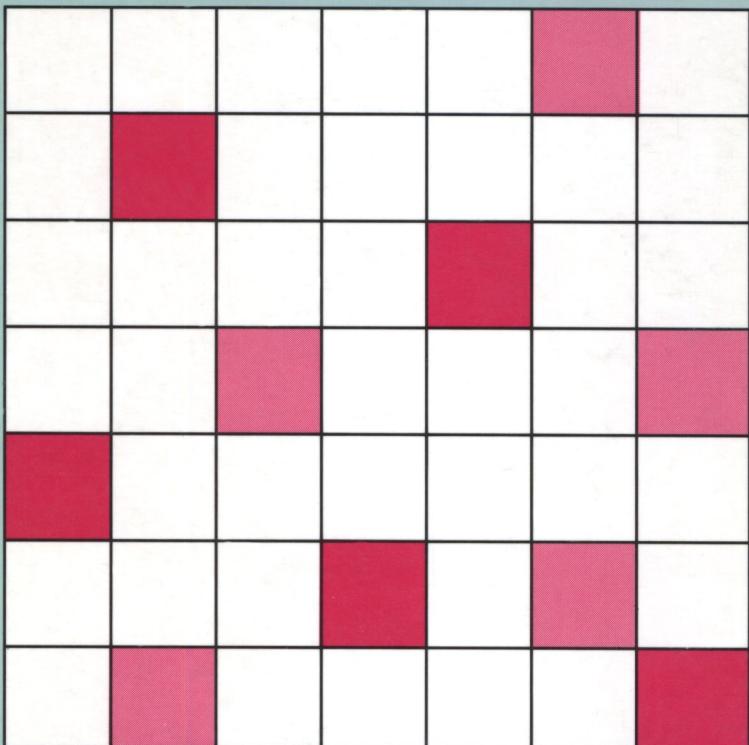


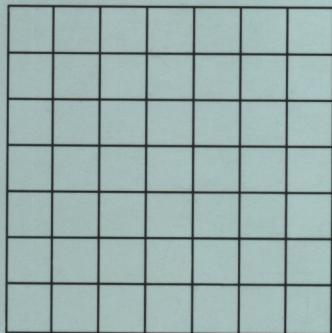
Zdzisław Hellwig

**ELEMENTY RACHUNKU
PRAWDOPODOBIEŃSTWA**

**STATYSTYKI
MATEMATYCZNEJ**



Wydawnictwo Naukowe PWN



Niniejsza książka jest kolejnym wydaniem podręcznika podstaw rachunku prawdopodobieństwa i statystyki matematycznej w ujęciu elementarnym. Zawiera także teorię regresji i korelacji oraz podstawy analizy procesów stochastycznych. Przystępny, zwięzły wykład uzupełniają liczne zadania i poglądowe przykłady.

Podręcznik przeznaczony jest dla studentów i wykładowców uczelni ekonomicznych, rolniczych i politechnik. Do swobodnego korzystania z niego wymagana jest znajomość podstaw rachunku różniczkowego i całkowego oraz ogólnej teorii statystyki (zwanej też statystyką teoretyczną).

ISBN 83-01-02137-3

9 788301 021375

ELEMENTY RACHUNKU
PRAWDOPODOBIEŃSTWA



**STATYSTYKI
MATEMATYCZNEJ**

Zdzisław Hellwig

**ELEMENTY RACHUNKU
PRAWDOPODOBIEŃSTWA**



**STATYSTYKI
MATEMATYCZNEJ**

Wydanie trzynaste



*Wydawnictwo Naukowe PWN
Warszawa 1998*

Okładkę projektowała

Anna Gogolewska

Tytuł dotowany przez Ministra Edukacji Narodowej

© Copyright by
 Państwowe Wydawnictwo Naukowe
 Warszawa 1967, 1970, 1974, 1987

Copyright © by
 Wydawnictwo Naukowe PWN Sp. z o.o.
 Warszawa 1993, 1995

Copyright © by
 Wydawnictwo Naukowe PWN SA
 Warszawa 1998

ISBN 83-01-02137-3

Wydawnictwo Naukowe PWN SA
 Wydanie trzynaste
 Arkuszy drukarskich 19,5
 Druk ukończono w marcu 1998 r.
 Druk i oprawa: Rzeszowskie Zakłady Graficzne
 35-025 Rzeszów, ul. płk. L. Lisa-Kuli 19
 Zam. 200/98

OD AUTORA

Niniejszy podręcznik zawiera wykład rachunku prawdopodobieństwa i statystyki matematycznej w ujęciu elementarnym. Po opanowaniu przedstawionego tu materiału czytelnik będzie mógł pogłębić swe wiadomości posługując się trudniejszymi opracowaniami.

Dla swobodnego korzystania z podręcznika wymagana jest znajomość podstaw rachunku różniczkowego i całkowego oraz przedmiotu znanego pod nazwą statystyki teoretycznej lub ogólnej teorii statystyki.

Ponieważ samodzielnego studiowania rachunku prawdopodobieństwa i statystyki matematycznej związane jest z poważnymi trudnościami natury matematycznej, przeto aby trudności te zmniejszyć, zastosowano następujące środki:

1. Część I podręcznika poświęcono wybranym zagadnieniom z matematyki, których znajomość jest niezbędna do opanowania treści wykładu.
2. Przy powoływaniu się na jakiekolwiek twierdzenie z matematyki, którego dowodu w książce nie przytoczono, podano w nawiasie kwadratowym pozycję literatury i stronę, na której czytelnik to twierdzenie i dowód z łatwością może znaleźć.
3. Dowody twierdzeń przytoczone w podręczniku zaopatrzone są w niezbędne wyjaśnienia, tak że ich przerobienie nie przedstawia poważniejszych trudności.
4. Ponieważ w praktyce wiadomo, że czytelników o słabszym przygotowaniu z matematyki odstręczają od samodzielnego studiowania rachunku prawdopodobieństwa i statystyki matematycznej skomplikowane wzory, nieznane znaki i symbole, przeto te ustępy książki, w których nie udało się uniknąć zastosowania trudniejszej aparatury matematycznej, ujęto klamrą znaków ► ◀. Ustępy te można pominąć bez szkody dla zrozumienia całości wykładu.
5. Podręcznik podzielony został na trzy części: I wybrane zagadnienia z algebry i analizy matematycznej, II rachunek prawdopodobieństwa, III statystyka matematyczna. Każda część dzieli się na rozdziały, a te z kolei na paragrafy i punkty. Na ich oznaczenie zastosowano dziesiętny system numeracji. W systemie tym pierwsza liczba oznacza rozdział, druga – paragraf, trzecia – punkt. I tak np. 2.2.1 oznacza pierwszy

punkt drugiego paragrafu w 2 rozdziale. Numeracja wzorów, tablic i rysunków biegnie oddzielnie w każdej numerowanej części pracy.

6. Liczby w nawiasach kwadratowych [] podają numery pozycji w spisie cytowanej literatury.

Część I

**WYBRANE ZAGADNIENIA Z ALGEBRY
I ANALIZY MATEMATYCZNEJ**

1.1. KOMBINATORYKA⁽¹⁾

1.1.1. Pojęcie silni

Obierzmy jakąś całkowitą dodatnią liczbę n , $n > 1$. Iloczyn kolejnych liczb naturalnych

$$1 \cdot 2 \cdot \dots \cdot n$$

oznaczać będziemy symbolem $n!$ i nazywać n *silnią*.

Zgodnie z tym określeniem $2! = 1 \cdot 2 = 2$, $3! = 1 \cdot 2 \cdot 3 = 6$ itd. Wartości silni wzrastają bardzo szybko ze wzrostem n . Łatwo się o tym przekonać patrząc na następującą tablicę, która zawiera wartości silni od 1 do 14:

Tablica 1

n	$n!$	n	$n!$
1	1	8	40 320
2	2	9	362 880
3	6	10	3 628 800
4	24	11	39 916 800
5	120	12	479 001 600
6	720	13	6 227 020 800
7	5 040	14	87 178 291 200

Uwaga. Przyjmuje się umownie, że $0! = 1! = 1$. Konwencja ta, jak zobaczymy, okaże się w dalszych rozważaniach bardzo wygodna.

W praktycznych obliczeniach często mamy do czynienia z ułamkami, w których liczniku i mianowniku występują silnie. Obliczenie takich ułamków staje się znacznie łatwiejsze, gdy zamiast obliczania występujących w tych ułamkach wartości silni i dzielenia ich przez siebie, dokonamy uprzednio wszelkich możliwych uproszczeń. Mamy np. obliczyć, czemu się równa

$$\frac{5!}{4!}.$$

Postępujemy w sposób następujący. Rozpisujemy odpowiednie silnie jako iloczyny kolej-

⁽¹⁾ Pełny wykład kombinatoryki znajdzie czytelnik w pracy [21].

nych czynników, a następnie skreślamy jednakowe czynniki w liczniku i mianowniku:

$$\frac{5!}{4!} = \frac{1 \cdot 2 \cdot 3 \cdot 4 \cdot 5}{1 \cdot 2 \cdot 3 \cdot 4} = 5.$$

Takie postępowanie jest szczególnie wygodne w przypadku, gdy mamy obliczyć wartość ułamka, w którego liczniku i mianowniku znajdują się silnie dwóch dużych, bliskich siebie liczb naturalnych. Obliczenie takiego ułamka bez dokonania uprzednio odpowiednich uproszczeń byłoby kłopotliwe. Musimy obliczyć np. $1000! : 998!$. Rozwijamy licznik i mianownik w iloczyn kolejnych liczb naturalnych. Mamy

$$\frac{1000!}{998!} = \frac{1 \cdot 2 \cdot \dots \cdot 997 \cdot 998 \cdot 999 \cdot 1000}{1 \cdot 2 \cdot \dots \cdot 997 \cdot 998} = 999 \cdot 1000 = 999\,000.$$

Przedstawimy proces upraszczania silni na liczbach ogólnych. Dane są dwie liczby naturalne n i m , przy czym $m < n$. Niech

$$(1) \quad n - m = k.$$

W takim razie

$$(2) \quad \frac{n!}{m!} = \frac{1 \cdot 2 \cdot \dots \cdot (m-1) m(m+1) \dots (m+k-1)(m+k)}{1 \cdot 2 \cdot \dots \cdot (m-1) \cdot m} = \\ = (m+1)(m+2) \dots (m+k-1)(m+k).$$

Z równości (1) wynika, że $m = n - k$. Podstawiając tę wielkość zamiast m do wzoru (2) otrzymujemy

$$(3) \quad \frac{n!}{m!} = (n-k+1)(n-k+2) \dots (n-1)n.$$

1.1.2. Permutacje

Wyobraźmy sobie, że mamy kilka przedmiotów, np. pióro, ołówek i ekierkę. Przedmioty te można ułożyć obok siebie w pewnej kolejności. Przypuśćmy, że ułożyliśmy je tak:

ołówek, pióro, ekierka.

Nazwijmy takie ułożenie przedmiotów permutacją. Zamieniając miejscami dwa przedmioty otrzymamy nową permutację. Łatwo przekonać się, że trzy przedmioty, występujące w naszym doświadczeniu, można ułożyć sześcioma różnymi sposobami.

Zajmijmy się zbadaniem, na ile sposobów można rozmieścić dowolną ilość przedmiotów. Założymy w tym celu, że dany jest pewien zbiór H , liczący n różnych elementów. Elementy te oznaczać będziemy małymi literami alfabetu a, b, c, \dots

OKREŚLENIE 1. Zbiór H nazywać będziemy *zbiorem uporządkowanym*, jeśli ma on następujące własności:

1° ma element pierwszy i element ostatni;

2° po każdym elemencie, z wyjątkiem ostatniego, następuje inny, określony element;

3º elementom zbioru mogą być przyporządkowane liczby ciągu liczb naturalnych, przy czym pierwszemu elementowi przyporządkowuje się liczbę 1, a ostatniemu liczbę n . Jeśli k -temu elementowi przyporządkowano liczbę k , to następnemu po nim elementowi przyporządkowuje się liczbę $k+1$.

OKREŚLENIE 2. Każdy uporządkowany zbiór H nazywa się *permutacją* ⁽¹⁾ utworzoną z jego elementów.

TWIERDZENIE 1. Ogólna ilość wszystkich możliwych permutacji z n elementów wynosi $n!$.

