad (5.54)$$

to odrzuca się hipotezę zerową na korzyść hipotezy alternatywnej,

– jeśli natomiast:

$$|t_{r,0}| < t_{r,\alpha/2}, \quad (5.55)$$

to nie ma podstaw do odrzucenia weryfikowanej hipotezy zerowej.

W przypadku testu lewostronnego mamy:

$$D_k = (-\infty; -t_{r,\alpha}] \quad (5.56)$$

przy czym

$$P(t_r \leq -t_{r,\alpha}) = \alpha, \quad (5.57)$$

gdzie: $-t_{r,\alpha} < 0$.

Jeśli natomiast test jest prawostronny, to:

$$D_k = [t_{r,\alpha}; \infty) \quad (5.58)$$

przy czym

$$P(t_r \geq t_{r,\alpha}) = \alpha, \quad (5.59)$$

gdzie: $t_{r,\alpha} > 0$.

Tak więc jeśli wartość t_r obliczona według wzoru (5.49) albo (5.50) spełnia nierówność:

$$|t_{r,0}| \geq t_{r,\alpha}, \quad (5.60)$$

to odrzuca się hipotezę zerową na korzyść hipotezy alternatywnej. Jeśli natomiast:

$$|t_{r,0}| < t_{r,\alpha}, \quad (5.61)$$

to nie ma podstaw do odrzucenia weryfikowanej hipotezy zerowej.

Wartości krytyczne $t_{r,\alpha}$ i $t_{r,\alpha/2}$ należy odczytać z tablicy II, zamieszczonej w aneksie.

5.2.2. Weryfikacja istotności różnicy między wartościami oczekiwanymi dwóch zmiennych losowych

Niech X_1 i X_2 będą normalnymi zmiennymi losowymi o nieznanymi wartościach oczekiwanych μ_1 i μ_2 oraz wariancjach σ_1^2 i σ_2^2 , które mogą być znane lub nieznanne.

Weryfikacji poddawana jest najczęściej hipoteza zerowa:

$$H_0: \mu_1 = \mu_2 \quad (5.62)$$

wobec hipotezy alternatywnej:

$$H_1: \mu_1 \neq \mu_2. \quad (5.63)$$

W takiej sytuacji mamy do czynienia z testem dwustronnym. Można oczywiście stosować również testy jednostronne, jeśli wymaga tego rozważana sytuacja doświadczalna. Weryfikuje się wówczas hipotezę zerową (5.62) wobec hipotezy alternatywnej $H_1: \mu_1 < \mu_2$ albo $H_1: \mu_1 > \mu_2$. Hipoteza zerowa może być również hipotezą złożoną postaci $H_0: \mu_1 \leq \mu_2$ albo $H_0: \mu_1 \geq \mu_2$. Dalsze rozważania ograni-

czymy do przypadku, gdy hipoteza zerowa ma postać (5.62) i gdy jest ona weryfikowana wobec hipotezy alternatywnej (5.63).

W celu sprawdzenia hipotezy zerowej, ze zbiorowości, w których określone są zmienne losowe X_1 i X_2 , pobiera się proste próby losowe o liczebnościach n_1 i n_2 . Jeśli wariancje σ_1^2 i σ_2^2 są znane, to na podstawie zbiorów realizacji \mathbf{X}_1 i \mathbf{X}_2 wyznacza się średnie arytmetyczne \bar{x}_1 i \bar{x}_2 , a następnie – za pomocą testu u – rozstrzyga się, czy zaobserwowana wartość różnicy $\bar{x}_1 - \bar{x}_2$ świadczy na korzyść hipotezy zerowej, czy też na korzyść hipotezy alternatywnej. Jeśli wariancje σ_1^2 i σ_2^2 nie są znane, to na podstawie zbiorów realizacji \mathbf{X}_1 i \mathbf{X}_2 wyznacza się dodatkowo wartości wariancji z próby, a o istotności różnicy $\bar{x}_1 - \bar{x}_2$ rozstrzyga się za pomocą testu t . Do szacowania wariancji σ_1^2 i σ_2^2 można wykorzystać zarówno estymator \underline{S}^2 , jak i estymator S^2 .

5.2.2.1. Test u

Przypomnijmy, że – przy przyjętych powyżej założeniach – zmienna losowa:

$$U = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

o wartościach

$$u = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

ma rozkład $N(0; 1)$. Wynika stąd bezpośrednio, że jeśli prawdziwa jest hipoteza (5.62), to uzyskana empirycznie wartość:

$$u_0 = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \quad (5.64)$$

jest realizacją zmiennej losowej $U \sim N(0; 1)$.

Statystykę postaci (5.64) można więc wykorzystać jako sprawdzian hipotezy zerowej (5.62). Jeśli jest ona weryfikowana wobec hipotezy alternatywnej (5.63), to obszar krytyczny ma postać (5.26). Reguły decyzyjne są takie same jak w punkcie 5.2.1.1 (zob. wzory (5.33) i (5.34)).

5.2.2.2. Test t

Zwróćmy uwagę na przypadek, gdy do oszacowania wariancji σ_1^2 i σ_2^2 wykorzystuje się statystykę S^2 , o wartościach wyznaczonych według wzorów (4.27). Sprawdzian hipotezy zerowej (5.62) ma w takiej sytuacji postać:

$$t_{r,0} = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{n_1 + n_2}{n_1 n_2}}} = \frac{\bar{x}_1 - \bar{x}_2}{s_p} \sqrt{\frac{n_1 n_2}{n_1 + n_2}}, \quad (5.65)$$

przy czym s_p wynika ze wzoru (4.80):

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)}}.$$

Jeśli $n_1 = n_2 = n$, to:

$$t_{r,0} = \frac{\bar{x}_1 - \bar{x}_2}{s_p} \sqrt{\frac{n}{2}}, \quad (5.66)$$

$$s_p = \sqrt{\frac{s_1^2 + s_2^2}{2}}. \quad (5.67)$$

Wartości $t_{r,0}$ uzyskiwane według wzorów (5.65) i (5.66) są realizacjami zmiennej losowej t -Studenta o $r = (n_1 - 1) + (n_2 - 1) = n_1 + n_2 - 2$ stopniach swobody. Jeśli hipoteza alternatywna ma postać (5.63), to obszar krytyczny dla wartości $t_{r,0}$ wyraża się wzorem (5.51), a reguły decyzyjne są takie same jak w punkcie 5.2.1.2; zob. wzory (5.54) i (5.55).

Występująca we wzorach (5.65) i (5.66) wartość s_p – czyli tzw. uogólnione odchylenie standardowe z próbki – może być obliczana tylko wówczas, gdy nie ma podstaw do odrzucenia hipotezy o równości wariancji σ_1^2 i σ_2^2 (zob. punkt 5.3.2). Stanowi to istotne ograniczenie pola zastosowań testu t w postaci (5.65), (5.66).

Jeśli nie można utrzymać hipotezy o równości wariancji σ_1^2 i σ_2^2 , to jedną z możliwych dróg postępowania jest zastosowanie techniki tzw. połączonych obserwacji². To połączenie obserwacji polega w istocie na przekształceniu dwóch jednowymiarowych zmiennych losowych X_1 i X_2 w dwuwymiarową zmienną losową (X_1, X_2) . Każda para realizacji $(x_{1,i}, x_{2,i})$ powinna bowiem dotyczyć tego samego elementu A_i próby \underline{A} . W takiej sytuacji definiuje się zmienną losową:

$$Y = X_1 - X_2 \quad (5.68)$$

o realizacjach:

$$y_i = x_{1,i} - x_{2,i} \quad (5.69)$$

$$i = 1, 2, \dots, n.$$

Hipotezę zerową (5.62) zastępuje się równoważną hipotezą:

$$H_0: \mu(Y) = 0, \quad (5.70)$$

którą weryfikuje się wobec hipotezy alternatywnej:

$$H_1: \mu(Y) \neq 0. \quad (5.71)$$

Sprawdzian hipotezy (5.70) ma postać:

$$t_{r,0} = \frac{\bar{y}_n}{s(Y)} \sqrt{n} = \frac{\bar{y}_n}{\underline{s}(Y)} \sqrt{n-1}, \quad (5.72)$$

gdzie \bar{y}_n jest średnią arytmetyczną z próby, natomiast $s(Y)$ i $\underline{s}(Y)$ są oszacowaniami odchylenia standardowego zmiennej losowej Y , uzyskanymi za pomocą estymatora S albo \underline{S} .

Wartości $t_{r,0}$ uzyskiwane według wzoru (5.72) są realizacjami zmiennej losowej t -Studenta o $r = n - 1$ stopniach swobody. Obszar krytyczny ma postać (5.51), a reguły decyzyjne są takie same jak w punkcie 5.2.1.2; zob. wzory (5.54) i (5.55).

² Postępowanie takie jest możliwe tylko w odniesieniu do zdarzeniowych zbiorowości generalnych, i to pod warunkiem, że metody badania obu cech są nieniszczące. Ogólniejszym sposobem postępowania jest zastosowanie testu Aspin i Welcha. Stosowanie tego testu omówiono – między innymi – w książce: J.B. Czermiński, A. Iwasiewicz, Z. Paszek, A. Sikorski, *Metody statystyczne dla chemików*, PWN, Warszawa 1992, s. 130 i 131.

Przykład 5.2

W czasie kontroli długości sworzni metalowych (por. przykład 5.1) każdy sworznień mierzono dwukrotnie, za pomocą dwóch różnych przyrządów pomiarowych. Zmierzono w ten sposób $n = 10$ sworzni. Uzyskane wyniki zestawiono w kolumnach 1, 2, 3 tabeli 5.1. Celem badania było rozstrzygnięcie, czy obydwa przyrządy pomiarowe mają taką samą dokładność.

Definiujemy następujące zmienne losowe:

L_1 – długość sworzni mierzona za pomocą przyrządu nr 1,

L_2 – długość sworzni mierzona za pomocą przyrządu nr 2.

Jeśli dokładność obu przyrządów pomiarowych byłaby taka sama, to wartości oczekiwane obu zmiennych powinny być takie same: $\mu(L_1) = \mu(L_2)$. Ponieważ każda para pomiarów dotyczy tego samego sworzni, zatem istnieje możliwość zastosowania techniki połączonych obserwacji. Definiujemy zmienną losową $Y = L_1 - L_2$. Jeśli $\mu(L_1) = \mu(L_2)$, to $\mu(Y) = 0$. Hipoteza zerowa ma więc postać $H_0: \mu(Y) = 0$, natomiast hipoteza alternatywna przedstawia się następująco: $H_1: \mu(Y) \neq 0$.

Wyniki pośrednich obliczeń, niezbędnych w celu wyznaczenia średniej arytmetycznej \bar{y}_n oraz odchylenia standardowego $s(Y)$, zestawiono w kolumnach 4 i 5 tabeli 5.1.

Tabela 5.1

Obliczenia pomocnicze do przykładu 5.2

Nr sworzni i	Wynik pomiaru za pomocą przyrządu nr		$y_i = l_{1,i} - l_{2,i}$	y_i^2
	1	2		
	$l_{1,i}$	$l_{2,i}$		
1	2	3	4	5
1	102,7	101,9	0,8	0,64
2	101,8	102,4	-0,6	0,36
3	100,4	99,9	0,5	0,25
4	100,9	101,1	-0,2	0,04
5	99,8	100,4	-0,6	0,36
6	101,1	99,8	1,3	1,69
7	98,9	98,1	0,8	0,64
8	102,3	101,8	0,5	0,25
9	99,8	100,9	1,1	1,21
10	105,5	101,3	-0,8	0,64
Σ	-	-	0,6	6,08

Źródło: badania własne.

Wykorzystując wzory (4.19) i (4.27b) otrzymujemy:

$$y_n = \frac{1}{n} \sum_{i=1}^n y_i = \frac{1}{10} 0,6 = 0,06$$

$$s(Y) = \sqrt{\frac{1}{n-1} \left[\sum_{i=1}^n y_i^2 - \frac{1}{n} \left(\sum_{i=1}^n y_i \right)^2 \right]} = \sqrt{\frac{1}{9} (6,08 - \frac{1}{10} 0,36)} \approx 0,82.$$

Podstawiając te wartości do wzoru (5.61) otrzymujemy:

$$t_{9,0} = \frac{0,06}{0,82} \sqrt{10} = 0,2314.$$

Przyjmujemy poziom istotności testu $\alpha = 0,05$. W tablicy II (zamieszczonej w aneksie) znajdujemy wartość $t_{9,\alpha/2} = t_{9,0,025} = 2,262$. Zachodzi nierówność $t_9 = 0,2314 < t_{9,0,025} = 2,262$. Nie ma więc podstaw do odrzucenia hipotezy o jednakowej dokładności obu przyrządów pomiarowych. ■

5.3. Weryfikacja hipotez dotyczących wariancji

5.3.1. Weryfikacja istotności różnicy między wariancją a ustaloną wartością

Niech X oznacza normalną zmienną losową o nieznanymi parametrach μ i σ^2 . Ze zbiorowości generalnej \mathcal{A} , w której określona jest zmienna losowa X , pobiera się prostą próbę losową \underline{A} . Na podstawie zbioru realizacji X_n szacuje się wariancję σ^2 . Można do tego celu wykorzystać statystykę \underline{S}^2 albo S^2 . Mając wartości tych charakterystyk (\underline{s}^2 albo s^2), można poddać weryfikacji hipotezy dotyczące relacji między wariancją σ^2 a pewną wartością σ_0^2 , ustaloną najczęściej na podstawie przesłanek pozastatystycznych.

Weryfikacji poddaje się jedną z następujących hipotez zerowych:

$$H_0: \sigma^2 = \sigma_0^2, \quad (5.73)$$

$$H_0: \sigma^2 \leq \sigma_0^2, \quad (5.74)$$

$$H_0: \sigma^2 \geq \sigma_0^2. \quad (5.75)$$

Hipoteza (5.73) jest hipotezą prostą, jednoznacznie bowiem specyfikuje postać funkcyjną dystrybuanty zmiennej losowej X . Hipotezy (5.74) i (5.75) są natomiast hipotezami złożonymi.

Hipotezy alternatywne formułuje się następująco:

$$H_1: \sigma^2 \neq \sigma_0^2, \quad (5.76)$$

$$H_1: \sigma^2 < \sigma_0^2, \quad (5.77)$$

$$H_1: \sigma^2 > \sigma_0^2. \quad (5.78)$$

Każda z tych hipotez może być alternatywą hipotezy zerowej (5.73). Jeśli natomiast hipoteza zerowa ma postać (5.74), to alternatywą jest hipoteza (5.78). Jest to przypadek szczególnie często występujący w statystycznej kontroli jakości, gdzie wartość σ_0^2 traktuje się jako miarę najwyższego dopuszczalnego poziomu zróżnicowania zasobu albo strumienia produkowanego masowo wyrobu. Im mniejsza jest rzeczywista wartość wariancji σ^2 , tym wyższa jest jakość wykonania. Jeśli hipoteza zerowa ma postać (5.75), to alternatywą jest hipoteza (5.77).

Przypomnijmy (zob. rozdział 4), że jeśli $X \sim N(\mu, \sigma)$, to statystyka:

$$\frac{\sum_{i=1}^n (x_i - \bar{x}_n)^2}{\sigma^2} \quad (5.79)$$

jest zmienną losową o rozkładzie chi-kwadrat, o $r = n - 1$ stopniach swobody. Wykorzystując definicyjne wzory statystyki \underline{S}^2 i S^2 możemy napisać:

$$\sum_{i=1}^n (x_i - \bar{x}_n)^2 = n\underline{s}^2 = (n-1)s^2. \quad (5.80)$$

Wzór (5.79) można więc zapisać dwojako:

$$\frac{n\underline{s}^2}{\sigma^2} \quad (5.81)$$

albo

$$\frac{(n-1)s^2}{\sigma^2}. \quad (5.82)$$

Podstawiając $\sigma^2 = \sigma_0^2$ otrzymujemy:

$$\chi_{r,0}^2 = \frac{ns^2}{\sigma_0^2}, \quad (5.83)$$

$$\chi_{r,0}^2 = \frac{(n-1)s^2}{\sigma_0^2}. \quad (5.84)$$

Każda z tych charakterystyk może być wykorzystana jako sprawdzian hipotez zerowych (5.73)–(5.75). W przypadku gdy sprawdzana jest hipoteza zerowa (5.73) wobec hipotezy alternatywnej (5.76), obszar krytyczny dla uzyskiwanych empirycznie wartości $\chi_{r,0}^2$ ma postać:

$$\mathbf{D}_k = (0; \chi_{r,1-\alpha/2}^2] \cup [\chi_{r,\alpha/2}^2; \infty). \quad (5.85)$$

Reguły decyzyjne:

– jeśli wartość $\chi_{r,0}^2$ – obliczona według wzoru (5.83) albo (5.84) – spełnia nierówność:

$$\chi_{r,0}^2 \leq \chi_{r,1-\alpha/2}^2 \quad (5.86)$$

albo nierówność:

$$\chi_{r,0}^2 \geq \chi_{r,\alpha/2}^2, \quad (5.87)$$

to odrzuca się hipotezę $H_1: \sigma^2 = \sigma_0^2$ na korzyść hipotezy $H_1: \sigma^2 \neq \sigma_0^2$,

– jeśli natomiast uzyskana empirycznie wartość χ_r^2 spełnia nierówność:

$$\chi_{r,1-\alpha/2}^2 < \chi_{r,0}^2 < \chi_{r,\alpha/2}^2, \quad (5.88)$$

to nie ma podstaw do odrzucenia hipotezy $H_1: \sigma^2 = \sigma_0^2$.

Jeśli weryfikacji podlega hipoteza zerowa $H_1: \sigma^2 = \sigma_0^2$ albo $H_1: \sigma^2 \leq \sigma_0^2$ wobec hipotezy alternatywnej $H_1: \sigma^2 > \sigma_0^2$, to obszar krytyczny ma postać:

$$\mathbf{D}_k = [\chi_{r,\alpha}^2 ; \infty) \quad (5.89)$$

Reguły decyzyjne:

– jeśli obliczona wartość $\chi_{r,0}^2$ spełnia nierówność:

$$\chi_{r,0}^2 \geq \chi_{r,\alpha}^2, \quad (5.90)$$

to odrzuca się weryfikowaną hipotezę zerową na korzyść hipotezy alternatywnej,

– jeśli natomiast:

$$\chi_{r,0}^2 < \chi_{r,\alpha}^2, \quad (5.91)$$

to nie ma podstaw do odrzucenia hipotezy zerowej.

W przypadku gdy weryfikacji poddawana jest hipoteza $H_1: \sigma^2 = \sigma_0^2$ albo $H_1: \sigma^2 \geq \sigma_0^2$ wobec hipotezy alternatywnej $H_1: \sigma^2 < \sigma_0^2$, obszar krytyczny ma postać:

$$\mathbf{D}_k = (0; \chi_{r,1-\alpha}^2]. \quad (5.92)$$

Reguły decyzyjne:

– jeśli obliczona wartość $\chi_{r,0}^2$ spełnia nierówność:

$$\chi_{r,0}^2 \leq \chi_{r,1-\alpha}^2, \quad (5.93)$$

to odrzuca się hipotezę zerową na korzyść hipotezy alternatywnej,

– jeśli natomiast:

$$\chi_{r,0}^2 > \chi_{r,1-\alpha}^2, \quad (5.94)$$

to nie ma podstaw do odrzucenia hipotezy zerowej.

We wszystkich wymienionych przypadkach poziom istotności testu wynosi α . Wartości krytyczne należy odczytać z tablicy III, zamieszczonej w aneksie, dla $r = n - 1$ stopni swobody. Należy przy tym przyjąć – w zależności od konkretnej sytuacji – ε na poziomie $1 - \alpha/2$, $1 - \alpha$, $\alpha/2$ albo α .

Przykład 5.3

Konstruktor linii technologicznej twierdzi, że największe odchylenia długości (L) produkowanych elementów od wartości nominalnej (l_0) nie powinny przekraczać 12 mm (co do wartości bezwzględnej). Wiadomo przy tym, że rozkład zmiennej losowej L jest normalny.

W celu sprawdzenia twierdzenia konstruktora z bieżącej produkcji pobrano próbkę losową o liczebności $n = 20$ i oszacowano wariancję zmiennej losowej L . Wykorzystano do tego celu estymator S^2 i uzyskano $s^2 = 28,5 \text{ mm}^2$. Czy uzyskany wynik pozwala zaakceptować twierdzenie konstruktora?

Ponieważ mamy do czynienia z rozkładem normalnym, można przyjąć, że maksymalne odchylenie deklarowane przez konstruktora jest 3-krotnym odchyleniem standardowym w hipotetycznym rozkładzie obserwowanej zmiennej losowej L . Postępując w ten sposób dopuszczamy wprawdzie możliwość pojawienia się odchyłeń przekraczających 12 mm, ale tylko z prawdopodobieństwem 0,0027 (zob. rozdział 1). Hipotetyczne odchylenie standardowe wynosi więc $\sigma_0 = 12/3 = 4 \text{ mm}$, a stąd $\sigma_0^2 = 16 \text{ mm}^2$. Hipoteza zerowa ma postać $H_0: \sigma^2 \leq 16$ i jest weryfikowana przy alternatywie $H_1: \sigma^2 > 16$. Przyjmujemy poziom istotności $\alpha = 0,01$. Po podstawieniu wartości do wzoru (5.84) otrzymujemy:

$$\chi_{r,0}^2 = \frac{(n-1)s^2}{\sigma_0^2} = \chi_{19,0}^2 = \frac{19 \cdot 28,5}{16} = 33,8438.$$

Z zamieszczonej w aneksie tablicy III odczytujemy wartość krytyczną $\chi_{19;0,01}^2 = 36,191$.

Ponieważ obliczona wartość $\chi_{19,0}^2$ jest mniejsza od wartości krytycznej $\chi_{19;0,01}^2$, zatem nie ma podstaw do odrzucenia hipotezy zerowej $H_0: \sigma^2 \leq 16$, na poziomie istotności $\alpha = 0,01$. Nie ma więc również podstaw do odrzucenia twierdzenia konstruktora linii technologicznej. ■

Omówiony powyżej test chi-kwadrat jest podstawowym narzędziem sprawdzania hipotez dotyczących relacji między wariancją normalnej zmiennej losowej a ustaloną wartością. Nie jest to jednak narzędzie jedyne. Jeśli próbka jest liczna, to do weryfikacji hipotez (5.73)–(5.75) można zastosować test u . Przypomnijmy (zob. rozdział 4), że jeśli obserwowana zmienna losowa ma normalny rozkład prawdopodobieństwa, to statystyki $\underline{\underline{S}}^2$, \underline{S}^2 i S^2 mają asymptotyczne rozkłady normalne o parametrach (4.53)–(4.55). Wykorzystując ten fakt można

skonstruować testy u , funkcjonujące analogicznie jak testy u stosowane do weryfikacji hipotez dotyczących wartości oczekiwanej (zob. punkt 5.2).

5.3.2. Weryfikacja istotności różnicy między dwiema wariancjami

Niech X_1 i X_2 będą normalnymi zmiennymi losowymi o nieznanach wariancjach σ_1^2 i σ_2^2 . Zwróćmy uwagę na przypadek, gdy do estymacji wartości σ_1^2 i σ_2^2 stosowana jest statystyka S^2 i gdy mając wartości s_1^2 i s_2^2 chcemy rozstrzygnąć, czy wariancje obu zmiennych są równe, czy różne. W takiej sytuacji weryfikacji poddaje się hipotezę zerową:

$$H_0: \sigma_1^2 = \sigma_2^2 \quad (5.95)$$

wobec hipotezy alternatywnej

$$H_1: \sigma_1^2 \neq \sigma_2^2. \quad (5.96)$$

Zauważmy, że jeśli $\sigma_1^2 = \sigma_2^2$, to statystyka (4.90) przyjmuje postać:

$$F_{r_1, r_2, 0} = \frac{\frac{S_1^2}{\sigma_1^2}}{\frac{S_2^2}{\sigma_2^2}} = \frac{S_1^2}{S_2^2}. \quad (5.97)$$

Statystyka ta ma rozkład F -Snedecora, przy czym szczegółowa postać dystrybuanty zmiennej losowej F_{r_1, r_2} zależy od wartości $r_1 = n_1 - 1$; $r_2 = n_2 - 1$, czyli od liczby stopni swobody przy szacowaniu wariancji σ_1^2 i σ_2^2 .

Jeśli poziom istotności testu (α) jest ustalony, to w oparciu o dystrybuantę zmiennej losowej F_{r_1, r_2} można wyznaczyć kwantyle $F_{r_1, r_2, 1-\alpha/2}$ i $F_{r_1, r_2, \alpha/2}$, takie że:

$$P(F_{r_1, r_2} \leq F_{r_1, r_2, 1-\alpha/2}) = \frac{\alpha}{2}, \quad (5.98)$$

$$P(F_{r_1, r_2} \geq F_{r_1, r_2, \alpha/2}) = \frac{\alpha}{2}, \quad (5.99)$$

przy czym $F_{r_1, r_2, 1-\alpha/2} < 1$ i dotyczy to przypadku, gdy $s_1^2 < s_2^2$, natomiast $F_{r_1, r_2, \alpha/2} > 1$ i odnosi się do przypadku, gdy $s_1^2 > s_2^2$.

Obszar krytyczny ma więc postać:

$$\mathbf{D}_k = (0; F_{r_1, r_2, 1-\alpha/2}] \cup [F_{r_1, r_2, \alpha/2}; \infty), \quad (5.100)$$

przy czym

$$F_{r_1, r_2, 1-\alpha/2} = \frac{1}{F_{r_2, r_1, \alpha/2}}.$$

Statystykę (5.97) – o wartościach w przedziale $(0; \infty)$ – można zastąpić statystyką:

$$F_{r_1, r_2, 0} = \frac{S_1^2}{S_2^2}; \quad S_1^2 > S_2^2 \quad (5.97a)$$

o wartościach w przedziale $[1; \infty)$. Statystyka ta jest sprawdzianem hipotezy zerowej (5.95) i nosi nazwę testu F .

Numeracja zmiennych (X_1, X_2) jest tu wymuszona przez nierówność $s_1^2 > s_2^2$. Obszar krytyczny dla statystyki (5.97a) ma więc postać:

$$\mathbf{D}_k = [F_{r_1, r_2, \alpha/2}; \infty), \quad (5.101)$$

przy czym poziom istotności testu wynosi α . Wartości kwantyli $F_{r_1, r_2, \varepsilon}$, dla $\varepsilon = 0,01$ i $\varepsilon = 0,05$, zestawione są w tablicach IV/1 i IV/2 zamieszczonych w aneksie. Przy odczytywaniu wartości krytycznych z tych tablic należy przyjmować $\alpha/2 = \varepsilon$.

Reguły decyzyjne:

– jeśli wartość $F_{r_1, r_2, 0}$ obliczona według wzoru (5.97a) spełnia nierówność:

$$F_{r_1, r_2, 0} \geq F_{r_1, r_2, \alpha/2}, \quad (5.102)$$

to należy odrzucić hipotezę $H_0: \sigma_1^2 = \sigma_2^2$ na korzyść hipotezy alternatywnej $H_1: \sigma_1^2 \neq \sigma_2^2$, a poziom istotności takiej decyzji wynosi α ,

– jeśli natomiast:

$$F_{r_1, r_2, 0} < F_{r_1, r_2, \alpha/2}, \quad (5.103)$$

to nie ma podstaw do odrzucenia weryfikowanej hipotezy zerowej.

Przy porównywaniu wariancji σ_1^2 i σ_2^2 można także stosować jednostronny test F . W takiej sytuacji weryfikacji poddaje się hipotezę zerową postaci:

$$H_0: \sigma_1^2 = c \sigma_2^2 \quad (5.104)$$

albo postaci

$$H_0: \sigma_1^2 \leq c \sigma_2^2 \quad (5.105)$$

wobec hipotezy alternatywnej

$$H_1: \sigma_1^2 > c \sigma_2^2, \quad (5.106)$$

gdzie $c > 0$, przy czym σ_1^2 oznacza tę wariancję, której odpowiada większa wariancja próbkowa (s_1^2), natomiast σ_2^2 jest tą wariancją, której odpowiada mniejsza wariancja próbkowa (s_2^2). Numeracja zmiennych jest tu więc wymuszona – analogicznie jak powyżej – przez nierówność $s_1^2 > s_2^2$.

Sprawdzianem hipotezy zerowej (5.104) albo (5.105) jest statystyka (5.98). Przy wyborze wartości krytycznych z tablic IV/1 i IV/2 należy przyjmować $\alpha = \varepsilon$ (zob. aneks). Obszar krytyczny dla uzyskiwanych empirycznie wartości $F_{r_1, r_2, 0}$ ma więc postać:

$$D_k = [c F_{r_1, r_2, \alpha}; \infty). \quad (5.107)$$

Reguły decyzyjne:

– jeśli obliczona wartość $F_{r_1, r_2, 0}$ spełnia nierówność:

$$F_{r_1, r_2, 0} \geq c F_{r_1, r_2, \alpha}, \quad (5.108)$$

to odrzuca się hipotezę zerową $H_0: \sigma_1^2 = c \sigma_2^2$ albo $H_0: \sigma_1^2 \leq c \sigma_2^2$ na korzyść hipotezy alternatywnej $H_1: \sigma_1^2 > c \sigma_2^2$, a poziom istotności takiej decyzji wynosi α ,

– jeśli natomiast:

$$F_{r_1, r_2, 0} < c F_{r_1, r_2, \alpha}, \quad (5.109)$$

to nie ma podstaw do odrzucenia weryfikowanej hipotezy zerowej.

Przyjmując $c = 1$ otrzymujemy standardową (podstawową) postać testu, najczęściej wykorzystywaną w praktyce.

5.4. Weryfikacja hipotez dotyczących wskaźnika struktury

Hipotezy dotyczące wskaźnika struktury (p) weryfikuje się najczęściej za pomocą testów u . Wykorzystuje się przy tym fakt, że częstość względna z próby (W) jest zmienną losową o asymptotycznym rozkładzie normalnym, o wartości oczekiwanej $E(W) = p$ i wariancji $D^2(W) = p(1 - p)/n$. Ponieważ do konstrukcji testu wykorzystuje się asymptotyczne własności rozkładu, zatem próbka – będąca podstawą postępowania weryfikacyjnego – powinna być odpowiednio liczna i powinna obejmować co najmniej kilkadziesiąt losowo wybranych elementów zbiorowości generalnej. Jest to konieczny warunek aproksymacji dwumianowego rozkładu zmiennej losowej W za pomocą rozkładu normalnego.

5.4.1. Weryfikacja istotności różnicy między wskaźnikiem struktury a ustaloną wartością

Niech X oznacza zero-jedynkową zmienną losową o parametrze $p = P(X = 1)$. Parametr p będziemy traktować jako wskaźnik struktury zbiorowości generalnej A , w której jest określona zmienna X . Jeśli A jest zbiorem nieskończonym, to p należy interpretować jako prawdopodobieństwo pojawienia się tych stanów badanej cechy B , którym przypisuje się wartość $x^\circ = 1$. Jeśli natomiast A jest zbiorem skończonym, to p można również interpretować jako frakcję owych wyróżnionych elementów $A \in A$. Weryfikacji (sprawdzeniu) poddaje się prostą hipotezę zerową postaci:

$$H_0: p = p_0, \quad (5.110)$$

albo też jedną z następujących hipotez złożonych:

$$H_0: p \leq p_0, \quad (5.111)$$

$$H_0: p \geq p_0, \quad (5.112)$$

gdzie p_0 oznacza pewną liczbę z przedziału $(0; 1)$, ustaloną – najczęściej – na podstawie przesłanek pozastatystycznych.

Hipotezy alternatywne przedstawiają się następująco:

$$H_1: p \neq p_0, \quad (5.113)$$

$$H_1: p < p_0, \quad (5.114)$$

$$H_1: p > p_0. \quad (5.115)$$

Każda z tych hipotez może być alternatywą hipotezy zerowej (5.110). Alternatywą hipotezy zerowej (5.111) jest (5.115), natomiast w przypadku hipotezy zerowej (5.112) alternatywą jest (5.114).

Przypomnijmy (zob. rozdział 4), że jeśli liczebność próbki (n) jest dostatecznie duża, to:

$$u = \frac{\frac{z}{n} - p}{\sqrt{\frac{\frac{z}{n} \left(1 - \frac{z}{n}\right)}{n}}} = \frac{w - p}{\sqrt{\frac{w(1-w)}{n}}}$$

jest realizacją zmiennej losowej $U \sim N(0; 1)$, przy czym:

$$w = \frac{z}{n} = \frac{1}{n} \sum_{i=1}^n x_i,$$

gdzie x_i ($i = 1, 2, \dots, n$) jest realizacją zero-jedynkowej zmiennej losowej X o parametrze p .

W konsekwencji, jeśli $p = p_0$, to również

$$u_0 = \frac{w - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \quad (5.116)$$

jest realizacją zmiennej losowej $U \sim N(0; 1)$.

Charakterystyka ta jest sprawdzianem każdej spośród wymienionych powyżej hipotez zerowych (5.110)–(5.112). Obszary krytyczne i reguły decyzyjne są takie same jak w przypadku weryfikacji istotności różnicy między wartością oczekiwaną (μ) normalnej zmiennej losowej a ustaloną wartością μ_0 (zob. punkt 5.2.1.1).

Przykład 5.4

Niech A będzie strumieniem pewnego wyrobu sztukowego A , generowanym przez agregat produkcyjny. Mamy więc do czynienia z przedmiotową zbiorowością generalną, którą formalnie należy traktować jako zbiorowość nieskończoną, przeliczalną. W zbiorowości tej określona jest zero-jedynkowa zmienna losowa X , taka że:

$$X = \begin{cases} 0 & \text{– gdy element } A_i \in A \text{ spełnia wymagania} \\ & \text{jakościowe,} \\ 1 & \text{– gdy element } A_i \in A \text{ nie spełnia wymagań} \\ & \text{jakościowych.} \end{cases}$$

W rozważanej sytuacji wskaźnik struktury p określany jest jako wadliwość strumienia wyrobu. Jest to unormowana miara jakości wykonania.

Założmy, że najwyższa dopuszczalna wadliwość obserwowanego strumienia wyrobu A wynosi $p_0 = 0,1$ (10%). Należy sprawdzić, czy kontrolowany proces produkcyjny spełnia warunek $p \leq p_0 = 0,1$. Mamy więc $H_0: p \leq 0,1$ oraz $H_1: p > 0,1$.

W celu sprawdzenia sformułowanej hipotezy zerowej z bieżącej produkcji pobrano próbkę losową \underline{A} o liczebności $n = 200$. W rezultacie przeprowadzonych badań otrzymano $z = 28$. Mamy więc $w = 28/200 = 0,14$. Czy na poziomie istotności $\alpha = 0,01$ można utrzymać hipotezę $H_0: p \leq 0,1$ głoszącą, że kontrolowany proces produkcyjny jest uregulowany? Po podstawieniu wartości do wzoru (5.116) otrzymujemy:

$$u_0 = \frac{w - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0,14 - 0,10}{\sqrt{\frac{0,10 \cdot 0,90}{200}}} = \frac{0,04}{0,0212} = 1,87.$$

Z tablicy II (zamieszczonej w aneksie) odczytujemy wartość krytyczną $u_\alpha = u_{0,01} = 2,326$. Ponieważ spełniona jest nierówność $u_0 = 1,87 < u_{0,01} = 2,326$, nie ma więc podstaw do odrzucenia weryfikowanej hipotezy zerowej. ■

5.4.2. Weryfikacja istotności różnicy między dwoma wskaźnikami struktury

Niech X_1 i X_2 będą zero-jedynkowymi zmiennymi losowymi o nieznanach parametrach p_1 i p_2 . Założmy, że zmienne te są określone w zbiorowościach generalnych A_1 i A_2 . W konsekwencji, nieznanne wartości p_1 i p_2 będziemy traktować jako wskaźniki struktury tych zbiorowości ze względu na pewną cechę B , której liczbowymi obrazami są zmienne losowe X_1 i X_2 . Oszacowaniami wartości p_1 i p_2 są częstości względne z próbek losowych, a mianowicie $w_1 = z_1/n_1$ i $w_2 = z_2/n_2$. Mając do dyspozycji te informacje można dokonać weryfikacji hipotez dotyczących relacji między wskaźnikami struktury p_1 i p_2 . Podstawowe znaczenie ma weryfikacja hipotezy zerowej:

$$H_0: p_1 = p_2 \quad (5.117)$$

wobec hipotezy alternatywnej:

$$H_1: p_1 \neq p_2. \quad (5.118)$$

Przypomnijmy (zob. rozdział 4), że jeśli obydwie próbki (\underline{A}_1 , \underline{A}_2) są dostatecznie liczne, to:

$$u = \frac{(w_1 - w_2) - (p_1 - p_2)}{\sqrt{\frac{w_1(1-w_1)}{n_1} + \frac{w_2(1-w_2)}{n_2}}}$$

jest realizacją zmiennej losowej $U \sim N(0; 1)$.

Tak więc jeśli $p_1 = p_2$, to również wartości u uzyskiwane empirycznie według wzoru:

$$u_0 = \frac{w_1 - w_2}{\sqrt{\frac{w_1(1-w_1)}{n_1} + \frac{w_2(1-w_2)}{n_2}}} \quad (5.119)$$

są realizacjami zmiennej losowej $U \sim N(0; 1)$.

Charakterystyka (5.119) jest sprawdzianem hipotezy zerowej (5.117). Jeśli hipoteza alternatywna ma postać (5.119), to mamy do czynienia z testem dwustronnym. Obszar krytyczny ma postać (5.26). Reguły decyzyjne są takie same jak w przypadku weryfikacji hipotezy zerowej $H_0: \mu = \mu_0$ wobec hipotezy alternatywnej $H_1: \mu \neq \mu_0$; zob. wzory (5.33) i (5.34).

Przykład 5.5

Do odbiorczej kontroli jakości przedstawiono dwie partie (A_1 i A_2) produktu A . Należy sprawdzić, czy wadliwości obu partii produktu (p_1 i p_2) są jednakowe. W tym celu pobrano próbki losowe \underline{A}_1 i \underline{A}_2 o licznosciach $n_1 = n_2 = 250$. W rezultacie przeprowadzonych badań stwierdzono $z_1 = 15$ i $z_2 = 50$ sztuk wadliwych. Czy można utrzymać na poziomie istotności $\alpha = 0,05$ hipotezę zerową $H_0: p_1 = p_2$, jeśli konkurencyjna hipoteza alternatywna ma postać $H_1: p_1 \neq p_2$?

Po podstawieniu wartości do wzoru (5.119) mamy:

$$\begin{aligned}
 u_0 &= \frac{\frac{25}{250} - \frac{50}{250}}{\sqrt{\frac{\frac{25}{250} \left(1 - \frac{25}{250}\right)}{250} + \frac{\frac{25}{250} \left(1 - \frac{50}{250}\right)}{250}}} = \frac{0,1 - 0,2}{\sqrt{3,6 \cdot 10^{-4} + 6,4 \cdot 10^{-4}}} = \\
 &= \frac{-0,1}{3,1623 \cdot 10^{-2}} = -3,1623.
 \end{aligned}$$

Z zamieszczonej w aneksie tablicy II odczytujemy wartość krytyczną $u_{\alpha/2} = u_{0,05/2} = u_{0,025} = 1,96$. Ponieważ $|-3,1623| = 3,1623 > 1,96$, zatem weryfikowaną hipotezę zerową $H_0: p_1 = p_2$ odrzucamy na korzyść hipotezy alternatywnej $H_0: p_1 \neq p_2$. Badane partie towaru mają różne wadliwości. ■

Hipoteza (5.117) nie jest jedyną postacią hipotezy zerowej wykorzystywaną przy porównywaniu dwóch wskaźników struktury o nieznanymi wartościami p_1 i p_2 . W pewnych sytuacjach doświadczalnych bardziej racjonalnym sposobem postępowania jest formułowanie złożonych hipotez zerowych postaci:

$$H_0: p_1 \leq p_2, \quad (5.120)$$

albo

$$H_0: p_1 \geq p_2. \quad (5.121)$$

Alternatywą hipotezy zerowej (5.120) jest:

$$H_1: p_1 > p_2, \quad (5.122)$$

natomiast w przypadku hipotezy zerowej (5.121) mamy:

$$H_1: p_1 < p_2. \quad (5.123)$$

Hipotezy (5.122) i (5.123) mogą być oczywiście również alternatywami hipotezy zerowej (5.117).