Dowód. Oznaczmy ilość permutacji z m elementów symbolem P_m . Oczywiście

$$P_1 = 1 .$$

Wypiszmy permutacje dla zbioru składającego się z dwóch elementów a, b :

$$ab, \quad ba .$$

Widzimy więc, że

$$P_2 = 2 .$$

Aby znaleźć ilość permutacji z trzech elementów, wypiszmy te elementy kolejno jeden pod drugim i obok każdego elementu permutujmy dwa pozostałe. Oto co otrzymamy:

$$\begin{array}{ll} ab\ c, & a\ c\ b, \\ b\ a\ c, & b\ c\ a, \\ c\ a\ b, & c\ b\ a . \end{array}$$

Mnożąc ilość elementów, biorących udział w permutowaniu (tzn. 3) przez ilość permutacji z pomniejszonej o 1 liczby elementów (tzn. przez P_2), gdyż elementy nie mogą być, rzecz oczywista, kojarzone same ze sobą, otrzymujemy, że

$$P_3 = 2 \cdot 3 = 1 \cdot 2 \cdot 3 = 3! .$$

Podobne rozumowanie prowadzi nas do wzoru rekurencyjnego

$$(1) \quad P_1 = 1 , \quad P_{k+1} = P_k(k+1) .$$

Ze wzoru (1) wynika, że

$$(2) \quad P_n = n! .$$

PRZYKŁAD 1. Obliczyć, ile liczb sześciocyfrowych można utworzyć z cyfr liczby 324589.

Rozwiązanie. Nowe liczby otrzymamy przestawiając cyfry liczby danej. Musimy więc obliczyć ilość permutacji z 6 elementów, czyli

$$P_6 = 6! = 720 .$$

PRZYKŁAD 2. Obliczyć, iloma różnymi sposobami czterech studentów może zająć miejsca w ławce. Odpowiedź otrzymamy obliczając ilość permutacji z czterech elementów:

$$P_4 = 4! = 24 .$$

⁽¹⁾ Od francuskiego słowa *permutation*.

PRZYKŁAD 3. Danych jest n punktów na płaszczyźnie. Punkty te należy połączyć łamanaą zamkniętą, przechodzącą przez wszystkie punkty, tak aby łamana była najkrótsza.

Zagadnienie to znane jest w ekonometrii (a ściślej mówiąc w badaniach operacyjnych) pod nazwą *problemu komiwojażera*. Jak dotąd nie znaleziono ogólnego sposobu rozwiązania tego problemu tak, aby nie trzeba było badać długości różnych tras, jakie się otrzyma przez permutowanie punktów trasy. Ogólna liczba wszystkich tras, jakie można otrzymać za pomocą permutowania punktów trasy, wynosi

$$\frac{1}{2} P_{n-1} = \frac{(n-1)!}{2}.$$

Chcąc znaleźć najkrótszą trasę lotniczą łączącą 17 miast wojewódzkich Polski należałoby rozpatrzyć $\frac{16!}{2}$ sytuacji, co daje astronomiczną liczbę 10461394944000.

Odmianą problemu komiwojażera jest tzw. *problem marszruty (routing problem)*, który polega na znalezieniu najkrótszej łamanej łączącej dwa wybrane punkty, z których jeden jest początkiem, a drugi końcem łamanej, i przechodzącej przez wszystkie pozostałe $n-2$ punkty.

1.1.3. Wariacje

Dany jest pewien zbiór H elementów a, b, c, \dots Zakładamy, że zbiór ten nie ma elementów jednakowych. Niech ilość elementów w tym zbiorze będzie równa n .

Obierzmy jakąś liczbę naturalną $m \leq n$ i niech ta liczba oznacza ilość elementów podzbioru utworzonego z elementów zbioru H . Zmieniając elementy i ich porządek można z elementów zbioru H utworzyć pewną liczbę podzbiorów liczących m elementów. Liczba ta zależy od n i m . Oznaczać ją będziemy symbolem A_n^m ⁽¹⁾. Spotyka się także symbol V_n^m , którego tu jednak używać nie będziemy.

OKREŚLENIE 1. Uporządkowane podzbiory, utworzone z m elementów wybranych z n -elementowego zbioru H , różniące się między sobą bądź elementami, bądź ich porząkiem, nazywają się *wariacjami z n po m* elementów.

Oczywiście

$$(1) \quad A_n^1 = n,$$

gdyż tyle różnych zbiorów jednoelementowych można utworzyć z n elementów zbioru H .

Aby otrzymać wariacje z n elementów po dwa, należy kojarzyć każdy element ze wszystkimi pozostałymi elementami. W takim razie

$$(2) \quad A_n^2 = n(n-1).$$

Podobnie, aby otrzymać wariacje z n elementów po trzy, należy kojarzyć każdy element z wariacjami z $n-1$ elementów po dwa. Prowadzi to do następującego wzoru rekurencji

⁽¹⁾ Od pierwszej litery francuskiego słowa *arrangement*.

cyjnego:

$$(3) \quad A_n^3 = nA_{n-1}^2 = n(n-1)(n-2).$$

Wzór ten zapisany w postaci ogólnej przedstawia się następująco:

$$(4) \quad A_n^m = nA_{n-1}^{m-1} \quad \text{dla } m \geq 2.$$

Na mocy wzorów (2) i (4) mamy (patrz 1.1.1, wzór (3)):

$$(5) \quad A_n^m = n(n-1)(n-2)\dots(n-m+1) = \frac{n!}{(n-m)!}.$$

PRZYKŁAD 1. Obliczyć A_4^4 , A_4^3 , A_4^2 , A_4^1 i wypisać wszystkie te wariacje.

Rozwiążanie. Mamy

$$A_4^4 = \frac{4!}{(4-4)!} = 4! = 24;$$

<i>abcd</i>	<i>bacd</i>	<i>cabd</i>	<i>dabc</i>
<i>abdc</i>	<i>badc</i>	<i>cadb</i>	<i>dacb</i>
<i>acdb</i>	<i>bcda</i>	<i>cbad</i>	<i>dbac</i>
<i>acb</i>	<i>bca</i>	<i>cba</i>	<i>dba</i>
<i>abd</i>	<i>bad</i>	<i>cad</i>	<i>dac</i>
<i>adb</i>	<i>bda</i>	<i>cda</i>	<i>dca</i>
<i>acd</i>	<i>bcd</i>	<i>cbd</i>	<i>dbc</i>
<i>adc</i>	<i>bdc</i>	<i>cdb</i>	<i>dcb</i>

$$A_4^3 = \frac{4!}{(4-3)!} = 4 \cdot 3 \cdot 2 = 24;$$

<i>abc</i>	<i>bac</i>	<i>cab</i>	<i>dab</i>
<i>acb</i>	<i>bca</i>	<i>cba</i>	<i>dba</i>
<i>abd</i>	<i>bad</i>	<i>cad</i>	<i>dac</i>
<i>adb</i>	<i>bda</i>	<i>cda</i>	<i>dca</i>
<i>acd</i>	<i>bcd</i>	<i>cbd</i>	<i>dbc</i>
<i>adc</i>	<i>bdc</i>	<i>cdb</i>	<i>dcb</i>

$$A_4^2 = \frac{4!}{(4-2)!} = 4 \cdot 3 = 12;$$

<i>ab</i>	<i>ba</i>	<i>ca</i>	<i>da</i>
<i>ac</i>	<i>bc</i>	<i>cb</i>	<i>db</i>
<i>ad</i>	<i>bd</i>	<i>cd</i>	<i>dc</i>

$$A_4^1 = \frac{4!}{(4-1)!} = 4;$$

$$a \quad b \quad c \quad d.$$

Uwaga. Przyjmujemy, że

$$A_4^0 = \frac{4!}{(4-0)!} = 1$$

i ogólnie

$$A_n^0 = \frac{n!}{(n-0)!} = 1,$$

jakkolwiek takiej wariacji, rzecz prosta, wypisać nie możemy.

PRZYKŁAD 2. Obliczyć, ile liczb naturalnych można utworzyć z cyfr 1, 2, 3, 4.

Rozwiązanie. Z tych cyfr można utworzyć liczby czterocyfrowe, trzycyfrowe, dwucyfrowe i jedno-cyfrowe. Należy przeto obliczyć, czemu się równa

$$A_4^4 + A_4^3 + A_4^2 + A_4^1.$$

Ponieważ

$$A_4^4 = 24, \quad A_4^3 = 24, \quad A_4^2 = 12, \quad A_4^1 = 4,$$

więc

$$24 + 24 + 12 + 4 = 64.$$

Posługując się przykładem 1 czytelnik zechce wypisać wszystkie te liczby.

PRZYKŁAD 3. Przy montażu podzespołów pewnego typu odbiornika radiowego wyróżnia się 5 operacji. W celu usprawnienia organizacji i podniesienia wydajności pracy postanowiono utworzyć pięcio-osobową brygadę montażową, przy czym, aby wydajność brygady była jak największa, zdecydowano w drodze doświadczalnej wybrać 5 robotników spośród 12 kandydatów. Za pomocą eksperymentów chciano ustalić nie tylko optymalny sposób przydzielenia robotników do wykonywania poszczególnych operacji, ale również optymalny skład brygady, gwarantujący dobrą współpracę robotników. Dla sporządzenia planu eksperymentów należało wypisać wszystkie możliwe sposoby przyporządkowania robotników do poszczególnych operacji, tzn. wszystkie wariacje z 12 elementów po 5. Liczba tych wariacji wynosi

$$A_{12}^5 = \frac{12!}{(12-5)!} = \frac{12!}{7!} = 8 \cdot 9 \cdot 10 \cdot 11 \cdot 12 = 95040.$$

Liczba ta przesądza nieopłacalność przeprowadzania eksperymentów i uzasadnia celowość tworzenia brygad w oparciu o intuicję i doświadczenie majstra.

PRZYKŁAD 4. W statystyce spotykamy się często z wnioskowaniem o pewnych właściwościach zbioru zwanej *zbiorowością* lub *populacją generalną* w oparciu o znajomość części tego zbioru, tzw. *próbki*. Zbadamy, iloma sposobami można wybrać ze zbioru n -elementowego próbkę m -elementową.

Przypuśćmy, że elementy wylosowane wracają po wylosowaniu do populacji. Wtedy mamy do czynienia z tzw. *losowaniem ze zwracaniem*. Liczba sposobów wybrania próbki ze zbiorowością generalną, jak łatwo sprawdzić, wynosi wtedy n^m . Na przykład jeżeli $n=2$, a $m=4$, to liczba różnych sposobów wybrania 4-elementowych próbek ze zbiorów 2-elementowych wynosi $2^4=16$. Oznaczając elementy populacji literami a, b można łatwo sporządzić listę wszystkich sposobów wylosowania ze zwracaniem próbek czteroelementowych z dwuelementowego zbioru. Oto te sposoby⁽¹⁾:

	$a \ a \ b \ b$
	$a \ a \ a \ b$
	$a \ a \ b \ a$
	$a \ b \ a \ a$
$a \ a \ a \ a$	$a \ b \ b \ a$
	$b \ b \ a \ a$
	$b \ b \ b \ a$
	$b \ a \ b \ b$
	$a \ b \ b \ b$
	$b \ a \ b \ a$

Zauważmy, że jeżeli wylosowane do próbki elementy nie wracają do populacji, to $m \leq n$. Wtedy liczba sposobów wylosowania m -elementowych próbek z n -elementowej populacji wynosi A_n^m .

1.1.4. Kombinacje

Gdy była mowa o wariacjach, tworzyliśmy z n -elementowego zbioru H m -elementowe uporządkowane podzbiory, które różniły się między sobą bądź elementami, bądź ich porządkiem. O zbiorze H założyliśmy, że nie ma on elementów jednakowych. Obecnie zaj-

⁽¹⁾ Nazywa się je niekiedy *wariacjami z powtórzeniem*.

miemy się zbadaniem, ile m -elementowych podzbiorów można utworzyć z n -elementowego zbioru H , jeśli podzbiory te mają się różnić między sobą tylko elementami. Porządek elementów nie odgrywa roli. Założenie, że zbiór H nie zawiera elementów jednakowych, pozostaje w mocy.

OKREŚLENIE 1. Podzbiory, utworzone z m elementów zbioru H , różniące się między sobą przynajmniej jednym elementem, nazywają się *kombinacjami z n po m elementów*, przy czym n oznacza liczbę elementów zbioru H , a m liczbę elementów tych podzbiorów.

Z określenia tego wynika, że ilość kombinacji z n po m elementów, którą oznaczać będziemy symbolem⁽¹⁾ C_n^m lub $\binom{n}{m}$, jest mniejsza od ilości wariacji z n po m elementów. Z każdej m -elementowej kombinacji, permutując jej elementy, można otrzymać $m!$ m -elementowych wariacji. Oznacza to, że kombinacje i wariacje są ze sobą powiązane następującą relacją:

$$(1) \quad C_n^m = \frac{A_n^m}{m!} = \frac{n!}{m!(n-m)!}.$$

Wzór (1) zasługuje na specjalną uwagę. Nie tylko bowiem podaje liczbę kombinacji z n elementów po m , lecz wiąże również wszystkie poznane pojęcia kombinatoryczne: permutacje, wariacje i kombinacje. Stanie się to oczywiste, gdy wzór (1) przepiszemy w nieco innej postaci:

$$(2) \quad C_n^m = \frac{A_n^m}{P_m}.$$

PRZYKŁAD 1. Obliczyć

$$C_4^4, \quad C_4^3, \quad C_4^2, \quad C_4^1,$$

wypisać wszystkie te kombinacje przyjmując, że zbiór H składa się z elementów a, b, c, d .

Rozwiązanie. Mamy

$$C_4^4 = \frac{4!}{4!(4-4)!} = 1,$$

$$abcd;$$

$$C_4^3 = \frac{4!}{3!(4-3)!} = 4,$$

$$abc, \quad abd, \quad acd, \quad bcd;$$

$$C_4^2 = \frac{4!}{2!(4-2)!} = 6,$$

$$ab, \quad ac, \quad ad, \quad bc, \quad bd, \quad cd;$$

$$C_4^1 = \frac{4!}{1!(4-1)!} = 4,$$

$$a, \quad b, \quad c, \quad d.$$

Uwaga. Przyjmujemy, że

$$C_4^0 = \frac{4!}{0!(4-0)!} = 1$$

(1) Od pierwszej litery francuskiego słowa *combinaison*.

i ogólnie

$$C_n^0 = \frac{n!}{0!(n-0)!} = 1 .$$

PRZYKŁAD 2. Z grona 12 osób należy wybrać komisję składającą się z 5 osób. Iloma sposobami można to uczynić?

Aby otrzymać odpowiedź, należy obliczyć C_{12}^5 :

$$C_{12}^5 = \frac{12!}{5!(12-5)!} = 792 .$$

PRZYKŁAD 3. Wśród 15 osiedli należy usytuować 3 sklepy, przy czym każdy sklep ma znajdować się w jednym z tych osiedli, a w osiedlu może być usytuowany najwyżej jeden sklep. Należy tak zlokalizować sklepy, aby suma odległości poszczególnych osiedli od osiedli ze sklepami była najmniejsza.

Ponieważ liczba możliwych sposobów lokalizacji sklepów jest niewielka, wynosi bowiem

$$C_{15}^3 = \frac{15 \cdot 14 \cdot 13}{1 \cdot 2 \cdot 3} = 455$$

więc znając odległości między osiedlami zadanie można łatwo rozwiązać.

PRZYKŁAD 4. Udowodnić, że

$$C_n^m = C_n^{n-m} .$$

Dowód. Mamy

$$C_n^{n-m} = \frac{n!}{(n-m)![n-(n-m)]!} = \frac{n!}{m!(n-m)!} = C_n^m .$$

PRZYKŁAD 5. Obliczyć, iloma sposobami z populacji n -elementowej można wybrać m -elementową próbę, jeżeli 1° elementy wylosowane nie wracają z powrotem do populacji, 2° próbki mogą różnić się tylko elementami, a nie porządkiem.

Odpowiedź jest prosta: liczba sposobów wynosi C_n^m .

Zwracamy uwagę, że w statystyce spotykamy się często z zadaniem nieco ogólniejszej natury: znaleźć liczbę sposobów podziału n -elementowego zbioru na k grup odpowiednio o licznosciach m_1, m_2, \dots, m_k , gdzie $m_1 + m_2 + \dots + m_k = n$.

Można wykazać, że szukana liczba sposobów wynosi⁽¹⁾

$$(3) \quad \frac{n!}{m_1! m_2! \dots m_k!} .$$

PRZYKŁAD 6. Rozważmy następujące zadanie: iloma sposobami można rozmieścić n kul w m szufladach? Zadanie to znajduje liczne, praktycznie ważne interpretacje, np. iloma sposobami można rozdzielić partię n samochodów ciężarowych wśród m baz transportowych?

Można wykazać, że odpowiedź na te pytania daje wzór

$$(4) \quad C_{n+m-1}^m = C_{n+m-1}^{n-1} .$$

⁽¹⁾ Zwie się je niekiedy *permutacjami z powtórzeniami*.

Przypuśćmy, że liczba kul wynosi $n=3$, a liczba szuflad wynosi $m=3$. W takim razie liczba sposobów, jakimi można kule rozmieszczać w szufladach, jest równa

$$C_5^3 = 10.$$

Oznaczamy szuflady literami a , b , c . Oto możliwe sytuacje⁽¹⁾

$$\begin{array}{lll}
 a\ a\ a & a\ a\ b & b\ b\ c \\
 b\ b\ b & a\ a\ c & c\ c\ a & a\ b\ c \\
 c\ c\ c & b\ b\ a & c\ c\ b
 \end{array}$$

1.1.5. Trójkąt Pascala

Zbudujemy kwadratową tablicę dwudzielną (tablica 1). Wiersze i kolumny tej tablicy ponumerujemy liczbami $0, 1, 2, \dots, n$. Na przecięciu k -tego wiersza oraz m -tej kolumny ($k=0, 1, 2, \dots, n$; $m=0, 1, 2, \dots, n$) wypiszemy C_k^m . Wszystkie prostokąty leżące na przecięciu k -tego wiersza oraz m -tej kolumny zostawimy puste, jeśli $k < m$. Nasza tablica w tych warunkach przybierze następującą postać:

Tablica 1

Nr kolumny \ Nr wiersza	0	1	2	...	m	...	n
0	C_0^0						
1	C_1^0	C_1^1					
2	C_2^0	C_2^1	C_2^2				
...		
k	C_k^0	C_k^1	C_k^2	...	C_k^m		
...	
n	C_n^0	C_n^1	C_n^2	...	C_n^m	...	C_n^n

Widzimy, że zawarte w tablicy symbole kombinacji uformowały trójkąt. Trójkąt ten można również przedstawić nieco inaczej:

$$(1) \quad \begin{matrix} & & C_0^0 \\ & & C_1^0 & C_1^1 \\ & C_2^0 & C_2^1 & C_2^2 \\ C_3^0 & C_3^1 & C_3^2 & C_3^3 \\ \dots & \dots & \dots & \dots \\ C_n^0 & C_n^1 & \dots & C_n^m & \dots & C_n^{n-1} & C_n^n \end{matrix}$$

⁽¹⁾ Zwie się je niekiedy *kombinacjami z powtórzeniami*.

Gdy zamiast symboli kombinacji wpiszemy odpowiednie liczby, trójkąt przybierze postać następującą:

$$(2) \quad \begin{array}{c} 1 \\ & 1 & 1 \\ & 1 & 2 & 1 \\ & 1 & 3 & 3 & 1 \\ & 1 & 4 & 6 & 4 & 1 \\ & 1 & 5 & 10 & 10 & 5 & 1 \\ & \dots & \dots & \dots & \dots & \dots & \dots \end{array}$$

Trójkąt (2) nosi nazwę *trójkąta Pascala*. Liczby w nim zawarte otrzymaliśmy posługując się wzorem na ilość kombinacji. W miarę rozwijania kolejnych wierszy trójkąta posługiwanie się tym wzorem stawało się coraz bardziej uciążliwe.

Oczywiście moglibyśmy zaniechać korzystania z tego wzoru przy budowaniu trójkąta Pascala, gdyby liczby trójkąta przejawiały jakąś regularność, która by nam pozwoliła, znając liczby jednego wiersza w trójkącie, wypisać liczby wiersza następnego. Regularność taka istnieje i jest bardzo prosta. Zauważmy, że w trójkącie Pascala liczby wiersza następnego stoją między liczbami wiersza poprzedniego. Regularność polega na tym, że każda z takich liczb, z wyjątkiem pierwszej i ostatniej, które są zawsze jedynkami, równa się sumie liczb, stojących nad nią w wierszu poprzednim.

Czytelnik sprawdzi bez trudu, że

$$C_{k-1}^{m-1} + C_{k-1}^m = C_k^m.$$

Uczynione spostrzeżenie pozwala zbudować trójkąt Pascala bez potrzeby uciekania się do wzoru, podającego liczbę kombinacji. Znaczenie praktyczne trójkąta Pascala polega na tym, że prosta i łatwa do zapamiętania umiejętność budowania tego trójkąta uwalnia nas od wykonywania uciążliwych rachunków, związanych ze stosowaniem wzoru

$$C_k^m = \frac{k!}{m!(k-m)!}.$$

1.1.6. Uwagi o obliczaniu dużych wartości silni. Wzór Stirlinga

Jak wiemy (patrz 1.1.1), wartości $n!$ wzrastają bardzo szybko ze wzrostem n . Dla dużych n obliczanie $n!$ jest przeto sprawą nader kłopotliwą. Przybliżoną wartość $n!$ możemy znaleźć za pomocą logarytmowania. Mamy bowiem

$$(1) \quad \log n! = \log 1 + \log 2 + \log 3 + \dots + \log n = \sum_{k=1}^n \log k.$$

Jakkolwiek metoda logarytmowania może oddać usługi przy obliczaniu dużych wartości $n!$, jednak ze względu na konieczność sumowania logarytmów metoda ta stosowana jest

rzadko. Na ogół korzysta się ze wzoru

$$(2) \quad n! = \sqrt{2\pi n} n^n e^{-n} e^{\theta/12n},$$

gdzie $0 < \theta \leq 1$. Wzór ten znany jest pod nazwą wzoru Stirlinga⁽¹⁾. Logarytmując obie strony równości (2) otrzymamy logarytmiczną postać wzoru Stirlinga

$$(3) \quad \ln n! = \ln \sqrt{2\pi} + \left(n + \frac{1}{2} \right) \ln n - n + \frac{\theta}{12n}.$$

Ponieważ $\frac{\theta}{12n} \rightarrow 0$, gdy $n \rightarrow \infty$, przeto

$$(4) \quad \ln n! \approx \ln \sqrt{2\pi} + \left(n + \frac{1}{2} \right) \ln n - n.$$

Pytania kontrolne i zadania⁽²⁾

1. Uprościć ułamki:

a) $\frac{100!}{98!}$;

b) $\frac{n!}{(n-1)!}, \quad \frac{(n+1)!}{(n-1)!}, \quad \frac{(n-1)!(n+1)!}{(n!)^2}$.

2. Obliczyć P_3 , P_5 , P_{12} .

3. Ile liczb większych od 7 milionów można utworzyć przedstawiając cyfry liczby 3708925?

4. Ile liczb będących wielokrotnościami 5 można utworzyć przedstawiając cyfry liczby 102534?

5. Liczba permutacji z $n+3$ elementów jest 210 razy większa niż liczba permutacji z n elementów.

Znaleźć n .

6. $P_n : P_{n+1} = 1 : 9$. Znaleźć n .

7. Znaleźć: A_{10}^3 , A_{12}^4 , A_{20}^{15} .

8. Ile można utworzyć liczb a) 4-cyfrowych, b) 5-cyfrowych z cyfr 1, 2, ..., 9 w ten sposób, aby żadna z cyfr nie powtarzała się?

9. $A_n^2 = 132$. Znaleźć n .

10. Znaleźć ogólną ilość wszystkich możliwych wariacji (bez powtórzeń) z n elementów po 1, 2, ..., n .

11. Obliczyć C_8^3 , C_{10}^{10} , C_{12}^8 .

12. Udowodnić, że

a) $C_{10}^4 = C_{10}^6$, $C_{15}^7 = C_{15}^8$;

b) $C_n^m = C_{n-1}^m + C_{n-1}^{m-1}$;

c) $C_n^0 = 1$, $C_n^1 = n$, $C_n^n = 1$;

d) $C_n^m = C_n^{n-m}$.

13. Sprawdzić słuszność następujących wzorów:

a) $C_n^0 + C_n^1 + \dots + C_n^n = 2^n$;

b) $C_n^0 - C_n^1 + \dots \pm C_n^n = 0$.

14. Wykonano rzut pięcioma monetami. Ile różnych układów orłów i reszek może wystąpić w takim doświadczeniu?

⁽¹⁾ Wyprowadzenie wzoru Stirlinga znajdzie czytelnik w pracy [19].

⁽²⁾ Zbiór ciekawych zadań z kombinatoryki, rachunku prawdopodobieństwa i statystyki matematycznej znajdzie czytelnik w [6], [9], [34].

15. Malarz ma pomalować trzy przedmioty, mając do dyspozycji farby w pięciu kolorach. Ile układów kolorów farb może malarz otrzymać, zakładając, że każdy przedmiot jest malowany wyłącznie na jeden kolor (patrz [9])?