W każdym z wymienionych przypadków mamy do czynienia z testem jednostronnym. Obszary krytyczne dla empirycznie wyznaczanych wartości u_0 (według wzoru (5.119)), a także odpowiednie reguły decyzyjne są takie same jak w przypadku weryfikacji hipotez dotyczących relacji między dwiema wartościami oczekiwanymi normalnych zmiennych losowych. Przypomnijmy, że obszary krytyczne mają w rozważanej sytuacji postać (5.29) albo (5.30), natomiast przy podejmowaniu decyzji w odniesieniu do hipotezy zerowej wykorzystuje się nierówności (5.35) i (5.36).

5.5. Weryfikacja istotności współczynnika korelacji

Współczynnik korelacji liniowej z próby (r_{xy}) – jak każda charakterystyka z próby – jest zmienną losową. Ma to ten skutek, iż nawet wówczas, gdy współczynnik korelacji w zbiorowości generalnej (ρ_{xy}) jest równy zeru, uzyskiwane empirycznie wartości r_{xy} odbiegają zwykle od zera, zarówno in plus, jak i in minus. Powstaje więc problem weryfikacji hipotezy:

$$H_0: \rho_{xy} = 0 \quad (5.124)$$

na podstawie informacji uzyskanych w rezultacie badania próby. Hipotezę tę weryfikuje się zwykle wobec hipotezy dwustronnej:

$$H_1: \rho_{xy} \neq 0. \quad (5.125)$$

Można też oczywiście formułować hipotezy jednostronne postaci:

$$H_1: \rho_{xy} > 0 \quad (5.126)$$

albo

$$H_1: \rho_{xy} < 0, \quad (5.127)$$

zwłaszcza jeśli istnieją przesłanki merytoryczne pozwalające wysuwać przypuszczenia w odniesieniu do znaku współczynnika korelacji ρ_{xy} .

W procesie weryfikacji hipotezy zerowej (5.124) wykorzystuje się fakt, że współczynnik korelacji liniowej z próby ma rozkład asymptotycznie normalny, o wartości oczekiwanej $E(R_{xy}) = \rho_{xy}$ i odchyleniu standardowym $D(R_{xy}) = (1 - \rho_{xy}^2)/\sqrt{n}$. Tak więc (zob. rozdział 4), jeśli próba jest liczna, to:

$$u = \frac{r_{xy} - \rho_{xy}}{\sqrt{1 - \rho_{xy}^2}} \sqrt{n}$$

jest wartością zmiennej losowej $U \sim N(0; 1)$. W konsekwencji, jeśli prawdziwa jest hipoteza zerowa (5.124), to charakterystyka:

$$t_{n-2,0} = \frac{r_{xy}}{\sqrt{1 - r_{xy}^2}} \sqrt{n-2} \quad (5.128)$$

ma rozkład t -Studenta o $n - 2$ stopniach swobody. Zastosowanie testu u nie jest tu możliwe, albowiem wariancja współczynnika korelacji szacowana jest z próby. Jeśli hipoteza alternatywna ma postać $H_0: \rho_{xy} \neq 0$, to obszar krytyczny wyraża się wzorem (5.51), natomiast reguły decyzyjne oparte są na nierównościach (5.54) i (5.55), przy czym $r = n - 2$. W przypadku testu jednostronnego reguły decyzyjne oparte są na nierównościach (5.60) i (5.61). Jeśli test jest lewostronny, to obszar krytyczny ma postać (5.56). Jeśli natomiast test jest prawostronny, to obszar krytyczny wyraża się wzorem (5.58).

Przykład 5.6

W tablicy 3.16 i na rys. 3.11 przedstawiono rezultaty badania wilgotności piasku dwiema metodami, a mianowicie metodą suszarkową oraz metodą przemycania acetonem. Wyniki tych badań potraktujemy obecnie jako próbkę losową (o liczności $n = 12$) z nieskończenie licznej, zdarzeniowej zbiorowości generalnej, obejmującej wszystkie możliwe powtórzenia doświadczenia w ustalonych warunkach. W przykładzie 3.14 uzyskaliśmy $r_{xy} = 0,9947$. Sprawdźmy obecnie hipotezę zerową $H_0: \rho_{xy} = 0$ wobec hipotezy alternatywnej $H_1: \rho_{xy} > 0$, na poziomie istotności $\alpha = 0,01$. Podstawiając wartości do wzoru (5.128) otrzymujemy:

$$t_{n-2,0} = \frac{r_{xy}}{\sqrt{1-r_{xy}^2}} \sqrt{n-2} = \frac{0,9947}{\sqrt{1-(0,9947)^2}} \sqrt{10} = 30,60 = t_{10,0}.$$

Z tablicy II zamieszczonej w aneksie odczytujemy wartość krytyczną $t_{n-2;\alpha} = t_{10;0,01} = 2,764$. Ponieważ $t_{10,0} \gg t_{10;0,01}$, odrzucamy hipotezę zerową $H_0: \rho_{xy} = 0$ na korzyść hipotezy alternatywnej $H_1: \rho_{xy} > 0$. Oznacza to, że współzależność między wynikami uzyskiwanymi obydwoma metodami nie została ukształtowana działaniem czynników losowych. ■

Za pomocą sprawdzianu postaci (5.128) można też zweryfikować istotność współczynnika korelacji dwuseryjnej ($\rho_{d,xy}$) w zbiorowości generalnej (por. punkt 3.3.2.2).

Oszacowania $\rho_{d,xy}$ uzyskuje się za pomocą wzoru:

$$r_{d,xy} = \frac{\bar{y}_1 - \bar{y}_0}{s} = \sqrt{\frac{n_1 n_0}{n(n-1)}}, \quad (5.129)$$

w którym symbole \bar{y}_0 i \bar{y}_1 mają takie samo znaczenie jak we wzorze (3.108), s jest oszacowaniem odchylenia standardowego ciągłej zmiennej losowej Y ,

n jest liczebnością próbki, natomiast symbole n_0 i n_1 oznaczają liczebności podzbiorów zer i jedynek w zbiorze realizacji zero-jedynkowej zmiennej losowej X , przy czym $n_0 + n_1 = n$.

Mając do dyspozycji wartość $r_{d.xy}$ można dokonać weryfikacji hipotez dotyczących $\rho_{d.xy}$. Podstawowe znaczenie ma hipoteza zerowa postaci (5.124), która może być weryfikowana wobec hipotez alternatywnych postaci (5.125)–(5.127). W rozważanym przypadku sprawdzian (5.128) ma postać:

$$t_{n-2,0} = r_{d.xy} \frac{\sqrt{n-2}}{\sqrt{1-r_{d.xy}^2}}. \quad (5.130)$$

Obszar krytyczny dla wartości tej charakterystyki, a także odpowiednie reguły decyzyjne, są takie same jak w przypadku weryfikacji hipotez dotyczących współczynnika korelacji liniowej ρ_{xy} .

Przykład 5.7

W tablicy 3.17 przedstawiono wyniki badania ośmiu kształtek poliamidowych ze względu na dwie zmienne losowe, z których jedna (X) jest zmienną zero-jedynkową, natomiast druga (Y) zmienną ciągłą.

Obecnie potraktujemy uzyskane wyniki jako zbiór realizacji dwuwymiarowej zmiennej losowej (X, Y) w próbie losowej o liczebności $n = 8$ i dokonamy weryfikacji istotności współczynnika korelacji dwuseryjnej $\rho_{d.xy}$. Formułujemy hipotezę zerową:

$$H_0: \rho_{d.xy} = 0$$

oraz dwustronną hipotezę alternatywną:

$$H_1: \rho_{d.xy} \neq 0.$$

Wykorzystując informacje z przykładu 3.15 otrzymujemy:

$$\begin{aligned} s &= \sqrt{\frac{1}{n-1} \left[\sum_{i=1}^n y_i^2 - \frac{1}{n} \left(\sum_{i=1}^n y_i \right)^2 \right]} = \sqrt{\frac{1}{7} \left[596,6667 - \frac{1}{8} (68,97)^2 \right]} = \\ &= \sqrt{\frac{1}{7} (596,6667 - 594,6076)} \approx 0,5424 \approx 0,54. \end{aligned}$$

Podstawiając tę wartość, a także $n = 8$, $n_0 = 4$, $n_1 = 4$, $\bar{y}_0 = 9,08$ i $\bar{y}_1 = 8,16$ do wzoru (5.129), otrzymujemy:

$$r_{d.xy} = \frac{8,16 - 9,08}{0,54} \sqrt{\frac{4 \cdot 4}{8 \cdot 7}} \approx -0,9107.$$

Podstawiamy wartości do wzoru (5.130):

$$t_{n-2,0} = -0,9107 \sqrt{\frac{8-2}{1-(-0,9107)^2}} = -5,4003 = t_{6,0}.$$

Przyjmujemy poziom istotności $\alpha = 0,05$. Z zamieszczonej w aneksie tablicy II odczytujemy wartość krytyczną $t_{n-2,\alpha/2} = t_{6,0,025} = 1,96$. Ponieważ $|t_{6,0}| > t_{6,0,025}$, zatem odrzucamy hipotezę zerową $H_0: \rho_{d.xy} = 0$ na korzyść hipotezy alternatywnej $H_1: \rho_{d.xy} \neq 0$. Zaobserwowana współzależność między zmiennymi losowymi X i Y nie jest więc przypadkowa. Merytoryczna interpretacja uzyskanej empirycznie wartości $r_{d.xy}$ jest taka sama jak w przykładzie 3.15. ■

5.6. Weryfikacja niezależności dwóch cech za pomocą testu chi-kwadrat

Załóżmy, że zbiorowość generalna A badana jest ze względu na dwie cechy $B_1, B_2 \in \mathbf{B}$, o następujących zbiorach rozróżnialnych stanów: $B_1^\circ = \{b_{1,i}; i = 1, \dots, k_1\}$; $B_2^\circ = \{b_{2,j}; j = 1, \dots, k_2\}$, przy czym $k_1, k_2 \geq 2$ (por. rozdział 2). Załóżmy też, że należy rozstrzygnąć, czy cechy B_1 i B_2 są niezależne.

W tym celu ze zbiorowości generalnej A pobiera się próbkę losową \underline{A} o liczności n . Każdy element $A \in \underline{A}$ poddaje się badaniu ze względu na B_1 i B_2 . Rezultaty badania wygodnie jest zestawić w postaci tzw. tablicy wielodzzielczej (albo tablicy kontyngencji), o wymiarach $k_1 \times k_2$. Tablica taka ułatwia analizę materiału doświadczalnego.

Zwróćmy uwagę na najprostszy przypadek, gdy $k_1 = k_2 = 2$. Przykład takiej tablicy wielodzzielczej o wymiarach 2×2 pokazano na rys. 5.5. W rezultacie doświadczenia uzyskuje się pewien rozkład skojarzeń stanów obu cech ($b_{1,1}^\circ, b_{1,2}^\circ, b_{2,1}^\circ, b_{2,2}^\circ$), scharakteryzowany liczebnościami n_{11}, n_{12}, n_{21} i n_{22} , umieszczonymi w górnej części każdej klatki tablicy przedstawionej na rys. 5.5. Zachodzą następujące zależności:

$$n_{11} + n_{12} = n_{1.} = n(b_{1,1}^\circ), \quad (5.131)$$

$$n_{21} + n_{22} = n_{2.} = n(b_{1,2}^{\circ}), \quad (5.132)$$

$$n_{11} + n_{21} = n_{.1} = n(b_{2,1}^{\circ}), \quad (5.133)$$

$$n_{12} + n_{22} = n_{.2} = n(b_{2,2}^{\circ}), \quad (5.134)$$

$$n_{.1} + n_{.2} = n, \quad (5.135)$$

$$n_{.1} + n_{.2} = n. \quad (5.136)$$

Weryfikacji poddaje się hipotezę zerową:

$$H_0: \text{cechy } B_1 \text{ i } B_2 \text{ są niezależne} \quad (5.137)$$

wobec hipotezy alternatywnej:

$$H_1: \text{cechy } B_1 \text{ i } B_2 \text{ są zależne.} \quad (5.138)$$

				Cecha B_2		Σ
				stan $b_{2,1}^{\circ}$	stan $b_{2,2}^{\circ}$	
				j		
				1	2	
Cecha B_1	stan $b_{1,1}^{\circ}$	i	1	n_{11} \hat{n}_{11}	n_{12} \hat{n}_{21}	$n_{.1}$
	stan $b_{1,2}^{\circ}$		2	n_{21} \hat{n}_{21}	n_{22} \hat{n}_{22}	$n_{.2}$
Σ			$n_{.1}$	$n_{.2}$	n	

Rys. 5.5. Tablica wielodzielcza 2×2

Zauważmy, że mamy tu do czynienia z hipotezami nieparametrycznymi, albowiem dotyczą one postaci rozkładu, a nie parametrów tego rozkładu. Zakładając prawdziwość hipotezy zerowej wyznacza się liczebności oczekiwane w poszczególnych kłatkach tablicy wielodzielczej. Stosuje się następujący algorytm:

– prawdopodobieństwo przynależności pojedynczego elementu $A \in \underline{A}$ do pierwszej wiersza tablicy wynosi:

$$\frac{n_{1.}}{n} = \frac{n(b_{1,1}^\circ)}{n}, \quad (5.139)$$

– analogiczne prawdopodobieństwo dla pierwszej kolumny rozważanej tablicy wynosi:

$$\frac{n_{.1}}{n} = \frac{n(b_{2,1}^\circ)}{n}, \quad (5.140)$$

– jeśli badane cechy B_1 i B_2 są niezależne, to prawdopodobieństwo skojarzenia się stanu $b_{1,1}^\circ$ ze stanem $b_{2,1}^\circ$ w pojedynczym elemencie $A \in \underline{A}$ wynosi³:

$$\frac{n_{1.}}{n} \cdot \frac{n_{.1}}{n} = \frac{n(b_{1,1}^\circ)}{n} \cdot \frac{n(b_{2,1}^\circ)}{n} \quad (5.141)$$

– tak więc liczba elementów $A \in \underline{A}$, w których należałoby oczekiwać takiego skojarzenia, wynosi:

$$\left\{ \begin{array}{l} \hat{n}_{11} = n \frac{n_{1.} \cdot n_{.1}}{n^2} = \frac{n_{1.} \cdot n_{.1}}{n} \\ \hat{n}_{11} = n \frac{n(b_{1,1}^\circ) \cdot n(b_{2,1}^\circ)}{n^2} = \frac{n(b_{1,1}^\circ) \cdot n(b_{2,1}^\circ)}{n} \end{array} \right. \quad (5.142)$$

Analogicznie wyznacza się liczebności oczekiwane \hat{n}_{12} , \hat{n}_{21} i \hat{n}_{22} .

Oczekiwane liczebności skojarzeń poszczególnych stanów cech B_1 i B_2 podane są w dolnych częściach kolejnych klatek tablicy wielodzzielczej, pokazanej na rys. 5.5. Zachodzą następujące zależności:

$$\hat{n}_{11} + \hat{n}_{12} = n_{1.} = n(b_{1,1}^\circ), \quad (5.143)$$

$$\hat{n}_{21} + \hat{n}_{22} = n_{.1} = n(b_{1,2}^\circ), \quad (5.144)$$

³ Przypomnijmy (zob. rozdział 1), że zdarzenia losowe A i B są niezależne, jeśli $P(AB) = P(A) \cdot P(B)$.

$$\hat{n}_{11} + \hat{n}_{21} = n_{.1} = n(b_{2,1}^{\circ}), \quad (5.145)$$

$$\hat{n}_{12} + \hat{n}_{22} = n_{.2} = n(b_{2,2}^{\circ}). \quad (5.146)$$

Jeśli prawdziwa jest hipoteza zerowa (5.137), to statystyka:

$$\chi_{r,0}^2 = \sum_{i=1}^{k_1} \sum_{j=1}^{k_2} \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}} \quad (5.147)$$

ma rozkład chi-kwadrat o $r = (k_1 - 1)(k_2 - 1)$ stopniach swobody.

W szczególnym przypadku tablicy wielozdzielczej 2×2 mamy jeden stopień swobody, bowiem $r = (k_1 - 1)(k_2 - 1) = (2 - 1)(2 - 1) = 1$, a sumowaniu podlegają cztery składniki postaci $(n_{ij} - \hat{n}_{ij})^2 / \hat{n}_{ij}$. Mamy mianowicie:

$$\chi_{1,0}^2 = \frac{(n_{11} - \hat{n}_{11})^2}{\hat{n}_{11}} + \frac{(n_{12} - \hat{n}_{12})^2}{\hat{n}_{12}} + \frac{(n_{21} - \hat{n}_{21})^2}{\hat{n}_{21}} + \frac{(n_{22} - \hat{n}_{22})^2}{\hat{n}_{22}}. \quad (5.148)$$

Obszar krytyczny dla wartości statystyki (5.147) ma postać (5.89), a reguły decyzyjne formułuje się zgodnie z nierównościami (5.90) i (5.91).

Przykład 5.8

Partię pewnego wyrobu o liczności $N = 100$ sztuk poddano wyczerpującym badaniom, mającym na celu wykrycie uszkodzeń transportowych B_1 i B_2 . Otrzymane wyniki przedstawiono w tablicy 5.2. Wyniki te pozwalają przypuszczać, że obserwowane uszkodzenia B_1 i B_2 nie są niezależne. Zauważmy bowiem, że na 40 uszkodzonych sztuk aż w 25 przypadkach obydwa uszkodzenia wystąpiły łącznie. Formułujemy hipotezę zerową:

H_0 : uszkodzenia transportowe B_1 i B_2 są niezależne

oraz hipotezę alternatywną:

H_1 : uszkodzenia transportowe B_1 i B_2 są zależne.

Ponieważ hipotezy te dotyczą pewnych zjawisk – a nie konkretnych obiektów (sztek produktu), w których te zjawiska zrealizowały się – zatem uzyskane wyniki badań potraktujemy jako próbkę losową o liczności $n = 100$ z nieskończonej licznej, zdarzeniowej zbiorowości generalnej.

Wyniki badania uszkodzeń transportowych

			Uszkodzenie B_2		Σ	
			jest	nie ma		
			j			
			1	2		
Uszkodzenie B_1	jest	i	1	25	10	35
	nie ma		2	5	60	65
Σ				30	70	100

Źródło: opracowanie własne.

Jeśli prawdziwa jest hipoteza zerowa, to:

$$\hat{n}_{11} = \frac{35}{100} \cdot \frac{30}{100} \cdot 100 = 10,5,$$

$$\hat{n}_{12} = \frac{35}{100} \cdot \frac{70}{100} \cdot 100 = 24,5,$$

$$\hat{n}_{21} = \frac{65}{100} \cdot \frac{30}{100} \cdot 100 = 19,5,$$

$$\hat{n}_{22} = \frac{65}{100} \cdot \frac{70}{100} \cdot 100 = 45,5.$$

Podstawiając wartości do wzoru (5.147) otrzymujemy:

$$\begin{aligned} \chi_{1,0}^2 &= \frac{(25-10,5)^2}{10,5} + \frac{(10-24,5)^2}{24,5} + \frac{(5-19,5)^2}{19,5} + \frac{(60-45,5)^2}{45,5} = \\ &= 20,024 + 8,582 + 10,782 + 4,621 = 44,009. \end{aligned}$$

Przyjmujemy $\alpha = 0,01$. Z zamieszczonej w aneksie tablicy III odczytujemy wartość $\chi^2_{1,0,01} = 6,635$. Ponieważ zachodzi nierówność $\chi^2_{1,0} \gg \chi^2_{1,0,01}$, odrzucamy więc hipotezę o niezależności uszkodzeń transportowych. ■

5.7. Weryfikacja zgodności rozkładu empirycznego z rozkładem hipotetycznym za pomocą testu chi-kwadrat

Badanie zgodności rozkładu empirycznego z rozkładem hipotetycznym odbywa się w dwóch etapach. Przede wszystkim należy dokonać wyboru klasy rozkładu hipotetycznego. Jeśli nie ma ku temu dostatecznie wyraźnych przesłanek teoretycznych, to wyboru dokonuje się biorąc pod uwagę typ zmiennej (skokowa, ciągła) oraz kształt histogramu (albo poligonu) rozkładu empirycznego w próbie⁴. I tak, na przykład, jeśli zmienna jest ciągła i jednocześnie histogram szeregu rozdzielczego jest w przybliżeniu symetryczny i jednomodalny (rys. 5.6 A), to przyjmujemy zwykle, że obserwowana zmienna losowa ma rozkład normalny. Jeśli zmienna jest ciągła, a histogram szeregu rozdzielczego ma kształt przedstawiony na rys. 5.6 B, to możemy przypuszczać, że mamy do czynienia z rozkładem wykładniczym. W przypadku takiego samego kształtu rozkładu, lecz w odniesieniu do zmiennej dyskretnej, będziemy postulować rozkład Poissona (rys. 5.6 C). Dokonanie wyboru klasy rozkładu umożliwia sformułowanie odpowiedniej hipotezy zerowej o braku różnic między rzeczywistą dystrybuantą badanej zmiennej a postulowaną dystrybuantą o tych samych parametrach. Pozwala to przejść do drugiego etapu postępowania, a mianowicie do weryfikacji hipotezy zerowej postaci:

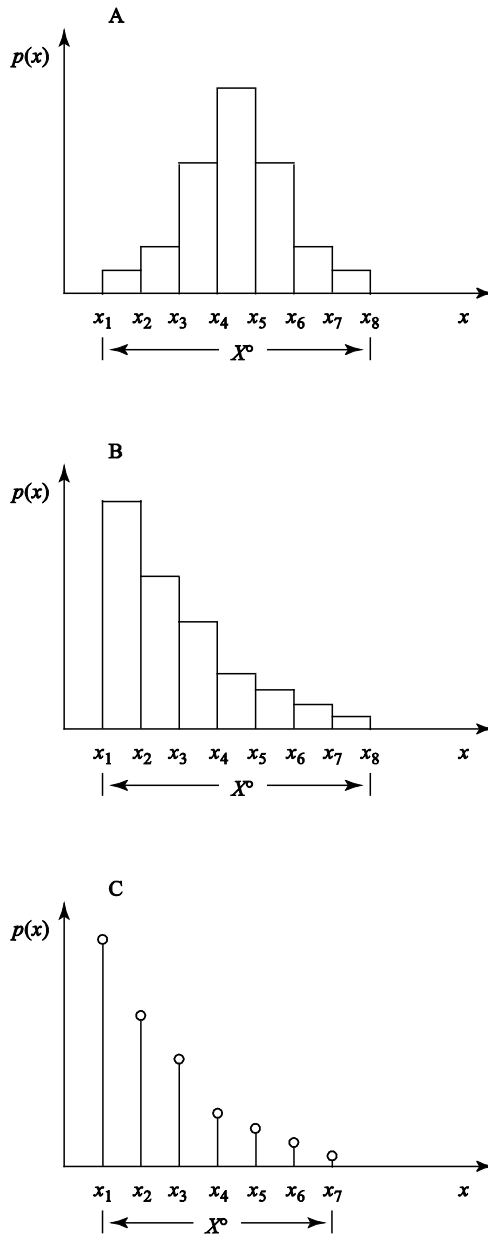
$$H_0: F(x) \in \mathbf{F}, \quad (5.149)$$

gdzie $F(x)$ jest nieznaną dystrybuantą obserwowanej zmiennej losowej, natomiast \mathbf{F} oznacza postulowaną klasę dystrybuant.

Hipotezę alternatywną formułuje się następująco:

$$H_1: F(x) \notin \mathbf{F}. \quad (5.150)$$

⁴ Zasady konstruowania szeregów rozdzielczych, a także ich graficznej prezentacji omówiono w rozdziale 3.



Rys. 5.6. Przykłady rozkładów empirycznych

Sprawdzianem hipotezy zerowej (5.149) jest statystyka postaci:

$$\chi_{r,0}^2 = \sum_{j=1}^k \frac{(f_j - \hat{f}_j)^2}{\hat{f}_j} = \sum_{j=1}^k \frac{(f_j - np_j)^2}{np_j}, \quad (5.151)$$

gdzie:

- f_j – liczba realizacji $x_i \in \mathbf{X}_n$ spełniających warunek $x_{d,j} < x_i \leq x_{g,j}$ (w przypadku ciągłej zmiennej losowej X) albo warunek $x_i = x_j^\circ$, gdy X jest dyskretną zmienną losową,
- \hat{f}_j – oczekiwana liczba realizacji $x_i \in \mathbf{X}_n$ spełniających warunek $x_{d,j} < x_i \leq x_{g,j}$ albo warunek $x_i = x_j^\circ$, przy założeniu, że hipoteza zerowa jest prawdziwa,
- p_j – prawdopodobieństwo zdarzenia losowego $x_{d,j} < X \leq x_{g,j}$ albo zdarzenia losowego $X = x_j^\circ$, przy założeniu, że hipoteza zerowa jest prawdziwa, przy czym $p_1 + \dots + p_j + \dots + p_k = 1$,
- n – licznosc próbki, przy czym $f_1 + \dots + f_j + \dots + f_k = n$,
- k – liczba elementów w zbiorze wartości X° albo liczba przedziałów klasowych, na które ten zbiór został rozcięty.

Jeśli hipoteza zerowa (5.149) jest prawdziwa, to statystyka (5.151) ma rozkład chi-kwadrat o $r = k - n_p - 1$ stopniach swobody, przy czym n_p oznacza liczbę parametrów funkcji rozkładu, które są szacowane z próby.

Obszar krytyczny dla uzyskiwanych empirycznie wartości statystyki (5.151) ma postać (5.89). Reguły decyzyjne oparte są na nierównościach (5.90) i (5.91).

Przykład 5.9

Wykonano $n = 76$ suszarkowych oznaczeń wilgotności pewnego gatunku tytoniu. Uzyskane wyniki zestawiono w postaci szeregu rozdzielczego, w kolumnach 2 i 3 tablicy 5.3. Rozkład licznosci empirycznych w poszczególnych przedziałach klasowych pozwala przypuszczać, że mamy do czynienia z rozkładem normalnym. Przemawia za tym ciągłość zmiennej losowej, jednododalność rozkładu empirycznego, a także jego symetria względem wartości modalnej. Formułujemy więc hipotezę zerową:

$$H_0: F(x) \in F_N$$

oraz hipotezę alternatywną:

$$H_1: F(x) \notin F_N.$$

Symbol F_N oznacza tu klasę dystrybuant normalnych, czyli zbiór wszystkich możliwych dystrybuant postaci:

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx.$$

Parametry μ i σ należy oszacować z próbki. Postępując według zasad omówionych w rozdziale 4, otrzymujemy:

$$\hat{\mu} = \bar{x}_n = \frac{1}{n} \sum_{j=1}^k f_j x_j^\circ = \frac{1417,20}{76} = 18,65,$$

$$\hat{\sigma} = s = \sqrt{\frac{1}{n-1} \sum_{j=1}^k f_j (\bar{x}_j - \bar{x}_n)^2} = \sqrt{\frac{34,55}{75}} = \sqrt{0,4607} = 0,6787.$$

Obliczenia pomocnicze zamieszczone są w kolumnach 4–8 tabelicy 5.3.

Celem dalszego postępowania jest rozstrzygnięcie, czy wyniki eksperymentu wykazują wystarczającą zgodność z dystrybuantą:

$$F(x) = \frac{1}{0,6787\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{1}{2}\left(\frac{x-18,65}{0,6787}\right)^2} dx.$$

W tym celu standaryzujemy górne kresy ($x_{g,j}$) kolejnych przedziałów klasowych ($j = 1, \dots, k = 6$). Uzyskane wartości zestawione są w kolumnie 9. Kres górny ostatniego przedziału klasowego przesuwamy do nieskończoności (∞). Odpowiednie wartości dystrybuanty w punktach $x_{g,j}$ zawiera kolumna 10. Obliczamy następnie wartości p_j i \hat{f}_j , zestawione w kolumnach 11 i 12. Dalsze kolumny (13–15) zawierają obliczenia pomocnicze niezbędne dla wyznaczenia wartości $\chi_{r,0}^2$ według wzoru (5.151). Wartość tę uzyskujemy sumując kolumnę 15. Liczba stopni swobody wynosi $r = k - n_p - 1 = 7 - 2 - 1 = 4$. Ponieważ przedstawione powyżej postępowanie wymagało oszacowania z próby dwóch parametrów (μ, σ), zatem $n_p = 2$. Otrzymaną wartość $\chi_{4,0}^2 = 0,693$ należy porównać z od-

Tablica 5.3

Weryfikacja normalności rozkładu za pomocą testu chi-kwadrat

j	$(x_{d,j}; x_{g,j}]$	f_j	x_j°	$f_j x_j^\circ$	$x_j^\circ - \bar{x}_n$	$(x_j^\circ - \bar{x}_n)^2$	$f_j (x_j^\circ - \bar{x}_n)^2$	$\frac{x_{g,j} - \bar{x}_n}{s}$	$\Phi\left(\frac{x_{g,j} - \bar{x}_n}{s}\right)$	p_j	$\hat{f}_j = np_j$	$f_j - \hat{f}_j$	$(f_j - \hat{f}_j)^2$	$\frac{(f_j - \hat{f}_j)^2}{\hat{f}_j}$
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	(17,2; 17,6]	6	17,4	104,40	-1,25	1,5625	9,3750	-1,55	0,0606	0,0606	4,41	1,39	1,9321	0,419
2	(17,6; 18,0]	8	17,8	142,40	-0,85	0,7225	5,7800	-0,86	0,1685	0,1079	8,20	-0,20	0,0400	0,005
3	(18,0; 18,4]	13	18,2	236,60	-0,45	0,2025	2,6325	-0,37	0,3557	0,1872	14,23	-1,23	1,5129	0,106
4	(18,4; 18,8]	17	18,6	316,20	-0,05	0,0025	0,0425	0,22	0,5871	0,2314	17,58	-0,58	0,3364	0,019
5	(18,8; 19,2]	15	19,0	285,00	0,35	0,1225	1,8375	0,81	0,7910	0,2039	15,50	-0,50	0,2500	0,016
6	(19,2; 19,6]	10	19,4	194,00	0,75	0,5625	5,6250	1,40	0,9192	0,1282	9,74	0,26	0,0676	0,007
7	(19,6; 20,0]	7	19,8	138,60	1,15	1,3225	9,2575	*	1*	0,0808	6,14	0,86	0,7396	0,121
Σ	-	76	-	1417,20	-	-	34,5500	-	-	1,0000	76,00	-	-	0,693

Uwaga: gwiazdką (*) oznaczono miejsca, w których przyjęty tok obliczeń należy zastąpić ogólniejszym rozumowaniem.

Źródło: opracowanie własne.

powiednią wartością krytyczną $\chi_{4,\alpha}^2$. Przyjmujemy $\alpha = 0,05$ i z tablicy III, zamieszczonej w aneksie, odczytujemy $\chi_{4;0,05}^2 = 9,488$. Ponieważ zachodzi nierówność $\chi_{4,0}^2 < \chi_{4;0,05}^2$, nie ma podstaw do odrzucenia hipotezy, że obserwowana zmienna losowa ma normalny rozkład prawdopodobieństwa. ■

5.8. Weryfikacja hipotezy o losowości wyników doświadczenia

Przypomnijmy (zob. rozdział 2), że doświadczenie można przedstawić w postaci następującego schematu:

$$A \rightarrow \underline{A} \rightarrow X_n,$$

w którym – analogicznie jak dotychczas – A jest zbiorowością generalną, \underline{A} oznacza próbkę, natomiast X_n jest zbiorem realizacji obserwowanej zmiennej losowej X . Zauważmy, że losowość transformacji $A \rightarrow \underline{A}$ nie implikuje losowego charakteru realizacji między kolejnymi wartościami $x_i \in X_n$, albowiem zastosowana metoda pomiaru – za pomocą której realizowana jest transformacja $\underline{A} \rightarrow X_n$ – może generować systematyczne zakłócenia. Z sytuacją taką mamy do czynienia – na przykład – wówczas, gdy zastosowany przyrząd pomiarowy zbyt wolno wraca do stanu równowagi po każdym powtórzeniu doświadczenia. Zakłócenia pojawiające się po pierwszym pomiarze przenoszą się na dalsze powtórzenia eksperymentu, a w konsekwencji realizacje między kolejnymi wartościami $x_i \in X_n$ przestają być losowe. Ów brak losowości może przejawiać się dwojako. Może to być przede wszystkim trend polegający na tym, że kolejne wartości x_i są przeciętnie coraz niższe albo przeciętnie coraz wyższe. Może też wystąpić zjawisko nadmiernej przemienności uzyskiwanych wyników, na przykład o charakterze nielosowej oscylacji wokół pewnej nieznannej wartości obserwowanej zmiennej losowej.

Weryfikacji poddaje się hipotezę zerową:

$$H_0: \text{ kolejne wartości } x_i \in X_n \text{ są niezależnymi realizacjami} \\ \text{ zmiennej losowej } X, \quad (5.152)$$

wobec hipotezy alternatywnej:

$$H_1: \text{ w zbiorze } X_n \text{ występują wartości } x_i, \text{ które nie są} \\ \text{ niezależnymi realizacjami zmiennej losowej } X. \quad (5.153)$$

Postępowanie:

1. Na podstawie zbioru realizacji X_n wyznacza się medianę $x_{me,n}$ według wzorów (4.22)–(4.24).

2. Zbiór realizacji X_n przekształca się w ciąg symboli a i b według następującej reguły:

– jeśli $x_i \leq x_{me,n}$, to wartość x_i zastępuje się symbolem a ,

– jeśli natomiast $x_i > x_{me,n}$, to wartości x_i zastępuje się symbolem b .

Każdy – nawet jednoelementowy – ciąg jednakowych symboli nazywany jest serią.

3. Liczbę liter a oznacza się symbolem n_a , natomiast liczbę liter b symbolem n_b . Liczbę serii oznacza się symbolami $k_{n_a n_b}$. Jest to realizacja zmiennej losowej $K_{n_a n_b}$ o rozkładzie zależnym od n_a i n_b .

4. Ustala się poziom istotności α , a następnie z zamieszczonej w aneksie tablicy VI odczytuje się krytyczne wartości liczby serii $k_{n_a n_b; \alpha/2}$ i $k_{n_a n_b; 1-\alpha/2}$. Jeśli prawdziwa jest hipoteza zerowa (5.152), to:

$$P(k_{n_a n_b} \leq k_{n_a n_b; \alpha/2}) = \frac{\alpha}{2}, \quad (5.154)$$

$$P(k_{n_a n_b} \leq k_{n_a n_b; 1-\alpha/2}) = 1 - \frac{\alpha}{2}. \quad (5.155)$$

5. Jeśli zachodzi nierówność:

$$k_{n_a n_b; \alpha/2} < k_{n_a n_b} \leq k_{n_a n_b; 1-\alpha/2}, \quad (5.156)$$

to nie ma podstaw do odrzucenia hipotezy zerowej (5.152).

Jeśli natomiast

$$k_{n_a n_b} \leq k_{n_a n_b; \alpha/2} \quad (5.157)$$

albo

$$k_{n_a n_b} > k_{n_a n_b; 1-\alpha/2}, \quad (5.158)$$

to hipotezę zerową (5.152) odrzuca się na korzyść hipotezy alternatywnej (5.153).

Przykład 5.10

Wykonano $n = 12$ pomiarów twardości płytki poliamidowej uzyskanej metodą odlewania ciśnieniowego. Pomiarzy wykonano metodą Brinella i otrzymano wyniki przedstawione w kolumnie 2 tablicy 5.4. W celu wyznaczenia mediany porządkujemy zbiór \mathbf{X}_n za pomocą relacji „ \leq ” i otrzymujemy:

7,961; 8,042; 8,208; 8,425; 8,701; 8,787;
8,797; 8,895; 8,945; 9,047; 9,257; 9,366.

Tablica 5.4

Wyniki pomiarów twardości płytki poliamidowej otrzymanej metodą odlewania ciśnieniowego

i	x_i	$a; b$	serie
1	2	3	4
1	9,366	b	1
2	8,208	a	2
3	8,895	b	3
4	8,787	a	} 4
5	8,042	a	
6	9,257	b	5
7	8,425	a	} 6
8	8,701	a	
9	9,047	b	} 7
10	8,945	b	
11	8,797	b	
12	7,961	a	8

Źródło: badania własne.

Medianę wyznaczamy według wzoru (4.23):

$$x_{me,n} = \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2} = \frac{x_{(6)} + x_{(7)}}{2} = \frac{8,787 + 8,797}{2} = 8,792.$$

Wykorzystując tę wartość zakodowano realizacje $x_i \in \mathbf{X}_n$ za pomocą symboli a i b (zob. kolumna 3). Mamy $n_a = n_b = 6$. W kolumnie 4 zaznaczono numery kolejnych serii. Otrzymaliśmy $k_{6,6} = 8$. Z zamieszczonej w aneksie tablicy VI/1

odczytujemy wartości krytyczne dla $\alpha = 0,05$. Mamy mianowicie: $k_{6,6;0,025} = 3$, $k_{6,6;0,975} = 10$. Ponieważ $k_{6,6;0,025} < 8 < k_{6,6;0,975}$, nie ma podstaw do odrzucenia hipotezy zerowej głoszącej, że relacje między kolejnymi realizacjami $x_i \in \mathbf{X}_n$ zostały ukształtowane tylko przez czynniki losowe funkcjonujące w doświadczeniu. ■

5.9. Sekwencyjne metody weryfikacji hipotez statystycznych

5.9.1. Ogólne zasady funkcjonowania testów sekwencyjnych

We wszystkich omówionych testach liczność próbki (n) jest ustalona i nie podlega modyfikacjom w procesie weryfikacji danej hipotezy zerowej. Ma to ten skutek, że w pewnych sytuacjach ustalona liczność próbki może okazać się za duża dla odróżnienia hipotezy prawdziwej od hipotezy fałszywej, przy zadanej wartości α i w postulowanym zakresie wartości β . W innych natomiast przypadkach liczność próbki może okazać się za mała i może nie gwarantować utrzymania β w pożądanym zakresie wartości (zob. rys. 5.3). W testach sekwencyjnych wartości α i β zadawane są *a priori*, a liczność próbki jest wartością wynikową. Podstawą tych testów jest następująca nierówność:

$$\frac{\beta}{1-\alpha} < K_{i_r} = \frac{\prod_{i=1}^{i_r} f(x_i; H_1)}{\prod_{i=1}^{i_r} f(x_i; H_0)} < \frac{1-\beta}{\alpha}; \quad i_r = 1, 2, \dots \quad (5.159)$$

gdzie:

- $f(x_i; H_0)$ – prawdopodobieństwo zdarzenia losowego $X = x_i$, gdy prawdziwa jest hipoteza H_0 ,
- $f(x_i; H_1)$ – prawdopodobieństwo zdarzenia losowego $X = x_i$, gdy prawdziwa jest hipoteza H_1 ,
- $1 - \alpha$ – prawdopodobieństwo przyjęcia hipotezy H_0 , gdy jest ona prawdziwa,
- α – prawdopodobieństwo przyjęcia hipotezy H_1 , gdy prawdziwa jest hipoteza H_0 ,
- $1 - \beta$ – prawdopodobieństwo przyjęcia hipotezy H_1 , gdy jest ona prawdziwa,
- β – prawdopodobieństwo przyjęcia hipotezy H_0 , gdy prawdziwa jest hipoteza H_1 .

Nierówność (5.159) wyraża warunek przedłużenia sekwencji (kontynuowania badań). Jeśli warunek ten jest spełniony w danym kroku badania (i_r), to przechodzi się do następnego kroku ($i_r + 1$), generując kolejne realizacje obserwowanej zmiennej losowej X .