16. Ile liczb sześciocyfrowych można utworzyć z cyfr 1, 2, 3, 4, 5, 6, 7, 8, 9?

1.2. DWUMIAN NEWTONA

Z algebry elementarnej wiadomo, że

$$(1) \quad \begin{aligned} (p+q)^0 &= 1, \quad p+q \neq 0, \\ (p+q)^1 &= p+q, \\ (p+q)^2 &= p^2 + 2pq + q^2, \\ (p+q)^3 &= p^3 + 3p^2q + 3pq^2 + q^3. \end{aligned}$$

Przyjrzyjmy się uważnie wzorom (1) i spróbujmy wykryć w nich jakąś prawidłowość, która pozwoliłaby nam znaleźć rozwinięcie dwumianu

$$(p+q)^n,$$

gdzie p i q są to liczby rzeczywiste, a n – liczba całkowita nieujemna, bez konieczności uciekania się do potęgowania.

Oczywiście wzory (1) można napisać w postaci następującej:

$$(2) \quad \begin{aligned} (p+q)^0 &= p^0q^0, \\ (p+q)^1 &= p^1q^0 + p^0q^1, \\ (p+q)^2 &= p^2q^0 + 2p^1q^1 + p^0q^2, \\ (p+q)^3 &= p^3q^0 + 3p^2q^1 + 3p^1q^2 + p^0q^3. \end{aligned}$$

Widzimy, że prawe strony równań (2) są sumami iloczynów typu

$$(3) \quad p^k q^{n-k},$$

gdzie n jest to wykładnik potęgi dwumianu, a $k = 0, 1, \dots, n$.

Przy poszczególnych iloczynach stoją współczynniki. Łatwo sprawdzić, że współczynniki stojące obok iloczynów $p^k q^{n-k}$ są liczbami trójkąta Pascala. Korzystając z tego spostrzeżenia ciąg (2) przedstawiamy nieco inaczej:

$$(4) \quad \begin{aligned} (p+q)^0 &= 1p^0q^0, \\ (p+q)^1 &= 1p^1q^0 + 1p^0q^1, \\ (p+q)^2 &= 1p^2q^0 + 2p^1q^1 + 1p^0q^2, \\ (p+q)^3 &= 1p^3q^0 + 3p^2q^1 + 3p^1q^2 + 1p^0q^3. \end{aligned}$$

Jeśli zamiast liczb trójkąta Pascala wpiszemy w miejsce współczynników symbole kombinacji, to z łatwością znajdziemy ogólny wzór na rozwinięcie dwumianu $(p+q)^n$.

Otrzymamy bowiem

$$\begin{aligned}
 (p+q)^0 &= C_0^0 p^0 q^0, \\
 (p+q)^1 &= C_1^0 p^1 q^0 + C_1^1 p^0 q^1, \\
 (p+q)^2 &= C_2^0 p^2 q^0 + C_2^1 p^1 q^1 + C_2^2 p^0 q^2, \\
 (5) \quad (p+q)^3 &= C_3^0 p^3 q^0 + C_3^1 p^2 q^1 + C_3^2 p^1 q^2 + C_3^3 p^0 q^3, \\
 &\dots \dots \dots \dots \dots \dots \dots \\
 (p+q)^n &= C_n^0 p^n q^0 + C_n^1 p^{n-1} q^1 + \dots + C_n^{n-1} p^1 q^{n-1} + C_n^n p^0 q^n.
 \end{aligned}$$

Dwumian $(p+q)^n$ nosi nazwę *dwumianu Newtona*. Stosując zapis skrócony możemy rozwinięcie dwumianu Newtona przedstawić w łatwej do zapamiętania postaci:

$$(6) \quad (p+q)^n = \sum_{k=0}^n C_n^k p^k q^{n-k}.$$

Wzór (6) został tu odgadnięty za pomocą rozważań intuicyjnych.

► Obecnie zajmiemy się wyprowadzeniem tego wzoru posługując się zasadą indukcji matematycznej.

Dla $n=1$ mamy

$$(p+q)^1 = p+q.$$

Zakładamy, że zachodzi równość

$$(7) \quad (p+q)^{n-1} = \sum_{k=0}^{n-1} C_{n-1}^k p^k q^{(n-1)-k}.$$

Zbadamy, czy przy tym założeniu zachodzi (6).

Oczywiście

$$(8) \quad (p+q)^n = (p+q)^{n-1}(p+q).$$

Stąd

$$(9) \quad (p+q)^n = (p+q) \sum_{k=0}^{n-1} C_{n-1}^k p^k q^{(n-1)-k}.$$

Wykonajmy mnożenie po prawej stronie równości (9):

$$\begin{aligned}
 (p+q)^n &= (p+q)(C_{n-1}^0 p^0 q^{n-1} + C_{n-1}^1 p^1 q^{n-2} + \dots + C_{n-1}^{n-2} p^{n-2} q^1 + C_{n-1}^{n-1} p^{n-1} q^0) = \\
 &= C_{n-1}^0 p^0 q^n + C_{n-1}^1 p^1 q^{n-1} + \dots + C_{n-1}^{n-2} p^{n-2} q^2 + C_{n-1}^{n-1} p^{n-1} q^1 + \\
 &\quad + C_{n-1}^0 p^1 q^{n-1} + C_{n-1}^1 p^2 q^{n-2} + \dots + C_{n-1}^{n-2} p^{n-1} q^1 + C_{n-1}^{n-1} p^n q^0 = \\
 &= C_{n-1}^0 p^0 q^n + p^1 q^{n-1} (C_{n-1}^0 + C_{n-1}^1) + \dots + p^{n-1} q^1 (C_{n-1}^{n-2} + C_{n-1}^{n-1}) + C_{n-1}^{n-1} p^n q^0 = \\
 &= C_n^0 p^0 q^n + C_n^1 p^1 q^{n-1} + \dots + C_n^{n-1} p^{n-1} q^1 + C_n^n p^n q^0 = \\
 &= \sum_{k=0}^n C_n^k p^k q^{n-k},
 \end{aligned}$$

gdyż

$$C_{n-1}^0 p^0 q^n = C_n^0 p^0 q^n, \quad C_{n-1}^k + C_{n-1}^{k-1} = C_n^k,$$

o czym czytelnik miał możliwość przekonać się rozwiązyując zadanie 12b poprzedniego paragrafu; łatwo również sprawdzić, że

$$C_{n-1}^{n-1} p^n q^0 = C_n^n p^n q^0.$$



Pytania kontrolne i zadania

1. Wykazać, że

- a) $(1+x)^n \geq 1+nx$,
- b) $\sum_{k=0}^n C_n^k = 2^n$ (patrz 1.1, zadanie 13a),
- c) $\sum_{k=0}^n C_n^k p^k q^{n-k} = 1$ przy założeniu, że $p+q=1$.

2. Zbadać, czemu się równa rozwinięcie następujących dwumianów:

- a) $(x-1)^n$,
- b) $(x+1)^n$,
- c) $(x-2)^n$.

1.3. CAŁKI EULERA

1.3.1. Funkcja gamma Eulera

Funkcją gamma lub całką Eulera drugiego rodzaju nazywa się funkcja

$$(1) \quad \Gamma(x) = \int_0^\infty e^{-t} t^{x-1} dt \quad (0 < x < \infty).$$

Funkcja ta ma szereg ważnych zastosowań praktycznych. Wynika to między innymi stąd, że można ją uważać za uogólnienie silni. Jeśli x jest liczbą dodatnią, to

$$(2) \quad \Gamma(x+1) = \int_0^\infty e^{-t} t^x dt = x\Gamma(x).$$

► Udowodnimy to za pomocą całkowania przez części. Oznaczmy w tym celu

$$\begin{aligned} t^x &= u, & e^{-t} dt &= dv, \\ xt^{x-1} dt &= du, & e^{-t} &= v. \end{aligned}$$

Otrzymujemy

$$\Gamma(x+1) = \int_0^\infty e^{-t} t^x dt = -t^x e^{-t} \Big|_0^\infty + x \int_0^\infty e^{-t} t^{x-1} dt.$$

Łatwo wykazać, że

$$\frac{-t^x}{e^t} \Big|_0^\infty = 0.$$

Stosując bowiem dostatecznie długo wzór de L'Hospitala [19] otrzymamy

$$\lim_{t \rightarrow \infty} \frac{t^x}{e^t} = 0.$$

Podobnie

$$\lim_{t \rightarrow 0} \frac{t^x}{e^t} = 0.$$

Wobec tego

$$(3) \quad \Gamma(x+1) = x \int_0^\infty e^{-t} t^{x-1} dt = x \Gamma(x).$$

Stąd, gdy x przybiera wartości liczb naturalnych,

$$(4) \quad \Gamma(n+1) = n \Gamma(n) = n(n-1) \Gamma(n-2) = n(n-1) \cdot \dots \cdot 2 \cdot 1 = n!,$$

gdzie, jak łatwo sprawdzić, $\Gamma(1) = 1$.

W kursie analizy ([7], t. II) dowodzi się, że

$$\Gamma(x) \Gamma(1-x) = \frac{\pi}{\sin \pi x}, \quad 0 < x < 1.$$

Stąd, gdy $x = \frac{1}{2}$,

$$\Gamma(\frac{1}{2}) \Gamma(\frac{1}{2}) = \frac{\pi}{\sin \frac{1}{2}\pi} = \pi,$$

czyli

$$\Gamma^2(\frac{1}{2}) = \pi.$$

Wobec tego

$$(5) \quad \Gamma(\frac{1}{2}) = \sqrt{\pi}.$$

Ponieważ

$$\Gamma(\frac{1}{2}) = \int_0^\infty e^{-t} t^{1/2-1} dt = \int_0^\infty e^{-t} t^{-1/2} dt = \sqrt{\pi},$$

przetoż podstawiając $t = z^2$, $dt = 2z dz$ otrzymamy

$$\int_0^\infty e^{-t} t^{-1/2} dt = \int_0^\infty e^{-z^2} (z^2)^{-1/2} 2z dz = 2 \int_0^\infty e^{-z^2} dz = \sqrt{\pi}.$$

Stąd

$$(6) \quad \int_0^\infty e^{-z^2} dz = \frac{\sqrt{\pi}}{2}.$$

Całka ta nosi nazwę *całki Eulera-Poissona* i ma duże znaczenie w rachunku prawdopodobieństwa i statystyce matematycznej.

Funkcja $\Gamma(x)$ jest ciągła w przedziale $0 < x < \infty$ i ma wszystkie pochodne w tym przedziale.

1.3.2. Funkcja beta Eulera

Funkcją beta lub całką Eulera pierwszego rodzaju nazywa się funkcja

$$(1) \quad B(x, y) = \int_0^1 t^{x-1} (1-t)^{y-1} dt.$$

Całka ta istnieje, gdy $x > 0$ i $y > 0$.

Funkcję beta można wyrazić za pomocą funkcji gamma. Między tymi funkcjami zachodzi bowiem następujący związek:

$$(2) \quad B(x, y) = \frac{\Gamma(x) \Gamma(y)}{\Gamma(x+y)}.$$

Dowód podany przez Dirichleta znajdzie czytelnik w podręczniku [7], t. II.

1.4. WZORY EULERA

Jak wiadomo, gdy spełnione są pewne warunki, funkcję $f(x)$ można rozwinać w szereg

$$(1) \quad f(x) = f(0) + f'(0) \frac{x}{1!} + \dots + f^{(n)}(0) \frac{x^n}{n!} + \dots,$$

zwany *szeregiem Maclaurina* (patrz [19]).

Posługując się szeregiem Maclaurina czytelnik sprawdzi bez trudu, że

$$(2) \quad e^x = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots + \frac{x^n}{n!} + \dots,$$

$$(3) \quad \sin x = \frac{x}{1!} - \frac{x^3}{3!} + \frac{x^5}{5!} - \dots + (-1)^n \frac{x^{2n+1}}{(2n+1)!} + \dots,$$

$$(4) \quad \cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \dots + (-1)^n \frac{x^{2n}}{(2n)!} + \dots$$

Korzystając ze wzoru (2) napiszemy rozwinięcie funkcji zmiennej zespolonej $f(z) = e^z$, gdzie $z = ix$, przy czym $i = \sqrt{-1}$, a x jest dowolną liczbą rzeczywistą.

Mamy

$$e^z = e^{ix} = 1 + \frac{ix}{1!} + \frac{(ix)^2}{2!} + \frac{(ix)^3}{3!} + \frac{(ix)^4}{4!} + \frac{(ix)^5}{5!} + \frac{(ix)^6}{6!} + \frac{(ix)^7}{7!} + \frac{(ix)^8}{8!} + \dots$$

Ponieważ

$$i^2 = -1, \quad i^3 = -i, \quad i^4 = 1, \quad i^5 = i, \quad \dots,$$

przeto

$$e^{ix} = 1 + i \frac{x}{1!} - \frac{x^2}{2!} - i \frac{x^3}{3!} + \frac{x^4}{4!} + i \frac{x^5}{5!} - \frac{x^6}{6!} - i \frac{x^7}{7!} + \frac{x^8}{8!} + \dots$$

Po rozdzieleniu szeregu na część rzeczywistą i część urojoną mamy

$$(5) \quad e^{ix} = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \frac{x^8}{8!} - \dots + i \frac{x}{1!} - i \frac{x^3}{3!} + i \frac{x^5}{5!} - i \frac{x^7}{7!} + \dots$$

Uwzględniając wzory (3) i (4) otrzymujemy

$$(6) \quad e^{ix} = \cos x + i \sin x.$$

Jeśli zamiast x podstawimy $-x$, to

$$(7) \quad e^{-ix} = \cos x - i \sin x.$$

Dodając stronami wyrażenia (6) i (7) otrzymujemy

$$(8) \quad \cos x = \frac{e^{ix} + e^{-ix}}{2},$$

natomiast odejmując je mamy

$$(9) \quad \sin x = \frac{e^{ix} - e^{-ix}}{2i}.$$

Jeśli zmienna zespolona $z = x + iy$, to

$$(10) \quad e^z = e^{x+iy} = e^x e^{iy}.$$

Korzystając ze wzoru (6) wzór (10) można przedstawić inaczej

$$(11) \quad e^{x+iy} = e^x (\cos y + i \sin y).$$

1.5. ZBIORY

1.5.1. Ogólne wiadomości o zbiorach

Jednym z podstawowych pojęć współczesnej matematyki jest pojęcie *zbioru*⁽¹⁾. Jest to pojęcie pierwotne, którego nie definiuje się przyjmując, że jest bezpośrednio zrozumiałe. Każdy zbiór składa się ze swoich *elementów*. Pojęcie elementu zbioru nie jest również definiowane. Zbiory oznacza się zwykle wielkimi literami, natomiast elementy zbiorów odpowiednimi małymi literami. Przypuśćmy, że rozważamy pewien zbiór A składający

⁽¹⁾ Zamiast słowa *zbiór* używamy również słowa *mnogość*, stąd dział matematyki traktujący o zbiorach i ich własnościach nazywa się *teorią mnogości*.

się z n różnych elementów a_1, a_2, \dots, a_n . Możemy to zapisać następująco:

$$(1) \quad a_i \in A, \quad i=1, 2, \dots, n,$$

przy czym znak \in czytamy „jest elementem”.

Zbiór A może być podzielony na części, które nazywamy *podzbiorami* zbioru A . Przypuśćmy, że A_1 jest podzbiorem zbioru A . Zapisujemy to następująco:

$$(2) \quad A_1 \subset A.$$

Znak \subset nazywa się znakiem *inkluzji*.

Przypuśćmy, że dla jakiegoś zbioru A i jego części A_1 zachodzą relacje

$$(3) \quad A_1 \subset A, \quad A \subset A_1.$$

Oznacza to, że część równa jest całości, czyli że zbiory A i A_1 są *identyczne*. Identyczność lub równość zbiorów należy rozumieć w ten sposób, że zbiory te mają jednakowe elementy.

Z definicji inkluzji wynika również, że jeżeli

$$(4) \quad A \subset B \quad \text{i} \quad B \subset C, \quad \text{to} \quad A \subset C.$$

Oznacza to, że relacja inkluzji jest przechodnia.

Umówimy się, że jeżeli wszystkie nasze rozważania dotyczą pewnego wyróżnionego zbioru X , to zbiór ten nazywać będziemy *przestrzenią*. Ta niecô dziwna nazwa, zapożyczona z geometrii, ma nam przypominać, że w naszych rozważaniach X stanowi całość, poza którą nic nie istnieje i jeżeli coś istnieje, to jest częścią X .

Obok zbioru X , który zawiera wszystkie elementy, wyróżnia się także zbiór \emptyset , nie zawierający żadnego elementu, zwany *zbiorem pustym*.

Zbiory ze względu na ich licznosć dzieli się na *skończone*, zawierające skońzoną liczbę elementów, *przeliczalne*, tj. równoliczne ze zbiorem liczb naturalnych, oraz zbiory *nieprzeliczalne*. Zbiory przeliczalne i nieprzeliczalne są, oczywiście, przykładami zbiorów nieskończonych.

Licznośc zbioru, czyli tzw. *moc zbioru*, określamy za pomocą tzw. *liczb kardynalnych*.

W przypadku zbiorów skończonych mocą zbioru jest po prostu liczba elementów tego zbioru. Moc zbiorów przeliczalnych oznacza się (za Cantorem) literą hebrajską \aleph_0 (*alef zero*), natomiast moc zbiorów równej mocy ze zbiorem liczb rzeczywistych oznaczamy literą c (od słowa *continuum*).

Zbiór, którego elementami są zbiory, nazywać będziemy *klasą zbiorów*.

Przypuśćmy, że \mathcal{A} jest klasą, której elementami są wszystkie podzbiory danego zbioru A . Jeżeli zbiór A jest skończony, liczba tych podzbiiorów jest równa liczbie kombinacji, jakie można utworzyć z elementów tego zbioru. Jak wiemy (patrz 1.1.5, zadanie 13a), liczba kombinacji, utworzonych ze zbiorem n -elementowego, wynosi 2^n . Oczywiście $2^n > n$. Nierówność ta daje się uogólnić na dowolne liczby kardynalne. Między innymi jest

$$2^{\aleph_0} > \aleph_0 \quad \text{oraz} \quad 2^c > c.$$

Nie wiadomo, czy istnieją liczby kardynalne między \aleph_0 i c . Przypuszczenie, że takich liczb nie ma, nosi nazwę *hipotezy continuum* (patrz [29]). Można dowieść, że $2^{\aleph_0} = c$.

Oznacza to, między innymi, że moc klasy podzbiorów, jakie można utworzyć ze zbioru przeliczalnego, jest continuum, tzn. taka sama jak moc zbioru liczb rzeczywistych.

1.5.2. Algebra zbiorów

OKREŚLENIE 1. Sumą $A_1 + A_2 + \dots + A_n$ zbiorów A_1, A_2, \dots, A_n nazywa się zbiór utworzony z elementów należących przynajmniej do jednego z tych zbiorów. Sumę zbiorów oznacza się również symbolami $A_1 \cup A_2 \cup \dots \cup A_n$ oraz $\bigcup_{i=1}^n A_i$.

OKREŚLENIE 2. Iloczynem $A_1 \cdot A_2 \cdot \dots \cdot A_n$ zbiorów A_1, A_2, \dots, A_n nazywa się zbiór utworzony z elementów należących do każdego ze zbiorów A_1, A_2, \dots, A_n . Iloczyn zbiorów oznacza się również symbolami $A_1 \cap A_2 \cap \dots \cap A_n$ oraz $\bigcap_{i=1}^n A_i$.

OKREŚLENIE 3. Jeżeli jakieś dwa zbiory A, B mają tę własność, że ich iloczyn $A \cdot B = \emptyset$, tzn. jest zbiorem pustym, to o zbiorach tych mówimy, że są *rozłączne*.

OKREŚLENIE 4. Różnicą $A - B$ dwóch zbiorów A, B nazywa się zbiór elementów należących do A , lecz nie należących do B .

OKREŚLENIE 5. Dopełnieniem zbioru A w przestrzeni X nazywa się różnica $X - A$.

Czytelnik zwróci uwagę, że zbiór dopełniający $X - A$ zawiera wszystkie elementy, które nie należą do A . Wygodnie jest więc na oznaczenie różnicy $X - A$ wprowadzić symbol \bar{A} , który czytamy „zbior nie A ”.

W zastosowaniach często występują dwie relacje znane pod nazwą *praw De Morgana*:

$$(1) \quad \overline{A_1 + A_2 + \dots + A_n} = \bar{A}_1 \cdot \bar{A}_2 \cdot \dots \cdot \bar{A}_n,$$

$$(2) \quad \overline{A_1 \cdot A_2 \cdot \dots \cdot A_n} = \bar{A}_1 + \bar{A}_2 + \dots + \bar{A}_n.$$

1.5.3. Przedziały, produkty kartezjańskie

Oznaczmy symbolem R zbiór liczb rzeczywistych. Niech $a, b \in R$ oraz $a < b$.

OKREŚLENIE 1. Zbiory liczb, określonych jedną z nierówności

$$(1) \quad a < x < b,$$

$$(2) \quad a \leq x < b,$$

$$(3) \quad a < x \leq b,$$

$$(4) \quad a \leq x \leq b$$

nazywamy *przedziałami*.

W szczególności przedział (1) będziemy nazywać *przedziałem otwartym* i oznaczać symbolem (a, b) ; przedział (2) nazywa się *przedziałem domkniętym z lewa*, a przedział (3) – *przedziałem domkniętym z prawa*; oznacza się je odpowiednio symbolami $\langle a, b \rangle$ oraz $(a, b]$; przedział (4) nosi nazwę *przedziału domkniętego* i oznacza się symbolem $[a, b]$.

Uwaga. Jeżeli $a = -\infty$ lub $b = \infty$, mamy do czynienia z tzw. *przedziałem nieskończonym*.

OKREŚLENIE 2. *Kresem górnym (dolnym) zbioru $A \subset R$* nazywa się najmniejszą (największą) liczbę $x_0 \in R$, taką że $x \leq x_0$ ($x \geq x_0$) dla każdego $x \in A$. Kres górny zbioru A oznacza się symbolem $\sup_{x \in A} x$, natomiast kres dolny – symbolem $\inf_{x \in A} x$.

Niech X_1, X_2, \dots, X_n będzie ciągiem zbiorów.

OKREŚLENIE 3. *Produktem (iloczynem) kartezjańskim* zbiorów X_1, X_2, \dots, X_n nazywamy zbiór ciągów postaci $\{x_1, x_2, \dots, x_n\}$, przy czym $x_j \in X_j$ dla $j = 1, 2, \dots, n$.

Produkt kartezjański oznaczany jest symbolem $X_1 \times X_2 \times \dots \times X_n$. Zauważmy w szczególności, że jeżeli X_1 jest przedziałem $\langle a, b \rangle$, a X_2 jest przedziałem $\langle c, d \rangle$, to $X_1 \times X_2$ jest prostokątem określonym nierównościami

$$a \leq x_1 \leq b, \quad c \leq x_2 \leq d, \quad x_1 \in X_1, \quad x_2 \in X_2.$$

Analogicznie produkt kartezjański $R \times R = R^2$ punktów na osi liczbowej jest płaszczyzną. W tym sensie można powiedzieć, że $X_1 \times X_2$ jest przedziałem w R^2 . Podobnie, jeżeli X_3 jest przedziałem $\langle e, f \rangle$, to $X_1 \times X_2 \times X_3$ jest przedziałem w R^3 . Takie uogólnienie pojęcia przedziału może być podane dla dowolnej ilości czynników produktu kartezjańskiego.

1.5.4. Ciało zbiorów

OKREŚLENIE 1. Niepustą klasę zbiorów \mathcal{L} utworzonych z podzbiorów przestrzeni X nazywamy *ciałem zbiorów*, jeżeli relacja $A, B \in \mathcal{L}$ implikuje $A + B \in \mathcal{L}$ i $\bar{A} \in \mathcal{L}$ lub – co na jedno wychodzi – jeżeli relacja $A, B \in \mathcal{L}$ implikuje $A \cdot B \in \mathcal{L}$ oraz $\bar{A} \in \mathcal{L}$.

OKREŚLENIE 2. Niepustą klasę zbiorów \mathcal{L} nazywa się *σ -ciąłem* lub *ciąłem przeliczalnie addytywnym*, jeżeli z relacji $A, A_1, A_2, \dots \in \mathcal{L}$ wynika $A_1 + A_2 + \dots \in \mathcal{L}$ i $\bar{A} \in \mathcal{L}$ lub – co na jedno wychodzi – jeżeli z relacji $A, A_1, A_2, \dots \in \mathcal{L}$ wynika $A_1 \cdot A_2 \cdot \dots \in \mathcal{L}$ i $\bar{A} \in \mathcal{L}$, przy czym $A = A_1 + A_2 + \dots$.

Z określenia ciała zbiorów widać, że każde ciało przestrzeni X zawiera zbiór pusty \emptyset i przestrzeń X .

OKREŚLENIE 3. O zbiorze mówimy, że jest *najmniejszym zbiorem* o wyróżnionej własności W , jeżeli ten zbiór ma własność W i jeżeli każdy zbiór o własności W zawiera ten zbiór.

OKREŚLENIE 4. Najmniejsze przeliczalnie addytywne ciało zbiorów o własności W nosi nazwę *borełowskiego ciała zbiorów* i jest oznaczane symbolem \mathcal{B} . Zbiory $A \in \mathcal{B}$ nazywają się *zbiorami borełowskimi*.