Jeśli w kolejnym kroku postępowania zostanie spełniona nierówność:

$$K_{i_r} \leq \frac{\beta}{1 - \alpha}, \quad (5.160)$$

to przyjmuje się hipotezę H_0 , a prawdopodobieństwo zdarzenia losowego polegającego na tym, że w rzeczywistości prawdziwa jest hipoteza H_1 , nie przekracza wartości β . Jeśli natomiast:

$$K_{i_r} \geq \frac{1 - \beta}{\alpha}, \quad (5.161)$$

to przyjmuje się hipotezę H_1 , a prawdopodobieństwo zdarzenia losowego polegającego na tym, że w rzeczywistości prawdziwa jest hipoteza H_0 , nie przekracza wartości α . Spełnienie warunku (5.160) albo (5.161) oznacza zakończenie sekwencji badań, a wynikowa liczność próbki wynosi

$$n = \sum_{i=1}^{i_r} m_i, \quad (5.162)$$

gdzie m_i jest licznnością próbki cząstkowej, badanej w i -tym kroku badania.

Testy sekwencyjne konstruowane są zwykle w ten sposób, że:

$$\bigwedge_i m_i = m, \quad (5.163)$$

przy czym najbardziej racjonalna jest taka procedura, w której $m = 1$. Mamy wówczas $i_r = n$.

5.9.2. Weryfikacja hipotez dotyczących wartości oczekiwanej normalnej zmiennej losowej

Niech X będzie zmienną losową o nieznannej wartości oczekiwanej μ oraz o znanym odchyleniu standardowym σ . Zwróćmy uwagę na przypadek, gdy weryfikacji poddawana jest hipoteza zerowa (5.18) postaci $H_0: \mu = \mu_0$ wobec hipotezy alternatywnej:

$$H_1: \mu = \mu_1, \quad (5.164)$$

przy czym $\mu_1 > \mu_0$. Zauważmy, że hipoteza (5.164) jest szczegółową postacią hipotezy (5.12). Jest ona równoważna sformułowanej wcześniej hipotezie (5.43).

Założmy, że w kolejnych krokach postępowania sekwencyjnego generowane są pojedyncze realizacje (x_i) zmiennej losowej X , a obserwowana charakterystyka z próbki skumulowanej o liczności $n = i_r$ ma postać:

$$z_n = \sum_{i=1}^n x_i; \quad n = i_r = 1, 2, \dots \quad (5.165)$$

Wykorzystując przyjęte założenia, nierówność (5.159) zapiszemy następująco:

$$\frac{\beta}{1-\alpha} < \frac{\prod_{i=1}^n f(x_i; \mu_1)}{\prod_{i=1}^n f(x_i; \mu_0)} < \frac{1-\beta}{\alpha}. \quad (5.166)$$

Po podstawieniu funkcji gęstości prawdopodobieństwa rozkładu normalnego mamy:

$$\frac{\beta}{1-\alpha} < \frac{\prod_{i=1}^n \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x_i - \mu_1}{\sigma}\right)^2\right]}{\prod_{i=1}^n \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x_i - \mu_0}{\sigma}\right)^2\right]} < \frac{1-\beta}{\alpha},$$

$$\frac{\beta}{1-\alpha} < \exp\left[\frac{1}{2\sigma^2}\left(\sum_{i=1}^n (x_i - \mu_0)^2 - \sum_{i=1}^n (x_i - \mu_1)^2\right)\right] < \frac{1-\beta}{\alpha}.$$

Po zlogarytmowaniu i dalszych przekształceniach otrzymujemy:

$$\frac{\sigma^2 \ln \frac{\beta}{1-\alpha}}{\mu_1 - \mu_0} + \frac{\mu_0 + \mu_1}{2} n < z_n < \frac{\sigma^2 \ln \frac{1-\beta}{\alpha}}{\mu_1 - \mu_0} + \frac{\mu_0 + \mu_1}{2} n. \quad (5.167)$$

Jest to warunek kontynuowania badań (przedłużania sekwencji), w przypadku gdy obserwowana charakterystyka z próby skumulowanej ma postać (5.165). Jak łatwo zauważyć, warunek ten można zapisać następująco:

$$a + cn = z'_n < z_n < z''_n = b + cn, \quad (5.168)$$

przy czym

$$a = \frac{\sigma^2}{\mu_1 - \mu_0} \ln \frac{\beta}{1 - \alpha}, \quad (5.169)$$

$$b = \frac{\sigma^2}{\mu_1 - \mu_0} \ln \frac{1 - \beta}{\alpha}, \quad (5.170)$$

$$c = \frac{\mu_0 + \mu_1}{2}. \quad (5.171)$$

Jeśli w kolejnym kroku postępowania charakterystyka z_n spełni nierówność:

$$z_n \leq z'_n = a + cn, \quad (5.172)$$

to przerywa się kumulację wartości x_i w postaci sumy (5.165) i przyjmuje się hipotezę zerową $H_0: \mu = \mu_0$. Jeśli natomiast zostanie spełniona nierówność:

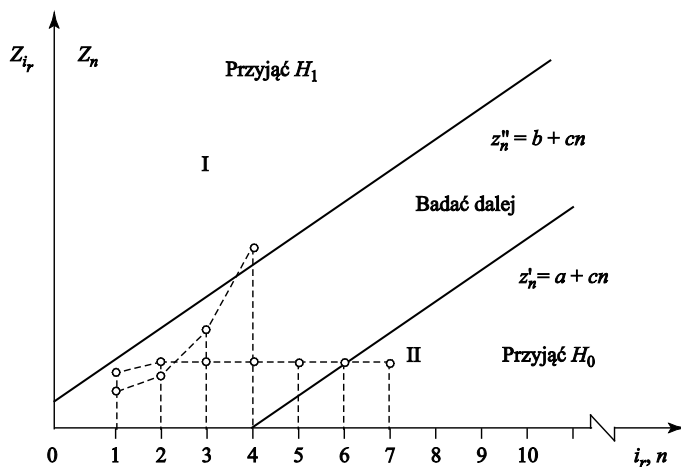
$$z_n \geq z''_n = b + cn, \quad (5.173)$$

to również przerywa się kumulację wartości x_i generowanych w kolejnych krokach postępowania i przyjmuje się hipotezę alternatywną $H_1: \mu = \mu_1$.

Przedstawiona procedura może być również realizowana w formie algorytmu graficznego. Stosowany w takich przypadkach diagram przeglądowy przedstawiono na rys. 5.7. Na rysunku tym pokazano schematycznie dwie sekwencje obserwacji. Sekwencja I ma postać:

$$\begin{aligned} z_1 &= x_1 \\ z_2 &= x_1 + x_2 = z_1 + x_2 \\ z_3 &= x_1 + x_2 + x_3 = z_2 + x_3 \\ z_4 &= x_1 + x_2 + x_3 + x_4 = z_3 + x_4 \end{aligned}$$

i zakończona jest przyjęciem hipotezy $H_1: \mu = \mu_1$. Sekwencja II, zrealizowana w kolejnych 7 krokach, zakończona jest przyjęciem hipotezy $H_0: \mu = \mu_0$.



Rys. 5.7. Funkcjonowanie sekwencyjnej procedury weryfikacji hipotez statystycznych

Omówiony powyżej sekwencyjny test ilorazowy jest testem prawostronnym. W analogiczny sposób można skonstruować test lewostronny, służący do weryfikacji hipotezy zerowej $H_0: \mu = \mu_0$ wobec hipotezy alternatywnej:

$$H_{-1}: \mu = \mu_{-1} \quad (5.174)$$

takiej, że $\mu_{-1} < \mu_0$; zob. wzór (5.13).

Łącząc obydwie te schematy otrzymujemy test dwustronny, w którym hipoteza zerowa $H_0: \mu = \mu_0$ jest weryfikowana wobec dwóch hipotez alternatywnych, a mianowicie $H_1: \mu = \mu_1$ i $H_{-1}: \mu = \mu_{-1}$, przy czym $\mu_{-1} < \mu_0 < \mu_1$.

5.9.3. Weryfikacja hipotez dotyczących odchylenia standardowego normalnej zmiennej losowej

Niech X będzie normalną zmienną losową o znanej wartości oczekiwanej (μ) i nieznanym odchyleniu standardowym σ . Zwróćmy uwagę na przypadek, gdy weryfikacji podlega hipoteza zerowa:

$$H_0: \sigma = \sigma_0 \quad (5.175)$$

wobec hipotezy alternatywnej:

$$H_1: \sigma = \sigma_1, \quad (5.176)$$

przy czym $\sigma_1 > \sigma_0$. Zauważmy, że hipoteza (5.175) jest szczególną postacią hipotezy (5.1) i że jest ona równoważna sformułowanej wcześniej hipotezie (5.73). Hipoteza (5.176) jest szczególną postacią hipotezy (5.12).

Założmy – analogicznie jak w punkcie 5.9.2 – że w kolejnych krokach postępowania sekwencyjnego pobierane są próbki o liczności $m = 1$. W konsekwencji mamy $i_r = n$. Nierówność (5.165) zapiszemy obecnie następująco:

$$\frac{\beta}{1-\alpha} < \frac{\prod_{i=1}^n f(x_i; \sigma_1)}{\prod_{i=1}^n f(x_i; \sigma_0)} < \frac{1-\beta}{\alpha}. \quad (5.177)$$

Po podstawieniu funkcji gęstości prawdopodobieństwa rozkładu normalnego mamy:

$$\frac{\beta}{1-\alpha} < \frac{\prod_{i=1}^n \frac{1}{\sigma_1 \sqrt{2\pi}} \exp\left[-\frac{1}{2} \left(\frac{x_i - \mu}{\sigma_1}\right)^2\right]}{\prod_{i=1}^n \frac{1}{\sigma_0 \sqrt{2\pi}} \exp\left[-\frac{1}{2} \left(\frac{x_i - \mu}{\sigma_0}\right)^2\right]} < \frac{1-\beta}{\alpha}.$$

Nierówność tę przekształcimy w taki sposób, by jako charakterystykę z próby skumulowanej można było przyjąć:

$$z_n = \sum_{i=1}^n (x_i - \mu)^2; \quad i_r = n = 1, 2, \dots \quad (5.178)$$

W rezultacie tych przekształceń otrzymujemy:

$$\ln \frac{\beta}{1-\alpha} < n \ln \frac{\sigma_0}{\sigma_1} + \frac{1}{2} \left(\frac{\sigma_1^2 - \sigma_0^2}{\sigma_0^2 \cdot \sigma_1^2} \right) \sum_{i=1}^n (x_i - \mu)^2 < \ln \frac{1-\beta}{\alpha}.$$

Warunek kontynuowania badań (przedłużania sekwencji) można więc przedstawić w postaci (5.168), przy czym:

$$a = \frac{2\sigma_0^2 \sigma_1^2}{\sigma_1^2 - \sigma_0^2} \ln \frac{\beta}{1-\alpha}, \quad (5.179)$$

$$b = \frac{2\sigma_0^2\sigma_1^2}{\sigma_1^2 - \sigma_0^2} \ln \frac{1-\beta}{\alpha}, \quad (5.180)$$

$$c = \frac{2\sigma_0^2\sigma_1^2}{\sigma_1^2 - \sigma_0^2} \ln \frac{\sigma_0}{\sigma_1}. \quad (5.181)$$

Decyzje o przyjęciu hipotezy $H_0: \sigma = \sigma_0$ albo hipotezy $H_1: \sigma = \sigma_1$ podejmowane są zgodnie z nierównościami (5.172) i (5.173). Jeśli procedura realizowana jest w formie algorytmu graficznego, to diagram przeglądowy ma postać pokazaną na rys. 5.10. W teście lewostronnym hipoteza zerowa $H_0: \sigma = \sigma_0$ weryfikowana jest wobec hipotezy alternatywnej:

$$H_{-1}: \sigma = \sigma_{-1}, \quad (5.182)$$

przy czym $\sigma_0 > \sigma_{-1}$.

Hipoteza (5.182) jest szczegółową postacią hipotezy (5.13). W teście dwustronnym mamy: $H_0: \sigma = \sigma_0$, $H_1: \sigma = \sigma_1$, $H_{-1}: \sigma = \sigma_{-1}$, przy czym $\sigma_{-1} < \sigma_0 < \sigma_1$. Zasady konstruowania testu lewostronnego są takie same jak w przypadku omówionego powyżej testu prawostronnego. Test dwustronny jest złożeniem testu lewostronnego i prawostronnego.

5.9.4. Weryfikacja hipotez dotyczących wskaźnika struktury

Niech X oznacza zero-jedynkową zmienną losową o nieznanym parametrze p . Rozważmy problem weryfikacji hipotezy zerowej (5.110) postaci $H_0: p = p_0$ wobec hipotezy alternatywnej:

$$H_1: p = p_1, \quad (5.183)$$

przy czym $p_1 > p_0$.

Do weryfikacji hipotezy $H_0: p = p_0$ wykorzystamy charakterystykę (5.165). Ponieważ w rozważanym przypadku realizacje x_i są zerami albo jedynkami, zatem wartości z_n oznaczają liczbę jedynek w n -elementowym słowie binarnym, jakim jest – w każdym kroku postępowania sekwencyjnego – zbiór X_n . Przypomnijmy (zob. przykład 5.4), że w statystycznej kontroli jakości wartość „1” przypisuje się elementowi wadliwemu, natomiast wartość „0” elementowi spełniającemu wymagania jakościowe. W takiej sytuacji z_n jest liczbą elementów

wadliwych w n -elementowej, skumulowanej próbce. Nierówność (5.159) zapiszemy następująco:

$$\frac{\beta}{1-\alpha} < \frac{\prod_{i=1}^n p_1^{x_i} (1-p_1)^{1-x_i}}{\prod_{i=1}^n p_0^{x_i} (1-p_0)^{1-x_i}} < \frac{1-\beta}{\alpha}. \quad (5.184)$$

Po niezbędnych przekształceniach mamy:

$$\frac{\beta}{1-\alpha} < \left(\frac{p_1}{p_0}\right)^{z_n} \left(\frac{1-p_1}{1-p_0}\right)^{n-z_n} < \frac{1-\beta}{\alpha},$$

$$\ln \frac{\beta}{1-\alpha} < z_n \ln \frac{p_1}{p_0} + n \ln \frac{1-p_1}{1-p_0} - z_n \ln \frac{1-p_1}{1-p_0} < \ln \frac{1-\beta}{\alpha},$$

$$\ln \frac{\beta}{1-\alpha} + n \ln \frac{1-p_0}{1-p_1} < z_n \left(\ln \frac{p_1}{p_0} - \ln \frac{1-p_1}{1-p_0} \right) < \ln \frac{1-\beta}{\alpha} + n \ln \frac{1-p_0}{1-p_1}.$$

W dalszych przekształceniach należy uwzględnić to, że:

$$\ln \frac{p_1}{p_0} > \ln \frac{1-p_1}{1-p_0},$$

ponieważ $p_1 > p_0$.

Tak więc zapisując warunek kontynuowania badań (przedłużania sekwencji) w postaci nierówności (5.168), mamy:

$$a = \frac{\ln \frac{\beta}{1-\alpha}}{\ln \frac{p_1(1-p_0)}{p_0(1-p_1)}}, \quad (5.185)$$

$$b = \frac{\ln \frac{1-\beta}{\alpha}}{\ln \frac{p_1(1-p_0)}{p_0(1-p_1)}}, \quad (5.186)$$

$$c = \frac{\ln \frac{1-p_0}{1-p_1}}{\ln \frac{p_1(1-p_0)}{p_0(1-p_1)}}. \quad (5.187)$$

Decyzje o przyjęciu hipotezy $H_0: p = p_0$ albo hipotezy $H_1: p = p_1$ podejmowane są zgodnie z nierównościami (5.172) i (5.173).

Omówiony powyżej test prawostronny ma podstawowe znaczenie w statystycznej kontroli jakości, gdzie p_0 traktuje się jako najwyższą dopuszczalną wadliwość strumienia albo partii wyrobu, natomiast p_1 jako najniższą dyskwalifikującą wadliwość. W teście lewostronnym weryfikacji poddaje się hipotezę $H_0: p = p_0$ wobec hipotezy alternatywnej $H_{-1}: p = p_{-1}$, przy czym $p_0 > p_{-1}$. Test dwustronny – analogicznie jak poprzednio – jest złożeniem testu lewostronnego i prawostronnego. W przypadku tego testu mamy: $H_0: p = p_0$; $H_1: p = p_1$; $H_{-1}: p = p_{-1}$; przy czym $p_{-1} < p_0 < p_1$.

5.9.5. Weryfikacja hipotez dotyczących parametru rozkładu Poissona

Niech X będzie zmienną losową o rozkładzie Poissona z nieznanym parametrem λ . Zwróćmy uwagę na przypadek, gdy w odniesieniu do tego parametru formułowana jest hipoteza zerowa:

$$H_0: \lambda = \lambda_0 \quad (5.188)$$

wobec hipotezy alternatywnej:

$$H_1: \lambda = \lambda_1, \quad (5.189)$$

przy czym $\lambda_1 > \lambda_0$. Zauważmy, że hipoteza zerowa (5.188) jest szczegółową postacią hipotezy (5.1), natomiast hipoteza alternatywna (5.189) jest szczegółową postacią hipotezy (5.12).

Założmy – analogicznie jak poprzednio – że w kolejnych krokach postępowania sekwencyjnego pobierane są próbki o liczności $m = 1$. W konsekwencji mamy $i_r = n$. Nierówność (5.159) zapiszemy następująco:

$$\frac{\beta}{1-\alpha} < \frac{\prod_{i=1}^n \frac{\lambda_1^{x_i}}{x_i!} e^{-\lambda_1}}{\prod_{i=1}^n \frac{\lambda_0^{x_i}}{x_i!} e^{-\lambda_0}} < \frac{1-\beta}{\alpha}. \quad (5.190)$$

Po wykonaniu mnożenia mamy:

$$\frac{\beta}{1-\alpha} < \frac{\frac{\sum_{i=1}^n x_i}{\lambda_1^{x_1} \dots x_i! \dots x_n!} e^{-n\lambda_1}}{\frac{\sum_{i=1}^n x_i}{\lambda_0^{x_1} \dots x_i! \dots x_n!} e^{-n\lambda_0}} < \frac{1-\beta}{\alpha}. \quad (5.191)$$

Wyrażenie to przekształcimy w taki sposób, by obserwowana charakterystyka z próby miała postać sumy (5.165), w której realizacji (x_i) mogą przyjmować wartości 0, 1, 2, ... Przekształcając wzór (5.191) otrzymujemy:

$$\frac{\beta}{1-\alpha} < \left(\frac{\lambda_1}{\lambda_0} \right)^{\sum_{i=1}^n x_i} e^{n(\lambda_0 - \lambda_1)} < \frac{1-\beta}{\alpha},$$

$$\ln \frac{\beta}{1-\alpha} - (\lambda_0 - \lambda_1)n < \ln \frac{\lambda_1}{\lambda_0} \sum_{i=1}^n x_i < \ln \frac{1-\beta}{\alpha} - (\lambda_0 - \lambda_1)n.$$

Ponieważ $\lambda_1 > \lambda_0$, więc $\ln \frac{\lambda_1}{\lambda_0} > 0$. Możemy zatem napisać:

$$\frac{\ln \frac{\beta}{1-\alpha}}{\ln \frac{\lambda_1}{\lambda_0}} + \frac{\lambda_1 - \lambda_0}{\ln \frac{\lambda_1}{\lambda_0}} n < \sum_{i=1}^n x_i < \frac{\ln \frac{1-\beta}{\alpha}}{\ln \frac{\lambda_1}{\lambda_0}} + \frac{\lambda_1 - \lambda_0}{\ln \frac{\lambda_1}{\lambda_0}} n.$$

Otrzymaliśmy więc – analogicznie jak poprzednio – wyrażenie postaci (5.168), w którym:

$$a = \frac{\ln \frac{\beta}{1-\alpha}}{\ln \lambda_1 - \ln \lambda_0}, \quad (5.192)$$

$$b = \frac{\ln \frac{1-\beta}{\alpha}}{\ln \lambda_1 - \ln \lambda_0}, \quad (5.193)$$

$$c = \frac{\lambda_1 - \lambda_0}{\ln \lambda_1 - \ln \lambda_0}. \quad (5.194)$$

Powiększenie próbki (o jeden element w każdym kroku) trwa tak długo, jak długo spełniona jest nierówność (5.168). Decyzje o przyjęciu hipotezy $H_0: \lambda = \lambda_0$ albo hipotezy $H_1: \lambda = \lambda_1$ podejmowane są zgodnie z nierównościami (5.172) i (5.173). Jeśli procedura realizowana jest w formie algorytmu graficznego, to diagram przeglądowy ma postać pokazaną na rys. 5.7.

W teście lewostronnym hipoteza zerowa (5.188) weryfikowana jest wobec hipotezy alternatywnej:

$$H_{-1}: \lambda = \lambda_{-1}, \quad (5.195)$$

przy czym $\lambda_{-1} < \lambda_0$. Zasady konstruowania testu lewostronnego są takie same jak w przypadku omówionego powyżej testu prawostronnego. Łącząc obydwa testy jednostronne otrzymujemy test dwustronny, w którym hipoteza zerowa $H_0: \lambda = \lambda_0$ weryfikowana jest wobec dwóch hipotez alternatywnych, a mianowicie $H_1: \lambda = \lambda_1$ i $H_{-1}: \lambda = \lambda_{-1}$, przy czym $\lambda_{-1} < \lambda_0 < \lambda_1$.

W statystycznej kontroli jakości podstawowe znaczenie ma test prawostronny. Test ten wykorzystywany jest do kontroli liczby wad w jednostce produktu, która może być najczęściej traktowana jako zmienna losowa w rozkładzie Poissona. W takiej sytuacji parametr λ reprezentuje przeciętny, rzeczywisty poziom wad w jednostce produktu. Wartość λ_0 oznacza najwyższy dopuszczalny poziom wad, natomiast λ_1 jest najniższym dyskwalifikującym poziomem wad. Wartości λ_0 i λ_1 ustalane są na podstawie przesłanek pozastatystycznych.

Rozdział 6

ANALIZA SZEREGÓW CZASOWYCH

6.1. Wprowadzenie

Szeregiem czasowym (dynamicznym, rozwojowym, chronologicznym) nazywamy uporządkowany ze względu na kryterium czasu ciąg wyników obserwacji (t, y_t) , w którym t oznacza numer kolejnej jednostki czasu (okresu, momentu), y_t – wynik obserwacji w czasie $t = 1, 2, \dots$

Szereg czasowy przedstawia się więc następująco:

t	1	2	3	4	...
y_t	y_1	y_2	y_3	y_4	...

Zadaniem analizy szeregów czasowych jest – przede wszystkim – określanie zmian zachodzących w poziomie badanego procesu, czyli wyznaczanie tendencji rozwojowej oraz wyodrębnianie wahań cyklicznych i sezonowych.

Przykład 6.1

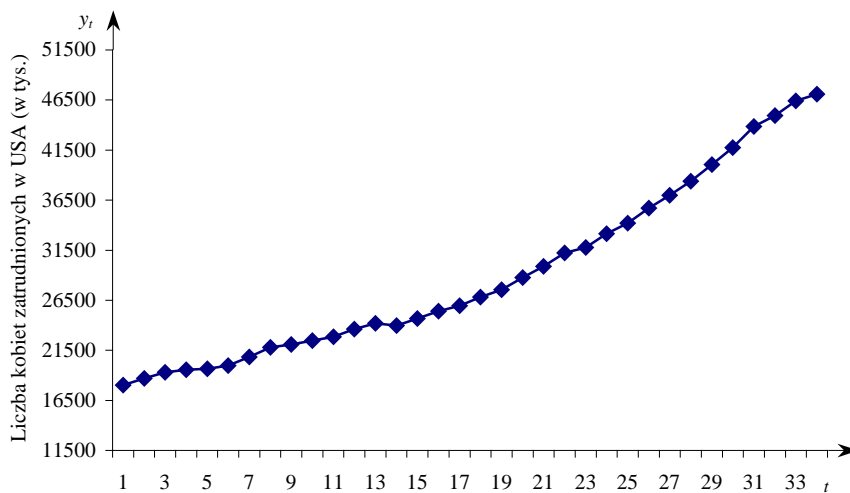
W tabelicy 6.1 przedstawiono rozwój zatrudnienia kobiet w USA w latach 1949–1982. Jak wynika z tabelicy, w roku 1949 było zatrudnionych 18 mln 30 tys. kobiet, natomiast w roku 1982 zatrudnienie to wzrosło do 47 mln 95 tys. osób. Na podstawie danych z tabelicy 6.1 sporządzono wykres pokazany na rys. 6.1. Analiza wykresu pozwala ocenić tendencję rozwojową obserwowanego zjawiska. Jak łatwo zauważyć, jest to wyraźna tendencja wzrostowa, z lekko zaznaczonym odchyleniem od liniowości w drugiej połowie obserwowanego przedziału czasu. Do problemu tego powrócimy w punkcie 6.4.2. □

Tablica 6.1

Liczba kobiet zatrudnionych w USA

t	Rok	Zatrudnienie (w tys.) y_t	t	Rok	Zatrudnienie (w tys.) y_t
1	1949	18030	18	1966	26820
2	1950	18680	19	1967	27545
3	1951	19309	20	1968	28778
4	1952	19559	21	1969	29898
5	1953	19668	22	1970	31233
6	1954	19970	23	1971	31778
7	1955	20842	24	1972	33152
8	1956	21808	25	1973	34195
9	1957	22097	26	1974	35708
10	1958	22482	27	1975	36981
11	1959	22865	28	1976	38399
12	1960	23619	29	1977	40053
13	1961	24199	30	1978	41747
14	1962	23978	31	1979	43844
15	1963	24675	32	1980	44934
16	1964	25399	33	1981	46415
17	1965	25952	34	1982	47095

Źródło: *Abstract of US Statistics, 1946–1982.*



Rys. 6.1. Rozwój zatrudnienia kobiet w USA

6.2. Wskaźnikowa analiza szeregów czasowych

Podstawowymi narzędziami stosowanymi w procesie analizy szeregów czasowych są wskaźniki. Jeśli badane zjawisko opisywane jest ze względu na jedno kryterium, to do analizy szeregu czasowego wykorzystuje się tzw. wskaźniki indywidualne. Jeśli natomiast mamy do czynienia z wielokryterialnym opisem badanego zjawiska, to do analizy takiego szeregu czasowego stosowane są wskaźniki agregatowe (zespołowe).

6.2.1. Wskaźniki indywidualne

Najprostszymi narzędziami analitycznymi zaliczanymi do wskaźników indywidualnych są tzw. przyrosty absolutne. Wyróżnia się przy tym przyrosty jednopodstawowe i przyrosty zmiennopodstawowe. Jednopodstawowy przyrost absolutny dany jest wzorem:

$$\Delta_{t/t_b} = y_t - y_{t_b}, \quad (6.1)$$

w którym y_t oznacza poziom obserwowanego zjawiska w okresie badanym (t), natomiast y_{t_b} jest poziomem tego samego zjawiska w okresie bazowym (t_b), stanowiącym punkt odniesienia.

Zmiennopodstawowy przyrost absolutny definiowany jest następująco:

$$\Delta_{t/t-k} = y_t - y_{t-k}, \quad (6.2)$$

gdzie y_{t-k} oznacza poziom analizowanego zjawiska w okresie o k jednostek wcześniejszym ($t-k$) w stosunku do okresu badanego (t), przy czym $k = 1, 2, \dots$

Przyjmując $k = 1$, uzyskujemy szczególną i w praktyce najważniejszą postać zmiennopodstawowego przyrostu absolutnego, a mianowicie tzw. absolutny przyrost łańcuchowy:

$$\Delta_{t/t-1} = y_t - y_{t-1}, \quad (6.2a)$$

za pomocą którego porównuje się aktualny stan badanego zjawiska ze stanem o jeden okres wcześniejszym.

Przyrosty absolutne, zarówno jednopodstawowe jak i zmiennopodstawowe, mogą być dodatnie, ujemne lub zerowe. Są to wielkości mianowane, wyrażone w tych samych jednostkach miary, w których wyrażany jest poziom obserwowanego

nego zjawiska. Uniemożliwia to stosowanie tych przyrostów w analizie porównawczej zjawisk, których natężenie wyrażane jest w różnych jednostkach miary.

W celu uniknięcia skutków tego ograniczenia można posłużyć się przyrostami względnymi, które są wielkościami niemianowanymi. Jednospodstawowy przyrost względny zdefiniowany jest wzorem:

$$\delta_{t/t_b} = \frac{y_t - y_{t_b}}{y_{t_b}} = \frac{\Delta_{t/t_b}}{y_{t_b}}. \quad (6.3)$$

Zmiennospodstawowy przyrost względny ma postać następującą:

$$\delta_{t/t-k} = \frac{y_t - y_{t-k}}{y_{t-k}} = \frac{\Delta_{t/t-k}}{y_{t-k}}. \quad (6.4)$$

Przyjmując – analogicznie jak poprzednio – $k = 1$ otrzymujemy łańcuchowy przyrost względny:

$$\delta_{t/t-1} = \frac{y_t - y_{t-1}}{y_{t-1}} = \frac{\Delta_{t/t-1}}{y_{t-1}}. \quad (6.4a)$$

Osobną grupę wskaźników indywidualnych stanowią indeksy dynamiki, które – analogicznie jak omówione powyżej przyrosty – mogą być jednospodstawowe i zmiennospodstawowe. Jednospodstawowy wskaźnik (indeks) dynamiki zdefiniowany jest następującym wzorem:

$$\gamma_{t/t_b} = \frac{y_t}{y_{t_b}}. \quad (6.5)$$

Indeks zmiennospodstawowy przedstawia się następująco:

$$\gamma_{t/t-k} = \frac{y_t}{y_{t-k}}. \quad (6.6)$$

Jeśli przyjmiemy $k = 1$, to otrzymamy łańcuchowy wskaźnik (indeks) dynamiki:

$$\gamma_{t/t-1} = \frac{y_t}{y_{t-1}}. \quad (6.6a)$$

Jak łatwo zauważyć, między przyrostami względnymi i wskaźnikami dynamiki zachodzą następujące związki:

$$\delta_{t/t_b} = \frac{y_t - y_{t_b}}{y_{t_b}} = \gamma_{t/t_b} - 1, \quad (6.7)$$

$$\delta_{t/t-k} = \frac{y_t - y_{t-k}}{y_{t-k}} = \gamma_{t/t-k} - 1, \quad (6.8)$$

$$\delta_{t/t-1} = \frac{y_t - y_{t-1}}{y_{t-1}} = \gamma_{t/t-1} - 1. \quad (6.8a)$$

Podczas wskaźnikowej analizy szeregu czasowego pojawia się niekiedy potrzeba oceny średniego tempa zmian poziomu obserwowanego zjawiska. W takiej sytuacji znajduje zastosowanie średnia geometryczna łańcuchowych indeksów dynamiki. O problemie tym wspomniano już w rozdziale 3. W punkcie 3.2.2.4 podano przykład 3.7, w którym rozważany jest sześćelementowy szereg czasowy ($t = 1, 2, \dots, 6$). Daje to podstawę do obliczenia pięciu ($6 - 1 = 5$) kolejnych wartości łańcuchowego wskaźnika (indeksu) dynamiki zdefiniowanego wzorem (6.6a). Średnią wartość tego wskaźnika obliczamy według wzoru:

$$g_{\gamma_{t/t-1}} = \sqrt[5]{\gamma_{2/1} \cdot \gamma_{3/2} \cdot \gamma_{4/3} \cdot \gamma_{5/4} \cdot \gamma_{6/5}},$$

w którym $g_{\gamma_{t/t-1}}$ oznacza średnią geometryczną łańcuchowego wskaźnika dynamiki; zob. wzór (3.49).

Wykorzystując wzór (6.6a) powyższe wyrażenie można zapisać następująco:

$$g_{\gamma_{t/t-1}} = \sqrt[5]{\frac{y_2}{y_1} \cdot \frac{y_3}{y_2} \cdot \frac{y_4}{y_3} \cdot \frac{y_5}{y_4} \cdot \frac{y_6}{y_5}} = \sqrt[5]{\frac{y_6}{y_1}}.$$

Końcowa postać tego wyrażenia jest szczególnym przypadkiem ogólnego wzoru:

$$g_{\gamma_{t/t-1}} = \sqrt[t_n-1]{\frac{y_{t_n}}{y_1}}, \quad (6.9)$$

w którym t_n oznacza numer ostatniego elementu badanego szeregu czasowego.

Wykorzystując wzór (6.8a) można obliczyć średnią wartość łańcuchowego wskaźnika przyrostu względnego:

$$g_{\delta_{t/t-1}} = g_{\gamma_{t/t-1}} - 1. \quad (6.10)$$

Uzyskany wynik można wyrazić w procentach:

$$g_{\delta_{t/t-1}} = 100 \cdot g_{\gamma_{t/t-1}} - 100 = 100 \cdot (g_{\gamma_{t/t-1}} - 1) \quad [\%]. \quad (6.11)$$

W procentach można też oczywiście wyrażać przyrosty względne i wskaźniki dynamiki.

Przykład 6.2

Liczbę abonentów telefonii przewodowej przypadającą na 1000 mieszkańców przedstawia tablica 6.2. Na podstawie tych danych obliczono przyrosty absolutny i względny, indeksy indywidualne oraz średnie roczne tempo przyrostu liczby abonentów (tablica 6.3).

Tablica 6.2

Abonenci telefonii przewodowej na 1000 mieszkańców

t	Rok	Abonenci y_t	t	Rok	Abonenci y_t
1	1970	32,8	14	1983	60,3
2	1971	34,4	15	1984	63,4
3	1972	36,2	16	1985	66,4
4	1973	38,2	17	1986	69,9
5	1974	40,5	18	1987	73,5
6	1975	42,9	19	1988	78,2
7	1976	45,4	20	1989	82,1
8	1977	47,9	21	1990	86,2
9	1978	50,2	22	1991	93,1
10	1979	52,3	23	1992	102,5
11	1980	54,4	24	1993	114,7
12	1981	56,7	25	1994	129,8
13	1982	57,9			

Źródło: Rocznik Statystyczny GUS, 1975, 1985, 1995.

Obliczenia pomocnicze do przykładu 6.2

t	Rok	Przyrost absolutny		Przyrost względny		Indeksy indywidualne	
		jedno- podstawowy	łańcu- chowy	jedno- podstawowy	łańcu- chowy	jedno- podstawowe	łańcu- chowe
1	1970	–	–	–	–	100,00	–
2	1971	1,6	1,6	4,88	4,88	104,88	104,88
3	1972	3,4	1,8	10,37	5,23	110,37	105,23
4	1973	5,4	2,0	16,46	5,52	116,46	105,52
5	1974	7,7	2,3	23,48	6,02	123,48	106,02
6	1975	10,1	2,4	30,76	5,93	130,79	105,93
7	1976	12,6	2,5	38,41	5,83	138,41	105,83
8	1977	15,1	2,5	46,04	5,51	146,04	105,51
9	1978	17,4	2,3	53,05	4,80	153,05	104,80
10	1979	19,5	2,1	59,45	4,18	159,45	104,18
11	1980	21,6	2,1	65,85	4,02	165,85	104,02
12	1981	23,9	2,3	72,87	4,23	172,87	104,23
13	1982	25,1	1,2	76,52	2,12	176,52	102,12
14	1983	27,5	2,4	83,84	4,15	183,84	104,15
15	1984	30,6	3,1	93,29	5,14	193,29	105,14
16	1985	33,6	3,0	102,44	4,73	202,44	104,73
17	1986	37,1	3,5	113,11	5,27	213,11	105,27
18	1987	40,7	3,6	124,09	5,15	224,09	105,15
19	1988	45,4	4,7	138,41	6,39	238,41	106,39
20	1989	49,3	3,9	150,30	4,99	250,30	104,99
21	1990	53,4	4,1	162,80	4,99	262,80	104,99
22	1991	60,3	6,9	183,84	8,00	283,84	108,00
23	1992	69,7	9,4	212,50	10,10	312,50	110,10
24	1993	81,9	12,2	249,70	11,90	349,70	111,90
25	1994	97,0	15,1	295,73	13,16	395,73	113,16

Źródło: obliczenia własne.

Średnie roczne tempo przyrostu wyrażone w procentach wynosi $\delta = 5,90\%$. Zgodnie ze wzorem (6.10) odpowiada to wartości $\delta = 0,059$. ■

6.2.2. Wskaźniki agregatowe

Wśród agregatowych wskaźników dynamiki wyróżnia się:

- agregatowy indeks wartości,
- agregatowy indeks masy towarowej (ilości),
- agregatowy indeks cen.

Agregatowy indeks wartości (I_w) definiuje się jako stosunek sumy wartości określonego zespołu dóbr w okresie badanym do sumy wartości tych samych dóbr w okresie podstawowym:

$$I_w = \frac{\sum_{i=1}^N q_{1,i} \cdot p_{1,i}}{\sum_{i=1}^N q_{0,i} \cdot p_{0,i}}, \quad i = 1, \dots, N \quad (6.12)$$

gdzie:

- $q_{0,i}$ – ilość dobra i w okresie podstawowym,
- $q_{1,i}$ – ilość dobra i w okresie badanym,
- $p_{0,i}$ – cena dobra i w okresie podstawowym,
- $p_{1,i}$ – cena dobra i w okresie badanym.

Agregatowy indeks wartości (I_w) informuje, o ile wzrosła lub zmalała wartość pewnego agregatu dóbr w okresie badanym w stosunku do okresu podstawowego. Natomiast nie wyjaśnia on, w jakim stopniu zostało to spowodowane zmianami w ilościach, a w jakim – zmianami cen. Odpowiedzi na te pytania dostarczają agregatowe indeksy ilości oraz cen.

Konstrukcja agregatowych indeksów masy towarowej (ilości) oraz indeksów cen sprowadza się do zabiegu standaryzacji wskaźników dynamiki. Polega ona na „unieruchomieniu” jednego z dwóch czynników (q lub p) na poziomie jednego i tego samego okresu.

Agregatowy indeks masy towarowej (ilości):

$$I_q = \frac{\sum_{i=1}^N q_{1,i} \cdot p_{const.}}{\sum_{i=1}^N q_{0,i} \cdot p_{const.}}, \quad i = 1, \dots, N, \quad (6.13)$$

przy czym w zależności od sposobu standaryzacji otrzymuje się:

– agregatowy indeks ilości według Laspeyresa:

$$I_{q.L} = \frac{\sum_{i=1}^N q_{1,i} \cdot p_{0,i}}{\sum_{i=1}^N q_{0,i} \cdot p_{0,i}}, \quad i = 1, \dots, N, \quad (6.14)$$

w którym wagami są ceny z okresu podstawowego;

– agregatowy indeks ilości według Paaschego:

$$I_{q.P} = \frac{\sum_{i=1}^N q_{1,i} \cdot p_{1,i}}{\sum_{i=1}^N q_{0,i} \cdot p_{1,i}}, \quad i = 1, \dots, N, \quad (6.15)$$

w którym wagami są ceny z okresu badanego;

– agregatowy indeks ilości według Fishera:

$$I_{q.F} = \sqrt{I_{q.L} \cdot I_{q.P}}, \quad i = 1, \dots, N, \quad (6.16)$$

który jest średnią geometryczną z agregatowych indeksów ilości Laspeyresa i Paaschego.

Agregatowy indeks cen ma postać:

$$I_P = \frac{\sum_{i=1}^N p_{1,i} \cdot q_{const.}}{\sum_{i=1}^N p_{0,i} \cdot q_{const.}}, \quad i = 1, \dots, N, \quad (6.17)$$

przy czym w zależności od sposobu standaryzacji otrzymuje się:

– agregatowy indeks cen według Laspeyresa:

$$I_{P.L} = \frac{\sum_{i=1}^N q_{0,i} \cdot p_{1,i}}{\sum_{i=1}^N q_{0,i} \cdot p_{0,i}}, \quad i = 1, \dots, N, \quad (6.18)$$

w którym wagami są ilości z okresu podstawowego;

– agregatowy indeks cen według Paaschego:

$$I_{p.P} = \frac{\sum_{i=1}^N q_{1.i} \cdot p_{1.i}}{\sum_{i=1}^N q_{1.i} \cdot p_{0.i}}, \quad i = 1, \dots, N, \quad (6.19)$$

w którym wagami są ilości z okresu badanego;

– agregatowy indeks cen według Fishera:

$$I_{p.F} = \sqrt{I_{p.L} \cdot I_{p.P}}, \quad i = 1, \dots, N, \quad (6.20)$$

który jest średnią geometryczną z agregatowych indeksów cen Laspeyresa i Paaschego.