Wyjaśniamy, że borełowskie ciało zbiorów jest addytywną klasą zbiorów, do której należą wszystkie podzbiory przestrzeni X oraz wszystkie zbiory, jakie można utworzyć z tych podzbiorów drogą wykonywania w dowolnym porządku przeliczalnego ciągu działań dodawania, odejmowania i mnożenia zbiorów.

OKREŚLENIE 5. Przestrzeń X nazywamy *ośrodkową*, jeżeli istnieje w niej ciąg punktów

x_1, x_2, \dots taki, że każdy punkt $x \in X$ spełnia relację

$$(1) \quad x = \lim_{n \rightarrow \infty} x_{k_n}.$$

Dla przestrzeni ośrodkowych prawdziwe jest

TWIERDZENIE 1. Klasa wszystkich podzbiorów borelowskich przestrzeni ośrodkowej ma moc nie większą od \mathfrak{c} (patrz [18], str. 120).

Zauważymy, że ponieważ klasa podzbiorów zbioru mocy \mathfrak{c} jest mocy $2^{\mathfrak{c}}$, więc moc klasy podzbiorów nieborelowskich jest większa od \mathfrak{c} . Zauważymy dalej, że przestrzeń liczb rzeczywistych jest ośrodkowa. Wynika stąd, że klasa podzbiorów borelowskich zbioru liczb rzeczywistych (mającego moc \mathfrak{c}) jest mocy nie większej od \mathfrak{c} , natomiast klasa wszystkich podzbiorów tej przestrzeni jest większa od \mathfrak{c} .

1.5.5. Miara

Weźmy pod uwagę któryś z podzbiorów określonych nierównościami (1) - (4) w 1.5.3. Liczbę $b-a$ nazywamy *długością przedziału*. Długość przedziału nie zależy od tego, czy mamy do czynienia z przedziałem otwartym czy domkniętym. Podobnie, jeżeli weźmiemy pod uwagę prostokąt, będący w R^2 uogólnieniem przedziału, to liczbę $(b-a)(d-c)$ nazywamy *powierzchnią przedziału*. Jeżeli rozważania nasze dotyczą przedziału w R^3 , to liczba $(b-a)(d-c)(f-e)$ nazywa się *objętością*. Z przykładów tych widać, że zbiorom mogą być przyporządkowane liczby lub, mówiąc inaczej, że na elementach pewnej klasy zbiorów może być określona *niewjemna funkcja zbioru*.

OKREŚLENIE 1. Niech \mathcal{L} oznacza przeliczalnie addytywne ciało zbiorów. Mówimy, że funkcja rzeczywista $P(A)$ określona na elementach A ciała \mathcal{L} jest *przeliczalnie addytywną* (lub σ -*addytywną*) *funkcją zbioru*, jeżeli

$$(1) \quad P(\emptyset) = 0,$$

$$(2) \quad P\left(\bigcup_{i=1}^n A_i\right) = \sum_{i=1}^n P(A_i)$$

dla każdego ciągu zbiorów parami rozłącznych $A_i \in \mathcal{L}$.

OKREŚLENIE 2. Dowolna przeliczalnie addytywna funkcja zbioru przybierająca wyłącznie niewjemne wartości nazywa się *miarą* i jest oznaczana symbolem μ .

Zapis $\mu(A)$ czytamy „miara zbioru A ”. Zbiory należące do takiego ciała zbiorów, na którym określona została miara μ , nazywają się *zbiorami mieralnymi*. Jeżeli $\mu < \infty$, to miara jest *skończona*.

Umówimy się, że mówiąc o funkcji zbioru $P(A)$ będziemy mieli zawsze na myśli funkcję niewjemną. Wtedy funkcję tę można utożsamić z pojęciem miary.

Można udowodnić, że jeżeli $A \subset B$, gdzie $A, B \subset \mathcal{L}$, to

$$(3) \quad P(A) \leq P(B).$$

Oznacza to, że miara μ jest niemalejącą funkcją zbioru. Jeżeli $P(X) = 1$, gdzie X jak zwykle jest przestrzenią, to miara μ jest *unormowana*.

Na tle rozważań dotyczących miary i jej własności rodzi się pytanie, czy miarę da się określić na dowolnym ciele zbiorów. Nasuwa się podejrzenie, że tak nie jest, już choćby dlatego, że określenie miary na elementach jakiegoś ciała zbiorów polega na przyporządkowaniu tym elementom liczb rzeczywistych, a to prowadzi do wniosku, że moc ciała powinna być co najwyżej równa mocy zbioru liczb rzeczywistych. Stąd jeżeli moc ciała jest większa niż c , to takie ciało może nie być mierzalne.

Podejrzenie to jest słuszne. W pracy [1] na str. 27 zacytowane jest twierdzenie, z którego wynika, że nie na każdym ciele zbiorów można określić miarę; jeżeli jednak ciało jest ciałem zbiorów borełowskich, to określenie na nim miary jest zawsze możliwe. Właśnie dlatego tyle uwagi poświęciliśmy zbiorom borełowskim⁽¹⁾.

Z kolei można postawić pytanie, czy istnieje jakiś wygodny sposób określenia miary na ciałach zbiorów borełowskich, jeżeli zgodziliśmy się ograniczyć naszą uwagę tylko do takich zbiorów. Czyni się to za pomocą tzw. całki Lebesgue'a-Stieltjesa. Całka ta stanowi uogólnienie dobrze czytelnikowi znanej całki Riemanna. Definiując całkę Riemanna rozpatruje się tzw. sumy Darboux

$$\sum_{k=1}^n m_k \Delta x_k \quad \text{oraz} \quad \sum_{k=1}^n M_k \Delta x_k,$$

gdzie Δx_k są to długości rozłącznych podprzedziałów danego przedziału $\langle a, b \rangle$, a m_k i M_k są to odpowiednio dolny i górny kres danej funkcji $f(x)$ (patrz [19] str. 228 - 230). Jeżeli obie sumy są sobie równe, to ich wspólna wartość nazywa się całką Riemanna i jest oznaczana symbolem $\int_a^b f(x) dx$. Uogólnienie tej całki znane pod nazwą całki Lebesgue'a-Stieltjesa otrzymuje się przez zastąpienie w sumach Darboux długości Δx_k podprzedziałów przedziału $\langle a, b \rangle$ uogólnieniami miary $P(A_i)$ podzbiorów $A_i \in \mathcal{B}$. Tak więc całkę Lebesgue'a-Stieltjesa nazywa się wspólną wartość sum Darboux zdefiniowanych następująco:

$$\sum_i m_i P(A_i), \quad \sum_i M_i P(A_i).$$

⁽¹⁾ Przypominamy, że moc ciała zbiorów borełowskich będących podzbiorami przestrzeni liczb rzeczywistych jest nie większa od c .

Część II

RACHUNEK PRAWDOPODOBIEŃSTWA

2.1. WIADOMOŚCI Z ZAKRESU HISTORII RACHUNKU PRAWDOPODOBIĘŃSTWA⁽¹⁾

Rachunek prawdopodobieństwa jest gałęzią matematyki. Ciekawe jest powstanie tej nauki. Podwaliny jej zostały położone przez dwóch wybitnych matematyków francuskich B. Pascala (1623 - 1662) i P. Fermata (1601 - 1661). Do zajęcia się zagadnieniami probabilistycznymi miała rzekomo skłonić Pascala korespondencja, którą prowadził z Chevalierem de Méré. Był to namiętny gracz. Tryb życia, jaki prowadził, był bardzo kosztowny. Szukając łatwej drogi wzbogacenia się pragnął odkryć system, który pozwoliłby mu ujarzmić kapryśną Fortunę, kierującą wynikami gry w karty lub kości.

Grając dużo i często Chevalier de Méré poczynił szereg wnikliwych spostrzeżeń. Nie mogąc sobie poradzić z rozwiązyaniem zagadnień, które się na tle tych spostrzeżeń wyłoniły, zwrócił się o pomoc do Pascala. Rozwiązujeć zadania de Méré i wymieniając przy tej okazji swe poglądy z Fermatem, Pascal dał początek nowej dyscyplinie matematycznej zwanej rachunkiem prawdopodobieństwa.

Ogromne zasługi na polu rozwoju teorii prawdopodobieństwa położył matematyk szwajcarski J. Bernoulli (1654 - 1705), który pierwszy w swym dziele *De Arte Coniectandi Tractatus (Traktat o sztuce przewidywania)* w sposób wyraźny sformułował i udowodnił twierdzenie znane w rachunku prawdopodobieństwa pod nazwą prawa wielkich liczb. Prawo to dało matematyczną interpretację dobrze z doświadczenia znanego faktu, że zdarzenia przypadkowe, występujące masowo, wykazują pewne prawidłowości. Paradoks, tkwiący w tym, że przypadek przeobraża się w swoje przeciwieństwo – prawidłowość, tłumaczy, być może, zainteresowanie zagadnieniami probabilistycznymi, jakie wykazywała większość wybitnych matematyków, podpatrujących z pasją przejawy tej prawidłowości i przybierających rezultaty swych obserwacji w szatę matematyczną.

Wyniki osiągnięte przez Bernoulliego przejął i rozwinał matematyk francuski A. de Moivre (1667 - 1754). Zajmował się on zagadnieniami kombinatorycznymi i teorią gier losowych. Rozważaniom swym poświęcił dwie prace *The doctrine of chance* i *Miscellanea analytica*. W drugiej z tych prac zawarty jest dowód jednego z podstawowych twierdzeń rachunku prawdopodobieństwa, znanego pod nazwą twierdzenia Moivre'a-Laplace'a.

Szybkie postępy, jakie czyniła teoria prawdopodobieństwa, były niewątpliwie związane z rozwojem statystyki. Trudno określić dokładnie okres narodzin tej nauki, można jednak twierdzić, że powstała ona w epoce kapitalizmu wolnokonkurencyjnego.

⁽¹⁾ Szerszy wykład historii rachunku prawdopodobieństwa znajdzie czytelnik w pracach [11], [30].

Za ojca statystyki uważa się W. Petty'ego. Co było w owych czasach przedmiotem zainteresowań tej nauki – wymienia nieco przydługi podtytuł dzieła, które napisał Petty: *Rozważania, dotyczące rozmiarów cen ziemi, ludności, zabudowań, gospodarki rolnej, manufaktury, handlu, przemysłu rybnego, rzemieślników, marynarzy, żołnierzy oraz dochodów państwowych, procentów, podatków, sposobu powiększania dochodów.*

Okresem szybkiego rozwoju rachunku prawdopodobieństwa jest wiek XIX. Wielkie zasługi dla rozwoju teorii prawdopodobieństwa położył w tym okresie P. S. Laplace (1749 - 1827), który w swym dziele *Théorie analytique des probabilités* daje już bardzo rozwinięty, jak na owe czasy, wykład teorii prawdopodobieństwa. Dzieło to jest poprzedzone wstępem znanym pod nazwą *Essai philosophique sur des probabilités*. We wstępie tym zebrana jest treść wykładów z rachunku prawdopodobieństwa, które Laplace wygłosił w École Normale w 1795 r.

Nie mniejsze zasługi niż Laplace położył na polu rozwoju myśli probabilistycznej K. F. Gauss (1777 - 1855), twórca teorii błędów obserwacji i metody najmniejszych kwadratów.

Również A. L. Cauchy (1789 - 1857), ten niezmiernie pracowity, wszechstronny i płodny matematyk francuski, wniosł duży wkład do rachunku prawdopodobieństwa. Tuż obok Cauchy'ego wymienić należy również nazwisko S. D. Poissona (1781 - 1840), którego imieniem nazwany został jeden z najważniejszych rozkładów statystycznych.

Omawiając rozwój teorii prawdopodobieństwa osobne miejsce należy poświęcić matematykom rosyjskim. Na czoło wysuwają się tu prace członka Petersburskiej Akademii Nauk, Szwajcara z pochodzenia, L. Eulera (1707 - 1783). Imię tego wybitnego uczonego złotymi zgłoskami upamiętniło się w historii wielu dyscyplin matematycznych. Tutaj interesują nas prace Eulera z zakresu demografii i ubezpieczeń. Całkami Eulera nazywa się tzw. funkcję gamma i funkcję beta. Funkcje te mają szerokie zastosowanie w statystyce matematycznej.

Nad zagadnieniami, wchodzącyymi w zakres rachunku prawdopodobieństwa, pracował twórca geometrii nieeuklidesowej M. Łobaczewski (1792 - 1856).

Rachunkiem prawdopodobieństwa interesowali się również znani matematycy rosyjscy Ostrogradski i Buniakowski. Pozostawili oni w swej spuściźnie naukowej szereg cennych prac z zakresu teorii prawdopodobieństwa.

Za twórcę rosyjskiej szkoły probabilistycznej uznać niewątpliwie należy P. Czebyszewa (1821 - 1894). Napisał on wprawdzie jedynie cztery publikacje z zakresu rachunku prawdopodobieństwa, lecz prace te wywarły ogromny wpływ na dalszy rozwój tej nauki. Ze szkoły Czebyszewa wyszli uczeni tej miary, co A. Markow (1856 - 1922), A. Lapunow (1857 - 1918) oraz E. Ślucki (1880 - 1948).

Wybitni matematycy radzieccy: S. Bernsztejn, A. Kołmogorow, N. Smirnow, A. Chinczyn, B. Gniedenko, W. Romanowski, W. Gliwenko i ich uczniowie stworzyli radziecką szkołę teorii prawdopodobieństwa, która należy do czołowych w świecie.

Osiągnięcia współczesnej probabilistyki w Polsce są związane z imieniem profesora Uniwersytetu Wrocławskiego H. Steinhausa i jego uczniów.

Głównym ośrodkiem naukowym, zajmującym się rozwojem rachunku prawdopodobieństwa w Polsce, jest Instytut Matematyczny PAN.

2.2. O ZDARZENIACH

2.2.1. Klasyfikacja zdarzeń

Podając definicję jakiegoś nowego pojęcia posługujemy się pojęciami znanymi, zdefiniowanymi uprzednio. W ten sposób tłumaczymy sens pojęć bardziej złożonych za pomocą pojęć prostszych. Oczywiście przy takim postępowaniu muszą istnieć pojęcia najprostsze, których się nie definiuje, gdyż przyjmuje się, że są one zrozumiałe bezpośrednio. Takim pojęciem jest np. punkt w geometrii euklidesowej.

Gdy mowa o zdarzeniach – odpowiednikiem pojęcia punkt z geometrii jest *zdarzenie elementarne*. Pojęcie to jest kategorią aprioryczną, nie może być przeto zdefiniowane. Oto kilka przykładów zdarzeń elementarnych: wyrzucenie orła przy rzucie monetą, wy ciągnięcie z urny – zawierającej kule biale i czarne – kuli białej, urodzenie się chłopca, zgon w wieku 47 lat, wystąpienie braku w produkcji, natrafienie na ziarno pszenicy o wadze 0,75 g itp. Za pomocą zdarzeń elementarnych można zdefiniować inne rodzaje zdarzeń, będących kombinacją zdarzeń elementarnych. Oto co pisze na ten temat A. Rényi w artykule pt. *Podstawowe problemy rachunku prawdopodobieństwa* (patrz [25], str. 100):

„Oznaczmy przez a_1, a_2, \dots, a_n możliwe wyniki jakiegoś doświadczenia i nazwijmy je zdarzeniami elementarnymi. Utwórzmy ze zdarzeń elementarnych a_1, a_2, \dots, a_n wszelkie możliwe kombinacje i oznaczmy je wielkimi literami łacińskimi, np. $A = (a_2, a_1, a_5)$, $B = (a_3, a_4)$ itd. Niech A będzie jakąś kombinacją elementów a_1, a_2, \dots, a_n , $A = (a_{i_1}, a_{i_2}, \dots, a_{i_k})$. Oznaczmy przez A^+ zdarzenie, które polega na tym, że w doświadczeniu zrealizuje się pewna możliwość, zawarta w kombinacji A ; innymi słowy, zdarzenie A^+ zajdzie, gdy wynikiem doświadczenia będzie realizacja któregoś ze zdarzeń $a_{i_1}, a_{i_2}, \dots, a_{i_k}$. Przekładając to na język teorii mnogości, przez zdarzenie złożone lub, krótko mówiąc, przez zdarzenie rozumiemy dowolny podzbiór rozpatrywanego zbioru”.

Cytat powyższy przytoczyliśmy dlatego, że zawiera on współczesną wykładnię pojęcia *zdarzenie*, które dziś niemal powszechnie interpretuje się w sensie mnogościowym. Czytelnik znajdzie w rozdziale 1 te wybrane wiadomości o zbiorach, które teraz okażą się przydatne przy studiowaniu klasyfikacji zdarzeń i poznawaniu działań na zdarzeniach.

Umówmy się, że zbiór zdarzeń elementarnych będziemy zawsze oznaczać literą E , natomiast elementy tego zbioru literą e z odpowiednimi wskaźnikami. Jeżeli zbiór E jest skończony i liczy m elementów, to zapisujemy to symbolicznie w sposób następujący: $e_i \in E$, $i = 1, 2, \dots, m$. Podobnie wygląda ten zapis w przypadku, gdy E jest nieskończony, ale przeliczalny. Piszymy wtedy $e_i \in E$, $i = 1, 2, \dots$. Jeżeli natomiast zbiór E jest mocą continuum, to możemy dać temu wyraz pisząc np. $e_t \in E$, gdzie e_t jest funkcją rzeczywistego argumentu t . Zbiór E nazywać będziemy *przestrzenią zdarzeń elementarnych*.

Przechodzimy obecnie do wprowadzenia podstawowego pojęcia rachunku prawdopodobieństwa, a mianowicie pojęcia *zdarzenia losowego* lub krótko *zdarzenia*. Rozważmy na początek klasę wszystkich możliwych podzbiorów zbioru E . Jeżeli E jest zbiorem skończonym, to liczba elementów tej klasy wynosi, jak wiadomo, 2^n (jako suma kombinacji, jakie można utworzyć z elementów zbioru n -elementowego). Analogicznie, jeżeli zbiór E jest zbiorem przeliczalnym, to moc tej klasy wynosi 2^{\aleph_0} , czyli zgodnie z hipo-

tezą continuum moc ta wynosi c . Jeżeli jednak E jest zbiorem nieprzeliczalnym, to klasa podzbiorów zbioru E ma moc $2^c > c$. Oznacza to, że nie można na elementach tej klasy określić funkcji rzeczywistej, która mogłaby pełnić rolę miary, gdyż moc zbioru wartości, jakie ta funkcja mogłaby przybierać, byłaby mniejsza od mocy tej klasy. Właśnie z tych względów zrezygnujemy z rozważania całej klasy podzbiorów, jakie można utworzyć z elementów zbioru E i ograniczamy się do klasy zbiorów borełowskich. Jak wiadomo, na cieле zbiorów borełowskich określenie miary jest zawsze możliwe (patrz 1.5.5). W następnym paragrafie pokażemy, jak się konstruuje to ciało, tu już jednak sygnalizujemy, że właśnie elementy tego ciała będziemy interpretować jako *zdarzenia losowe*.

2.2.2. Algebra zdarzeń

Czytelnik spostrzegł niewątpliwie, że wszystkie nasze rozważania poświęcone zdarzeniom miały interpretację mnogościową, tzn. były oparte na teorii zbiorów. Spostrzeżenie to jest słuszne i będzie obowiązywało nadal. Przypomnijmy sobie, że określając borełowskie ciało zbiorów (patrz 1.5.4, określenie 4) korzystaliśmy z operacji dodawania, mnożenia i odejmowania zbiorów. Przechodzimy obecnie do zdefiniowania tych operacji w odniesieniu do zdarzeń. W całkowitej analogii do pojęcia sumy, iloczynu i różnic zbiorów wprowadzamy obecnie pojęcie sumy, iloczynu i różnicy zdarzeń.

OKREŚLENIE 1. *Sumą dwóch zdarzeń E_1 i E_2* nazywać będziemy zdarzenie polegające na tym, że zajdzie przynajmniej jedno z tych zdarzeń. Sumę zdarzeń oznaczać będziemy symbolem $E_1 + E_2$.

Określenie to można rozciągnąć na dowolną ilość zdarzeń. Dla ilustracji przytaczamy kilka przykładów na sumę zdarzeń. Rzucamy kością do gry. Zdarzenie polegające na tym, że wyrzucimy jedno oczko albo dwa, albo trzy, albo cztery, albo pięć, albo sześć, jest sumą zdarzeń. Czekamy na tramwaj. Zdarzenie polegające na tym, że przyjedzie „dwójka” albo „pięiątka”, albo „szóstka”, jest także sumą zdarzeń. Sumą zdarzeń będą również następujące zdarzenia: urodzenie się chłopca albo dziewczynki, trafienie albo nietrafienie do tarczy, wyrzucenie orła albo reszki, znalezienie w partii towaru, liczącej 100 sztuk, 0 albo 1, albo 2, albo 3 sztuk złych itd.

OKREŚLENIE 2. *Iloczynem dwóch zdarzeń E_1 i E_2* nazywać będziemy zdarzenie polegające na tym, że zajdzie zarówno zdarzenie E_1 jak i zdarzenie E_2 .

Określenie iloczynu zdarzeń można rozciągnąć na dowolną ilość zdarzeń. Iloczyn dwóch zdarzeń będziemy oznaczać symbolem $E_1 \cdot E_2$ lub $E_1 E_2$.

Podajemy kilka przykładów iloczynu zdarzeń, które wyjaśnią czytelnikowi treść tego pojęcia. Z talii kart wyciągamy jedną kartę. Symbolem E_1 oznaczamy zdarzenie, polegające na wyciągnięciu kiera, natomiast symbolem E_2 zdarzenie, polegające na wyciągnięciu damy. Iloczynem zdarzeń $E_1 \cdot E_2$ jest w takim razie wyciągnięcie z talii kart damy kier.

Nie należy sądzić, że iloczynem zdarzeń E_1 i E_2 jest zdarzenie polegające na jednoczesnym zajściu tych zdarzeń. Wyobraźmy sobie, że rzucamy dwa razy monetą. Niech E_1 oznacza wyrzucenie orła w pierwszym rzucie, natomiast E_2 – wyrzucenie orła w drugim rzucie. Oczywiście iloczynem zdarzeń E_1 i E_2 będzie zdarzenie, polegające na wyrzuceniu orła w obu rzutach. W przykładzie tym zdarzenia E_1 i E_2 nie zachodzą jednocześnie.

Kontrolujemy jakość produkcji zapałek. Kontrola polega na tym, że z partii wyprodukowanych zapałek wybieramy na chybił trafił 10 pudełek i badamy jakość wszystkich zapałek, zawartych w tych pudełkach. Oznaczmy symbolem E_i ($i = 1, 2, \dots, 10$) zdarzenie polegające na tym, że w i -tym pudełku znajdzie się więcej niż 20 złych zapałek. W takim razie zdarzenie $E_1 E_3 E_6 E_7$ polega na znalezieniu ponad 20 złych zapałek zarówno w pudełku pierwszym, jak i w trzecim, szóstym i siódmym.

OKREŚLENIE 3. *Różnicą dwóch zdarzeń E_1 i E_2* nazywać będziemy zdarzenie polegające na tym, że zdarzenie E_1 zachodzi, a zdarzenie E_2 nie zachodzi. Różnicę zdarzeń E_1 i E_2 oznaczać będziemy symbolem $E_1 - E_2$.

Oto przykład, który objaśni znaczenie tego terminu. W sortowni znajduje się partia jabłek dostarczonych przez dostawców. Jabłka te podlegają sortowaniu, które polega na tym, że jabłka dorodne, zdrowe, bez skaz, nadające się na eksport, pakuje się do skrzyń, a pozostałe przeznacza się na rynek wewnętrzny.

Jeżeli oznaczymy symbolem E_1 zdarzenie polegające na wybraniu dowolnego jabłka dostarczonego do sortowni, a symbolem E_2 – wybranie jabłka przeznaczonego na eksport, to różnicą zdarzeń $E_1 - E_2$ będzie zdarzenie polegające na tym, że wybrane zostało jabłko spośród jabłek dostarczonych do sortowni, ale nie było to jabłko eksportowe. Oczywiście oznacza to, że wybrano jabłko przeznaczone na rynek wewnętrzny. Zdarzenie to jest równoważne zajściu różnicy zdarzeń $E_1 - E_2$.

Do skonstruowania borełowskiego ciała zbiorów potrzebne były nie tylko suma, iloczyn i różnica zbiorów, lecz również zbiór pusty i przestrzeń. Dlatego właśnie wprowadzamy dwie dalsze definicje:

OKREŚLENIE 4. Podzbiór zbioru E nie zawierający żadnego elementu E nazywać będziemy *zdarzeniem niemożliwym* i oznaczać literą V .

OKREŚLENIE 5. Podzbiór zbioru E zawierający wszystkie elementy zbioru E nazywać będziemy *zdarzeniem pewnym* i oznaczać symbolem U .

Z pomocą zdarzenia pewnego i niemożliwego definiuje się zdarzenia przeciwnie.