Przykład 6.3

Spżycie wybranych artykułów (na 1 mieszkańca) przedstawia tablica 6.4. Na podstawie tych danych obliczyć wartości agregatowych indeksów dynamiki Laspeyresa, Paaschego i Fishera oraz skomentować uzyskane wyniki.

Tablica 6.4

Spżycie wybranych artykułów na jednego mieszkańca

<i>i</i>	Produkt	Spżycie		Cena	
		1990	1985	1990	1985
		$q_{1.i}$	$q_{0.i}$	$p_{1.i}$	$p_{0.i}$
1	Ziemniaki (kg)	148	118	0,5	0,35
2	Masło (kg)	7,8	8,5	1,4	1,1
3	Warzywa (kg)	119	105	0,7	0,55
4	Cukier (kg)	44,1	41,3	1,2	1,1
5	Jaja (szt.)	190	220	0,5	0,41

Źródło: Rocznik Statystyczny GUS, 1994.

Obliczenia pomocnicze do przykładu 6.3

<i>i</i>	Produkt	$q_{0i}p_{0i}$	$q_{1i}p_{1i}$	$q_{0i}p_{1i}$	$q_{1i}p_{0i}$
1	Ziemniaki	41,3	74	59	51,8
2	Masło	9,35	10,92	11,9	8,58
3	Warzywa	57,75	83,3	73,5	65,45
4	Cukier	45,43	52,95	49,56	48,51
5	Jaja	90,2	95	110	77,9
Σ	*	244,03	316,14	303,96	252,24

Źródło: obliczenia własne.

Po podstawieniu wartości liczbowych z tablicy 6.5 do odpowiednich wzorów otrzymujemy:

– agregatowy indeks wartości:

$$I_w = 1,295496,$$

– agregatowe indeksy ilości:

$$I_{q-L} = 1,033643,$$

$$I_{q-P} = 1,040071,$$

$$I_{q-F} = 1,036852,$$

– agregatowe indeksy cen:

$$I_{p-L} = 1,245585,$$

$$I_{p-P} = 1,253330,$$

$$I_{p-F} = 1,249451.$$

Uzyskane wartości agregatowych wskaźników masy towarowej (ilości) wskazują na niewielki wzrost spożycia, rzędu 3–4%. W tym samym okresie zaobserwowano bardzo wyraźny wzrost cen, rzędu 24–25%. Wskazują na to agregatowe indeksy cen. ■

6.3. Procedury wyznaczania tendencji rozwojowej

Wyodrębnienie funkcji trendu z szeregu czasowego przeprowadza się na zasadzie eliminacji wpływu wahań okresowych i losowych. Stosuje się tutaj zwykle metodę średnich ruchomych i metodę najmniejszych kwadratów.

6.3.1. Metoda średnich ruchomych

Istota tej metody polega na zastąpieniu zaobserwowanych wartości szeregu czasowego (y_t) przez wartości średnie (\bar{y}_t). Zabieg ten pozwala na częściowe wyeliminowanie wahań okresowych i przypadkowych. W zależności od tego, z ilu kolejnych lat oblicza się średnią, mówi się o trzy-, cztero-, pięcioletniej itd. średniej ruchomej.

Niech

$$y_1, y_2, \dots, y_t, \dots \quad (6.21)$$

oznaczają kolejne wartości szeregu czasowego. Wtedy średnią ruchomą z nieparzystej liczby okresów (momentów) – na przykład z trzech – oblicza się w następujący sposób:

$$\begin{aligned} \bar{y}_2 &= \frac{y_1 + y_2 + y_3}{3}, \\ \bar{y}_3 &= \frac{y_2 + y_3 + y_4}{3}, \\ &\vdots \quad \quad \quad \vdots \end{aligned}$$

i ogólnie

$$\bar{y}_t = \frac{y_{t-1} + y_t + y_{t+1}}{3}. \quad (6.22)$$

W przypadku parzystej liczby okresów (momentów) uzyskaną wartość średniej przyporządkowuje się nie istniejącym w rzeczywistości punktom środkowym.

Liczba okresów, z których oblicza się wartości średnich ruchomych, zależy od celu prowadzonych badań, długości szeregu czasowego oraz od dającego się zaobserwować cyklu wahań okresowych.

Zaletą metody średnich ruchomych jest prostota procedury obliczeniowej oraz jej interpretacja. Natomiast jej wadą – skracanie szeregów czasowych (utrata początkowych i końcowych wyrazów), a także brak możliwości przedstawienia funkcji trendu w postaci analitycznej.

Przykład 6.4

Technikę obliczania średnich ruchomych trzy- i siedmiookresowych przedstawiono na przykładzie dotyczącym liczby zgonów na raka płuc w Polsce na 1000 zgonów ogółem (rozdział 2, tablica 2.2, lata 1967–1991).

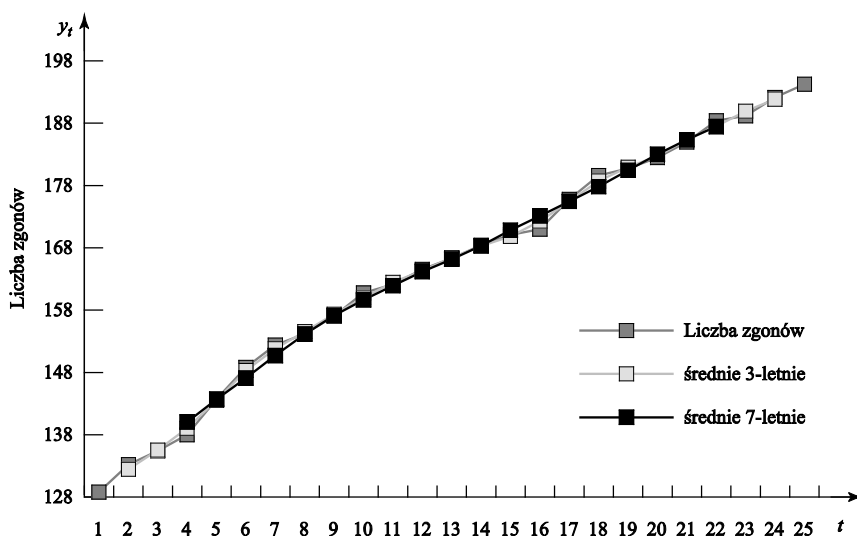
Tablica 6.6

Obliczenia pomocnicze do przykładu 6.4

<i>t</i>	Rok	Szereg empiryczny Liczba zgonów na raka na tysiąc zgonów ogółem	Szereg wyrównany	
			Średnie ruchome 3-letnie	Średnie ruchome 7-letnie
1	1967	128,8	–	–
2	1968	133,2	132,47	–
3	1969	135,4	135,53	–
4	1970	138	139,05	140,05
5	1971	143,75	143,52	143,68
6	1972	148,8	148,32	147,09
7	1973	152,4	151,80	150,71
8	1974	154,2	154,55	154,16
9	1975	157,05	157,33	157,13
10	1976	160,75	159,98	159,64
11	1977	162,15	162,48	161,92
12	1978	164,55	164,35	164,19
13	1979	166,35	166,43	166,19
14	1980	168,4	168,28	168,34
15	1981	170,1	169,84	170,83
16	1982	171	172,30	173,14
17	1983	175,8	175,47	175,45
18	1984	179,6	178,72	177,82
19	1985	180,75	180,95	180,44
20	1986	182,5	182,75	183,04
21	1987	185	185,30	185,37
22	1988	188,4	187,53	187,46
23	1989	189,2	189,92	–
24	1990	192,15	191,87	–
25	1991	194,25	–	–

Źródło: obliczenia własne.

Na podstawie tych danych wyznaczono tendencję rozwojową za pomocą średniej ruchomej trzyletniej i siedmioletniej. Wyrównane szeregi czasowe przedstawiono w tabelicy 6.6 i na rys. 6.2.



Rys. 6.2. Wyrównywanie szeregów czasowych za pomocą średnich ruchomych

Wartość średniej ruchomej trzyletniej – na przykład dla roku 1968 – obliczona na podstawie wzoru (6.26) wynosi:

$$\bar{y}_{1968} = \frac{128,8 + 133,2 + 135,4}{3} = 132,47.$$

Natomiast wartość średniej siedmioletniej – na przykład dla roku 1970 – wynosi:

$$\bar{y}_{1970} = \frac{128,8 + 133,2 + 135,4 + 138 + 143,75 + 148,8 + 152,4}{7} = 140,05. \blacksquare$$

6.3.2. Analityczne wygładzanie szeregów czasowych

Procesy badane w określonym przedziale czasu kształtują się pod wpływem wielu powiązanych ze sobą przyczyn głównych (systematycznych) i przypadkowych (losowych). Przyczyny główne działają na proces w sposób systematyczny, wpływając na jego tempo i kierunek rozwoju, a przyczyny losowe (tzw. wahania losowe) – nieregularnie.

Zespół przyczyn, które kształtują tendencje rozwojową procesu, w zależności od tego, czy są one zależne czy niezależne oraz z jaką funkcją mamy do czynienia, można przedstawić w postaci modelu addytywnego lub modelu multiplikatywnego. Najczęściej wykorzystuje się model addytywny postaci:

$$Y_t = F(t) + G(t) + E(t), \quad (6.23)$$

gdzie:

- Y_t – poziom badanego procesu,
- $F(t)$ – funkcja trendu,
- $G(t)$ – funkcja wahań okresowych,
- $E(t)$ – składnik losowy.

Metoda najmniejszych kwadratów (por. rozdział 3) jest jedną z wielu metod analitycznego wygładzania szeregów czasowych. Istota jej – podobnie jak w analizie regresji i korelacji – polega na tym, że krzywą empiryczną, opisującą rozwój procesu w czasie, aproksymuje się funkcją, która spełnia kryterium minimalizacji sumy kwadratów odchyleń występujących między punktami wyznaczającymi daną funkcję a odpowiadającymi im punktami wyznaczonymi przez krzywą empiryczną.

Niech obrazem tendencji rozwojowej będzie funkcja liniowa:

$$\hat{y}_t = at + b, \quad (6.24)$$

gdzie:

- t – zmienna czasowa $t = 1, \dots, N$,
- a i b – parametry funkcji trendu.

Wykorzystując metodę najmniejszych kwadratów otrzymujemy:

$$a = \frac{N \sum_{t=1}^N y_t t - \sum_{t=1}^N y_t \sum_{t=1}^N t}{N \sum_{t=1}^N t^2 - \left(\sum_{t=1}^N t \right)^2}, \quad (6.25)$$

$$b = \frac{\sum_{t=1}^N y_t - a \sum_{t=1}^N t}{N}, \quad (6.26)$$

gdzie:

- y_t – zaobserwowany poziom badanego zjawiska w okresie (momencie) t ,
- t – kolejny okres (moment) w ciągu $t = 1, 2, \dots, N$.

Parametr a w liniowej funkcji (6.24) nazywamy współczynnikiem trendu. Oznacza on przeciętny wzrost lub spadek poziomu badanego procesu (y_t) w jednostce czasu t . Natomiast parametr b oznacza poziom procesu (y_t) w okresie (momencie) $t = 0$.

Przykład 6.5

Technikę obliczania wartości parametrów liniowej funkcji trendu metodą najmniejszych kwadratów przedstawiono na przykładzie danych zawartych w tabelicy 2.2 (zob. rozdział 2). Niezbędne obliczenia przedstawiono w tabelicy 6.7.

Tabelica 6.7

Obliczenia pomocnicze do przykładu 6.5

Rok	Liczba zgonów na raka na tysiąc zgonów ogółem	t	$y_t \cdot t$	t^2	\hat{y}_t
1967	128,8	1	128,8	1	136,02
1968	133,2	2	266,4	4	138,65
1969	135,4	3	406,2	9	141,28
1970	138,0	4	552,0	16	143,91
1971	143,75	5	718,75	25	146,54
1972	148,8	6	892,8	36	149,17
1973	152,4	7	1066,8	49	151,80
1974	154,2	8	1233,6	64	154,43
1975	157,05	9	1413,45	81	157,06
1976	160,75	10	1607,5	100	159,69
1977	162,15	11	1783,65	121	162,32
1978	164,55	12	1974,6	144	164,95
1979	166,35	13	2162,55	169	167,58
1980	168,4	14	2357,6	196	170,21
1981	170,1	15	2551,5	225	172,84
1982	171,0	16	2736,0	256	175,47
1983	175,8	17	2988,6	289	178,10
1984	179,6	18	3232,8	324	180,73
1985	180,75	19	3434,25	361	183,36
1986	182,5	20	3650,0	400	185,99
1987	185,0	21	3885,0	441	188,62
1988	188,4	22	4144,8	484	191,25
1989	189,2	23	4351,6	529	193,88
1990	192,15	24	4611,6	576	196,51
1991	194,25	25	4856,25	625	199,14
Ogółem	4122,55	325	57007,1	5525	4189,5

Źródło: obliczenia własne.

Otrzymane wyniki wstawiamy do wzorów (6.25) oraz (6.26) i otrzymujemy:

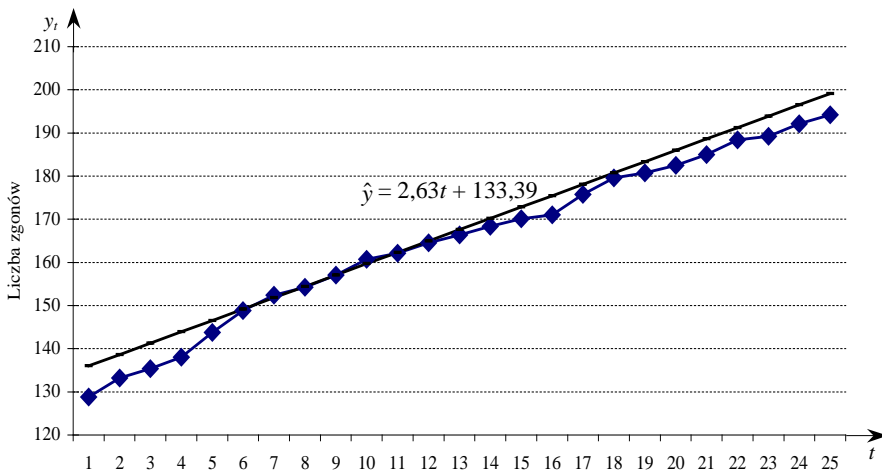
$$a = \frac{25 \cdot 52884,6 - 4122,55 \cdot 300}{25 \cdot 4900 - 300^2} = 2,63,$$

$$b = \frac{4122,55 - 2,63 \cdot 300}{25} = 133,39.$$

Oszacowana liniowa funkcja trendu ma postać:

$$\hat{y}_t = 2,63t + 133,39, \quad t = 0, 1, \dots, 24.$$

Liczbowa ocena współczynnika kierunkowego $a = 2,63$ jest dodatnia, co oznacza, że w latach 1967–1991, pod wpływem działania czystego trendu, liczba zgonów na raka zwiększała się o 2,63. Natomiast $b = 133,39$ jest oceną wyrazu wolnego funkcji trendu i oznacza, że w okresie $t = 0$, tzn. roku 1967, pod wpływem czystego trendu liczba zgonów na raka wynosiłaby 133,39 na tysiąc zgonów ogółem.

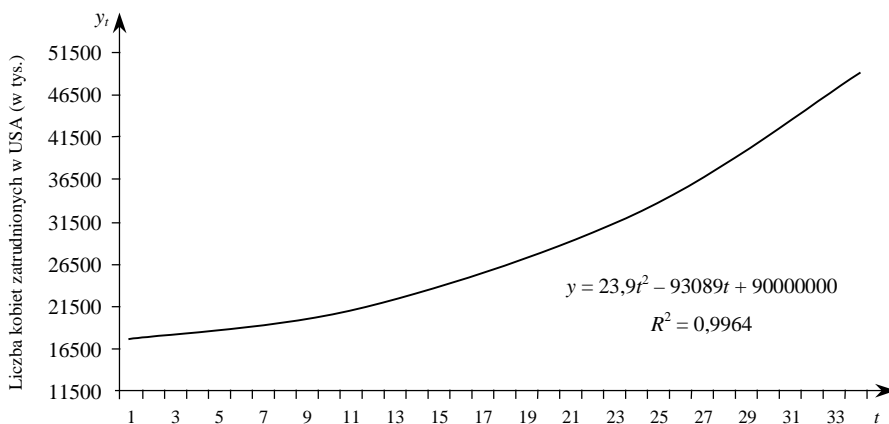


Rys. 6.3. Tendencja rozwojowa zgonów na raka

Graficznym obrazem omawianego szeregu czasowego jest wykres na rys. 6.3. Do poruszanych tu problemów powrócimy w dalszym wykładzie, w punkcie 6.6. □

Przykład 6.1 (zakończenie)

Powróćmy do problemu zatrudnienia kobiet w USA, który omawialiśmy w punkcie 6.1 w celu zilustrowania istoty szeregu czasowego. Dane źródłowe podane są w tabelicy 6.1, a na rys. 6.1 przedstawiono graficznie rozwój tego zjawiska. Analizując ten wykres stwierdziliśmy, że w drugiej połowie badanego okresu można zauważyć odchylenie od liniowego rozwoju zjawiska. Takie spostrzeżenie daje podstawę do podjęcia próby wygładzenia szeregu za pomocą wielomianu wyższego stopnia albo za pomocą funkcji innej klasy. Obszerny przegląd najczęściej stosowanych funkcji Czytelnik znajdzie w tabelicy 3.18. W rozważanym przypadku najbardziej odpowiednie wydało się zastosowanie paraboli drugiego stopnia. Do wyznaczenia parametrów tej paraboli wykorzystano równania podane w tabelicy 3.18, przy czym w miejsce wartości zmiennej objaśniającej (x_i) podstawiono wartości zmiennej czasowej t . Efekty wykonanych obliczeń przedstawiono na rys. 6.4.



Rys. 6.4. Tendencja rozwojowa zatrudnienia kobiet w USA w latach 1949–1982

Dla celów interpretacyjnych obliczono również wartość współczynnika determinacji R^2 . Charakterystykę tę omówiono dokładnie w punkcie 3.3.4. W analizie szeregów czasowych R^2 charakteryzuje dobroć dopasowania wybranej funkcji trendu do danych empirycznych. Funkcja ta jest tym lepiej dobrana, im bliższa jedności jest wartość R^2 . W rozważanym przypadku parabola drugiego stopnia bardzo dobrze odwzorowuje tendencję rozwojową badanego zjawiska (zatrudnienia kobiet w USA). O wysokiej wartości R^2 decyduje również to, że dane empiryczne reprezentują bardzo wysoki stopień agregacji. Przypomnijmy

(zob. tablica 6.1), że są to dane roczne. Nie są więc one obciążone zmiennością sezonową w ciągu roku. ■

6.4. Wyodrębnianie wahań sezonowych (okresowych)

Do wyodrębnienia i analizy wahań sezonowych występujących w szeregach czasowych wykorzystuje się wskaźniki sezonowości.

W pierwszym etapie postępowania, zmierzającego do wyznaczenia wartości tych wskaźników, każdą wartość empiryczną y_t ($t = 1, 2, \dots, N$) zastępuje się wartością z_t obliczaną według wzoru:

$$z_t = \frac{y_t}{\hat{y}_t}, \quad (6.27)$$

w którym \hat{y}_t oznacza wartość funkcji zastosowanej do aproksymacji trendu (tendencji rozwojowej), odpowiadającą momentowi (okresowi) t .

W drugim etapie postępowania dokonuje się zmiany indeksów określających położenie każdej wartości z_t (a tym samym również y_t) w ogólnym zbiorze wartości empirycznych. Dokonuje się mianowicie transformacji:

$$z_t \rightarrow z_{ij}, \quad (6.28)$$

przy czym: $t = 1, 2, \dots, N$; $i = 1, 2, \dots, l$; $j = 1, 2, \dots, k$; $N = l \cdot k$.

Indeks i oznacza numer podokresu, natomiast j jest numerem kolejnej realizacji y_t w danym podokresie. Jeśli na przykład obserwowano jakieś zjawisko w ciągu pięciu lat rejestrując kolejne miesięczne poziomy tego zjawiska, to:

– $N = 60$, albowiem uzyskany szereg czasowy składa się z sześćdziesięciu elementów. Mamy mianowicie:

$$y_1, y_2, \dots, y_l, \dots, y_{60},$$

– $k = 5$, albowiem każda jednoimienna obserwacja powtórzona jest pięć razy. Mamy mianowicie $k = 5$ obserwacji dotyczących stycznia, $k = 5$ obserwacji dotyczących lutego itd.,

– $l = 12$, albowiem wyróżniono dwanaście podokresów (miesiące) w każdym rocznym cyklu obserwowanego zjawiska. Tak więc, na przykład, z_{24} oznacza wartość z_t uzyskaną według wzoru (6.27), dotyczącą lutego ($i = 2$) w czwartym roku ($j = 4$) badania danego zjawiska.

W trzecim etapie postępowania wyznacza się wartości wskaźników sezonowości. Wartości te oblicza się według wzoru:

$$p_i = \frac{\sum_{j=1}^k z_{ij}}{k}. \quad (6.29)$$

Dodajmy, że p_i jest wskaźnikiem sezonowości dotyczącym i -tego podokresu. Należy więc obliczyć łączne l wskaźników sezonowości.

Wskaźniki obliczane według wzoru (6.29) nazywane są surowymi wskaźnikami sezonowości. Są one obciążone wahaniami przypadkowymi. W celu wyeliminowania tych wahań należy podzielić każdą wartość p_i ($i = 1, 2, \dots, l$) przez ich średnią arytmetyczną. W rezultacie otrzymuje się tzw. oczyszczone wskaźniki sezonowości p'_i . Mamy zatem:

$$p'_i = \frac{p_i}{\frac{\sum_{i=1}^l p_i}{l}} = p_i \frac{l}{\sum_{i=1}^l p_i}, \quad (6.30)$$

przy czym wartość:

$$u = \frac{l}{\sum_{i=1}^l p_i} \quad (6.31)$$

nazywana jest współczynnikiem korygującym.

Przykład 6.6

Przypomnijmy (zob. rozdział 2), że jedną z metod pozyskiwania informacji jest ciągła rejestracja. W taki sposób gromadzone są w szczególności informacje o podstawowych zjawiskach demograficznych, takich jak urodziny i zgony, zawierane małżeństwa, rozwody.

W tablicy 6.8 przedstawiono dane dotyczące zgonów zarejestrowanych w Polsce, w kolejnych miesiącach (od stycznia do grudnia) pięciu lat, od 1990 do 1994 roku. Na podstawie tych danych sporządzono wykres przedstawiony na rys. 6.5. Położenie punktów empirycznych na wykresie wskazuje na wyraźną sezonowość obserwowanego zjawiska. Liczba zgonów w miesiącach letnich jest niższa niż w pozostałych miesiącach. Największa liczba zgonów ma miejsce

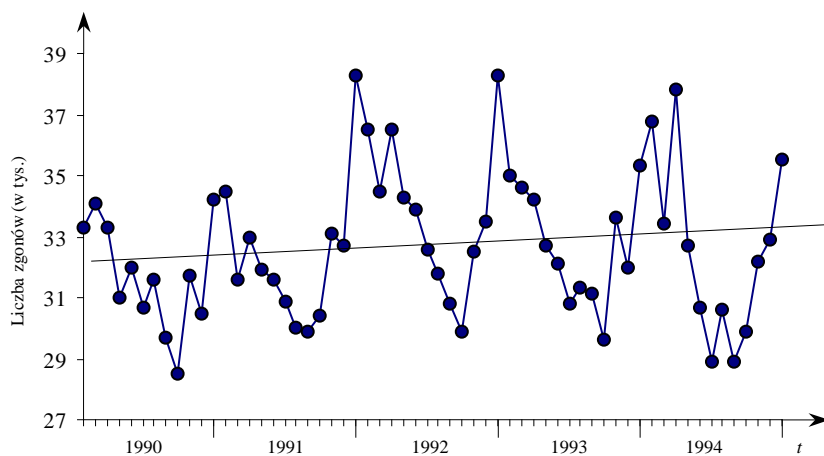
w miesiącach zimowych. Ta prawidłowość powtarza się systematycznie w kolejnych latach.

Tablica 6.8

Liczba zgonów w Polsce

Miesiąc	Zgony (w tys.)				
	1990	1991	1992	1993	1994
I	33,3	34,5	36,5	35,0	36,8
II	34,1	31,6	34,5	34,6	33,4
III	33,3	33,0	36,5	34,2	37,8
IV	31,0	31,9	34,3	32,7	32,7
V	32,0	31,6	33,9	32,1	30,7
VI	30,7	30,9	32,6	30,8	28,9
VII	31,6	30,0	31,8	31,3	30,6
VIII	29,7	29,9	30,8	31,1	28,9
IX	28,5	30,4	29,9	29,6	29,9
X	31,7	33,1	32,5	33,6	32,2
XI	30,5	32,7	33,5	32,0	32,9
XII	34,2	38,3	38,3	35,3	35,5
Σ	380,6	387,9	405,1	392,3	390,3

Źródło: Roczniki Statystyczne GUS, 1990–1994.



Rys. 6.5. Liczba zgonów w Polsce w kolejnych miesiącach lat 1990–1994

Na podstawie przedstawionych powyżej danych obliczymy wartości wskaźników sezonowości oraz wyznaczmy poziom wahań sezonowych. Pierwszy etap niezbędnych obliczeń przedstawiono w tabelicy 6.9. Obliczenia te miały na celu wyeliminowanie trendu z obserwowanego zjawiska. W tym celu wyznaczono równanie tendencji rozwojowej:

$$\hat{y}_t = 0,01085t + 32,2723.$$

Wykorzystując to równanie obliczono wartości \hat{y}_t dla $t = 1, 2, \dots, 60$, a następnie wyznaczono wartości z_t według wzoru (6.27).

Tabela 6.9

Obliczenia pomocnicze do przykładu 6.6

Rok	Miesiąc	t	y_t	t^2	$y_t \cdot t$	\hat{y}_t	y_t / \hat{y}_t
1	2	3	4	5	6	7	8
1990	1	1	33,3	1	33,3	32,28	1,0316
	2	2	34,1	4	68,2	32,29	1,0561
	3	3	33,3	9	99,9	32,30	1,0310
	4	4	31,0	16	124	32,32	0,9592
	5	5	32,0	25	160	32,33	0,9898
	6	6	30,7	36	184,2	32,34	0,9493
	7	7	31,6	49	221,2	32,35	0,9768
	8	8	29,7	64	237,6	32,36	0,9178
	9	9	28,5	81	256,5	32,37	0,8804
	10	10	31,7	100	317	32,38	0,9790
	11	11	30,5	121	335,5	32,39	0,9416
	12	12	34,2	144	410,4	32,40	1,0556
1991	1	13	34,5	169	448,5	32,41	1,0645
	2	14	31,6	196	442,4	32,42	0,9747
	3	15	33,0	225	495	32,44	1,0173
	4	16	31,9	256	510,4	32,45	0,9831
	5	17	31,6	289	537,2	32,46	0,9735
	6	18	30,9	324	556,2	32,47	0,9516
	7	19	30,0	361	570	32,48	0,9236
	8	20	29,9	400	598	32,49	0,9203
	9	21	30,4	441	638,4	32,50	0,9354
	10	22	33,1	484	728,2	32,51	1,0181
	11	23	32,7	529	752,1	32,52	1,0055
	12	24	38,3	576	919,2	32,53	1,1774

cd. tablicy 6.9

1	2	3	4	5	6	7	8
1992	1	25	36,5	625	912,5	32,54	1,1217
	2	26	34,5	676	897	32,55	1,0599
	3	27	36,5	729	985,5	32,57	1,1207
	4	28	34,3	784	960,4	32,58	1,0528
	5	29	33,9	841	983,1	32,59	1,0402
	6	30	32,6	900	978	32,60	1,0000
	7	31	31,8	961	985,8	32,61	0,9752
	8	32	30,8	1024	985,6	32,62	0,9442
	9	33	29,9	1089	986,7	32,63	0,9163
	10	34	32,5	1156	1105	32,64	0,9957
	11	35	33,5	1225	1172,5	32,65	1,0260
	12	36	38,3	1296	1378,8	32,66	1,1727
1993	1	37	35,0	1369	1295	32,67	1,0713
	2	38	34,6	1444	1314,8	32,68	1,0588
	3	39	43,2	1521	1333,8	32,70	1,0459
	4	40	32,7	1600	1308	32,71	0,9997
	5	41	32,1	1681	1316,1	32,72	0,9811
	6	42	30,8	1764	1293,6	32,73	0,9410
	7	43	31,3	1849	1345,9	32,74	0,9560
	8	44	31,1	1936	1368,4	32,75	0,9496
	9	45	29,6	2025	1332	32,76	0,9035
	10	46	33,6	2116	1545,6	32,77	1,0253
	11	47	32,0	2209	1504	32,78	0,9762
	12	48	35,3	2304	1694,4	32,79	1,0765
1994	1	49	36,8	2401	1803,2	32,80	1,1220
	2	50	33,4	2500	1670	32,81	1,0180
	3	51	37,8	2601	1927,8	32,83	1,1514
	4	52	32,7	2704	1700,4	32,84	0,9957
	5	53	30,7	2809	1627,1	32,85	0,9346
	6	54	28,9	2916	1560,6	32,86	0,8795
	7	55	30,6	3025	1683	32,87	0,9307
	8	56	28,9	3136	1618,4	32,88	0,8790
	9	57	29,9	3249	1704,3	32,89	0,9091
	10	58	32,2	3364	1867,6	32,90	0,9787
	11	59	32,9	3481	1941,1	32,91	0,9997
	12	60	35,5	3600	2130	32,92	1,0784

Źródło: obliczenia własne.

Dalszy ciąg obliczeń przedstawiono w tabelicy 6.10, której ogólny układ odpowiada tabelicy 6.8, zawierającej surowe wyniki obserwacji. Każda zaobserwowana wartość y_t w tabelicy 6.10 reprezentowana jest przez odpowiadającą jej wartość $z_t = y_t/\hat{y}_t$, przy czym – zgodnie z (6.28) – zastosowano tu podwójne ideksowanie (z_{ij}), albowiem zachodziła konieczność rozróżnienia miesięcy (i) i lat (j), a także konieczność wyróżnienia miesięcy jednoimiennych (ustalone i). □

Tabelica 6.10

Obliczenia pomocnicze do przykładu 6.6

Miesiąc (i)	Lata (j)					$\sum_{j=1}^5 \frac{y_t}{\hat{y}_t}$	P_i
	1	2	3	4	5		
	1990	1991	1992	1993	1994		
	y_t / \hat{y}_t						
1	2	3	4	5	6	7	8
1	1,0316	1,0645	1,1217	1,0713	1,1220	5,4111	1,0822
2	1,0561	0,9747	1,0599	1,0588	1,0180	5,1675	1,0335
3	1,0310	1,0173	1,1207	1,0459	1,1514	5,3663	1,0733
4	0,9592	0,9831	1,0528	0,9997	0,9957	4,9905	0,9981
5	0,9898	0,9735	1,0402	0,9811	0,9346	4,9192	0,9838
6	0,9493	0,9516	1,0000	0,9410	0,8795	4,7214	0,9443
7	0,9768	0,9236	0,9752	0,9560	0,9307	4,7623	0,9525
8	0,9178	0,9203	0,9442	0,9496	0,8790	4,6109	0,9222
9	0,8804	0,9354	0,9163	0,9035	0,9091	4,5447	0,9089
10	0,9790	1,0181	0,9957	1,0253	0,9787	4,9968	0,9994
11	0,9416	1,0055	1,0260	0,9762	0,9997	4,949	0,9898
12	1,0556	1,1774	1,1727	1,0765	1,0784	5,5606	1,1121
$\sum_{i=1}^{12}$	–	–	–	–	–	–	12,0001

Źródło: obliczenia własne.

W analizowanym przykładzie występuje bardzo słaby trend wzrostowy. Jeśli uznamy, że zaobserwowany trend jest praktycznie pomijalny, to przedstawione powyżej rozumowanie można znacznie uprościć. W takiej sytuacji wskaźniki sezonowości p_i można wyznaczyć według wzoru:

$$p_i = \frac{\bar{y}_i}{\bar{y}}, \quad (6.32)$$

gdzie:

$$\bar{y}_i = \frac{\sum_{j=1}^k y_{ij}}{k}, \quad (6.33)$$

natomiast \bar{y} oznacza ogólny średni poziom zjawiska i wynika ze wzoru:

$$\bar{y} = \frac{\sum_{i=1}^l \sum_{j=1}^k y_{ij}}{k \cdot l}. \quad (6.34)$$

Dodajmy, iż mając wskaźniki sezonowości p_i można wyznaczyć absolutne poziomy wahań sezonowych Q_i ($i = 1, 2, \dots, l$).

W tym celu należy posłużyć się wzorem:

$$Q_i = \bar{y}_i - \bar{y} = p_i \bar{y} - \bar{y} = \bar{y}(p_i - 1). \quad (6.35)$$

Zauważmy, że:

$$\sum_{i=1}^l Q_i = 0. \quad (6.36)$$

Przykład 6.6 (zakończenie)

W tabelicy 6.11 przedstawiono obliczenia niezbędne do wyznaczenia wskaźników sezonowości według uproszczonego wzoru (6.32).

Wykorzystując wzór (6.34) mamy:

$$\bar{y} = \frac{1956,2}{60} = 32,60.$$

Mając ogólną średnią (wyrażoną w tysiącach zgonów) i kolejne średnie miesięczne \bar{y}_i ($i = 1, 2, \dots, 12$) możemy wyznaczyć wartości wskaźników sezonowości P_i . Dla stycznia ($i = 1$) mamy:

$$P_1 = \frac{\bar{y}_1}{\bar{y}} \cdot 100 = \frac{35,22}{32,60} \cdot 100 = 108,04 \text{ [%]}.$$

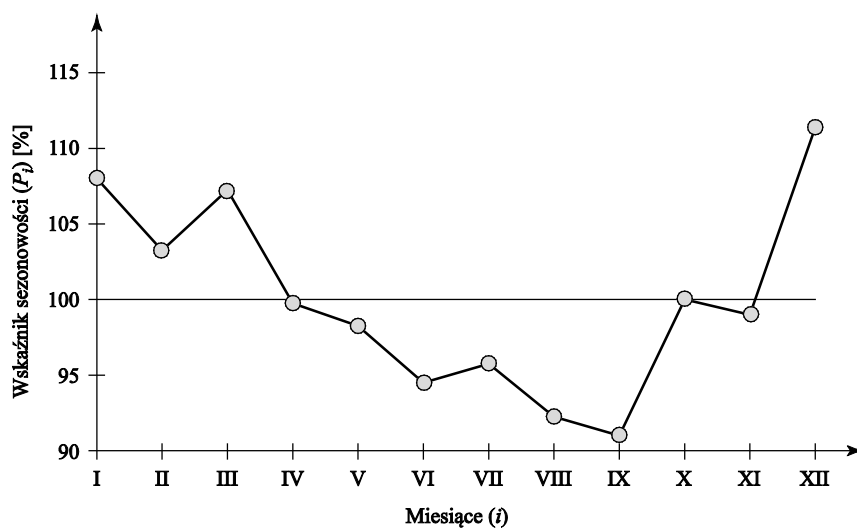
Postępując analogicznie w odniesieniu do pozostałych miesięcy ($i = 2, 3, \dots, 12$) uzyskujemy wartości zestawione w kolumnach 4 i 5 tabelicy 6.11 i na rys. 6.6.

Tablica 6.11

Obliczenia pomocnicze do przykładu 6.6

Miesiąc (i)	$\sum_{j=1}^5 y_j$	\bar{y}_{iy}	P_i (w %)	Zaokrąglone wskaźniki sezonowości (w %)	Q_i
1	2	3	4	5	6
1	176,1	35,22	108,04	8	2,62
2	168,2	33,64	103,19	3	1,04
3	174,8	34,96	107,24	7	2,36
4	162,6	32,52	99,75	0	-0,07
5	160,3	32,06	98,34	-2	-0,54
6	153,9	30,78	94,42	-6	-1,82
7	155,3	31,06	95,28	-5	-1,54
8	150,4	30,08	92,27	-8	-2,52
9	148,3	29,66	90,98	-9	-2,94
10	163,1	32,62	100,06	0	0,02
11	161,6	32,32	99,14	-1	-0,28
12	181,6	36,32	111,41	11	3,72
$\sum_{i=1}^{12}$	1956,2	391,24	1200,00	0,00	0,00

Źródło: obliczenia własne.



Rys. 6.6. Sezonowość zgonów w latach 1990–1994

Interpretacja charakterystyk przedstawionych w kolumnach 4 i 5 tablicy 6.11 oraz na rys. 6.6 jest następująca: liczba zgonów w styczniu jest wyższa od przeciętnej miesięcznej o 8,04%, w lutym o 3,19%, w marcu o 7,24% i w grudniu o 11,42%. W pozostałych miesiącach, z wyjątkiem października, liczba zgonów ogółem jest niższa od przeciętnej miesięcznej, w kwietniu o 0,25%, w maju o 1,66%, w czerwcu o 5,58%, w lipcu o 4,72%, w sierpniu o 7,73%, we wrześniu o 9,02% i w listopadzie o 0,86%. Jak łatwo zauważyć, rezultaty tych przybliżonych obliczeń w nieznacznym stopniu odbiegają od wyników przedstawionych w tablicy 6.10.

Obliczając wartości różnicy:

$$Q_i = \bar{y}_i - \bar{y}, \quad i = 1, 2, \dots, 12$$

uzyskujemy oszacowania absolutnych wahań sezonowych przedstawione w kolumnie 6 tablicy 6.11. I tak na przykład w styczniu liczba zgonów jest wyższa od przeciętnej miesięcznej ($\bar{y} = 32,6$ tys.) o 2,62 tys. osób, w lutym o 1,04 tys., w marcu o 2,36 tys., w grudniu o 3,72 tys. ■

6.5. Wyodrębnianie wahań przypadkowych (losowych)

Jeśli mamy do czynienia z szeregiem czasowym odpowiadającym modelowi addytywnemu (6.23), to każdą zrealizowaną wartość y_t należy rozważać jako następującą sumę:

$$y_t = \hat{y}_t + g_t + e_t, \quad (6.37)$$

gdzie:

- \hat{y}_t – składowa systematyczna, wynikająca z trendu,
- g_t – składowa uwarunkowana wahaniami okresowymi (sezonowymi),
- e_t – składowa losowa (przypadkowa).

Wynika stąd bezpośrednio, że:

$$e_t = y_t - \hat{y}_t - g_t = (y_t - \hat{y}_t) - g_t. \quad (6.38)$$

Jeśli w obserwowanym szeregu czasowym nie występują wahania okresowe (sezonowe), to $g_t = 0$.

W takiej sytuacji:

$$y_t = \hat{y}_t + e_t \quad (6.39)$$

i w konsekwencji

$$e_t = y_t - \hat{y}_t. \quad (6.40)$$

W procesie pomiaru intensywności losowych odchyleń szeregu czasowego od funkcji trendu można wykorzystać pojęcie zmienności resztowej, zdefiniowane w punkcie 3.3.4, przy omawianiu regresji i korelacji. W rozważanej tu sytuacji owa zmienność resztowa może być mierzona za pomocą sumy:

$$\sum_{t=1}^N e_t^2, \quad (6.41)$$

gdzie e_t wynika ze wzoru (6.38) lub (6.40). Czym mniejsza jest wartość tej sumy, tym mniejszy jest udział czynników losowych (przypadkowych) w kształtowaniu się obserwowanego szeregu czasowego.

Wartość sumy (6.41) zależy nie tylko od intensywności pojawiania się losowych odchyleń od funkcji trendu, ale również od liczby elementów (N) w szeregu czasowym. W celu uniezależnienia się od tego czynnika, do pomiaru intensywności wahań losowych (przypadkowych) w badanym szeregu czasowym wykorzystuje się wariancję resztową (s_e^2), którą definiuje się następująco:

$$s_e^2 = \frac{\sum_{t=1}^N e_t^2}{N - k}, \quad (6.42)$$

gdzie k oznacza liczbę parametrów funkcji trendu, których wartości wyznacza się na podstawie dostępnego materiału empirycznego. W przypadku trendu liniowego $k = 2$. Resztowe odchylenie standardowe:

$$s_e = \sqrt{s_e^2} \quad (6.43)$$

informuje, o ile przeciętnie odchyła się pojedynczy element szeregu czasowego (y_t) od odpowiadającej mu wartości \hat{y}_t , wyznaczonej na podstawie funkcji trendu. Jeśli w badanym szeregu czasowym występują wahania okresowe (sezonowe), to s_e wyraża przeciętną odległość punktu $y_t + g_t$ od punktu \hat{y}_t .

Jak już podkreślono powyżej (zob. punkt 6.4.2), do oceny intensywności losowych (przypadkowych) fluktuacji szeregu czasowego można wykorzystać współczynniki determinacji (R^2) i indeterminacji ($1 - R^2$), zdefiniowane w punkcie 3.3.4. Po odpowiedniej zmianie symboli mamy:

$$1 - R^2 = \frac{\sum_{t=1}^N (y_t - \hat{y}_t)^2}{\sum_{t=1}^N (y_t - \bar{y})^2} \quad (6.44)$$

i w konsekwencji

$$R^2 = 1 - \frac{\sum_{t=1}^N (y_t - \hat{y}_t)^2}{\sum_{t=1}^N (y_t - \bar{y})^2}, \quad (6.45)$$

gdzie \bar{y} oznacza średni poziom badanego zjawiska w obserwowanym szeregu czasowym, natomiast znaczenie pozostałych symboli jest takie samo jak dotychczas.

Przykład 6.5 (zakończenie)

Powróćmy do danych przedstawionych w tabelicy 2.2, opisujących częstość zgonów spowodowanych chorobami nowotworowymi. Na rys. 6.3 pokazano rozwój tego zjawiska na przestrzeni 25 lat, w latach 1967–1991. Na rysunku tym zaznaczono położenie funkcji trendu, a niezbędne obliczenia przedstawiono w tabelicy 6.7.

Położenie punktów empirycznych na rys. 6.3 upoważnia do przyjęcia założenia, że w obserwowanym szeregu czasowym nie występują wahania okresowe (sezonowe). Pozwala to traktować wszystkie odchylenia od funkcji trendu jako odchylenia przypadkowe, zgodnie ze wzorem (6.40).

W tabelicy 6.12 przedstawiono obliczenia niezbędne do wyznaczenia wariancji resztowej i resztowego odchylenia standardowego. Wykorzystując wzory (6.42) i (6.43) otrzymujemy:

$$s_e^2 = \frac{125,53}{25 - 2} = \frac{125,53}{23} = 5,4578,$$

$$s_e = \sqrt{5,4578} = 2,3362.$$

Tablica 6.12

Analiza odchyłeń przypadkowych w szeregu czasowym

Rok	t	y_t	\hat{y}_t	$e_t = y_t - \hat{y}_t$	e_t^2
1967	1	128,8	133,39	-4,59	21,0681
1968	2	133,2	136,01	-2,81	7,8961
1969	3	135,4	138,64	-3,24	10,4976
1970	4	138	141,27	-3,27	10,6929
1971	5	143,75	143,89	-0,14	0,0196
1972	6	148,8	146,52	2,28	5,1984
1973	7	152,4	149,15	3,25	10,5625
1974	8	154,2	151,77	2,43	5,9049
1975	9	157,05	154,40	2,65	7,0225
1976	10	160,75	157,02	3,73	13,9129
1977	11	162,15	159,65	2,50	6,2500
1978	12	164,55	162,28	2,27	5,1529
1979	13	166,35	164,90	1,45	2,1025
1980	14	168,4	167,53	0,87	0,7569
1981	15	170,1	170,15	-0,05	0,0025
1982	16	171	172,78	-1,78	3,1684
1983	17	175,8	175,41	0,39	0,1521
1984	18	179,6	178,03	1,57	2,4649
1985	19	180,75	180,66	0,09	0,0081
1986	20	182,5	183,28	-0,78	0,6084
1987	21	185	185,91	-0,91	0,8281
1988	22	188,4	188,54	-0,14	0,0196
1989	23	189,2	191,16	-1,96	3,8416
1990	24	192,15	193,79	-1,64	2,6896
1991	25	194,25	196,42	-2,17	4,7089
Σ	*	4122,55	4122,55	0,00	125,5300

Źródło: obliczenia własne.

Zatem liczba zgonów na raka w pojedynczym roku, w okresie 1967–1991, odchyła się losowo od funkcji trendu o około 2,34. Jeśli przyjąć, że odchylenia losowe podlegają rozkładowi normalnemu, to można sformułować następujące wnioski:

– rzeczywiście zarejestrowana liczba zgonów na raka w roku t z prawdopodobieństwem 0,6826 będzie należeć do przedziału $(\hat{y}_t - 2,34; \hat{y}_t + 2,34)$,

– z prawdopodobieństwem 0,9545 zarejestrowana liczba zgonów będzie należeć do przedziału:

$$(\hat{y}_t - 2 \cdot 2,34; \hat{y}_t + 2 \cdot 2,34),$$

$$(\hat{y}_t - 4,68; \hat{y}_t + 4,68),$$

– z prawdopodobieństwem 0,9973, czyli niemal na pewno, zarejestrowana liczba zgonów na raka w roku t znajdzie się w przedziale:

$$(\hat{y}_t - 3 \cdot 2,34; \hat{y}_t + 3 \cdot 2,34),$$

$$(\hat{y}_t - 7,02; \hat{y}_t + 7,02).$$

Przedstawione powyżej wnioski wynikają z elementarnych właściwości rozkładu normalnego omówionych w punkcie 1.4.2.2.

Na podstawie danych zawartych w tablicach 6.7 i 6.12 obliczamy wartości współczynników R^2 i $1 - R^2$, wykorzystując w tym celu wzory (6.44) i (6.45). Z tablicy 6.12 odczytujemy:

$$\sum_{t=1}^{25} (y_t - \hat{y}_t)^2 = \sum_t e_t^2 = 125,53.$$

Obliczamy wartość:

$$\sum_{t=1}^{25} (y_t - \bar{y})^2 = \sum_{t=1}^{25} y_t^2 - \frac{\left(\sum_{t=1}^{25} y_t \right)^2}{25} = 688907,7225 - \frac{(4122,55)^2}{25} = 9090,9824.$$

Wykorzystując wzór (6.44) otrzymujemy:

$$1 - R^2 = \frac{125,53}{9090,9824} = 0,0138 \text{ (1,38\%)}.$$

W konsekwencji, na podstawie wzoru (6.45), mamy:

$$R^2 = 1 - 0,0138 = 0,9862 \text{ (98,62\%)}.$$

Tak więc 98,62% zmienności obserwowanej w ciągu wartości y_t wyjaśnia liniowa funkcja trendu $\hat{y}_t = 2,63t + 133,39$. Tylko 1,38% ogólnej zmienności obserwowanej w ciągu wartości $y_1, y_2, \dots, y_t, \dots, y_{25}$ nie można wyjaśnić przyjętą postacią funkcji trendu. Owe 1,38% ogólnej zmienności przypisuje się działaniu czynników przypadkowych (losowych). ■

6.6. Monitorowanie procesów za pomocą kart kontrolnych

W połowie lat dwudziestych ubiegłego stulecia opracowano procedury statystyczne służące do monitorowania procesów, przede wszystkim produkcyjnych, dla potrzeb zarządzania jakością. Ich twórcą był W.A. Shewhart¹. Procedury te nazwano kartami kontrolnymi (ang. *control chart*), co miało podkreślać fakt, iż były one pomyślane jako procedury graficzne. Wiązało się to z bardzo ograniczonymi możliwościami ówczesnych środków obliczeniowych. Należy pamiętać, że w tamtych latach inżynier kierujący procesem produkcyjnym nie dysponował współczesnymi środkami obliczeniowymi. W przypadku konieczności wykonania niezbędnych obliczeń mógł on posłużyć się co najwyżej mechanicznym dwudziałaniowym arytmetrem wykonującym operacje dodawania i odejmowania, a także suwakiem logarytmicznym, za pomocą którego mógł wykonywać operacje mnożenia i dzielenia, przy czym uzyskiwane w ten sposób wyniki obciążone były niekiedy znacznymi błędami. Jak już powiedziano powyżej, podstawowym przeznaczeniem opracowanych przez W.A. Shewharta kart kontrolnych było sterowanie procesami dla potrzeb zarządzania jakością. Nie zmienia to jednak faktu, że procedury te mogą być z powodzeniem wykorzystywane do monitorowania dowolnego procesu, jeśli tylko monitorowanie to polega na śledzeniu zmian mierzalnych charakterystyk tego procesu. Charakterystyki te nazywane są zmiennymi diagnostycznymi, albowiem na podstawie poczynionych obserwacji formułuje się diagnozy o stanie obserwowanego procesu. Idea kart kontrolnych powstała i rozwijała się równoległe z klasyczną teorią weryfikacji hipotez statystycznych, omówioną w poprzednim rozdziale. Karty kontrolne Shewharta są w istocie odpowiednio zorganizowanymi sekwencjami testów istotności, przedstawionych w punktach 5.2.1 i 5.3.1. W przypadku kart kontrolnych opracowanych przez W.A. Shewharta zmienne diagnostyczne są więc jednowymiarowymi zmiennymi losowymi, najczęściej o rozkładzie normalnym albo o rozkładzie zbieżnym do normalnego. W latach późniejszych opracowano również karty kontrolne dla dwuwymiarowych zmiennych diagnostycznych. Ten typ kart kontrolnych pominiemy w dalszych rozważaniach. Omówimy natomiast karty kontrolne sum skumulowanych, oparte na sekwencyjnych metodach weryfikacji hipotez statystycznych, omówionych w podrozdziale 5.9.

¹ Walter Andrew Shewhart (1891–1967), amerykański matematyk i statystyk. Pierwszą pracę na temat stosowania kart kontrolnych w bieżącej kontroli jakości opublikował w 1926 r.

6.6.1. Karty kontrolne Shewharta

6.6.1.1. Karta kontrolna \bar{x}

W poprzednim rozdziale rozważaliśmy problemy związane z weryfikacją istotności różnicy między uzyskaną empirycznie średnią arytmetyczną z próby (\bar{x}) a pewną wartością μ_0 , ustalaną najczęściej na podstawie przesłanek pozastatystycznych. W postępowaniu zmierzającym do sprawdzenia (zweryfikowania) istotności tej różnicy formuluje się jedną spośród hipotez zerowych (5.18)–(5.20), a następnie weryfikuje się ją wobec jednej spośród hipotez alternatywnych (5.21)–(5.23). Jeśli zachodzi potrzeba powtarzania tego postępowania w odniesieniu do pewnego ciągu zmiennych losowych $X_1, X_2, \dots, X_t, \dots$, mających normalne rozkłady prawdopodobieństwa, o jednakowych i znanych odchyleniach standardowych, to można w tym celu posłużyć się tzw. *kartą kontrolną \bar{x}* . Podstawowym elementem takiej karty jest tzw. *tor kontrolny*, którego zasadnicze elementy przedstawiono schematycznie na rys. 6.7. Na osi odciętych prostokątnego układu współrzędnych odkłada się kolejne wartości t . Na osi rzędnych natomiast – wartości \bar{x}_t , wynikające ze wzoru:

$$\bar{x}_t = \frac{\sum_{i=1}^{n_t} x_{t,i}}{n_t}. \quad (6.46)$$

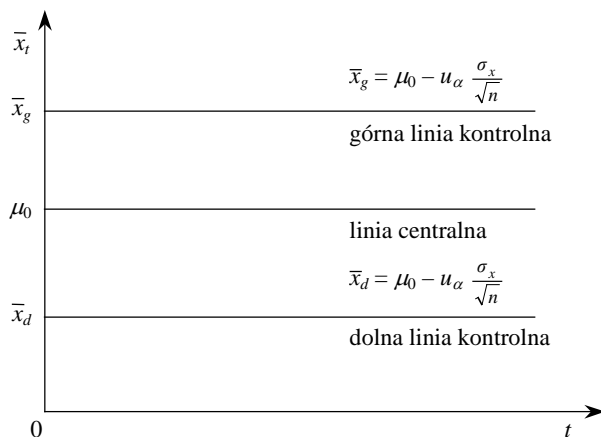
Z punktu $\bar{x}_t = \mu_0$ wystawiamy półprostą, równoległą do osi t . Jest to tzw. *linia centralna*. Powyżej i poniżej linii centralnej zaznaczamy wartości krytyczne dla statystyki \bar{x}_t . Górne wartości krytyczne $\bar{x}_{g,t}$ wynikają z równania:

$$\bar{x}_{g,t} = \mu_0 + u_\alpha \frac{\sigma_x}{\sqrt{n_t}}. \quad (6.47)$$

Jeśli poziom istotności (α) nie zmienia się w czasie badania i jeśli liczebność pobieranych do badania próbek jest stała ($\bigwedge_t n_t = n$), lub przynajmniej przedziałami stała, to wartości krytyczne dla statystyki \bar{x}_t nie zależą od t w czasie całego badania lub przynajmniej w pewnych przedziałach wyróżnionych na osi t . Mamy więc wówczas:

$$\bar{x}_g = \mu_0 + u_\alpha \frac{\sigma_x}{\sqrt{n}}. \quad (6.48)$$

Jest to równanie *górnjej linii kontrolnej* zaznaczonej na rys. 6.7.



Rys. 6.7. Schemat karty kontrolnej \bar{x}

Analogicznie przedstawia się problem dolnych wartości krytycznych dla statystyki \bar{x}_t . Mamy więc:

$$\bar{x}_{d,t} = \mu_0 - u_\alpha \frac{\sigma_x}{\sqrt{n_t}} \quad (6.49)$$

oraz

$$\bar{x}_d = \mu_0 - u_\alpha \frac{\sigma_x}{\sqrt{n}}. \quad (6.50)$$

Jest to równanie *dolnej linii kontrolnej* wykreślonej na rys. 6.7. Przy sporządzaniu tego rysunku założono bowiem, że zarówno n_t jak i α pozostają stałe w czasie całego badania. Należy podkreślić, że tylko wówczas, gdy założenia te spełnione są w praktyce, zastosowanie karty kontrolnej \bar{x} , zamiast kolejnego stosowania zwykłego testu u (zob. rozdział 5), przynosi istotną oszczędność w pracochłonności badania.

Zauważmy, że przedstawione powyżej równania (6.47)–(6.50) wynikają z przekształcenia podstawowego wzoru (5.25). Biorąc kolejno prawą stronę tego wyrażenia ze znakiem plus (co odpowiada wartościom $\bar{x} > \mu_0$) i ze znakiem minus (co odpowiada wartościom $\bar{x} < \mu_0$), otrzymujemy:

$$\bar{x} = \mu_0 \pm u \frac{\sigma_x}{\sqrt{n}},$$

a ustalając u na poziomie u_α , gdzie $\alpha = P(U > u_\alpha)$, lewą stronę powyższego wzoru możemy traktować jako wartość krytyczną dla średniej arytmetycznej obliczonej na podstawie wyników badania próbki losowej o licznosci n .

Jeśli w każdym punkcie t weryfikujemy hipotezę zerową $H_0: \mu_{x,t} = \mu_0$ wobec hipotezy alternatywnej $H_1: \mu_{x,t} \neq \mu_0$, to wykorzystujemy zarówno dolną, jak i górną linię kontrolną. W przypadku gdy weryfikowane kolejno hipotezy zerowe mają postać $H_0: \mu_{x,t} \leq \mu_0$, a hipotezy alternatywne $H_1: \mu_{x,t} > \mu_0$, wykorzystywana jest tylko górna linia kontrolna \bar{x}_g . Natomiast dolna linia kontrolna (\bar{x}_d) wykorzystywana jest tylko wówczas, gdy w każdym punkcie pomiarowym t weryfikowana jest hipoteza zerowa $H_0: \mu_{x,t} \geq \mu_0$ wobec hipotezy alternatywnej $H_1: \mu_{x,t} < \mu_0$. Na diagram przeglądowy przedstawiony na rys. 6.7 nanosimy kolejne punkty (t, \bar{x}_t) . Tak długo nie ma podstaw do odrzucenia weryfikowanej hipotezy zerowej, jak długo żaden punkt (t, \bar{x}_t) nie przekroczy odpowiedniej linii kontrolnej. Weryfikowaną hipotezę odrzucamy natomiast, gdy pojawi się punkt (t, \bar{x}_t) wykraczający poza linię kontrolną. W statystycznej kontroli jakości punkt taki nosi nazwę *punktowego sygnału o rozregulowaniu procesu*. W tabelicy 6.13 zestawiono reguły generowania tych sygnałów w różnych wariantach karty kontrolnej \bar{x} .

Tabela 6.13

Karta kontrolna \bar{x} ; generowanie sygnałów o rozregulowaniu monitorowanego procesu

Hipoteza		Punktowy sygnał o rozregulowaniu ma miejsce wówczas, gdy	Prawdopodobieństwo fałszywego sygnału o rozregulowaniu
zerowa H_0	alternatywna H_1		
$\mu_{x,t} = \mu_0$	$\mu_{x,t} \neq \mu_0$	$\bar{x}_t \leq \bar{x}_{d,t}$ lub $\bar{x}_t \geq \bar{x}_{g,t}$	2α
$\mu_{x,t} \leq \mu_0$	$\mu_{x,t} > \mu_0$	$\bar{x}_t \geq \bar{x}_{g,t}$	α
$\mu_{x,t} \geq \mu_0$	$\mu_{x,t} < \mu_0$	$\bar{x}_t \leq \bar{x}_{d,t}$	α

Przykład 6.7

Wagę laboratoryjną wykorzystywano do seryjnego przygotowywania 10-gramowych odważek pewnej substancji. W celu sprawdzenia, czy waga ta funkcjonuje prawidłowo, wykonywano okresowo $n = 5$ niezależnych kontrolnych wa-

żeń, umieszczając każdorazowo na szalce 10-gramowy wzorcowy odważnik zamiast odważki substancji. Załóżmy też, że precyzja wagi jest znana i wynosi $\sigma_x = 0,2$ g. Podczas trzech początkowych kontroli uzyskano wyniki przedstawione w tabelicy 6.14. Dokonajmy analizy uzyskanych wyników za pomocą karty kontrolnej \bar{x} , przyjmując przy tym poziom istotności $\alpha = 0,05$. Formułujemy hipotezy:

$$\bigwedge_t H_0: \mu_{x,t} = \mu_0 = 10, \quad \bigwedge_t H_1: \mu_{x,t} \neq \mu_0 = 10.$$

Tabela 6.14

Wyniki doświadczenia

<i>i</i>	<i>t</i>			
	1	2	3	...
	x_{1i}	x_{2i}	x_{3i}	...
1	10,1	10,3	10,4	...
2	10,2	10,1	10,5	...
3	9,6	9,7	10,3	...
4	9,8	10,0	10,0	...
5	10,1	9,8	10,5	...
\sum_1^5	49,8	49,9	51,7	...
\bar{x}_t	9,96	9,98	10,34	...

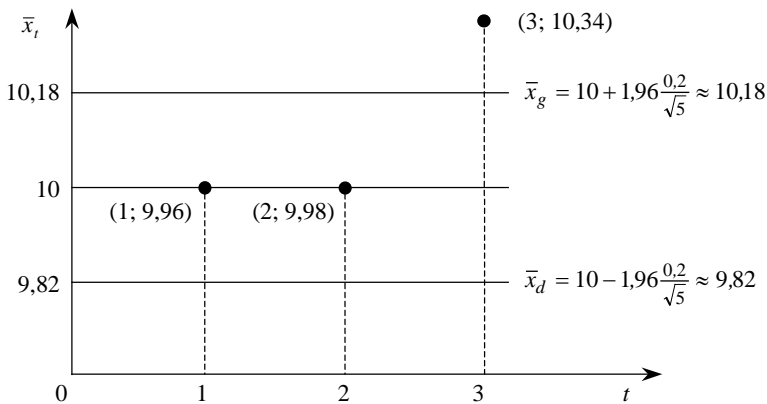
Źródło: badania własne.

Mamy więc do czynienia z dwustronnym schematem kontrolnym i w związku z tym należy wyznaczyć zarówno \bar{x}_d , jak i \bar{x}_g , przyjmując przy tym w równaniach (6.48) i (6.50) $u_{\alpha/2} = u_{0,025} = 1,96$. Po podstawieniu wartości do tych wzorów i wykonaniu obliczeń otrzymujemy:

$$\bar{x}_d = 10 - 1,96 \cdot \frac{0,2}{\sqrt{5}} \approx 10 - 0,18 = 9,82, \quad \bar{x}_g = 10 + 1,96 \cdot \frac{0,2}{\sqrt{5}} \approx 10 + 0,18 = 10,18.$$

Mając te wartości, możemy sporządzić diagram przeglądowy według schematu przedstawionego na rys. 6.7, na który będziemy kolejno nanosić wartości \bar{x}_t z ostatniego wiersza tabelicy 6.14. Diagram ten przedstawiono na rys. 6.8. Jak widać, zarówno $\bar{x}_1 = 9,96$, jak i $\bar{x}_2 = 9,98$ nie dają podstaw do odrzucenia wery-

fikowanej hipotezy zerowej. Sygnałem ostrzegającym o rozregulowaniu używanej wagi technicznej jest natomiast punkt o współrzędnych $t = 3$ i $\bar{x}_t = 10,34$, albowiem leży on ponad górną linią kontrolną. Prawdopodobieństwo zdarzenia losowego polegającego na tym, że sygnał ten jest fałszywy, wynosi $\alpha = 0,05$. W praktyce oznacza to, że wagę należy poddać regulacji, a odważki substancji przygotowane w okresie od $t = 2$ do $t = 3$ należy ponownie przeważyć.



Rys. 6.8. Diagram przegładowy karty kontrolnej \bar{x}

Karta kontrolna \bar{x} , która – jak to już powiedziano powyżej – jest w istocie niczym innym jak sekwencją testów u , stwarza pewne nowe możliwości w zakresie analizy tzw. sekwencyjnych sygnałów ostrzegawczych, których to możliwości nie daje standardowy test u . W celu omówienia tego problemu rozpatrzmy przypadek, gdy w kolejnych punktach t weryfikowana jest hipoteza $H_0: \mu_{x,t} = \mu_0$ wobec hipotezy alternatywnej $H_1: \mu_{x,t} \neq \mu_0$. Jeśli prawdziwa jest hipoteza zeroowa, to kolejno uzyskiwane wartości \bar{x}_t powinny się lokować losowo po obu stronach linii centralnej. Jeśli $\mu_{x,t} > \mu_0$, to powinno to doprowadzić do sygnału $\bar{x}_t > \bar{x}_g$, i analogicznie, jeśli $\mu_{x,t} < \mu_0$, to powinniśmy uzyskać sygnał $\bar{x}_t < \bar{x}_d$. Jeśli jednak odchylenia $\mu_{x,t}$ względem μ_0 są niewielkie, to mimo że kolejne hipotezy zerowe $H_0: \mu_{x,t} = \mu_0$ są fałszywe, prawdopodobieństwo uzyskania sygnału punktowego tylko nieznacznie wzrasta ponad α . W takiej sytuacji celowe może się okazać śledzenie położenia kolejnych punktów (t, \bar{x}_t) na diagramie przegładowym. Wyraźne grupowanie się tych punktów tylko powyżej albo tylko poniżej linii centralnej może być podstawą do odrzucenia hipotezy zerowej:

$$H_0: \bigwedge_t \mu_{x,t} = \mu_0 \quad (6.51)$$

na korzyść hipotezy alternatywnej:

$$H_1: \bigvee_r \bigwedge_{t>r} \mu_{x,t} > \mu_0 \quad (6.52)$$

albo też na korzyść hipotezy alternatywnej:

$$H_{-1}: \bigvee_r \bigwedge_{t>r} \mu_{x,t} < \mu_0. \quad (6.53)$$

Technika weryfikacji hipotezy (6.51) wobec hipotez alternatywnych (6.52), (6.53) może polegać, na przykład, na obliczaniu w każdym punkcie t wartości:

$$\bar{x}_{t,s} = \frac{1}{s}(\bar{x}_t + \bar{x}_{t-1} + \dots + \bar{x}_{t-s+1}), \quad (6.54)$$

gdzie s jest długością badanej sekwencji.

Linie kontrolne dla tej dodatkowej charakterystyki mają następujące równania:

$$\bar{x}_{g,s} = \mu_0 + u_\alpha \frac{\sigma_x}{\sqrt{s \cdot n}}, \quad (6.55)$$

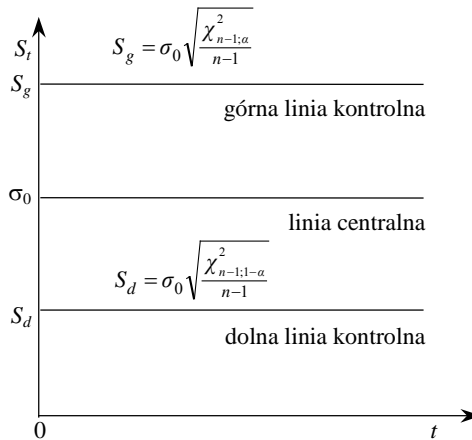
$$\bar{x}_{d,s} = \mu_0 - u_\alpha \frac{\sigma_x}{\sqrt{s \cdot n}}. \quad (6.56)$$

Zauważmy, że jeśli wartości α i n są ustalone, to $\bar{x}_{g,s} < \bar{x}_g$ i $\bar{x}_{d,s} > \bar{x}_d$. Jeśli $\bar{x}_{t,s} > \bar{x}_{g,s}$, to należy odrzucić hipotezę zerową (6.51) na korzyść hipotezy alternatywnej (6.52). Jeśli natomiast $\bar{x}_{t,s} < \bar{x}_{d,s}$, to hipotezę zerową (6.51) należy odrzucić na korzyść hipotezy alternatywnej (6.53).

6.6.1.2. Karta kontrolna $\bar{x} - s$

Karta kontrolna $\bar{x} - s$ została opracowana dla potrzeb statystycznej kontroli jakości do jednoczesnej kontroli wartości oczekiwanej (μ_x) i odchylenia standardowego (σ_x) zmiennej diagnostycznej X . Jest to karta dwutorowa. Tor \bar{x} służy do weryfikacji hipotez dotyczących wartości oczekiwanej μ_x . Jego funkcjonowanie jest analogiczne do funkcjonowania karty kontrolnej \bar{x} omówionej w punkcie 6.6.1.1. Obecnie omówimy konstrukcję i funkcjonowanie toru kontrolnego s , albowiem może on być efektywnie wykorzystany również poza staty-

styczną kontrolą jakości, gdy zachodzi potrzeba weryfikacji omówionych w rozdziale 5 hipotez zerowych (5.84), (5.85) i (5.86) w odniesieniu do pewnego ciągu zmiennych losowych $X_1, X_2, \dots, X_t, \dots$, o hipotetycznie jednorodnych odchyleniach standardowych. Konstrukcję toru kontrolnego s przedstawiono schematycznie na rys. 6.9. Na osi odciętych odkładamy kolejne chwile pobierania próbek z obserwowanego procesu stochastycznego $\{X_t\}$. Na osi rzędnych odkładamy natomiast wartości odchyłeń standardowych z próbki (s_t). Linię centralną wykreślamy na wysokości σ_0 i jest ona równoległa do osi t niezależnie od zmian liczebności próbek (n_t) i ewentualnych zmian poziomu istotności (α) w czasie badania.

Rys. 6.9. Tor kontrolny s

Równania linii kontrolnych przedstawiają się następująco:

$$s_{g,t} = \sigma_0 \sqrt{\frac{\chi_{n_t-1;x}^2}{n_t-1}}, \quad (6.57)$$

$$s_{d,t} = \sigma_0 \sqrt{\frac{\chi_{n_t-1;1-x}^2}{n_t-1}}. \quad (6.58)$$

W przypadku gdy $\bigwedge_t n_t = n$, mamy:

$$s_g = \sigma_0 \sqrt{\frac{\chi_{n-1, \alpha}^2}{n-1}}, \quad (6.59)$$

$$s_d = \sigma_0 \sqrt{\frac{\chi_{n-1, 1-\alpha}^2}{n-1}}. \quad (6.60)$$

Linie kontrolne wykreślone według tych równań są równoległe do linii centralnej, jak to pokazano na rys. 6.9.

W przypadku gdy pobierane do badania próbki są liczne ($n_t > 30$), linie kontrolne mogą być wykreślane według równań:

$$s_{g,t} = \sigma_0 \left(1 + \frac{u_\alpha}{\sqrt{2n_t}} \right), \quad (6.61)$$

$$s_{d,t} = \sigma_0 \left(1 - \frac{u_\alpha}{\sqrt{2n_t}} \right). \quad (6.62)$$

Po ustaleniu liczebności próbek ($\bigwedge_t n_t = n$) wzory te przybierają postać:

$$s_g = \sigma_0 \left(1 + \frac{u_\alpha}{\sqrt{2n}} \right), \quad (6.63)$$

$$s_d = \sigma_0 \left(1 - \frac{u_\alpha}{\sqrt{2n}} \right). \quad (6.64)$$

Linie kontrolne wykreślone według tych równań, podobnie jak linie kontrolne wykreślone według (6.59) i (6.60), są równoległe do linii centralnej. W przypadku gdy w każdym punkcie t weryfikujemy hipotezę zerową $H_0: \sigma_{x,t} = \sigma_0$ wobec hipotezy alternatywnej $H_1: \sigma_{x,t} \neq \sigma_0$, wykreślamy zarówno górną, jak i dolną linię kontrolną, a prawdopodobieństwo popełnienia błędu pierwszego rodzaju wynosi wówczas 2α . Na diagram przeglądowy (rys. 6.9) przygotowany według omówionych powyżej zasad nanosimy kolejne punkty (t, s_t). Weryfikowana hipoteza zerowa zostaje odrzucona wówczas, gdy pojawi się tzw. punktowy sygnał o rozregulowaniu procesu (tablica 6.15). Jeśli chcielibyśmy, by w schemacie dwustronnym prawdopodobieństwo fałszywego sygnału wynosiło α , to w odpowiednich równaniach linii kontrolnych należałoby przyjąć $1 - \alpha/2$ zamiast $1 - \alpha$ oraz $\alpha/2$ zamiast α .

Karta kontrolna $\bar{x} - s$; generowanie sygnałów o rozregulowaniu procesu na torze s

Hipoteza		Punktowy sygnał o rozregulowaniu ma miejsce wówczas, gdy	Prawdopodobieństwo fałszywego sygnału o rozregulowaniu
zerowa H_0	alternatywna H_1		
$\sigma_{x,t} = \sigma_0$	$\sigma_{x,t} \neq \sigma_0$	$S_{x,t} \geq S_g$ lub $S_{x,t} \leq S_d$	2α
$\sigma_{x,t} \leq \sigma_0$	$\sigma_{x,t} > \sigma_0$	$S_{x,t} \geq S_g$	α
$\sigma_{x,t} \geq \sigma_0$	$\sigma_{x,t} < \sigma_0$	$S_{x,t} \leq S_d$	α

Zwróćmy jeszcze uwagę na to, że karta kontrolna $\bar{x} - s$ nie jest prostym złożeniem toru kontrolnego \bar{x} i toru kontrolnego s_x . Ujawnia się to szczególnie wyraźnie w procesie analizy sekwencyjnych sygnałów o rozregulowaniu procesu. W omawianym schemacie kontrolnym z sygnałem sekwencyjnym mamy do czynienia nie tylko wówczas, gdy kolejne punkty (t, s_t) grupują się w pobliżu linii kontrolnej, ale także wówczas, gdy kolejne punkty (t, \bar{x}_t) wykazują nadmierny rozrzut na torze kontrolnym \bar{x} .

6.6.1.3. Karta kontrolna $\bar{x} - r$

Zakres stosowania, konstrukcja i zasady funkcjonowania karty kontrolnej $\bar{x} - r$ są takie same jak w przypadku karty $\bar{x} - s$. Różnica polega jedynie na tym, że do weryfikacji hipotez zerowych dotyczących odchyłeń standardowych ($\sigma_{x,t}$) kolejnych zmiennych losowych $X_1, X_2, \dots, X_t, \dots$ wykorzystuje się rozstęp z próbki ($r_{n,t}$).

Przypomnijmy, że jeśli realizacje zmiennej losowej $x_1, \dots, x_i, \dots, x_n$ uporządkujemy w ten sposób, że $x_{(1)} \leq \dots \leq x_{(i)} \leq \dots \leq x_{(n)}$, to *rozstępem z próbki* nazywamy różnicę $r_n = x_{(n)} - x_{(1)}$. Uzyskiwane w ten sposób wartości są realizacjami zmiennej losowej R_n . Indeks n umieszczamy przy symbolu zmiennej dla podkreślenia, że jej rozkład zależy od liczebności próbki. Wykorzystując podane w rozdziale 4 zależności między odchyleniem standardowym (σ_x) a rozstępem z próbki (r_n), możemy napisać równanie linii centralnej i równania linii kontrolnych. Linia centralna wykreślana jest na poziomie:

$$r_{n,0} = \sigma_0 d_n. \quad (6.65)$$

Położenie linii kontrolnych wyznacza się według następujących wzorów:

$$r_{n,g} = \sigma_0(d_n + u_\alpha k_n), \quad (6.66)$$

$$r_{n,d} = \sigma_0(d_n - u_\alpha k_n). \quad (6.67)$$

Tablica 6.16

Karta kontrolna $\bar{x} - r$; generowanie sygnałów na torze r

Hipoteza		Punktowy sygnał o rozregulowaniu ma miejsce wówczas, gdy	Prawdopodobieństwo fałszywego sygnału o rozregulowaniu
zerowa H_0	alternatywna H_1		
$\sigma_{x,t} = \sigma_0$	$\sigma_{x,t} \neq \sigma_0$	$r_{n,t} \geq r_{n,g}$ lub $r_{n,t} \leq r_{n,d}$	2α
$\sigma_{x,t} \leq \sigma_0$	$\sigma_{x,t} > \sigma_0$	$r_{n,t} \geq r_{n,g}$	α
$\sigma_{x,t} \geq \sigma_0$	$\sigma_{x,t} < \sigma_0$	$r_{n,t} \leq r_{n,d}$	α

Wartości współczynników d_n i k_n podane są w tablicy V, zamieszczonej w aneksie. Zauważmy, że w przypadku karty kontrolnej $\bar{x} - r$ liczebność próbki powinna być ustalona ($\bigwedge_t n_t = n$). Zmiana liczebności próbki powoduje tu bowiem nie tylko zmianę położenia linii kontrolnych (jak w przypadku kart kontrolnych \bar{x} i $\bar{x} - s$), ale również zmianę położenia linii centralnej. Przyjmując w równaniach (6.66), (6.67) $u_{\alpha/2}$ (zamiast u_α), otrzymujemy dwustronny schemat kontrolny, w którym prawdopodobieństwo fałszywego sygnału wynosi α . Wygenerowanie sygnału punktowego powoduje odrzucenie weryfikowanej hipotezy zerowej i przyjęcie odpowiedniej hipotezy alternatywnej. Brak punktowego sygnału o rozregulowaniu oznacza brak podstaw do odrzucenia weryfikowanej hipotezy zerowej. Zasady generowania sygnałów podano w tablicy 6.16.

6.6.1.4. Karta kontrolna z i karta kontrolna w

Weźmy obecnie pod uwagę przypadek, gdy monitorowanie procesu polega na śledzeniu wskaźnika struktury albo frakcji. Weryfikacji poddaje się wówczas jedną spośród zdefiniowanych w poprzednim rozdziale hipotez zerowych (5.110)–(5.112). Wybraną hipotezę zerową weryfikuje się wobec jednej spośród hipotez alternatywnych (5.113)–(5.115). Jeśli zachodzi potrzeba powtarzania tych działań w odniesieniu do pewnego ciągu zmiennych losowych $Z_1, Z_2, \dots, Z_t, \dots$, zdefiniowanych w rozdziale 1 wzorem (1.62), albo też w odniesieniu do odpowiedniego ciągu zmiennych losowych $W_1, W_2, \dots, W_t, \dots$, (zdefiniowanych wzorem (1.67)), to można posłużyć się kartą kontrolną z albo kartą kontrolną w . W literaturze angielskojęzycznej karta kontrolna z oznaczana jest symbolem np , natomiast karta kontrolna w symbolem p . Konstrukcja i zasady funkcjonowania tych procedur kontrolnych są w ogólnych założeniach takie same jak w przy-

padku omówionych powyżej kart kontrolnych \bar{x} , $\bar{x} - s$ i $\bar{x} - r$. Dlatego też ograniczymy się tu do podania równań linii centralnych i linii kontrolnych oraz schematów decyzyjnych.

Karta kontrolna z

Z monitorowanego procesu pobiera się kolejne próbki losowe o stałej liczności n . Na podstawie wyników badania każdej próbki oblicza się wartość następującej charakterystyki:

$$z_t = \sum_{i=1}^n x_{t,i}, \quad (6.68)$$

gdzie $x_{t,i}$ jest i -tą realizacją zmiennej losowej zero-jedynkowej X_t . Obliczone wartości z_t nanosi się na diagram przeglądowy o konstrukcji pokazanej na rys. 6.7–6.9. Linię centralną wykreśla się na poziomie np_0 , przy czym n jest licznoscią próbki losowej pobieranej jednorazowo do badania, natomiast p_0 oznacza postulowaną albo graniczną wartość wskaźnika struktury (frakcji), ustaloną zazwyczaj na podstawie przesłanek pozastatystycznych. Położenie górnej linii kontrolnej (z_g) i dolnej linii kontrolnej (z_d) wyznacza się na podstawie następujących wzorów:

$$z_g = np_0 + u_\alpha \sqrt{np_0(1-p_0)}, \quad (6.69)$$

$$z_d = np_0 - u_\alpha \sqrt{np_0(1-p_0)}, \quad (6.70)$$

przy czym u_α jest taką wartością zmiennej U o rozkładzie $N(0; 1)$, że: $P(U > u_\alpha) = \alpha$.

Tablica 6.17

Karta kontrolna z ; generowanie sygnałów o rozregulowaniu procesu

Hipoteza		Punktowy sygnał o rozregulowaniu ma miejsce wówczas, gdy	Prawdopodobieństwo fałszywego sygnału o rozregulowaniu
zerowa H_0	alternatywna H_1		
$p = p_0$	$p \neq p_0$	$z_t \geq z_g$ lub $z_t \leq z_d$	2α
$p \leq p_0$	$p > p_0$	$z_t \geq z_g$	α
$p \geq p_0$	$p < p_0$	$z_t \leq z_d$	α

Przykład 6.8

Monitorowano proces produkcji pewnego wyrobu sztukowego. Z bieżącej produkcji pobierano próbki losowe o stałej licznosci $n = 50$. Każdą jednostkę produktu oceniano alternatywnie, kwalifikując ją do kategorii jednostek zgodnych z wymaganiami albo do kategorii jednostek niezgodnych z wymaganiami. Zgodnie ze wzorem (1.61), jednostkom zgodnym przyporządkowywano wartość $x_{t,i} = 0$, natomiast jednostkom niezgodnym przyporządkowywano wartość $x_{t,i} = 1$. Zrealizowane wartości obserwowanej zmiennej diagnostycznej Z obliczano według wzoru (6.68), przy czym $n = 50$. Uzyskiwane w ten sposób wartości z_t oznaczają liczbę jednostek produktu nie spełniających wymagań jakościowych, w próbie o licznosci $n = 50$. W początkowych 10 krokach postępowania kontrolnego uzyskano następujące rezultaty:

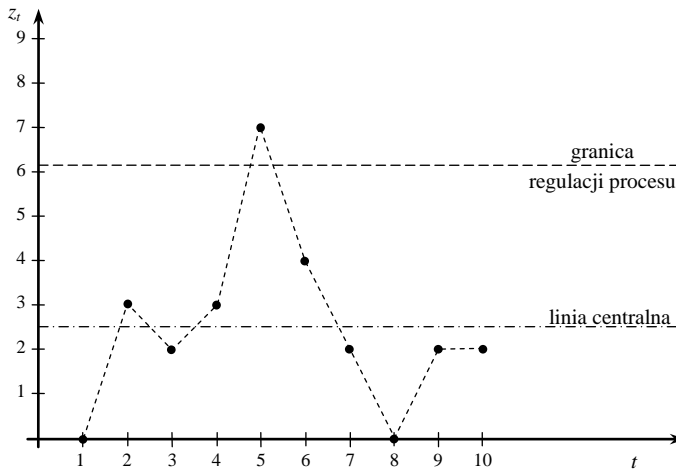
t	1	2	3	4	5	6	7	8	9	10	...
z_t	0	3	2	3	7	4	2	0	2	2	...

Kierownictwo przedsiębiorstwa ustaliło, że największa dopuszczalna frakcja jednostek produktu nie spełniających wymagań jakościowych może wynosić $p_0 = 5\%$. Celem monitorowania procesu było wykrycie punktów jego rozregulowania. W tym celu w każdym kroku postępowania weryfikowano hipotezę zerową (5.111) wobec hipotezy alternatywnej (5.115). Jeśli prawdziwa jest hipoteza zerowa (5.111), a więc jeśli rzeczywista frakcja niezgodnych jednostek produktu (p) nie przekracza ustalonej wartości $p_0 = 5\%$, to monitorowany proces jest uregulowany. Jeśli natomiast $p > p_0 = 5\%$, a więc jeśli prawdziwa jest hipoteza alternatywna (5.115), to monitorowany proces jest rozregulowany. Prawdopodobieństwo wyemitowania fałszywego sygnału o rozregulowaniu procesu ustalono na poziomie $\alpha = 0,01$. Diagram przeglądowy przedstawiono na rys. 6.10.

W przedstawionej sytuacji tylko górna linia kontrolna jest granicą regulacji procesu. Położenie tej linii na diagramie przeglądowym wyznaczono według równania (6.69):

$$z_g = 50 \cdot 0,05 + 2,326 \cdot \sqrt{50 \cdot 0,05 \cdot 0,95} = 2,5 + 2,326 \cdot 1,541 \approx 6,1.$$

Linie centralną wykreślmy na poziomie 2,5. Jest to oczekiwana liczba jednostek produktu nie spełniających wymagań jakościowych, w próbie o licznosci $n = 50$, gdy $p = p_0 = 0,05$.



Rys. 6.10. Diagram przeglądowy karty kontrolnej z

Karta kontrolna w

Z analizy podanych powyżej równań wynika, że karta kontrolna z zachowuje swą praktyczną użyteczność wówczas, gdy kolejne próbki losowe pobierane do badania mają taką samą liczebność. Jeśli warunek ten nie jest spełniony, to nie tylko wartości z_g i z_d , ale również położenie linii centralnej powinno być korygowane po każdej zmianie liczebności próbek. Konieczność dokonywania tylu zmian przekreśla użyteczność karty z jako graficznego algorytmu weryfikacji hipotez zerowych (5.110)–(5.112). Znacznie mniej pracochłonny może się wówczas okazać algorytm numeryczny, jakim jest test u , omówiony w podrozdziale 5.4. Lepszym rozwiązaniem jest w takiej sytuacji zastosowanie karty kontrolnej w. Karta ta jest mniej wrażliwa na zmiany liczebności próbek. Mniejsza wrażliwość polega na tym, że linia centralna wykreślana jest na poziomie p_0 i jej położenie nie zależy od liczebności próbki (n). Korekt wymagają tylko linie kontrolne. Obserwowana charakterystyka z próbki ma postać:

$$w_t = \frac{z_t}{n_t}, \quad (6.71)$$

gdzie wartość z_t wyznacza się według wzoru (6.68), natomiast n_t jest liczebnością próbki o numerze t . Obliczone w ten sposób wartości w_t nanosi się na diagram przeglądowy (rys. 6.7–6.10). Położenie górnej linii kontrolnej (w_g) i dolnej linii kontrolnej (w_d) wyznacza się według następujących wzorów:

$$w_{g,t} = p_0 + u_\alpha \sqrt{\frac{p_0(1-p_0)}{n_t}}, \quad (6.72)$$

$$w_{d,t} = p_0 - u_\alpha \sqrt{\frac{p_0(1-p_0)}{n_t}}. \quad (6.73)$$

We wzorach tych – analogicznie jak poprzednio – u_α jest kwantylem rzędu $1 - \alpha$ zmiennej losowej U o rozkładzie $N(0; 1)$, a więc taką wartością, że $P(U > u_\alpha) = \alpha$. Linie kontrolne wykreślone według tych równań są liniami łamanymi o wierzchołkach $(t, w_{g,t})$ oraz $(t, w_{d,t})$ i powinny być one stopniowo przedłużane w kolejnych krokach badania.

Tablica 6.18

Karta kontrolna w ; generowanie sygnałów o rozregulowaniu procesu

Hipoteza		Punktowy sygnał o rozregulowaniu ma miejsce wówczas, gdy	Prawdopodobieństwo fałszywego sygnału o rozregulowaniu
zerowa H_0	alternatywna H_1		
$p = p_0$	$p \neq p_0$	$w_t \geq w_g$ lub $w_t \leq w_d$	2α
$p \leq p_0$	$p > p_0$	$w_t \geq w_g$	α
$p \geq p_0$	$p < p_0$	$w_t \leq w_d$	α

Jeśli można zapewnić stałą licznosc w kolejnych próbkach pobieranych do badania, to kolejne wartości w_t oblicza się według wzoru:

$$w_t = \frac{z_t}{n}. \quad (6.74)$$

Uproszczeniu ulegają też oczywiście wzory, według których wyznacza się położenie linii kontrolnych. Mamy mianowicie:

$$w_g = p_0 + u_\alpha \sqrt{\frac{p_0(1-p_0)}{n}}, \quad (6.75)$$

$$w_d = p_0 - u_\alpha \sqrt{\frac{p_0(1-p_0)}{n}}. \quad (6.76)$$

Są to linie równoległe do linii centralnej, wykreślonej na poziomie p_0 .

6.6.1.5. Karta kontrolna c i karta kontrolna u

Karty kontrolne c i u znajdują zastosowanie wówczas, gdy monitorowaną charakterystyką procesu jest wskaźnik natężenia, a nie wskaźnik struktury (frakcja), jak w przypadku omówionych powyżej kart kontrolnych z i w . Przypomnijmy, że wskaźnik natężenia opisuje stosunek dwóch różnych, ale powiązanych ze sobą zjawisk. Wskaźnikiem natężenia jest – na przykład – gęstość zaludnienia mierzona liczbą mieszkańców przypadających na 1 km², liczba uczniów przypadających na jednego nauczyciela, liczba lekarzy przypadających na 10000 mieszkańców. Wskaźnikami natężenia często wykorzystywanymi podczas monitorowania różnego rodzaju procesów są np. wymienione poniżej charakterystyki:

- liczba wad (albo niezgodności) w określonej umownej jednostce produktu. Charakterystykę tę wykorzystuje się często podczas monitorowania procesów produkcyjnych dla potrzeb zarządzania jakością;

- liczba błędów technologicznych popełnianych w określonym, umownym przedziale czasu trwania procesu produkcji albo procesu świadczenia usługi. Stosowanie tej charakterystyki jest oczywiście możliwe tylko wówczas, gdy natężenie obserwowanego procesu jest wyrównane w całym okresie jego monitorowania.

W każdym z wymienionych przypadków, a także w innych podobnych sytuacjach, obserwowaną charakterystykę procesu traktuje się jako zmienną losową o rozkładzie Poissona. Rozkład ten omówiliśmy w rozdziale 1 (zob. punkt 1.4.1.3). Jeśli podczas monitorowania procesu śledzone są zmiany jednej z wymienionych powyżej albo innej podobnej charakterystyki, to najczęściej mamy do czynienia z prawostronnym schematem kontrolnym. Należy pamiętać, że monitorowanie procesu podejmuje się zwykle w celu wykrycia zaburzeń w jego przebiegu, a tylko zbyt duża liczba wad lub niezgodności przypadająca na umowną jednostkę produktu albo zbyt duża liczba błędów popełnianych w ustalonym przedziale czasu świadczą o rozregulowaniu procesu. Ma to ten skutek, że najczęściej w każdym kroku ($t = 1, 2, \dots$) postępowania kontrolnego weryfikacji poddaje się hipotezę zerową postaci:

$$H_0: \lambda_{m,t} \leq \lambda_{m,0} \quad (6.77)$$

wobec hipotezy alternatywnej

$$H_1: \lambda_{m,t} > \lambda_{m,0}. \quad (6.78)$$

Hipoteza zerowa odpowiada stanowi uregulowania, natomiast hipoteza alternatywna odpowiada stanowi rozregulowania monitorowanego procesu. W hipotezach tych symbol $\lambda_{m,t}$ może np. oznaczać:

- rzeczywistą (a nie postulowaną) przeciętną liczbę wad lub niezgodności w ustalonej umownej jednostce produktu (m), w chwili t ,
 - rzeczywistą (a nie postulowaną) przeciętną liczbę błędów technologicznych popełnianych w umownej jednostce czasu trwania monitorowanego procesu.
- Symbol $\lambda_{m,0}$ oznacza natomiast postulowaną, największą dopuszczalną wartość charakterystyki λ_m .

Zauważmy, że wymienione powyżej, a także inne podobne charakterystyki mogą być wykorzystywane nie tylko do wykrywania zaburzeń w przebiegu monitorowanego procesu, ale również do ujawniania korzystnych zmian w tym procesie. W takiej sytuacji w każdym kroku postępowania kontrolnego weryfikuje się hipotezę zerową:

$$H_0: \lambda_{m,t} \geq \lambda_{m,0} \quad (6.79)$$

wobec hipotezy alternatywnej:

$$H_1: \lambda_{m,t} < \lambda_{m,0}. \quad (6.80)$$

Mamy wówczas do czynienia z lewostronnym schematem kontrolnym. Wartość $\lambda_{m,0}$ – podobnie jak w przypadku schematu prawostronnego – jest tu interpretowana jako kres górny przedziału pożądanych (korzystnych) stanów procesu. Możliwy jest też oczywiście, chociaż w praktyce nie posiadający większego znaczenia, dwustronny schemat kontrolny.

Karta kontrolna c

Procedura ta może być stosowana do monitorowania procesu wówczas, gdy parametr m może być utrzymany na stałym poziomie podczas całego badania. Oznacza to więc np., że w każdym kroku ($t = 1, 2, \dots$) postępowania kontrolnego badana jest umowna jednostka produktu składająca się z m jednostek rzeczywistych albo że błędy technologiczne zliczane są zawsze w przedziale czasu o takiej samej długości m . Dodajmy, że jeśli m jest długością przedziału czasu, to wymaga się dodatkowo, by natężenie monitorowanego procesu było – jak to już powiedziano powyżej – wyrównane w całym okresie podlegającym badaniu. W rozważanej sytuacji obserwowana charakterystyka z próby ma postać:

$$c_t = \sum_{i=1}^m z_{ti}, \quad (6.81)$$

gdzie z_{ti} jest liczbą wad lub niezgodności stwierdzoną w i -tej elementarnej jednostce produktu, liczbą błędów popełnionych w i -tej elementarnej jednostce

czasu albo inną podobną charakterystyką odnoszącą się do elementarnej jednostki badanej próbki.

Wartości c_t obliczone według powyższego wzoru nanosi się na diagram przeglądowy, na którym wcześniej wykreślono linię centralną i odpowiednią linię kontrolną. W rozważanej tu sytuacji linię centralną wykreśla się zawsze na poziomie $\lambda_{m,0}$, natomiast położenie linii kontrolnej zależy od celu badania. Jeśli monitorowanie procesu ma na celu wykrycie momentów jego rozregulowania, a więc jeśli realizowany jest wspomniany powyżej prawostronny schemat kontrolny, to wykreśla się zwykle tylko jedną, górną linię kontrolną, której położenie na płaszczyźnie diagramu przeglądowego wynika z równania:

$$c_g = \lambda_{m,0} + u_\alpha \sqrt{\lambda_{m,0}}, \quad (6.82)$$

gdzie u_α – analogicznie jak poprzednio – jest kwantylem standaryzowanej normalnej zmiennej losowej U , spełniającym warunek $P(U > u_\alpha) = \alpha$.

Jeśli w konkretnej sytuacji stosowany jest lewostronny schemat kontrolny, to położenie linii kontrolnej wyznacza się według wzoru:

$$c_d = \lambda_{m,0} - u_\alpha \sqrt{\lambda_{m,0}}. \quad (6.83)$$

Sygnaly o zmianie stanu monitorowanego procesu generowane są według reguł przedstawionych w tablicy 6.19.

Tablica 6.19

Karta kontrolna c ; generowanie sygnałów o zmianie stanu procesu

Hipoteza		Punktowy sygnał o zmianie stanu monitorowanego procesu emitowany jest, gdy	Prawdopodobieństwo fałszywego sygnału o zmianie stanu procesu
zerowa H_0	alternatywna H_1		
$\lambda_{m,t} \leq \lambda_{m,0}$	$\lambda_{m,t} > \lambda_{m,0}$	$c_t \geq c_g$	α
$\lambda_{m,t} \geq \lambda_{m,0}$	$\lambda_{m,t} < \lambda_{m,0}$	$c_t \leq c_d$	α

Przykład 6.9

Monitorowano proces księgowania w banku, w którym w ciągu każdego dnia rejestrowano kilkaset operacji finansowych. W tym celu codziennie pobierano losowo próbkę o licznosci $m = 100$ dokumentów źródłowych i zliczano

błędy popełnione podczas księgowania. W rezultacie obserwacji poczynionych w kolejnych ośmiu dniach badania uzyskano następujące wyniki:

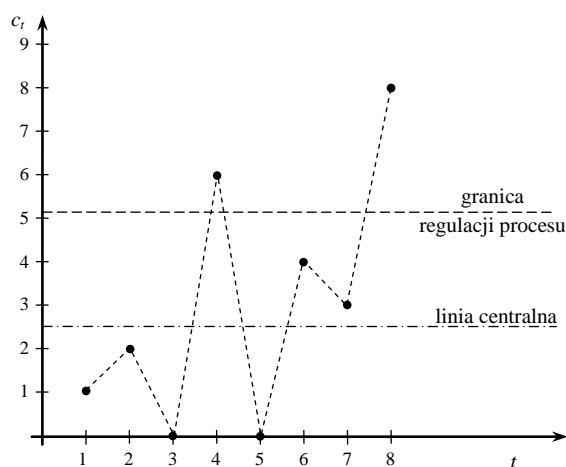
Numer dnia (t)	1	2	3	4	5	6	7	8	...
Liczba błędów (c_t)	1	2	0	6	0	4	3	8	...

Kierownictwo banku ustaliło, że proces księgowania przebiega zadowalająco, jeżeli przeciętna liczba popełnionych błędów przypadająca na $m = 100$ operacji nie przekracza $\lambda_{100,0} = 2,5$. Należy skonstruować odpowiednią kartę kontrolną do analizy powyższych danych. Celem podjętego monitorowania było wykrycie ewentualnych objawów pogorszenia jakości procesu księgowania. Prawdopodobieństwo zbędnej regulacji procesu ustalono na poziomie $\alpha = 0,05$. Mamy tu do czynienia z prawostronnym schematem kontrolnym. W każdym dniu weryfikowana była hipoteza zerowa $H_0: \lambda_{100,t} \leq 2,5$ wobec hipotezy alternatywnej $H_0: \lambda_{100,t} > 2,5$.

Przebieg monitorowania procesu przedstawiono na rys. 6.11. Linię centralną wykreślono na poziomie $\lambda_{100,0} = 2,5$, natomiast położenie górnej linii kontrolnej wyznaczono według równania (6.82). Po podstawieniu wartości do tego wzoru mamy:

$$c_g = 2,5 + 1,645\sqrt{2,5} = 2,5 + 2,6 = 5,1.$$

Wartość $u_{0,05} = 1,645$ odczytano z ostatniego wiersza tablicy II, zamieszczonej w aneksie.



Rys. 6.11. Diagram przeglądowy karty kontrolnej c

Jak wynika z przeprowadzonych badań, w dwóch dniach ($t = 4$ i $t = 8$) zaobserwowano rozregulowanie procesu księgowania. W tych dniach zaobserwowana liczba błędów księgowania przekroczyła wartość $c_g = 5,1$. Oznacza to konieczność odrzucenia hipotezy zerowej głoszącej, że $\lambda_{100,t} \leq \lambda_{100,0} = 2,5$, na korzyść hipotezy alternatywnej, według której $\lambda_{100,t} > \lambda_{100,0} = 2,5$. Prawdopodobieństwo zdarzenia losowego polegającego na tym, że podjęta decyzja jest błędna, nie przekracza $\alpha = 0,05$. ■

Karta kontrolna u

Jeśli w czasie całego badania nie można utrzymać na stałym poziomie liczności próbek albo długości obserwowanych przedziałów czasu, to zastosowanie karty kontrolnej c staje się niemożliwe, albowiem uzyskiwane empirycznie wartości c_t są nieporównywalne. W takiej sytuacji oblicza się wartości następującej charakterystyki z próby:

$$u_t = \frac{c_t}{m_t}, \quad (6.84)$$

gdzie m_t oznacza licznosc próby pobranej do badania albo dlugosc obserwowanego przedzialu czasu, w kolejnym kroku badania o numerze t . Odpowiedniego przekształcenia wymagają oczywiście również wszystkie parametry procedury kontrolnej. Linie centralną wyznacza się na poziomie:

$$\lambda_{1,0} = \frac{\lambda_{m,0}}{m}. \quad (6.85)$$

Obliczona w ten sposób wartość $\lambda_{1,0}$ jest największą dopuszczalną, przeciętną liczbą wad lub niezgodności przypadającą na elementarną jednostkę produktu, największą przeciętną liczbą błędów technologicznych przypadającą na elementarną jednostkę czasu trwania monitorowanego procesu albo inną podobną charakterystyką. W przypadku karty kontrolnej u , podobnie jak w przypadku omówionej powyżej karty kontrolnej c , najczęściej mamy do czynienia z prawostronnym schematem kontrolnym. W takim przypadku, w każdym kroku postępowania kontrolnego weryfikacji poddaje się hipotezę zerową:

$$H_0: \lambda_{1,t} \leq \lambda_{1,0} \quad (6.86)$$

wobec hipotezy alternatywnej:

$$H_1: \lambda_{1,t} > \lambda_{1,0}. \quad (6.87)$$

Położenie górnej linii kontrolnej wyznacza się według wzoru:

$$u_{g,t} = \lambda_{1,0} + u_\alpha \sqrt{\frac{\lambda_{1,0}}{m_t}}, \quad (6.88)$$

w którym m_t oznacza liczność próbki albo długość przedziału obserwacji w kolejnym kroku postępowania kontrolnego o numerze t . Jeśli m_t rośnie, to $u_{g,t}$ przybliża się do linii centralnej $\lambda_{1,0}$, i odwrotnie, jeśli m_t maleje, to $u_{g,t}$ oddala się od linii centralnej $\lambda_{1,0}$, zachowując względem niej równoległość.

W pewnych sytuacjach – analogicznie jak w przypadku karty kontrolnej c – znajduje zastosowanie również lewostronny schemat kontrolny. Weryfikacji poddaje się wówczas hipotezę zerową:

$$H_0: \lambda_{1,t} \geq \lambda_{1,0} \quad (6.89)$$

wobec hipotezy alternatywnej:

$$H_1: \lambda_{1,t} < \lambda_{1,0}. \quad (6.90)$$

Położenie dolnej linii kontrolnej wyznacza się według wzoru:

$$u_{d,t} = \lambda_{1,0} - u_\alpha \sqrt{\frac{\lambda_{1,0}}{m_t}}. \quad (6.91)$$

W tablicy 6.20 zestawiono reguły generowania sygnałów o zmianie stanu obserwowanego procesu.

Tablica 6.20

Karta kontrolna u ; generowanie sygnałów o zmianie stanu procesu

Hipoteza		Punktowy sygnał o zmianie stanu monitorowanego procesu emitowany jest, gdy	Prawdopodobieństwo fałszywego sygnału o zmianie stanu procesu
zerowa H_0	alternatywna H_1		
$\lambda_{1,t} \leq \lambda_{1,0}$	$\lambda_{1,t} > \lambda_{1,0}$	$u_t \geq u_{g,t}$	α
$\lambda_{1,t} \geq \lambda_{1,0}$	$\lambda_{1,t} < \lambda_{1,0}$	$u_t \leq u_{d,t}$	α

Przykład 6.10

Jakość produktu oceniano na podstawie przeciętnej liczby wad. Największą przeciętną liczbę wad w elementarnej jednostce produktu ustalono na poziomie $\lambda_{1,0} = 1,00$. Organizacja procesu nie zapewniała stałej licznosci próbek pobieranych do badania w kolejnych krokach procedury ($t = 1, 2, \dots$). Dlatego też do

monitorowania procesu zastosowano kartę kontrolną u . W każdym kroku postępowania kontrolnego weryfikowano hipotezę zerową:

$$H_0: \lambda_{1,t} \leq 1,00$$

wobec hipotezy alternatywnej:

$$H_1: \lambda_{1,t} > 1,00.$$

Przy wyznaczaniu położenia górnej linii kontrolnej, która w konkretnym przypadku była granicą regulacji procesu, przyjęto $\alpha = 0,01$. W ostatnim wierszu tablicy II, zamieszczonej w aneksie, znajdujemy wartość $u_{0,01} = 2,326$. Po podstawieniu wartości liczbowych do wzoru (6.88) mamy więc:

$$u_{g,t} = 1,00 + 2,326 \sqrt{\frac{1,00}{m_t}}.$$

Ze względu na duże zróżnicowanie licznosci próbek pobieranych do badania w kolejnych krokach postępowania kontrolnego zrezygnowano z wykreślania diagramu przeglądowego. Wyniki badań i ich analizę przedstawiono w tablicy 6.21. W kolumnach 1, 2 i 3 tej tablicy podano rezultaty badania kolejnych dziesięciu próbek. W kolumnie 4 zestawiono wartości u_t . W punkcie $t = 6$ pojawił się sygnał o rozregulowaniu monitorowanego procesu. Ponieważ wartość $u_6 = 2,21 > u_{g,6} = 1,62$, zatem należy odrzucić hipotezę H_0 na korzyść hipotezy H_1 , a prawdopodobieństwo zdarzenia losowego polegającego na tym, że podjęta decyzja jest błędna, nie przekracza $\alpha = 0,01$.

Tablica 6.21

Funkcjonowanie karty kontrolnej u

t	m_t	c_t	u_t	$u_{g,t}$	$u_{d,t}$
1	2	3	4	5	6
1	11	15	1,36	1,7	0,61
2	9	13	1,44	1,78	0,57
3	7	12	1,71	1,88	0,52
4	13	20	1,54	1,65	0,64
5	10	7	0,7	1,74	0,59
6	14	31	2,21	1,62	0,66
7	13	9	0,69	1,65	0,64
8	9	8	0,89	1,78	0,57
9	12	18	1,5	1,67	0,63
10	8	8	1	1,82	0,55

Źródło: opracowanie własne.

W kolumnie 6 tablicy 6.21 zestawiono wartości $u_{d,t}$ obliczone według wzoru (6.91). Przy obliczaniu tych wartości przyjęto $\alpha = 0,1$. Zastosowane równanie dolnej linii kontrolnej przedstawia się następująco:

$$u_{d,t} = 1,00 - 1,282 \sqrt{\frac{1,00}{m_t}}$$

Wartość $u_{0,1} = 1,282$ odczytano – analogicznie jak poprzednio – z ostatniego wiersza tablicy II zamieszczonej w aneksie. W rozważanym przypadku dolna linia kontrolna służy do wykrywania ewentualnych korzystnych zmian w monitorowanym procesie, będących skutkiem biernego postępu technologicznego. Jak wynika z wyników przeprowadzonych badań, zmiany takie nie pojawiły się. ■

6.6.2. Karty kontrolne sum skumulowanych

6.6.2.1. Monitorowanie wartości oczekiwanej normalnej zmiennej losowej

Skoncentrujmy uwagę na przypadku, gdy przedział tolerancji ograniczony jest tylko prawostronnie. W takiej sytuacji weryfikuje się następujące hipotezy:

$$H_0: \mu = \mu_0,$$

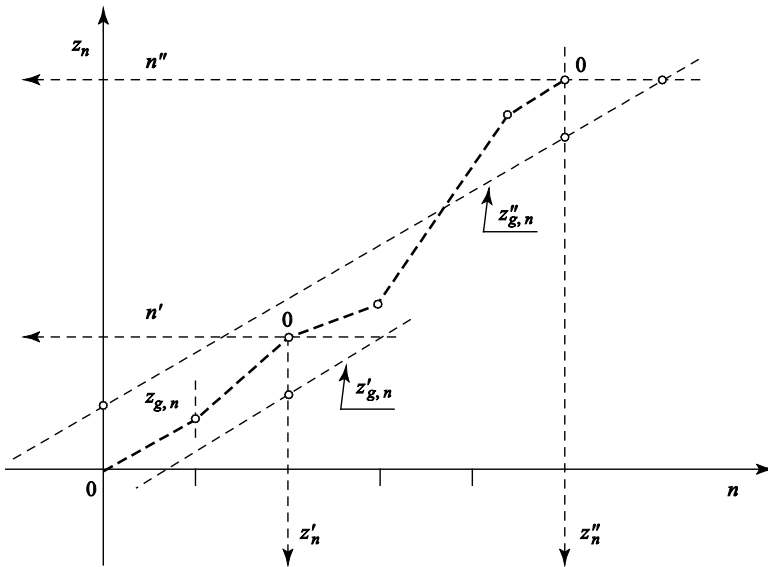
$$H_1: \mu = \mu_1,$$

przy czym $\mu_0 < \mu_1$. Hipotezy te są weryfikowane w kolejnych punktach próbkowania. Obserwowany proces jest uregulowany, jeżeli $\mu \leq \mu_0$. Jeżeli natomiast $\mu \geq \mu_1$, to proces ten uważa się za rozregulowany. Procedura kontrolna sum skumulowanych może być rozważana jako klasyczna procedura sekwencyjna realizowana wstecznie (rys. 6.12). Każdy punkt (n, z_n) kończący sekwencję traktowany jest jako początek nowego prostokątnego układu współrzędnych obróconego względem układu podstawowego o 180° . Na rys. 6.12 pokazano dwa takie pomocnicze układy współrzędnych, a mianowicie $n' 0z'_n$ i $n'' 0z''_n$. Wykorzystując te pomocnicze układy współrzędnych, w każdym kroku postępowania bada się, czy zaobserwowana dotychczas sekwencja:

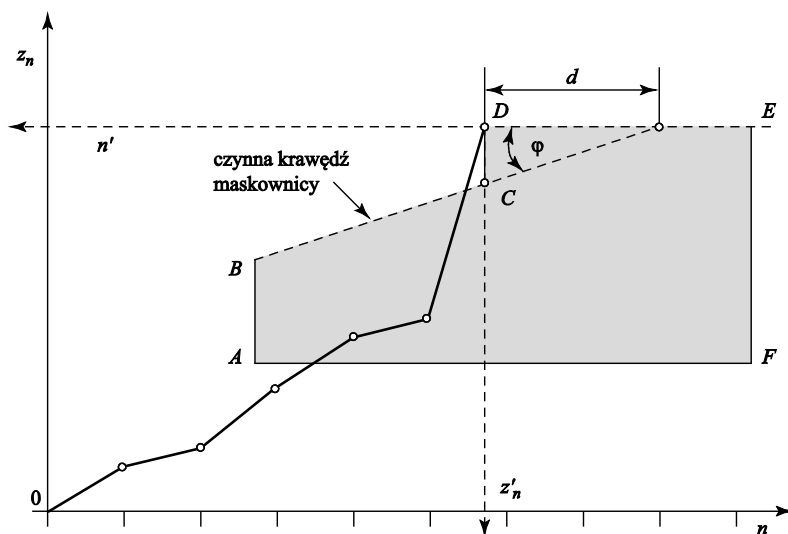
$$\begin{aligned} 0, z_0 &= 0 \\ 1, z_1 &= x_1 \\ 2, z_2 &= z_1 + x_2 = x_1 + x_2 \\ 3, z_3 &= z_2 + x_3 = x_1 + x_2 + x_3 \\ n, z_n &= z_{n-1} + x_n = x_1 + x_2 + x_3 + \dots + x_{n-1} + x_n \end{aligned} \quad (6.92)$$

wystarczy już do przyjęcia hipotezy alternatywnej. W procedurze kontrolnej sum skumulowanych łączy się obszar kontynuacji badań z obszarem przyjęć hipotezy zerowej. Przyjmuje się przy tym $\beta = 0$. Każdorazowe wykreślenie pomocniczego układu współrzędnych byłoby uciążliwe. Dlatego też pomocnicze układy współrzędnych zastępuje się ruchomą maskownicą, którą przesuwa się na diagramie przeglądowym w miarę wydłużania się obserwowanej sekwencji (6.92). Na rys. 6.13 przedstawiono sposób wyznaczania parametrów maskownicy. Jak łatwo zauważyć, parametr d jest liczbą przeciwną w stosunku do pierwiastka (n_0) równania $z_{g,n} = b + cn$, zmodyfikowanego przez przyjęcie $\beta = 0$. Możemy więc napisać:

$$z_{g,n} = \frac{\sigma^2}{\mu_1 - \mu_0} \ln \frac{1 - \beta}{\alpha} + \frac{\mu_0 + \mu_1}{2} n. \quad (6.93)$$



Rys. 6.12. Zależność między klasyczną procedurą sekwencyjną a procedurą sum skumulowanych



Rys. 6.13. Konstrukcja i funkcjonowanie maskownicy w schemacie kontrolnym z prawostronnym ograniczeniem przedziału tolerancji

Przyjmując $\beta = 0$ i przyrównując równanie (6.93) do zera, otrzymujemy:

$$\frac{\sigma^2}{\mu_1 - \mu_0} \ln \frac{1}{\alpha} + \frac{\mu_0 + \mu_1}{2} n = 0, \quad (6.94)$$

$$-\frac{\sigma^2 \ln \alpha}{\mu_1 - \mu_0} + \frac{\mu_0 + \mu_1}{2} n = 0. \quad (6.95)$$

Rozwiązując to równanie względem n , otrzymujemy:

$$n_0 = \frac{\sigma^2 \ln \alpha}{\mu_1 - \mu_0} \cdot \frac{2}{\mu_0 + \mu_1} = \frac{2\sigma^2 \ln \alpha}{\mu_1^2 - \mu_0^2}. \quad (6.96)$$

W konsekwencji:

$$d = -\frac{2\sigma^2 \ln \alpha}{\mu_1^2 - \mu_0^2}, \quad (6.97)$$

albowiem zachodzi zależność:

$$d = -n_0. \quad (6.98)$$

Drugim parametrem procedury jest kąt φ (zob. rys. 6.13), pod którym czynna krawędź maskownicy BC zastępująca linię kontrolną jest nachylona wzglę-

dem prostej DE , równoległej do osi odciętych n i przechodzącej przez punkt (n, z_n) , kończący w danym momencie obserwowaną sekwencję.

Parametr ten wynika wprost z równania:

$$c = \frac{\mu_1 + \mu_0}{2}.$$

Mamy mianowicie:

$$\varphi = \arctg c = \arctg \frac{\mu_0 + \mu_1}{2}. \quad (6.99)$$

Mając diagram przeglądowy oraz maskownicę o parametrach d i φ , postępujemy według podanego poniżej schematu:

– po naniesieniu na diagram przeglądowy każdego punktu (n, z_n) , umieszczamy na rysunku maskownicę w sposób pokazany na rys. 6.13. Istotne jest w szczególności to, by punkt D pokrywał się z punktem (n, z_n) , kończącym aktualnie sekwencję, oraz by krawędź DE była równoległa do osi odciętych n ;

– jeśli którykolwiek punkt należący do sekwencji (6.92) znajdzie się poniżej czynnej krawędzi maskownicy (lub jej przedłużenia), to przyjmujemy hipotezę alternatywną, a prawdopodobieństwo zdarzenia losowego polegającego na tym, że w rzeczywistości prawdziwa jest hipoteza zerowa, nie przekracza α ;

– jeśli żaden z punktów należących do sekwencji (6.92) nie jest „maskowany”, to przechodzimy do następnego kroku procedury. W rozważanym przypadku próbka skumulowana powiększona jest o jednostkę, a wartość sumy:

$$z_n = \sum_{i=1}^n x_i$$

powiększona jest o wartość x_i , odpowiadającą tej jednostce.

W literaturze przedmiotu omówiony wyżej algorytm graficzny określany jest najczęściej jako karta kontrolna sum skumulowanych (albo sum kumulacyjnych). Wadą tego algorytmu jest konieczność posługiwania się maskownicą, której czynna krawędź (zob. rys. 6.13) pełni rolę ruchomej linii kontrolnej. Niedogodność tę można usunąć, zastępując charakterystykę:

$$z_n = \sum_{i=1}^n x_i$$

charakterystyką postaci:

$$z_n^* = \sum_{i=1}^n (x_{ji} - \operatorname{tg} \varphi), \quad (6.100)$$

przy czym $n = 1, 2, \dots$. Wartość krytyczna dla tak zdefiniowanej charakterystyki z próby nie zależy od długości obserwowanej sekwencji (n) i przedstawia się następująco:

$$z_g^* = d \cdot \operatorname{tg} \varphi. \quad (6.101)$$

Obliczanie wartości z_n^* rozpoczyna się dopiero wówczas, gdy kolejna wartość x_j spełnia nierówność:

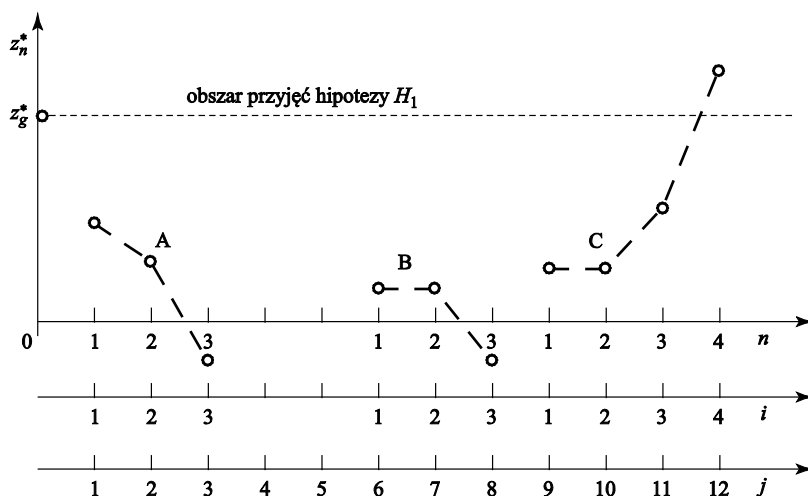
$$x_j > \operatorname{tg} \varphi. \quad (6.102)$$

Od takiej wartości rozpoczyna się podwójne indeksowanie realizacji zmiennej losowej X . Indeks i jest indeksem operacyjnym, funkcjonującym w ramach obserwowanej sekwencji, natomiast j jest indeksem funkcjonującym w całym cyklu badań. Jeśli w kolejnym kroku postępowania zostanie spełniona nierówność:

$$z_n^* \geq z_g^*, \quad (6.103)$$

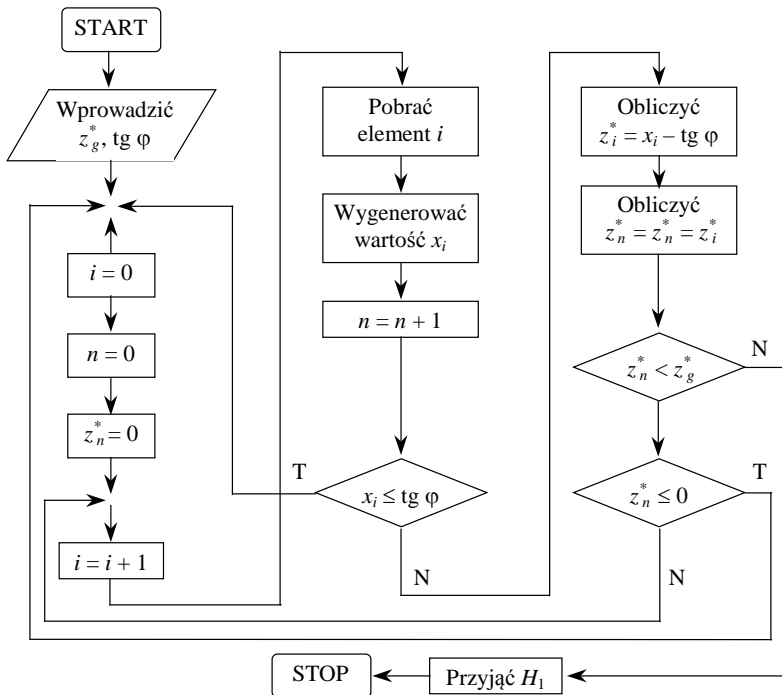
to przerywa się kumulację różnic ($x_{ji} - \operatorname{tg} \varphi$) i przyjmuje się hipotezę alternatywną. Następuje też zerowanie indeksu operacyjnego i . Kumulację różnic ($x_{ji} - \operatorname{tg} \varphi$) przerywa się także wówczas, gdy w kolejnym kroku procedury zostanie spełniona nierówność:

$$z_n^* \leq 0. \quad (6.104)$$



Rys. 6.14. Funkcjonowanie procedury sum skumulowanych ze stałą granicą regulacji, przy prawostronnym ograniczeniu przedziału tolerancji

Zeruje się wówczas indeks operacyjny i , przechodzi się do śledzenia znaku różnicy $x_j - \text{tg } \varphi$; zob. nierówność (6.102). Funkcjonowanie przedstawionego algorytmu ilustruje rys. 6.14. Na rysunku tym pokazano trzy sekwencje obserwacji (A, B, C). Sekwencje A i B zostały zakończone w efekcie spełnienia nierówności $z_{n=3}^* < 0$. Między tymi sekwencjami zaznaczono przerwę wynikającą z faktu, że $x_{j=4} - \text{tg } \varphi < 0$ oraz $x_{j=5} - \text{tg } \varphi < 0$. Sekwencja C została zakończona przyjęciem hipotezy alternatywnej. Algorytm graficzny ze stałą granicą regulacji można łatwo przekształcić w algorytm numeryczny. Funkcjonowanie tego algorytmu ilustruje schemat pokazany na rys. 6.15.



i	x_i	n	z_i^*	z_n^*
1	x_1	1	$x_1 - \text{tg } \varphi$	$x_1 - \text{tg } \varphi$
2	x_2	2	$x_2 - \text{tg } \varphi$	$(x_1 - \text{tg } \varphi) +$ $+ (x_2 - \text{tg } \varphi)$
\vdots	\vdots	\vdots	\vdots	\vdots

Rys. 6.15. Procedura sum skumulowanych ze stałą granicą regulacji; algorytm numeryczny przy prawostronnym ograniczeniu przedziału tolerancji

W dotychczasowych rozważaniach zakładaliśmy, że w każdym kroku postępowania sekwencyjnego dostępna jest jedna realizacja badanej zmiennej losowej, a także że obserwowaną charakterystyką z próby jest suma tych realizacji. Nie jest to rozwiązanie jedyne i nie zawsze najlepsze. Jeżeli np. wartości x_i wyrażają się dużymi liczbami, to wygodniejsze może się okazać operowanie realizacjami zmiennej losowej X transformowanymi do postaci $x_i - \mu_0$ lub $(x_i - \mu_0)/\sigma$. Charakterystyka z_n przyjmuje wówczas jedną z następujących postaci:

$$z_n = \sum_{i=1}^n (x_i - \mu_0), \quad (6.105)$$

$$z_n = \sum_{i=1}^n \frac{x_i - \mu_0}{\sigma}, \quad (6.106)$$

przy czym $n = 1, 2, \dots$

W pewnych sytuacjach doświadczalnych wyniki obserwacji są mechanicznie uśrednione. Jeżeli np. podczas kontroli jednostkowego ciężaru wyrobu jesteśmy zmuszeni do jednoczesnego pomiaru masy kilku lub kilkunastu sztuk, to dostępne są realizacje zmiennej losowej:

$$\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{m}}\right),$$

gdzie m jest liczebnością próbki pobieranej każdorazowo do badania. W takiej sytuacji podczas kontroli wykorzystuje się najczęściej jedną z następujących charakterystyk:

$$z_n = \sum_{i=1}^n \bar{x}_i, \quad (6.107)$$

$$z_n = \sum_{i=1}^n (\bar{x}_i - \mu_0), \quad (6.108)$$

$$z_n = \sum_{i=1}^n \frac{\bar{x}_i - \mu_0}{\sigma} \sqrt{m}, \quad (6.109)$$

gdzie:

$$\bar{x}_i = \frac{\sum_{r=1}^m x_{ir}}{m}. \quad (6.110)$$

W rozważanej sytuacji wartości x_{ir} pozostają nieznane. Znana jest natomiast ich suma.

Jeżeli przedział tolerancji jest ograniczony tylko lewostronnie, to wówczas weryfikacji poddaje się następujące hipotezy:

$$H_0: \mu = \mu_0,$$

$$H_{-1}: \mu = \mu_{-1},$$

przy czym $\mu_0 > \mu_{-1}$.

Uznaje się, że wyrób spełnia stawiane mu wymagania jakościowe, a proces produkcyjny jest uregulowany, gdy $\mu \geq \mu_0$. Jeżeli natomiast $\mu \leq \mu_{-1}$, to przyjmuje się, że wyrób nie spełnia wymagań jakościowych, proces produkcyjny jest rozregulowany.

Na podstawie diagramu przeglądowego stosowanego w procesie sekwencyjnej weryfikacji hipotez łatwo zauważyć, że wysokie wartości sumy $z_n = \sum_{i=1}^n x_i$

dają podstawę do przyjęcia hipotezy zerowej H_0 . Natomiast niskie wartości tej sumy przemawiają na korzyść hipotezy alternatywnej H_{-1} . Ma to swoje konsekwencje przy wyznaczaniu parametrów procedury kontrolnej, wykorzystującej technikę sum skumulowanych. W rozważanym przypadku parametry te można wyznaczyć przyjmując za punkt wyjścia równanie dolnej linii kontrolnej $z_{d,n}$, albowiem spełnienie nierówności $z_n^* \leq z_{d,n}$ jest warunkiem przyjęcia hipotezy H_{-1} . Parametr φ wyznaczamy wprost z zależności:

$$c = \frac{\mu_0 + \mu_{-1}}{2}.$$

Mamy mianowicie:

$$\varphi = \arctg \frac{\mu_0 - \mu_{-1}}{2}. \quad (6.111)$$

Parametr d uzyskujemy w efekcie podanych poniżej przekształceń:

$$z_{d,n} = \frac{\sigma^2}{\mu_{-1} - \mu_0} \ln \frac{1 - \beta}{\alpha} + \frac{\mu_0 + \mu_{-1}}{2} n,$$

$$\frac{\sigma^2}{\mu_{-1} - \mu_0} \ln \frac{1}{\alpha} + \frac{\mu_0 + \mu_{-1}}{2} n = 0,$$

$$n_0 = \frac{2\sigma^2 \ln \alpha}{\mu_{-1}^2 - \mu_0^2},$$

$$d = -n_0 = -\frac{2\sigma^2 \ln \alpha}{\mu_{-1}^2 - \mu_0^2} = \frac{2\sigma^2 \ln \alpha}{\mu_0^2 - \mu_{-1}^2}. \quad (6.112)$$

Wartość d wyznaczona na podstawie tego wzoru jest ujemna, z założenia bowiem $\mu_0^2 > \mu_{-1}^2$, a jednocześnie $\ln \alpha < 0$. W algorytmie graficznym wartość d powinna być więc odkładana w lewo od każdego punktu (n, z_n) kończącego sekwencję:

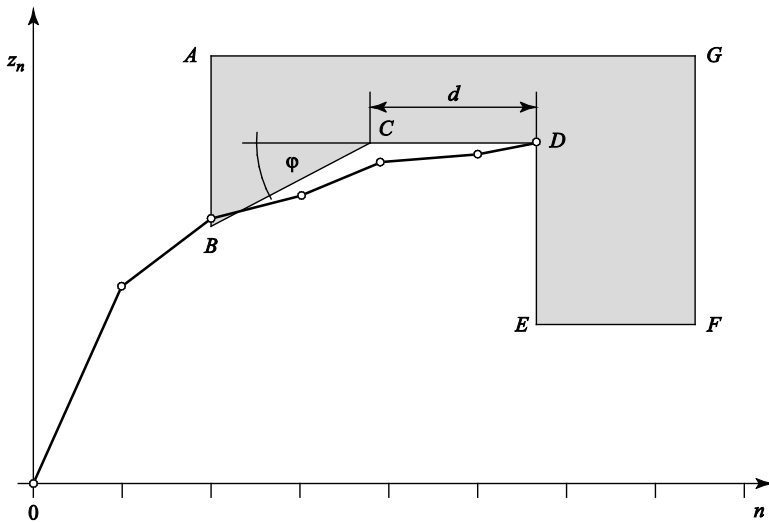
$$(0, 0), (1, z_1), (2, z_2), \dots, (n, z_n); \quad n = 1, 2, 3, \dots \quad (6.113)$$

w kolejnych krokach procedury.

W celu wyjaśnienia tego wniosku sporządzono rys. 6.16, na którym pokazano kształt maskownicy oraz mechanizm generowania sygnału o rozregulowaniu procesu produkcyjnego. W miarę wydłużania się sekwencji (6.113) maskownice należy przesuwać w taki sposób, aby w każdym kroku postępowania punkt D pokrywał się z punktem (n, z_n) oraz by odcinek CD był równoległy do osi odciętych n . Sygnał o rozregulowaniu procesu produkcyjnego jest generowany wówczas, gdy co najmniej jeden punkt należący do sekwencji (6.113) zostanie „zamaskowany”, a więc gdy znajdzie się powyżej czynnej krawędzi (BC) maskownicy albo jej przedłużenia. Sygnał jest więc generowany wówczas, gdy

przyrost sumy $z_n = \sum_{i=1}^n x_i$ jest zbyt powolny. Zbyt powolny przyrost sumy z_n

przemawia bowiem na korzyść hipotezy $H_{-1}: \mu = \mu_{-1}$. Mając wartość $c = \operatorname{tg} \varphi$ oraz d , można skonstruować algorytm graficzny ze stałymi granicami regulacji, a w dalszej kolejności również algorytm numeryczny. W obu algorytmach obserwowaną charakterystyką z próby jest suma z_n^* , której wartości wyznacza się według wzoru (6.100). Kumulację próby rozpoczyna się wówczas, gdy kolejna wartość x_j (gdzie j oznacza indeks bieżący) spełnia nierówność $x_j < \operatorname{tg} \varphi$. Od tego momentu uruchamia się indeks operacyjny (i), który funkcjonuje we wzorze (6.100). Sekwencję przerywa się bez podejmowania jakichkolwiek decyzji w odniesieniu do weryfikowanych hipotez, gdy zostanie spełniona nierówność $z_n^* > 0$. Powraca się wówczas do śledzenia znaku kolejnych różnic $x_j - \operatorname{tg} \varphi$. Obserwowaną sekwencję przerywa się także wówczas, gdy $z_n^* \leq z_d^* = d \cdot \operatorname{tg} \varphi$. W takiej sytuacji przyjmuje się hipotezę $H_{-1}: \mu = \mu_{-1}$. Dodajmy, że wartość krytyczna z_n^* jest ujemna, gdyż $\operatorname{tg} \varphi > 0$. Także d przyjmuje wartości ujemne.



Rys. 6.16. Konstrukcja i funkcjonowanie maskownicy w schemacie kontrolnym z lewostronnym ograniczeniem przedziału tolerancji

Jeżeli przedział tolerancji jest ograniczony dwustronnie, to wówczas weryfikacji podlega hipoteza zerowa:

$$H_0: \mu = \mu_0$$

wobec hipotez alternatywnych:

$$H_1: \mu = \mu_1,$$

$$H_{-1}: \mu = \mu_{-1},$$

przy czym $\mu_{-1} < \mu_0 < \mu_1$.

Obserwowany proces produkcyjny jest uregulowany, a wyrób spełnia stawiane mu wymagania jakościowe, gdy $\mu = \mu_0$. Jeżeli natomiast $\mu \leq \mu_{-1}$ lub $\mu \geq \mu_1$, to wyrób nie spełnia wymagań jakościowych, a obserwowany proces produkcyjny uważa się za rozregulowany.

Dwustronne procedury kontrolne są złożeniem odpowiednich procedur jednostronnych. Na podstawie równań dotyczących klasycznej procedury sekwencyjnej, omówionej w rozdziale 5, wyznaczamy obecnie parametry równoważnej formalnie procedury sum skumulowanych:

$$\varphi_{(1)} = \arctg \frac{\mu_1 - \mu_0}{2},$$

$$\begin{aligned}
\varphi_{(-1)} &= \operatorname{arctg} \frac{\mu_{-1} - \mu_0}{2}, \\
z_{g,n}^{(1)} &= \frac{\sigma^2}{\mu_1 - \mu_0} \ln \frac{1 - \beta}{\alpha} + \frac{\mu_1 - \mu_0}{2} n, \\
\frac{\sigma^2}{\mu_1 - \mu_0} \ln \frac{1 - \beta}{\alpha} + \frac{\mu_1 - \mu_0}{2} n &= 0, \\
-\frac{\sigma^2 \ln \alpha}{\mu_1 - \mu_0} + \frac{\mu_1 - \mu_0}{2} n &= 0, \\
n_0 &= \frac{2\sigma^2 \ln \alpha}{(\mu_1 - \mu_0)^2}, \\
d_{(1)} = -n_0 &= -\frac{2\sigma^2 \ln \alpha}{(\mu_1 - \mu_0)^2}. \tag{6.114}
\end{aligned}$$

W efekcie analogicznych przekształceń równania:

$$z_{d,n}^{(-1)} = \frac{\sigma^2}{\mu_{-1} - \mu_0} \ln \frac{1 - \beta}{\alpha} + \frac{\mu_{-1} - \mu_0}{2} n$$

otrzymujemy

$$d_{(-1)} = -\frac{2\sigma^2 \ln \alpha}{(\mu_{-1} - \mu_0)^2}. \tag{6.115}$$

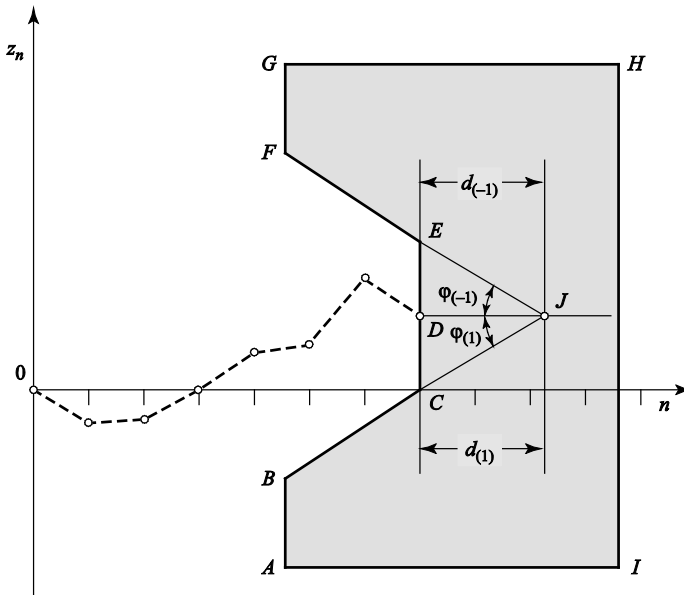
Łatwo zauważyć, że jeżeli wartość α jest ustalona i jeżeli $\mu_1 - \mu_0 = \mu_0 - \mu_{-1}$, to $d_{(1)} = d_{(-1)}$. Przy takim założeniu sporządzono rys. 6.17, na którym pokazano maskownicę stosowaną w dwustronnym schemacie kontrolnym, gdy suma z_n ma postać (6.105). Sposób posługiwania się maskownicą jest taki sam jak dotychczas. W miarę wydłużania się obserwowanej sekwencji postaci (6.113) maskownicę przesuwa się na płaszczyźnie diagramu przeglądowego. Umieszcza się ją każdorazowo w ten sposób, by punkt D pokrywał się z punktem (n, z_n) i by odcinek DJ (zob. rys. 6.17) był równoległy do osi n . Sygnał o rozregulowaniu procesu produkcyjnego generuje się wówczas, gdy chociaż jeden punkt należący do sekwencji (6.113) znajduje się powyżej górnej albo poniżej dolnej, czynnej krawędzi maskownicy. W sytuacji pokazanej na rys. 6.17 nie ma podstaw do generowania sygnału o rozregulowaniu obserwowanego procesu produkcyjnego. Mając parametry maskownicy, można skonstruować równoważny algorytm ze

stałymi granicami, a także algorytm numeryczny. W obu przypadkach charakterystyka z próby ma postać:

$$z_n^* = \sum_{i=1}^n (v_i - \operatorname{tg} \varphi), \quad (6.116)$$

gdzie:

$$v_i = x_i - \mu_0.$$



Rys. 6.17. Konstrukcja i funkcjonowanie maskownicy w schemacie kontrolnym sum skumulowanych z dwustronnym ograniczeniem przedziału tolerancji

Górna linia kontrolna, której osiągnięcie lub przekroczenie pociąga za sobą przyjęcie hipotezy H_1 , ma postać:

$$z_g^* = d_{(1)} \operatorname{tg} \varphi_{(1)} > 0. \quad (6.117)$$

Dolna linia kontrolna ma natomiast postać:

$$z_d^* = d_{(-1)} \operatorname{tg} \varphi_{(-1)} < 0. \quad (6.118)$$

Jeżeli wartość sumy z_n^* osiąga lub przekracza (*in minus*) linię z_d^* , to wtedy przyjmują się hipotezę H_{-1} .

6.6.2.2. Kontrola odchylenia standardowego normalnej zmiennej losowej

Weryfikacji poddaje się hipotezę zerową:

$$H_0: \sigma = \sigma_0$$

wobec hipotezy alternatywnej:

$$H_1: \sigma = \sigma_1,$$

gdzie σ_0 jest najwyższą dopuszczalną wartością odchylenia standardowego σ , natomiast σ_1 jest najniższą dyskwalifikującą wartością tego odchylenia. Wyrób spełnia stawiane mu wymagania jakościowe, a proces produkcyjny jest uregulowany, gdy zachodzi nierówność $\sigma \leq \sigma_0$. Jeżeli natomiast $\sigma \geq \sigma_1$, to wyrób nie spełnia wymagań jakościowych, a monitorowany proces produkcyjny jest rozregulowany. Na podstawie omówionych w rozdziale 5 zależności (5.179)–(5.181) można – analogicznie jak poprzednio – wyznaczyć parametry procedury sum skumulowanych. Parametr φ otrzymujemy wprost z równania (5.181). Mamy mianowicie:

$$\varphi = \arctg c = \arctg \frac{\sigma_0^2 \sigma_1^2}{\sigma_1^2 - \sigma_0^2} \ln \frac{\sigma_1^2}{\sigma_0^2}. \quad (6.119)$$

W celu wyznaczenia parametru d przyjmujemy $\beta = 0$ w równaniu górnej linii kontrolnej $z_{g,n} = b + cm$, a następnie znajdujemy pierwiastek (n_0) tak zmodyfikowanego równania. Biorąc $d = -n_0$ otrzymujemy parametr procedury sum skumulowanych. Mamy więc:

$$-\frac{2\sigma_0^2 \sigma_1^2 \ln \alpha}{\sigma_1 - \sigma_0^2} + \frac{\sigma_0^2 \sigma_1^2 \ln \left(\frac{\sigma_1^2}{\sigma_0^2} \right)}{\sigma_1 - \sigma_0^2} n = 0,$$

$$n_0 = \frac{2 \ln \alpha}{\ln \sigma_1^2 - \ln \sigma_0^2}$$

i w konsekwencji

$$d = \frac{2 \ln \alpha}{\ln \sigma_0^2 - \ln \sigma_1^2} = \frac{2 \ln \alpha}{\ln \left(\frac{\sigma_0^2}{\sigma_1^2} \right)}. \quad (6.120)$$

Na podstawie wyznaczonych w ten sposób wartości d i φ (albo $\operatorname{tg} \varphi$) można skonstruować trzy algorytmy kontrolne, oparte na technice sum skumulowanych. Są to:

– dwa algorytmy graficzne (z maskownicą albo ze stałą granicą regulacji), a także

– algorytm numeryczny.

Zasady funkcjonowania tych algorytmów są takie same jak w przypadku kontroli wartości oczekiwanej (μ), przy prawostronnym ograniczeniu przedziału tolerancji.

6.6.2.3. Monitorowanie parametru p zero-jedynkowej zmiennej losowej

Niech X oznacza zero-jedynkową zmienną losową o parametrze p , omówioną w rozdziale 1 (zob. punkt 1.4.1.1). Skupmy uwagę na przypadku, gdy weryfikacji poddaje się następujące hipotezy:

$$H_0: p = p_0,$$

$$H_1: p = p_1,$$

przy czym $p_1 > p_0$.

Jest to prawostronny schemat kontrolny, wykorzystywany do monitorowania różnego rodzaju procesów ze względu na wskaźnik struktury (frakcję) wyróżnionych stanów. Metoda ta ma podstawowe znaczenie podczas monitorowania procesów produkcyjnych, albo procesów świadczenia usług, dla potrzeb zarządzania jakością, przy alternatywnej ocenie właściwości jednostki produktu. Jeżeli p interpretuje się jako frakcję elementów wadliwych w strumieniu wyrobu, to proces produkcyjny uważa się za uregulowany, gdy $p \leq p_0$, natomiast za rozregulowany, gdy $p \geq p_1$.

Do weryfikacji sformułowanych powyżej hipotez wykorzystamy charakterystykę postaci:

$$z_n = \sum_{i=1}^n x_i, \quad (6.121)$$

w której x_i oznacza realizację zero-jedynkowej zmiennej losowej zdefiniowanej w rozdziale 1 wzorem (1.61), natomiast n jest licznością skumulowanej próbki; $n = 1, 2, 3, \dots$

Procedurę sekwencyjną opisaną w punkcie 5.9.4 można łatwo przekształcić w procedurę sum skumulowanych. Parametry maskownicy przyjmują tutaj następującą postać:

$$\phi = \arctg c = \arctg \frac{\ln \frac{1-p_0}{1-p_1}}{\ln \frac{p_1(1-p_0)}{p_0(1-p_1)}}, \quad (6.122)$$

$$d = -n_0 = -\frac{\ln \alpha}{\ln \frac{1-p_0}{1-p_1}}. \quad (6.123)$$

Obliczone według tych wzorów wartości d i ϕ można wykorzystać zarówno w graficznym jak i numerycznym algorytmie monitorowania procesu. W przypadku algorytmu graficznego przed rozpoczęciem monitorowania należy:

- ustalić wartości p_0 , p_1 i α ,
- sporządzić diagram przeglądowy w układzie współrzędnych (n, z_n) , gdzie n jest licznoscią skumulowanej próbki, natomiast z_n oznacza sumę obliczoną według wzoru (6.121),
- sporządzić maskownicę o parametrach określonych wzorami (6.122) i (6.123). Monitorowanie przebiega według następującego schematu:
 - na diagram przeglądowy nanosi się kolejne punkty o współrzędnych (n, z_n) , łącząc je ewentualnie linią łamaną,
 - po naniesieniu każdego punktu na diagram przeglądowy nakłada się maskownicę,
 - jeżeli którykolwiek z punktów danej sekwencji znajduje się poniżej krawędzi maskownicy lub jej przedłużenia, to przyjmujemy hipotezę alternatywną H_1 . Oznacza to konieczność przeprowadzenia regulacji monitorowanego procesu.
 - jeżeli natomiast wszystkie punkty aktualnie kontrolowanej sekwencji leżą powyżej ramienia maskownicy lub jego przedłużenia, to nie ma podstaw do odrzucenia hipotezy H_0 . W takiej sytuacji nie podejmuje się regulacji procesu.

Jeśli podjęto decyzję o zastosowaniu algorytmu numerycznego, to przed rozpoczęciem monitorowania należy obliczyć wartość parametru z_g , będącego granicą regulacji procesu. Wartość tego parametru oblicza się według wzoru:

$$z_g = d \cdot \operatorname{tg} \phi, \quad (6.124)$$

przy czym wartości d i ϕ wyznacza się na podstawie wzorów (6.122) i (6.123). Mając wartość parametru z_g można przystąpić do monitorowania procesu.

Przykład 6.11

Obserwowano staranność wypełniania pewnych formularzy związanych z odprawą celną towarów. W tym celu wybierano losowo pojedyncze dokumenty (wypełnione formularze) i szczegółowo sprawdzano staranność ich wypełnienia. Końcowe oceny formułowano alternatywnie z zastosowaniem następującej reguły:

$$x = \begin{cases} 0 & \text{– starannie wypełniony formularz} \\ 1 & \text{– niestarannie wypełniony formularz} \end{cases}$$

Na podstawie wcześniejszych obserwacji ustalono, że jeśli frakcja niestarannie wypełnionych formularzy nie przekracza 10%, to dalszy proces komputerowego przetwarzania informacji zawartych w tych dokumentach przebiega zadowalająco. Jeśli natomiast frakcja niestarannie wypełnionych formularzy przekracza 20%, to występują istotne zakłócenia w tym procesie. Mamy więc: $p_0 = 0,1$, $q_0 = 0,9$, $p_1 = 0,2$ i $q_1 = 0,8$. Prawdopodobieństwo zdarzenia losowego polegającego na tym, że sygnał ostrzegający o możliwości wystąpienia zakłóceń w procesie przetwarzania informacji okaże się sygnałem fałszywym, ustalono na poziomie $\alpha = 0,05$. Na podstawie tych wstępnych informacji obliczono wartości parametrów d i $\text{tg } \varphi$, a następnie z_g . Wykorzystano do tego celu odpowiednio zmodyfikowane wzory (6.122), (6.123) oraz wzór (6.124):

$$d = -\frac{\ln \alpha}{\ln q_0 - \ln q_1} = -\frac{\ln 0,05}{\ln 0,9 - \ln 0,8} = 25,4343 \approx 25,4,$$

$$\text{tg } \varphi = \frac{\ln \frac{q_0}{q_1}}{\ln \frac{p_1 q_0}{p_0 q_1}} = \frac{\ln \frac{0,9}{0,8}}{\ln \frac{0,2 \cdot 0,9}{0,1 \cdot 0,8}} = 0,1452 \approx 0,145,$$

$$z_g = 25,4343 \cdot 0,1452 = 3,6931 \approx 3,7.$$

Wyniki badania losowo wybranych dokumentów przedstawiono w kolumnach 1 i 2 tablicy 6.22. W dalszych kolumnach tej tablicy pokazano funkcjonowanie numerycznego algorytmu monitorowania procesu za pomocą karty kontrolnej sum skumulowanych. W początkowych dwóch krokach postępowania kontrolnego uzyskano $x_1 = 0 < 0,145$ oraz $x_2 = 0 < 0,145$. W takiej sytuacji nie podejmuje się dalszych działań analitycznych.

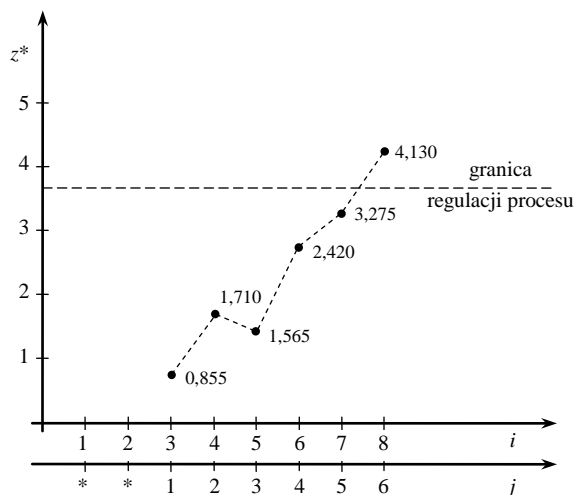
Tablica 6.22

Algorytm numeryczny karty kontrolnej sum skumulowanych

i	x_i	j	x_{ij}	$x_{ij} - 0,145$	$\sum(x_{ij} - 0,145)$	Uwagi
1	2	3	4	5	6	7
1	0	*	*	*	*	$< z_g$
2	0	*	*	*	*	$< z_g$
3	1	1	1	0,855	0,855	$< z_g$
4	1	2	1	0,855	1,710	$< z_g$
5	0	3	0	-0,145	1,565	$< z_g$
6	1	4	1	0,855	2,420	$< z_g$
7	1	5	1	0,855	3,275	$< z_g$
8	1	6	1	0,855	4,130	$> z_g$

Źródło: opracowanie własne.

W kolejnym kroku ($i = 3$) uzyskano $x_3 = 1 > 0,145$. W takiej sytuacji uruchamia się licznik indeksu operacyjnego (j) oraz rozpoczyna się obliczanie różnic $x_{ij} - 0,145$ (kolumna 5), a także kolejnych wartości skumulowanej sumy tych różnic (kolumna 6). Mamy mianowicie: $x_{31} - 0,145 = 1 - 0,145 = 0,855$. Jest to też pierwsza wartość sumy skumulowanej, przy czym $0,855 < z_g = 3,7$.



Rys. 6.18. Diagram przeglądowy karty kontrolnej sum skumulowanych ze stałą granicą regulacji

W kroku $i = 4$ uzyskano wartość $x_{42} = 1$. W konsekwencji wartość sumy skumulowanej wzrosła do poziomu $1,710 < z_g = 3,7$. W kolejnym kroku postępowania kontrolnego ($i = 5, j = 3$) nastąpiło obniżenie wartości sumy skumulowanej do poziomu $1,565$, bowiem $x_{53} = 0$. Kolejne wylosowane do kontroli dokumenty ($i = 6, 7, 8$) były niestarannie wypełnione ($x_{64} = 1, x_{75} = 1$ i $x_{86} = 1$). W konsekwencji, w kroku $i = 8$ postępowania kontrolnego, czyli – w rozważanej sytuacji – w kroku $j = 6$ obserwowanej sekwencji, suma skumulowana osiągnęła wartość $4,130 > z_g = 3,7$. Oznacza to emisję sygnału ostrzegającego o wzroście frakcji niestarannie wypełnionych formularzy do poziomu $p \geq 0,2$. Prawdopodobieństwo zdarzenia losowego polegającego na tym, że rzeczywista frakcja spełnia nierówność $p \leq 0,1$, nie przekracza $\alpha = 0,05$. ■

6.6.2.4. Monitorowanie parametru λ zmiennej losowej Poissona

W rozważanym przypadku najczęściej poddaje się weryfikacji hipotezę zerową:

$$H_0: \lambda = \lambda_0$$

wobec hipotezy alternatywnej:

$$H_1: \lambda = \lambda_1,$$

gdzie:

- λ – monitorowany parametr zmiennej losowej o rozkładzie Poissona,
- λ_0 – największa dopuszczalna wartość parametru λ ,
- λ_1 – najmniejsza niedopuszczalna wartość parametru λ , przy czym $\lambda_1 > \lambda_0$.

Przypomnijmy (zob. punkt 1.4.1.3), że parametr $\lambda = np$ jest wartością oczekiwaną i wariancją zmiennej losowej o rozkładzie Poissona. Jeśli w każdym kroku postępowania kontrolnego poddaje się weryfikacji sformułowane powyżej hipotezy, to mamy do czynienia z prawostronnym schematem kontrolnym. Monitorowany proces uważa się za uregulowany, jeżeli $\lambda \leq \lambda_0$. Proces jest natomiast rozregulowany, gdy $\lambda > \lambda_0$. W takiej sytuacji oczekujemy sygnału o rozregulowaniu. Żadna statystyczna procedura monitorowania procesów nie wykazuje jednak całkowitej selektywności. Dopiero wówczas, gdy $\lambda > \lambda_1$, prawdopodobieństwo braku sygnału o rozregulowaniu procesu (β) jest wystarczająco małe. Prawdopodobieństwo wygenerowania fałszywego sygnału o rozregulowaniu procesu nie przekracza α .

Prezentowana procedura jest szczególnie często wykorzystywana do monitorowania procesów dla potrzeb zarządzania jakością. Miary jakości produktu, albo procesu, są wówczas definiowane tak samo jak w przypadku omówionych powyżej (zob. punkt 6.6.1.5) kart kontrolnych c i u . Miarą jakości produktu jest

w takiej sytuacji przeciętna liczba wad (albo niezgodności) w odpowiednio zdefiniowanej umownej jednostce produktu. W przypadku procesu miarą jakości jest natomiast przeciętna liczba błędów popełnianych w ustalonej, umownej jednostce czasu trwania procesu.

Stosowanie omawianej tu procedury przynosi największe korzyści, w porównaniu z kartami kontrolnymi c i u , gdy monitorowanie procesu produkcji wyrobu, albo procesu świadczenia usługi, prowadzi się na podstawie prób jednoelementowych. Zmienne x_i oznaczające liczbę wad w jednej sztuce wyrobu przyjmują wartości ze zbioru liczb całkowitych dodatnich. Sumy skumulowane obliczane są według wzoru (6.121), a więc w taki sam sposób jak w przypadku procedury monitorowania parametru p zero-jedynkowej zmiennej losowej, omówionej w poprzednim punkcie.

Parametry maskownicy o jednej (dolnej) krawędzi czynnej wynikają z teoretycznych podstaw sekwencyjnego testu ilorazowego, omówionego w poprzednim rozdziale (zob. punkt 5.9.5). Parametr d zostanie wyznaczony z równania górnej linii kontrolnej $z_{g,n}$:

$$z_{g,n} = \frac{\ln \frac{1-\beta}{\alpha}}{\lambda_1 - \lambda_0} + \frac{\ln \lambda_1 - \ln \lambda_0}{\lambda_1 - \lambda_0} n.$$

Po podstawieniu $\beta = 0$ powyższe równanie rozwiązujemy względem n i otrzymujemy:

$$-\frac{\ln \alpha}{\ln \lambda_1 - \ln \lambda_0} + \frac{\lambda_1 - \lambda_0}{\ln \lambda_1 - \ln \lambda_0} n = 0,$$

$$n_0 = \frac{\ln \alpha}{\ln \lambda_1 - \ln \lambda_0} \cdot \frac{\ln \lambda_1 - \ln \lambda_0}{\lambda_1 - \lambda_0} = \frac{\ln \alpha}{\lambda_1 - \lambda_0}.$$

Ponieważ $d = -n_0$, więc:

$$d = -\frac{\ln \alpha}{\lambda_1 - \lambda_0}. \quad (6.125)$$

Parametr φ otrzymuje się bezpośrednio z równania:

$$c = \frac{\lambda_1 - \lambda_0}{\ln \lambda_1 - \ln \lambda_0}.$$

Mamy zatem:

$$\varphi = \arctg c = \arctg \frac{\lambda_1 - \lambda_0}{\ln \lambda_1 - \ln \lambda_0}. \quad (6.126)$$

Jeśli monitorowanie procesu ma być prowadzone z wykorzystywaniem algorytmu graficznego, a więc z zastosowaniem diagramu przeglądowego i maskownicy, to przed rozpoczęciem kontroli należy wykonać następujące czynności przygotowawcze:

- ustalić wartości $\lambda_0, \lambda_1, \alpha$,
 - sporządzić diagram przeglądowy w układzie współrzędnych (n, z_n) ,
 - sporządzić maskownicę o parametrach określonych wzorami (6.125) i (6.126).
- Przebieg monitorowania nie różni się od omówionego poprzednio:
- na diagram przeglądowy nanosi się kolejne punkty o współrzędnych (n, z_n) , łącząc je ewentualnie linią łamaną,
 - po naniesieniu każdego punktu na diagram przeglądowy nakłada się maskownicę,
 - jeżeli którykolwiek z punktów danej sekwencji znajduje się poniżej krawędzi maskownicy lub jej przedłużenia, to przyjmujemy hipotezę alternatywną H_1 ; oznacza to konieczność przeprowadzenia regulacji monitorowanego procesu,
 - jeżeli natomiast wszystkie punkty aktualnie kontrolowanej sekwencji leżą powyżej ramienia maskownicy lub jego przedłużenia, to nie ma podstaw do odrzucenia hipotezy H_0 ; w takiej sytuacji nie podejmuje się regulacji procesu.

Do monitorowania procesu można też zastosować algorytm numeryczny. Przed rozpoczęciem monitorowania należy wówczas:

- określić wartości $\lambda_0, \lambda_1, \alpha$,
- wyznaczyć wartość odniesienia (k) według wzoru:

$$k = \operatorname{tg} \phi \frac{\lambda_1 - \lambda_0}{\ln \lambda_1 - \ln \lambda_0},$$

- wyznaczyć granicę regulacji procesu (h) według wzoru:

$$z_g = d \cdot c = -\frac{\ln \alpha}{\lambda_1 - \lambda_0} \cdot \frac{\lambda_1 - \lambda_0}{\ln \lambda_1 - \ln \lambda_0} = \frac{-\ln \alpha}{\ln \lambda_1 - \ln \lambda_0}.$$

Monitorowanie procesu przebiega według schematu przedstawionego w przykładzie 6.11:

- w kolejnych krokach postępowania kontrolnego (i) pobiera się do badania jednoelementowe próbki produktu albo dokonuje się pojedynczych pomiarów obserwowanej cechy procesu,
- uzyskiwane empirycznie wartości x_i porównuje się z wartością k ,
- w tym kroku i , w którym zostanie spełniona nierówność $x_i > k$, uruchamia się licznik indeksu operacyjnego (j) i rozpoczyna się obliczanie wartości sumy:

$$v_{j_r} = \sum_{j=1}^{j_r} (x_{(i)j} - k),$$

- obliczone wartości v_{j_r} porównuje się z granicą regulacji procesu z_g ,
- w tym kroku j_r , w którym zostanie spełniona nierówność $v_{j_r} > z_g$, następuje emisja sygnału o rozregulowaniu,
- konsekwencją tego sygnału jest regulacja procesu, a także zerowanie licznika operacyjnego (j) oraz zerowanie sumy v_{j_r} ,
- zerowanie licznika operacyjnego (j), a także zerowanie sumy v_{j_r} następuje również wtedy, kiedy zostanie spełniona nierówność $v_{j_r} < 0$. W takiej sytuacji, w kolejnym kroku postępowania powraca się do porównywania wartości x_i z wartością k .

Przedstawiony tu algorytm numeryczny może być wzbogacony o diagram przeglądowy, o konstrukcji zbliżonej do karty kontrolnej Shewharta. Diagram taki – jak to pokazano na rys. 6.17 – wyposażony jest w dwie osie poziome. Na jednej z nich odkłada się kolejne wartości indeksu bieżącego (i), na drugiej natomiast wartości indeksu operacyjnego (j). Na osi pionowej odkłada się wartości sumy skumulowanej v_{j_r} . Na diagramie tym wykreśla się granicę regulacji procesu (z_g), a następnie – na tak przygotowany rysunek – nanosi się kolejne wartości v_{j_r} . Emisja sygnału ostrzegającego o przejściu procesu ze stanu uregulowania w stan rozregulowania następuje wówczas, gdy suma skumulowana v_{j_r} przekroczy *in plus* granicę regulacji z_g .

Aneks

TABLICE

Objaśnienia do tablicy I

Tablica I podaje wartości całki:

$$\Theta(u) = \frac{1}{\sqrt{2\pi}} \int_0^u \exp\left(-\frac{u^2}{2}\right) du. \quad (1)$$

Między funkcją $\Theta(u)$ a dystrybucją zmiennej losowej U , wyrażoną wzorem:

$$\Phi(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^u \exp\left(-\frac{u^2}{2}\right) du, \quad (2)$$

zachodzą następujące związki:

– jeżeli $u_0 < 0$, to:

$$\Phi(u_0) = P(U < u_0) = 0,5 - \Theta(-u_0), \quad (3)$$

– jeżeli $u_0 \geq 0$, to:

$$\Phi(u_0) = P(U < u_0) = 0,5 + \Theta(u_0). \quad (4)$$

Na podstawie związków (3) oraz (4), przy korzystaniu z tablicy można wyznaczać wartości u_α , takie że:

$$P(U > u_\alpha) = \alpha. \quad (5)$$

Przykład

Chcemy wyznaczyć wartość u_α , taką że $P(U > u_\alpha) = \alpha = 0,025$. Jak wiadomo, $P(U > u_\alpha) = 1 - P(U < u_\alpha) = 1 - \Phi(u_\alpha)$. Dokonujemy obliczeń: $1 - \Phi(u_\alpha) = 0,025$. Tak więc: $\Phi(u_\alpha) = 0,975$. Z kolei w tablicy I znajdujemy wartość u_α ,

której odpowiada $\Theta(u_\alpha) = 0,4750$, ponieważ na podstawie (4) mamy: $\Phi(u_\alpha) = 0,5 + \Theta(u_\alpha)$. Na podstawie tablicy I: $\Theta(1,96) = 0,4750$. Poszukiwana wartość $u_{0,025}$ wynosi więc 1,96. Zauważmy, że wartość tę można również odczytać z tablic (zamieszczonych w innych opracowaniach) zawierających całki (2). Obliczamy wówczas: $1 - 0,025 = \Phi(u_{0,025}) = 0,975$. Wtedy dystrybuanta zmiennej losowej U o rozkładzie $N(0; 1)$ przyjmuje tę wartość w punkcie $u_\alpha = 1,96$.

Objaśnienia do tablicy II

W tablicy tej zestawione są wartości $t_{r,\varepsilon}$, takie że:

$$P(t_r > t_{r,\varepsilon}) = \varepsilon, \quad (6)$$

gdzie r oznacza liczbę stopni swobody, przy czym $r = n - 1$.

Przykład

Dla $r = n - 1 = 4$ stopni swobody i przy $\alpha/2 = 0,025$ należy znaleźć wartość $t_{r;\alpha/2}$. Przyjmujemy $\alpha/2 = \varepsilon$ i z tablicy odczytujemy wartość $t_{4;0,025} = 2,776$, spełniającą warunek $P(t_r \geq t_{r,\varepsilon}) = \varepsilon$.

Objaśnienia do tablicy III

Tablica ta podaje wartości $\chi_{r,\varepsilon}^2$, takie że:

$$P(\chi_r^2 > \chi_{r,\varepsilon}^2) = \varepsilon, \quad (7)$$

gdzie r oznacza liczbę stopni swobody.

Przykład

Dla $r = n - 1 = 3$ stopni swobody, przy $\alpha = 0,05$, należy znaleźć wartości $\chi_{3;1-\alpha}^2$ i $\chi_{3;\alpha}^2$.

W przypadku $\chi_{3;1-\alpha}^2$ mamy $\varepsilon = 1 - \alpha = 1 - 0,05 = 0,95$. Z tablicy odczytujemy: $\chi_{3;0,95}^2 = 0,352$. Natomiast w przypadku $\chi_{3;\alpha}^2$ mamy $\varepsilon = 0,05$. W konsekwencji, z tablicy: $\chi_{3;0,05}^2 = 7,815$.

Objaśnienia do tablicy IV

W tablicy tej, dla wybranych poziomów istotności, zestawione są wartości $F_{r_1, r_2, \alpha}$, dla których:

$$P(F_{r_1, r_2} \geq F_{r_1, r_2, \alpha}) = \alpha, \quad (8)$$

gdzie r_1, r_2 oznaczają liczby stopni swobody, przy czym $r_1 = n_1 - 1, r_2 = n_2 - 1$.

Tablice są tak zbudowane, że w kolumnach podane są liczby stopni swobody odpowiadające licznikowi S_1^2 , natomiast w wierszach – odpowiadające mianownikowi S_2^2 .

Przykład

Dla $r_1 = n_1 - 1 = 9, r_2 = n_2 - 1 = 9$ stopni swobody i przy $\alpha = 0,05$ należy znaleźć wartość $F_{r_1, r_2, \alpha}$. Z tablicy odczytujemy wartość $F_{9;9;0,05} = 3,18$, spełniającą warunek $P(F_{9;9} \geq F_{9;9;0,05}) = 0,05$.

Objaśnienia do tablicy IX

Spośród nieskończenie wielu funkcji wykładniczych $y = a^x$, gdzie $a > 0, a \neq 1$, szczególnie często wykorzystywane są dwie funkcje:

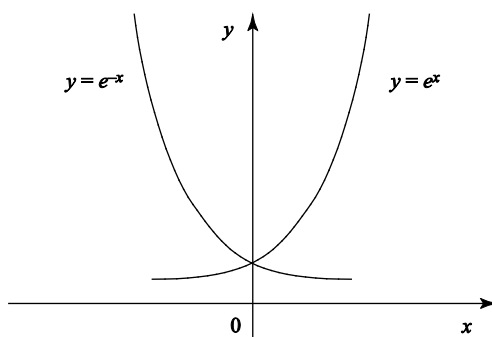
$$y = e^x \quad (10)$$

oraz

$$y = e^{-x}, \quad (11)$$

w których $e = 2,7182\dots$ jest podstawą logarytmów naturalnych.

Wykresy funkcji $y = e^x$ oraz $y = e^{-x}$ (otrzymanej przez obrót funkcji $y = e^x$ o 180° wokół osi y):



Tablica IX zawiera wartości funkcji wykładniczych e^x i e^{-x} dla wartości x od 0 do 10. Ponadto, wartości stałych e , $\lg e$, e^{-1} , a także $\lg e^{-1}$.

Tablica I

Wartości funkcji $\Theta(u)$

u	$\Theta(u)$														
	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09					
0,0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359					
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753					
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141					
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517					
0,4	0,1554	0,1591	0,1628	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,1879					
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224					
0,6	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2517	0,2549					
0,7	0,2580	0,2611	0,2642	0,2673	0,2704	0,2734	0,2764	0,2794	0,2823	0,2852					
0,8	0,2881	0,2910	0,2939	0,2967	0,2995	0,3023	0,3051	0,3078	0,3106	0,3133					
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365	0,3389					
1,0	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577	0,3599	0,3621					
1,1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810	0,3830					
1,2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015					
1,3	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147	0,4162	0,4177					
1,4	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279	0,4292	0,4306	0,4319					
1,5	0,4332	0,4345	0,4357	0,4370	0,4382	0,4394	0,4406	0,4418	0,4429	0,4441					

u	$\Theta(u)$									
	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
1,6	0,4452	0,4463	0,4474	0,4484	0,4495	0,4505	0,4515	0,4525	0,4535	0,4545
1,7	0,4554	0,4564	0,4573	0,4582	0,4591	0,4599	0,4608	0,4616	0,4625	0,4633
1,8	0,4641	0,4649	0,4656	0,4664	0,4671	0,4678	0,4686	0,4693	0,4699	0,4706
1,9	0,4713	0,4719	0,4726	0,4732	0,4738	0,4744	0,4750	0,4756	0,4761	0,4767
2,0	0,4772	0,4778	0,4783	0,4788	0,4793	0,4798	0,4803	0,4808	0,4812	0,4817
2,1	0,4821	0,4826	0,4830	0,4834	0,4838	0,4842	0,4846	0,4850	0,4854	0,4857
2,2	0,4861	0,4864	0,4868	0,4871	0,4875	0,4878	0,4881	0,4884	0,4887	0,4890
2,3	0,4893	0,4896	0,4898	0,4901	0,4904	0,4906	0,4909	0,4911	0,4913	0,4916
2,4	0,4918	0,4920	0,4922	0,4925	0,4927	0,4929	0,4931	0,4932	0,4934	0,4936
2,5	0,4938	0,4940	0,4941	0,4943	0,4945	0,4946	0,4948	0,4949	0,4951	0,4952
2,6	0,4953	0,4955	0,4956	0,4957	0,4959	0,4960	0,4961	0,4962	0,4963	0,4964
2,7	0,4965	0,4966	0,4967	0,4968	0,4969	0,4970	0,4971	0,4972	0,4973	0,4974
2,8	0,4974	0,4975	0,4976	0,4977	0,4977	0,4978	0,4979	0,4979	0,4980	0,4981
2,9	0,4981	0,4982	0,4982	0,4983	0,4984	0,4984	0,4985	0,4985	0,4986	0,4986
3,0	0,4987	0,4987	0,4987	0,4988	0,4988	0,4989	0,4989	0,4989	0,4990	0,4990

Kwantyle rozkładów zmiennej losowej t_r -Studenta

r	$t_{r,\varepsilon}$					r
	$t_{r,0,1}$	$t_{r,0,05}$	$t_{r,0,025}$	$t_{r,0,01}$	$t_{r,0,005}$	
1	3,078	6,314	12,706	31,821	63,657	1
2	1,886	2,920	4,303	6,965	9,925	2
3	1,638	2,353	3,182	4,541	5,841	3
4	1,533	2,132	2,776	3,747	4,604	4
5	1,476	2,015	2,571	3,365	4,032	5
6	1,440	1,943	2,447	3,143	3,707	6
7	1,415	1,895	2,365	2,998	3,499	7
8	1,397	1,860	2,306	2,896	3,355	8
9	1,383	1,833	2,262	2,821	3,250	9
10	1,372	1,812	2,228	2,764	3,169	10
11	1,363	1,796	2,201	2,718	3,106	11
12	1,356	1,782	2,179	2,681	3,055	12
13	1,350	1,771	2,160	2,650	3,012	13
14	1,345	1,761	2,145	2,624	2,977	14
15	1,341	1,753	2,131	2,602	2,947	15
16	1,337	1,746	2,120	2,583	2,921	16
17	1,333	1,740	2,110	2,567	2,898	17
18	1,330	1,734	2,101	2,552	2,878	18
19	1,328	1,729	2,093	2,539	2,861	19
20	1,325	1,725	2,086	2,528	2,845	20
21	1,323	1,721	2,080	2,518	2,831	21
22	1,321	1,717	2,074	2,508	2,819	22
23	1,319	1,714	2,069	2,500	2,807	23
24	1,318	1,711	2,064	2,492	2,797	24
25	1,316	1,708	2,060	2,485	2,787	25
26	1,315	1,706	2,056	2,479	2,779	26
27	1,314	1,703	2,052	2,473	2,771	27
28	1,313	1,701	2,048	2,467	2,763	28
29	1,311	1,699	2,045	2,462	2,756	29
∞	1,282	1,645	1,960	2,326	2,576	∞

Tablica III

Kwantyle rozkładów zmiennej losowej χ_r^2

r	$\chi_{r;\alpha}^2$										r
	$\chi_{r;0,0995}^2$	$\chi_{r;0,99}^2$	$\chi_{r;0,975}^2$	$\chi_{r;0,95}^2$	$\chi_{r;0,05}^2$	$\chi_{r;0,025}^2$	$\chi_{r;0,01}^2$	$\chi_{r;0,005}^2$			
1	0,0000393	0,000157	0,000982	0,00393	3,841	5,024	6,635	7,879	1		
2	0,0100	0,0201	0,0506	0,103	5,991	7,378	9,210	10,597	2		
3	0,0717	0,115	0,216	0,352	7,815	9,348	11,345	12,838	3		
4	0,207	0,297	0,484	0,711	9,488	11,143	13,277	14,860	4		
5	0,412	0,554	0,831	1,145	11,070	12,832	15,086	16,750	5		
6	0,676	0,872	1,237	1,635	12,592	14,449	16,812	18,548	6		
7	0,989	1,239	1,690	2,167	14,067	16,013	18,475	20,278	7		
8	1,344	1,646	2,180	2,733	15,507	17,535	20,090	21,955	8		
9	1,735	2,088	2,700	3,325	16,919	19,023	21,666	23,589	9		
10	2,156	2,558	3,247	3,940	18,307	20,483	23,209	25,188	10		
11	2,603	3,053	3,816	4,575	19,675	21,920	24,725	26,757	11		
12	3,074	3,571	4,404	5,226	21,026	23,337	26,217	28,300	12		
13	3,565	4,107	5,009	5,892	22,362	24,736	27,688	29,819	13		
14	4,075	4,660	5,629	6,571	23,685	26,119	29,141	31,319	14		
15	4,601	5,229	6,262	7,261	24,996	27,488	30,578	32,801	15		

cd. tablicy III

16	5,142	5,812	6,908	7,962	26,296	28,845	32,000	34,267	16
17	5,697	6,408	7,564	8,672	27,587	30,191	33,409	35,718	17
18	6,265	7,015	8,231	9,390	28,869	31,526	34,805	37,156	18
19	6,844	7,633	8,907	10,117	30,144	32,852	36,191	38,582	19
20	7,434	8,260	9,591	10,851	31,410	34,170	37,566	39,997	20
21	8,034	8,897	10,283	11,591	32,671	35,497	38,932	41,401	21
22	8,643	9,542	10,982	12,338	33,924	36,781	40,289	42,796	22
23	9,260	10,196	11,689	13,091	35,172	38,076	41,638	44,181	23
24	9,886	10,856	12,401	13,848	36,415	39,364	42,980	45,558	24
25	10,520	11,524	13,120	14,611	37,652	40,646	44,314	46,928	25
26	11,160	12,198	13,844	15,379	38,885	41,923	45,642	48,290	26
27	11,808	12,879	14,573	16,151	40,113	43,194	46,963	49,645	27
28	12,461	13,565	15,308	16,928	41,337	44,461	48,278	50,993	28
29	13,121	14,256	16,047	17,708	42,557	45,722	49,588	52,336	29
30	13,787	14,953	16,791	18,493	43,773	46,979	50,892	53,672	30

Kwantyle rozkładów zmiennej losowej F

		$F_{(n_1-1)(n_2-1);0,05}$																			
		stopnie swobody licznika (n_1-1)																			
		1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞	
stopnie swobody mianownika (n_2-1)	1	161	200	216	225	230	234	237	239	241	242	244	246	248	249	250	251	252	253	254	
	2	18,5	19,0	19,2	19,2	19,3	19,3	19,4	19,4	19,4	19,4	19,4	19,4	19,4	19,4	19,5	19,5	19,5	19,5	19,5	19,5
	3	10,1	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79	8,74	8,70	8,66	8,64	8,62	8,59	8,57	8,55	8,53	8,53
	4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,91	5,86	5,80	5,77	5,75	5,72	5,69	5,66	5,63	5,63
	5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74	4,68	4,62	4,56	4,53	4,50	4,46	4,43	4,40	4,37	4,37
	6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	4,00	3,94	3,87	3,84	3,81	3,77	3,74	3,70	3,67	3,67
	7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	3,57	3,51	3,44	3,41	3,38	3,34	3,30	3,27	3,23	3,23
	8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35	3,28	3,22	3,15	3,12	3,08	3,04	3,01	2,97	2,93	2,93
	9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14	3,07	3,01	2,94	2,90	2,86	2,83	2,79	2,75	2,71	2,71
	10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98	2,91	2,85	2,77	2,74	2,70	2,66	2,62	2,58	2,54	2,54
	11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85	2,79	2,72	2,65	2,61	2,57	2,53	2,49	2,45	2,40	2,40
	12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75	2,69	2,62	2,54	2,51	2,47	2,43	2,38	2,34	2,30	2,30
	13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67	2,60	2,53	2,46	2,42	2,38	2,34	2,30	2,25	2,21	2,21
	14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60	2,53	2,46	2,39	2,35	2,31	2,27	2,22	2,18	2,13	2,13
	15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54	2,48	2,40	2,33	2,29	2,25	2,20	2,16	2,11	2,07	2,07

		$F_{(n_1-1)(n_2-1);0,05}$																				
		stopnie swobody licznika (n_1-1)																				
		1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞		
stopnie swobody mianownika (n_2-1)	16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,42	2,35	2,28	2,24	2,19	2,15	2,11	2,06	2,01	2,00	
	17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45	2,45	2,38	2,31	2,23	2,19	2,15	2,10	2,06	2,01	1,96	1,96
	18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41	2,34	2,27	2,19	2,19	2,15	2,11	2,06	2,02	1,97	1,92	1,92
	19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38	2,31	2,23	2,16	2,16	2,11	2,07	2,03	1,98	1,93	1,88	1,88
	20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35	2,28	2,20	2,12	2,12	2,08	2,04	1,99	1,95	1,90	1,84	1,84
	21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32	2,25	2,18	2,10	2,10	2,05	2,01	1,96	1,92	1,87	1,81	1,81
	22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30	2,23	2,15	2,07	2,07	2,03	1,98	1,94	1,89	1,84	1,78	1,78
	23	4,28	3,42	3,03	2,80	2,64	2,53	2,44	2,37	2,32	2,27	2,20	2,13	2,05	2,05	2,01	1,96	1,91	1,86	1,81	1,76	1,76
	24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25	2,18	2,11	2,03	2,03	1,98	1,94	1,89	1,84	1,79	1,73	1,73
	25	4,24	3,39	2,99	2,76	2,60	2,49	2,40	2,34	2,28	2,24	2,16	2,09	2,01	2,01	1,96	1,92	1,87	1,82	1,77	1,71	1,71
	30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16	2,09	2,01	1,93	1,93	1,89	1,84	1,79	1,74	1,68	1,62	1,62
	40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08	2,00	1,92	1,84	1,84	1,79	1,74	1,69	1,64	1,58	1,51	1,51
	60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99	1,92	1,84	1,75	1,75	1,70	1,65	1,59	1,53	1,47	1,39	1,39
	120	3,92	3,07	2,68	2,45	2,29	2,18	2,09	2,02	1,96	1,91	1,83	1,75	1,66	1,66	1,61	1,55	1,50	1,43	1,35	1,25	1,25
	∞	3,84	3,00	2,60	2,37	2,21	2,10	2,01	1,94	1,88	1,83	1,75	1,67	1,57	1,57	1,52	1,46	1,39	1,32	1,22	1,00	1,00

Kwantyle rozkładów zmiennej losowej F

		$F_{(n_1-1)(n_2-1);0,01}$																			
		stopnie swobody licznika (n_1-1)																			
		1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞	
stopnie swobody mianownika (n_2-1)	1	4052	5000	5403	5625	5764	5859	5928	5982	6023	6056	6106	6157	6209	6235	6261	6287	6313	6339	6366	
	2	98,5	99,0	99,2	99,2	99,3	99,3	99,4	99,4	99,4	99,4	99,4	99,4	99,4	99,4	99,5	99,5	99,5	99,5	99,5	99,5
	3	34,1	30,8	29,5	28,7	28,2	27,9	27,7	27,5	27,5	27,3	27,2	27,1	26,9	26,7	26,6	26,5	26,4	26,3	26,2	26,1
	4	21,2	18,0	16,7	16,0	15,5	15,2	15,0	14,8	14,7	14,5	14,4	14,2	14,0	13,9	13,8	13,7	13,7	13,6	13,6	13,5
	5	16,3	13,3	12,1	11,4	11,0	10,7	10,5	10,3	10,2	10,1	9,89	9,72	9,55	9,47	9,38	9,29	9,20	9,11	9,02	9,02
	6	13,7	10,9	9,78	9,15	8,75	8,47	8,26	8,10	7,98	7,87	7,72	7,56	7,40	7,31	7,23	7,14	7,06	6,97	6,88	6,88
	7	12,2	9,55	8,45	7,85	7,46	7,19	6,99	6,84	6,72	6,62	6,47	6,31	6,16	6,07	5,99	5,91	5,82	5,74	5,65	5,65
	8	11,3	8,65	7,59	7,01	6,63	6,37	6,18	6,03	5,91	5,81	5,67	5,52	5,36	5,28	5,20	5,12	5,03	4,95	4,86	4,86
	9	10,6	8,02	6,99	6,42	6,06	5,80	5,61	5,47	5,35	5,26	5,11	4,96	4,81	4,73	4,65	4,57	4,48	4,40	4,31	4,31
	10	10,0	7,56	6,55	5,99	5,64	5,39	5,20	5,06	4,94	4,85	4,71	4,56	4,41	4,33	4,25	4,17	4,08	4,00	3,91	3,91
	11	9,65	7,21	6,22	5,67	5,32	5,07	4,89	4,74	4,63	4,54	4,40	4,25	4,10	4,02	3,94	3,86	3,78	3,69	3,60	3,60
	12	9,33	6,93	5,95	5,41	5,06	4,82	4,64	4,50	4,39	4,30	4,16	4,01	3,86	3,78	3,70	3,62	3,54	3,45	3,36	3,36
	13	9,07	6,70	5,74	5,21	4,86	4,62	4,44	4,30	4,19	4,10	3,96	3,82	3,66	3,59	3,51	3,43	3,34	3,25	3,17	3,17
	14	8,86	6,51	5,56	5,04	4,70	4,46	4,28	4,14	4,03	3,94	3,80	3,66	3,51	3,43	3,35	3,27	3,18	3,09	3,00	3,00
	15	8,68	6,36	5,42	4,89	4,56	4,32	4,14	4,00	3,89	3,80	3,67	3,52	3,37	3,29	3,21	3,13	3,05	2,96	2,87	2,87

		$F_{(n_1-1)(n_2-1);0,01}$																		
		stopnie swobody licznika (n_1-1)																		
stopnie swobody mianownika (n_2-1)		1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
		16	8,53	6,23	5,29	4,77	4,44	4,20	4,03	3,89	3,78	3,69	3,55	3,41	3,26	3,18	3,10	3,02	2,93	2,84
17	8,40	6,11	5,19	4,67	4,34	4,10	3,93	3,79	3,68	3,59	3,46	3,31	3,16	3,08	3,00	2,92	2,83	2,75	2,65	2,65
18	8,29	6,01	5,09	4,58	4,25	4,01	3,84	3,71	3,60	3,51	3,37	3,23	3,08	3,00	2,92	2,84	2,75	2,66	2,57	2,57
19	8,19	5,93	5,01	4,50	4,17	3,94	3,77	3,63	3,52	3,43	3,30	3,15	3,00	2,92	2,84	2,76	2,67	2,58	2,49	2,49
20	8,10	5,85	4,94	4,43	4,10	3,87	3,70	3,56	3,46	3,37	3,23	3,09	2,94	2,86	2,78	2,69	2,61	2,52	2,42	2,42
21	8,02	5,78	4,87	4,37	4,04	3,81	3,64	3,51	3,40	3,31	3,17	3,03	2,88	2,80	2,72	2,64	2,55	2,46	2,36	2,36
22	7,95	5,72	4,82	4,31	3,99	3,76	3,59	3,45	3,35	3,26	3,12	2,98	2,83	2,75	2,67	2,58	2,50	2,40	2,31	2,31
23	7,88	5,66	4,76	4,26	3,94	3,71	3,54	3,41	3,30	3,21	3,07	2,93	2,78	2,70	2,62	2,54	2,45	2,35	2,26	2,26
24	7,82	5,61	4,72	4,22	3,90	3,67	3,50	3,36	3,26	3,17	3,03	2,89	2,74	2,66	2,58	2,49	2,40	2,31	2,21	2,21
25	7,77	5,57	4,68	4,18	3,86	3,63	3,46	3,32	3,22	3,13	2,99	2,85	2,70	2,62	2,53	2,45	2,36	2,27	2,17	2,17
30	7,56	5,39	4,51	4,02	3,70	3,47	3,30	3,17	3,07	2,98	2,84	2,70	2,55	2,47	2,39	2,30	2,21	2,11	2,01	2,01
40	7,31	5,18	4,31	3,83	3,51	3,29	3,12	2,99	2,89	2,80	2,66	2,52	2,37	2,29	2,20	2,11	2,02	1,92	1,80	1,80
60	7,08	4,98	4,13	3,65	3,34	3,12	2,95	2,82	2,72	2,63	2,50	2,35	2,20	2,12	2,03	1,94	1,84	1,73	1,60	1,60
120	6,85	4,79	3,95	3,48	3,17	2,96	2,79	2,66	2,56	2,47	2,34	2,19	2,03	1,95	1,86	1,76	1,66	1,53	1,38	1,38
∞	6,63	4,61	3,78	3,32	3,02	2,80	2,64	2,51	2,41	2,32	2,18	2,04	1,88	1,79	1,70	1,59	1,47	1,32	1,00	1,00

Tablica V
Parametry rozkładu rozstepu

n	d_n	$1/d_n$	k_n
2	1,128	0,886	0,853
3	1,693	0,591	0,888
4	2,059	0,486	0,880
5	2,327	0,430	0,864
6	2,534	0,395	0,848
7	2,704	0,370	0,833
8	2,847	0,351	0,820
9	2,976	0,337	0,808
10	3,078	0,325	0,797

Tablica VI/2

Kwantyle rozkładu serii

n_a n_b		$P(k_{n_a n_b} \leq k_{n_a n_b; 0,05}) = 0,05$																				$P(k_{n_a n_b} \leq k_{n_a n_b; 0,95}) = 0,95$																			
2	2	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20		
3	3																				4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20				
4	4																				5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20					
5	5																				5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20					
6	6																				5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20					
7	7																				5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20					
8	8																				5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20					
9	9																				5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20					
10	10																				5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20					
11	11																				5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20					
12	12																				5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20					
13	13																				5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20					
14	14																				5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20					
15	15																				5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20					
16	16																				5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20					
17	17																				5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20					
18	18																				5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20					
19	19																				5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20					
20	20																				5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20					

Tabela VI/3

Kwantyle rozkładu serii $P(k \leq k_1) = \alpha$

$n_{\alpha} = h_b$	α					$n_{\alpha} = h_b$	α	α					$n_{\alpha} = h_b$	α	α				
	0,025	0,05	0,95	0,975	0,975			0,025	0,05	0,95	0,975	0,975			0,025	0,05	0,95	0,975	
10	6	6	15	15	42	33	56	46	47	66	67	79	67	69	90	92			
11	7	7	16	16	43	34	57	47	48	67	68	80	68	70	91	93			
12	7	8	17	18	44	35	58	47	49	68	70	81	69	71	92	94			
13	8	9	18	19	45	36	59	48	50	69	71	82	69	71	94	96			
14	9	10	19	20	46	37	60	49	51	70	72	83	70	72	95	97			
15	10	11	20	21	47	38	61	50	52	71	73	84	71	73	96	98			
16	11	11	22	22	49	39	62	51	53	72	74	85	72	74	97	99			
17	11	12	23	24	50	40	63	52	54	73	75	86	73	75	98	100			
18	12	13	24	25	51	41	64	53	55	74	76	87	74	76	99	101			
19	13	14	25	26	52	42	65	54	56	75	77	88	75	77	100	102			
20	14	15	26	27	53	43	66	55	57	76	78	89	76	78	101	103			
21	15	16	27	28	54	44	67	56	58	77	79	90	77	79	102	104			
22	16	17	28	29	55	45	68	57	58	79	80	91	78	80	103	105			
23	16	17	30	31	56	46	69	58	59	80	81	92	79	81	104	106			
24	17	18	31	32	57	47	70	58	60	81	83	93	80	82	105	107			
25	18	19	32	33	59	48	71	59	61	82	84	94	81	83	106	108			
26	19	20	33	34	60	49	72	60	62	83	85	95	82	84	107	109			
27	20	21	34	35	61	50	73	61	63	84	86	96	82	85	108	111			
28	21	22	35	36	62	51	74	62	64	85	87	97	83	86	109	112			
29	22	23	36	37	63	52	75	63	65	86	88	98	84	87	110	113			
30	22	24	37	39	64	53	76	64	66	87	89	99	85	87	112	114			
31	23	25	38	40	65	54	77	65	67	88	90	100	86	88	113	115			
32	24	25	40	41	66	55	78	66	68	89	91								

Tablica VII

Fragment tablicy liczb losowych

Blok 1										
Nr wiersza	Nr kolumny									
	1	2	3	4	5	6	7	8	9	10
1	87 35	67 44	51 49	18 98	97 84	75 22	53 29	10 52	26 87	54 92
2	25 52	29 67	35 99	48 88	40 68	63 68	82 39	38 47	91 39	11 00
3	87 17	83 31	25 59	87 48	25 80	24 08	81 45	21 32	90 08	44 31
4	05 04	40 35	72 95	48 56	77 57	63 19	80 16	48 52	06 47	64 98
5	81 16	09 24	91 71	29 76	54 01	53 47	30 67	62 95	56 58	10 91
6	54 85	79 88	57 91	11 69	10 22	71 87	24 92	52 64	42 82	78 95
7	44 78	19 18	35 40	27 66	89 72	21 17	71 69	95 17	97 17	62 60
8	97 20	98 97	37 33	93 75	18 88	35 85	46 05	07 20	08 17	66 24
9	98 77	57 51	40 41	76 24	18 54	60 61	79 13	94 57	50 73	89 68
10	78 12	77 30	83 30	59 28	73 33	47 07	60 07	45 38	82 10	73 19
11	41 19	70 62	43 46	06 13	22 38	31 18	64 60	07 14	49 16	28 16
12	70 64	30 55	67 46	95 79	63 66	82 56	67 10	76 77	03 22	42 18
13	06 56	09 89	68 87	79 19	35 94	66 18	17 94	72 81	72 77	92 39
14	29 46	18 28	08 88	48 56	49 44	67 82	72 67	28 83	10 26	58 13
15	42 14	55 51	72 95	29 25	15 18	25 68	48 92	87 16	78 43	17 47
16	33 75	87 15	15 23	13 79	62 73	76 09	09 77	82 65	72 47	59 56
17	09 80	99 61	98 08	34 11	88 79	08 32	46 78	33 58	44 16	12 23
18	98 31	57 50	85 80	53 39	05 92	54 42	29 01	35 23	09 84	96 64
19	51 70	52 55	83 12	95 02	79 11	49 79	87 95	98 48	88 68	64 77
20	27 83	61 07	49 05	46 20	35 78	31 34	42 50	68 11	42 14	29 77
21	78 84	69 15	64 42	92 39	36 08	56 39	35 02	92 78	48 63	82 98
22	22 12	89 66	49 09	99 10	62 53	19 31	81 83	50 43	37 42	10 00
23	69 41	59 54	82 72	44 66	64 03	76 59	12 12	41 56	34 90	26 06
24	54 99	46 54	51 38	59 07	64 21	81 17	88 47	23 05	63 43	08 67
25	99 91	82 79	92 62	44 24	01 34	45 16	33 56	17 78	42 86	70 94

Tablica VIII

Fragment tablicy liczb złotych

1	2	3	4	5	6	7	8	9	10
0,6180	31,5197	62,4214	93,3231	124,2248	155,1265	186,0282	216,9299	247,8316	278,7333
1,2361	32,1378	63,0395	93,9412	124,8429	155,7446	186,6463	217,5480	248,4497	279,3514
1,8541	32,7558	63,6575	94,5592	125,4609	156,3626	187,2643	218,1660	249,0677	279,9694
2,4721	33,3738	64,2755	95,1772	126,0789	156,9806	187,8823	218,7840	249,6857	280,5874
3,0902	33,9919	64,8936	95,7953	126,6970	157,5987	188,5004	219,4021	250,3038	281,2055
3,7082	34,6099	65,5116	96,4133	127,3150	158,2167	189,1184	220,0201	250,9218	281,8235
4,3262	35,2279	66,1296	97,0313	127,9330	158,8347	189,7364	220,6381	251,5398	282,4415
4,9443	35,8460	66,7477	97,6494	128,5511	159,4528	190,3545	221,2562	252,1579	283,0596
5,5623	36,4640	67,3657	98,2674	129,1691	160,0708	190,9725	221,8742	252,7759	283,6776
6,1803	37,0820	67,9837	98,8854	129,7871	160,6888	191,5905	222,4922	253,3939	284,2956
6,7984	37,7001	68,6018	99,5035	130,4052	161,3069	192,2086	223,1103	254,0120	284,9137
7,4164	38,3181	69,2198	100,1215	131,0232	161,9249	192,8266	223,7283	254,6300	285,5317
8,0344	38,9361	69,8378	100,7395	131,6412	162,5429	193,4446	224,3463	255,2480	286,1497
8,6525	39,5542	70,4559	101,3576	132,2593	163,1610	194,0627	224,9644	255,8661	286,7678
9,2705	40,1722	71,0739	101,9756	132,8773	163,7790	194,6807	225,5824	256,4841	287,3858
9,8885	40,7902	71,6919	102,5936	133,4953	164,3970	195,2987	226,2004	257,1021	288,0038
10,5066	41,4083	72,3100	103,2117	134,1134	165,0151	195,9168	226,8185	257,7202	288,6219
11,1246	42,0263	72,9280	103,8297	134,7314	165,6331	196,5348	227,4365	258,3382	289,2399
11,7426	42,6443	73,5460	104,4477	135,3494	166,2511	197,1528	228,0545	258,9562	289,8579
12,3607	43,2624	74,1641	105,0658	135,9675	166,8692	197,7709	228,6726	259,5743	290,4760

cd. tablicy VIII

1	2	3	4	5	6	7	8	9	10
12,9787	43,8804	74,7821	105,6838	136,5855	167,4872	198,3889	229,2906	260,1923	291,0940
13,5967	44,4984	75,4001	106,3018	137,2035	168,1052	199,0069	229,9086	260,8103	291,7120
14,2148	45,1165	76,0182	106,9199	137,8216	168,7233	199,6250	230,5267	261,4284	292,3301
14,8328	45,7345	76,6362	107,5379	138,4396	169,3413	200,2430	231,1447	262,0464	292,9481
15,4508	46,3525	77,2542	108,1559	139,0576	169,3593	200,8610	231,7627	262,6644	293,5661
16,0689	46,9706	77,8723	108,7740	139,6757	170,5774	201,4791	232,3808	263,2825	294,1842
16,6869	47,5886	78,4903	109,3920	140,2937	171,1954	202,0971	232,9988	263,9005	294,8022
17,3050	48,2067	79,1084	110,0100	140,9117	171,8134	202,7151	233,6168	264,5185	295,4202
17,9230	48,8247	79,7264	110,6281	141,5298	172,4315	203,3332	234,2349	265,1366	296,0383
18,5410	49,4427	80,3444	111,2461	142,1478	173,0495	203,9512	234,8529	265,7546	296,6563
19,1591	50,0608	80,9625	111,8642	142,7659	173,6676	204,5693	235,4709	266,3726	297,2743
19,7771	50,6788	81,5805	112,4822	143,3839	174,2856	205,1873	236,0890	266,9907	297,8924
20,3951	51,2968	82,1985	113,1002	144,0019	174,9036	205,8053	236,7070	267,6087	298,5104
21,0132	51,9149	82,8166	113,7183	144,6200	175,5217	206,4234	237,3251	268,2268	299,1285
21,6312	52,5329	83,4346	114,3363	145,2380	176,1397	207,0414	237,9431	268,8448	299,7465
22,2492	53,1509	84,0526	114,9543	145,8560	176,7577	207,6594	238,5611	269,4628	300,3645
22,8673	53,7690	84,6707	115,5724	146,4741	177,3758	208,2775	239,1792	270,0809	300,9826
23,4853	54,3870	85,2887	116,1904	147,0921	177,9938	208,8955	239,7972	270,6989	301,6006
24,1033	55,0050	85,9067	116,8084	147,7101	178,6118	209,5135	240,4152	271,3169	302,2186
24,7214	55,6231	86,5248	117,4265	148,3282	179,2299	210,1316	241,0333	271,9350	302,8367

cd. tablicy VIII

1	2	3	4	5	6	7	8	9	10
25,3394	56,2411	87,1428	118,0445	148,9462	179,8479	210,7496	241,6513	272,5530	303,4547
25,9574	56,8591	87,7608	118,6625	149,5642	180,4659	211,3676	242,2693	273,1710	304,0727
26,5755	57,4772	88,3789	119,2806	150,1823	181,0840	211,9857	243,8874	273,7891	304,6908
27,1935	58,0952	88,9969	119,8986	150,8003	181,7020	212,6037	243,5054	274,4071	305,3088
27,8115	58,7132	89,6149	120,5166	151,4183	182,3200	213,2217	244,1234	275,0251	305,9268
28,4296	59,3313	90,2330	121,1347	152,0364	182,9381	213,8398	244,7415	275,6432	306,5449
29,0476	59,9493	90,8510	121,7527	152,6544	183,5561	214,4578	245,3595	276,2612	307,1629
29,6656	60,5673	91,4690	122,3707	153,2724	184,1741	215,0758	245,9775	276,8792	307,7809
30,2837	61,1854	92,0871	122,9888	153,8905	184,7922	215,6939	246,5956	277,4973	308,3990
30,9017	61,8034	92,7051	123,6068	154,5085	185,4102	216,3119	247,2136	278,1153	309,0170

Tablica IX

Funkcja wykładnicza e^x i e^{-x}

x	e^x	e^{-x}	x	e^x	e^{-x}	x	e^x	e^{-x}	x	e^x	e^{-x}
0,00	1,0000	1,0000	0,50	1,6487	0,6065	1,0	2,7183	0,3679	6,0	403,43	0,0025
0,01	1,0101	0,9900	0,51	1,6653	0,6005	1,1	3,0043	0,3329	6,1	445,86	0,0022
0,02	1,0202	0,9802	0,52	1,6820	0,5945	1,2	3,3201	0,3012	6,2	492,75	0,0020
0,03	1,0305	0,9704	0,53	1,6989	0,5886	1,3	3,6693	0,2725	6,3	544,57	0,0018
0,04	1,0408	0,9608	0,54	1,7160	0,5827	1,4	4,0552	0,2466	6,4	601,85	0,0017
0,05	1,0513	0,9512	0,55	1,7333	0,5769	1,5	4,4817	0,2231	6,5	665,14	0,0015
0,06	1,0618	0,9418	0,56	1,7507	0,5712	1,6	4,9530	0,2019	6,6	735,10	0,0014
0,07	1,0725	0,9324	0,57	1,7683	0,5655	1,7	5,4739	0,1827	6,7	812,41	0,0012
0,08	1,0833	0,9231	0,58	1,7860	0,5599	1,8	6,0496	0,1653	6,8	897,85	0,0011
0,09	1,0942	0,9139	0,59	1,8040	0,5543	1,9	6,6859	0,1496	6,9	992,27	0,0010
0,10	1,1052	0,9048	0,60	1,8221	0,5488	2,0	7,3891	0,1353	7,0	1096,6	0,0009
0,11	1,1163	0,8958	0,61	1,8404	0,5434	2,1	8,1662	0,1225	7,1	1212,0	0,0008
0,12	1,1275	0,8869	0,62	1,8589	0,5379	2,2	9,0250	0,1108	7,2	1339,4	0,0007
0,13	1,1388	0,8781	0,63	1,8776	0,5326	2,3	9,9742	0,1003	7,3	1480,3	0,0007
0,14	1,1503	0,8694	0,64	1,8965	0,5273	2,4	11,023	0,0907	7,4	1636,0	0,0006
0,15	1,1618	0,8607	0,65	1,9155	0,5220	2,5	12,182	0,0821	7,5	1808,0	0,0006
0,16	1,1735	0,8521	0,66	1,9348	0,5169	2,6	13,464	0,0743	7,6	1998,2	0,0005
0,17	1,1853	0,8437	0,67	1,9542	0,5117	2,7	14,880	0,0672	7,7	2208,3	0,0005
0,18	1,1972	0,8353	0,68	1,9739	0,5066	2,8	16,445	0,0608	7,8	2440,6	0,0004
0,19	1,2092	0,8270	0,69	1,9937	0,5016	2,9	18,174	0,0550	7,9	2697,3	0,0004
0,20	1,2214	0,8187	0,70	2,0138	0,4966	3,0	20,086	0,0498	8,0	2981,0	0,0003
0,21	1,2337	0,8106	0,71	2,0340	0,4916	3,1	22,198	0,0450	8,1	3294,5	0,0003
0,22	1,2461	0,8025	0,72	2,0544	0,4868	3,2	24,533	0,0408	8,2	3641,0	0,0003
0,23	1,2586	0,7945	0,73	2,0751	0,4819	3,3	27,113	0,0369	8,3	4023,9	0,0002
0,24	1,2712	0,7866	0,74	2,0959	0,4771	3,4	29,964	0,0334	8,4	4447,1	0,0002

x	e^x	e^{-x}	x	e^x	e^{-x}	x	e^x	e^{-x}	x	e^x	e^{-x}	x	e^x	e^{-x}
0,25	1,2840	0,7788	0,75	2,1170	0,4724	3,5	33,115	0,0302	8,5	4914,8	0,0002			
0,26	1,2969	0,7711	0,76	2,1383	0,4677	3,6	36,598	0,0273	8,6	5431,7	0,0002			
0,27	1,3100	0,7634	0,77	2,1598	0,4630	3,7	40,447	0,0247	8,7	6002,9	0,0002			
0,28	1,3231	0,7558	0,78	2,1815	0,4584	3,8	44,701	0,0224	8,8	6634,2	0,0002			
0,29	1,3364	0,7483	0,79	2,2034	0,4538	3,9	49,402	0,0202	8,9	7332,0	0,0001			
0,30	1,3499	0,7408	0,80	2,2255	0,4493	4,0	54,598	0,0183	9,0	8103,1	0,0001			
0,31	1,3634	0,7334	0,81	2,2479	0,4449	4,1	60,340	0,0166	9,1	8955,3	0,0001			
0,32	1,3771	0,7261	0,82	2,2705	0,4404	4,2	66,686	0,0150	9,2	9897,1	0,0001			
0,33	1,3910	0,7189	0,83	2,2933	0,4360	4,3	73,700	0,0136	9,3	10938	0,0001			
0,34	1,4049	0,7118	0,84	2,3164	0,4317	4,4	81,451	0,0123	9,4	12088	0,0001			
0,35	1,4191	0,7047	0,85	2,3396	0,4274	4,5	90,017	0,0111	9,5	13360	0,0001			
0,36	1,4333	0,6977	0,86	2,3632	0,4232	4,6	99,484	0,0101	9,6	14765	0,0001			
0,37	1,4477	0,6907	0,87	2,3869	0,4190	4,7	109,95	0,0091	9,7	16318	0,0001			
0,38	1,4623	0,6839	0,88	2,4109	0,4148	4,8	121,51	0,0082	9,8	18034	0,0001			
0,39	1,4770	0,6771	0,89	2,4351	0,4107	4,9	134,29	0,0074	9,9	19930	0,0001			
0,40	1,4918	0,6703	0,90	2,4596	0,4066	5,0	148,41	0,0067	10,0	22026	0,0005			
0,41	1,5068	0,6637	0,91	2,4843	0,4025	5,1	164,02	0,0061						
0,42	1,5220	0,6570	0,92	2,5093	0,3985	5,2	181,27	0,0055						
0,43	1,5373	0,6505	0,93	2,5345	0,3946	5,3	200,34	0,0050						
0,44	1,5527	0,6440	0,94	2,5600	0,3906	5,4	221,41	0,0045						
0,45	1,5683	0,6376	0,95	2,5857	0,3867	5,5	244,69	0,0041						
0,46	1,5841	0,6313	0,96	2,6117	0,3829	5,6	270,43	0,0037						
0,47	1,6000	0,6250	0,97	2,6379	0,3791	5,7	298,87	0,0033						
0,48	1,6161	0,6188	0,98	2,6645	0,3753	5,8	330,30	0,0030						
0,49	1,6323	0,6126	0,99	2,6912	0,3716	5,9	365,04	0,0027						
0,50	1,6487	0,6065	1,00	2,7183	0,3679	6,0	403,43	0,0025						

$e = 2,71828183$
 $\lg e = 0,434229448$
 $e^{-1} = 0,36787944$
 $\lg e^{-1} = 0,56570552 - 1$