OKREŚLENIE 6. Dwa zdarzenia E_1 i E_2 nazywają się *zdarzeniami przeciwnymi*, jeżeli

$$E_1 + E_2 = U \quad \text{oraz} \quad E_1 E_2 = V.$$

Łatwo sprawdzić, że zdarzenie pewne i zdarzenie niemożliwe są zdarzeniami przeciwnymi. Czytelnik dostrzeże niewątpliwie, że pojęcie zdarzenia przeciwnego jest analogiem pojęcia dopełnienia zbioru. Otrzymywanie zdarzenia przeciwnego przez zaprzeczenie zdarzenia danego nazywać będziemy *negacją* i oznaczać kreską nad symbolem zdarzenia. Symbol \bar{E}_1 oznacza więc zdarzenie przeciwe do zdarzenia E_1 .

OKREŚLENIE 7. Powiadamy, że zdarzenie E_1 zawiera się w E_2 lub że E_1 implikuje (pociąga) E_2 , jeżeli zawsze, gdy zajdzie E_1 , zajście E_2 jest zdarzeniem pewnym. Na oznaczenie *implikacji* używa się zapisu $E_1 \subset E_2$ lub $E_1 \Rightarrow E_2$.

OKREŚLENIE 8. Dwa zdarzenia E_1 i E_2 nazywać będziemy *zdarzeniami równoważnymi* i oznaczać $E_1 = E_2$, jeśli

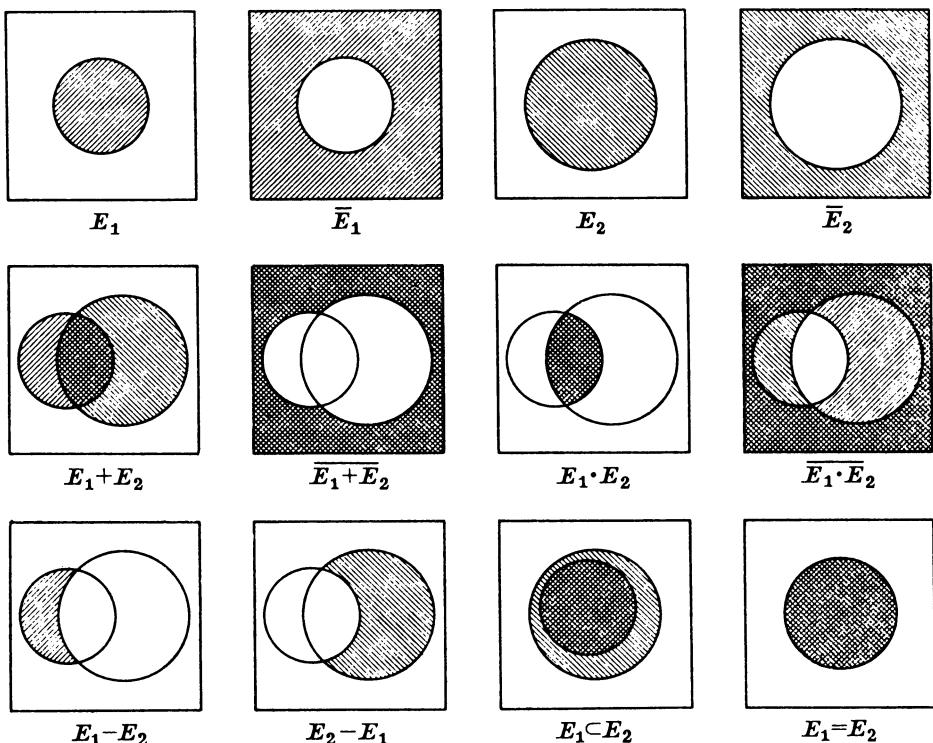
$$E_1 \subset E_2 \quad \text{oraz} \quad E_2 \subset E_1.$$

Jak widzimy, pojęcie równoważności zdarzeń zostało zdefiniowane za pomocą iloczynu i implikacji zdarzeń. Czytelnik sprawdzi, że pojęcie implikacji można zdefiniować za pomocą iloczynu i negacji

$$E_1 \subset E_2 = \bar{E}_1 \cdot E_2.$$

Tak więc zdefiniowaliśmy następujące działania i relacje dotyczące zdarzeń: sumę, iloczyn, różnicę oraz negację, implikację i równoważność. Są one odpowiednikami takich samych relacji i działań na zbiorach.

Bardzo często dla poglądowego przedstawienia sumy, iloczynu, różnicę, negacji, implikacji oraz równoważności korzysta się z tzw. *diagramów Eulera*.



Rys. 1

Przypuśćmy, że dany jest kwadrat o boku a . Wewnątrz tego kwadratu znajdują się dwa okręgi odpowiednio o promieniu s i S , gdzie $s \leq S$. W obrębie kwadratu obieramy na chybił trafił jakiś punkt. Symbolem E_1 oznaczono zdarzenie polegające na tym, że punkt ten znajdzie się wewnątrz okręgu o promieniu s , a symbolem E_2 zdarzenie polegające na tym, że punkt ten znajdzie się wewnątrz okręgu o promieniu S (rys. 1).

OKREŚLENIE 9. Dwa zdarzenia E_1 i E_2 noszą nazwę *zdarzeń wylczających się*, jeżeli

$$E_1 \cdot E_2 = V.$$

Zdarzeniami wyłączającymi się będą np. zdarzenia polegające na wyciągnięciu kiera, trefla lub pika, jeżeli z talii kart wyciągamy tylko jedną kartę lub zdarzenia polegające na wyrzuceniu orła lub reszki, jeżeli rzucamy tylko raz monetą.

Przedsiębiorstwo ma jeden wolny etat kierowcy. Trzech kandydatów ubiega się o tę pracę. Oczywiście zatrudnienie jednego z nich wyklucza możliwość zatrudnienia pozostałych. Są to więc także zdarzenia wyłączające się.

OKREŚLENIE 10. Niech zajście pewnego zdarzenia złożonego E_0 będzie równoważne z zajściem jednego ze zdarzeń E_1, E_2, \dots, E_n , tzn.

$$E_0 = E_1 + E_2 + \dots + E_n.$$

Jeśli zdarzenia E_1, E_2, \dots, E_n wyłączają się nawzajem, czyli jeżeli

$$E_i \cdot E_j = V \quad (i \neq j, i, j = 1, 2, \dots, n),$$

to mówimy, że zdarzenie E_0 rozkłada się na zdarzenia E_1, E_2, \dots, E_n .

Wynik rzutu kościami rozkłada się na następujących 6 zdarzeń: wyrzucenie jednego, dwóch, trzech, czterech, pięciu i sześciu oczek. Rzucając kościami możemy mieć jednak również do czynienia z innymi zdarzeniami. Możemy, dajmy na to, wyrzucić liczbę oczek podzielną przez dwa lub podzielną przez trzy. Te dwa zdarzenia nie są jednak zdarzeniami wyłączającymi się, przeto nie moglibyśmy powiedzieć, że wynik rzutu kościami do gry rozkłada się na wyrzucenie liczby oczek podzielnej przez dwa i liczby oczek podzielnej przez trzy.

OKREŚLENIE 11. Jeśli pewien zbiór \mathcal{B} zawiera zdarzenia E_1, E_2, \dots oraz jeśli zbiór ten zawiera również

- 1° zdarzenie pewne,
 - 2° zdarzenie niemożliwe,
 - 3° sumę zdarzeń E_1, E_2, \dots ,
 - 4° iloczyn zdarzeń E_1, E_2, \dots ,
 - 5° różnicę dowolnej pary zdarzeń ze zbioru E_1, E_2, \dots ,
- to zbiór \mathcal{B} nazywamy borelowskim ciałem zdarzeń.

Zwracamy uwagę, że definicję borelowskiego ciała zdarzeń można byłoby zredagować krócej wykorzystując fakt, że zdarzenie pewne, zdarzenie niemożliwe i iloczyn zdarzeń można wyrazić za pomocą sumy i różnicę zdarzeń. Definicja taka byłaby jednak mniej intuicyjna i dlatego rezygnujemy z niej, sygnalizując jednocześnie, że definicja taka byłaby sformułowana analogicznie do określenia 4 w 1.5.4.

OKREŚLENIE 12. Każdy element zbioru \mathcal{B} , utworzonego z podzbiorów zbioru zdarzeń elementarnych, nazywa się zdarzeniem losowym.

Z określenia 12 wynika, że przez zdarzenia losowe rozumieć będziemy takie zdarzenia, które w wyniku realizacji wyróżnionego zespołu warunków mogą zajść lub nie.

Oto garść przykładów zdarzeń losowych: wyrzucenie orła przy rzucie monetą, wyrzucenie pięciu oczek przy rzucie kościami, wyciągnięcie z talii kart karty czerwonej, urodzenie się chłopca, zgon człowieka w wieku 47 lat, pięciokrotne w ciągu dnia roboczego zerwanie

nici na warsztacie tkackim, znalezienie wśród 1000 puszek konserw 3 puszek dotkniętych bombażem itd.

Pytania kontrolne i zadania

1. Podać przykłady zdarzeń elementarnych i zdarzeń złożonych.
2. Jakie zdarzenie nazywamy zdarzeniem pewnym? Podać określenie zdarzenia pewnego.
3. Co tą są zdarzenia przeciwnie? Podać przykłady.
4. Co nazywamy zdarzeniami równoważnymi? Podać przykłady.
5. Wyjaśnić na przykładach, że jeśli $A \subset B$ i $B \subset C$, to $A \subset C$.
6. Podać określenie sumy, iloczynu i różniczki zdarzeń. Wyjaśnić treść tych pojęć na przykładach.
7. Uzasadnić i poprzeć przykładami słuszność następujących związków:
 - a) $A \subset A + B$; b) $B \subset A + B$; c) $A + AB = A$;
 - d) $A(B+C) = AB + AC$; e) $AB \subset A$; f) $AB \subset B$;
 - g) jeśli $C \subset A$ i $C \subset B$, to $C \subset AB$.
8. Zbadać, czy dla zdarzenia pewnego U i zdarzenia niemożliwego V zachodzą następujące związki:

$$U + V = U, \quad U \cdot V = V.$$

Co wynika ze słuszności tych związków?

9. Podać określenie i przykłady zdarzeń wyłączających się.
10. Wykazać podobieństwo i różnicę własności zdarzeń przeciwnych i zdarzeń wyłączających się.
11. Podać przykłady przestrzeni zdarzeń.
12. Podać określenie ciała zdarzeń.
13. Podać określenie i przykłady zdarzeń losowych.

2.3. POJĘCIE PRAWDOPODOBIEŃSTWA

2.3.1. Klasyczna definicja prawdopodobieństwa

Wyobraźmy sobie następujące doświadczenie. W urnie znajduje się 10 kul, 5 z nich jest koloru białego, a 5 koloru czerwonego. Założymy, że kule różnią się tylko barwą, poza tym są zupełnie jednakowe. Wyciągając na ślepo jedną kulę możemy przyjąć, że każda kula ma jednakową szansę wyciągnięcia. Gdyby kule były ponumerowane, to wyciągając jedną kulę moglibyśmy wyciągnąć kulę z numerem 1, 2, ..., 10.

Oznaczmy symbolem E zdarzenie, polegające na wyciągnięciu kuli, natomiast zdarzenie, polegające na wyciągnięciu kuli z numerem i – symbolem E_i ($i=1, 2, \dots, 10$). W takim razie

$$E = E_1 + E_2 + \dots + E_{10}$$

oraz

$$E_i \cdot E_j = V, \quad i \neq j, \quad i, j = 1, 2, \dots, 10.$$

Przypuśćmy, że kule o numerach od 1 do 5 są białe. Wyobraźmy sobie, że interesuje nas zdarzenie, polegające na wyciągnięciu kuli białej. Oznaczmy to zdarzenie symbolem A .

Wobec tego

$$A = E_1 + E_2 + \dots + E_5$$

oraz

$$E_i \cdot E_j = V.$$

Jak z tego widać, zdarzenie E rozkłada się na 10 zdarzeń, a zdarzenie A rozkłada się na 5 zdarzeń.

Wprowadzimy następujące

OKREŚLENIE 1. Jeśli w wyniku realizacji zdarzenia E , które rozkłada się na zdarzenia E_1, E_2, \dots, E_n , może zajść interesujące nas zdarzenie A , które rozkłada się na zdarzenia E_1, E_2, \dots, E_m , gdzie $0 \leq m \leq n$, to będziemy mówili, że zajściu zdarzenia A sprzyja m zdarzeń.

W naszym doświadczeniu zajściu zdarzenia A , tzn. wyciągnięciu kuli białej, sprzyja 5 zdarzeń, a mianowicie zdarzenia E_1, E_2, \dots, E_5 . Powiedzieliśmy poprzednio, że wyciągnięcie kuli białej jest w danych warunkach zdarzeniem losowym. Oznacza to, że zdarzenie to może wystąpić lub nie. Gdyby skład naszej urny uległ zmianie, np. gdyby w urnie tej znajdowało się 999 kul czerwonych, a tylko jedna biała, to wyciągnięcie kuli białej byłoby także zdarzeniem losowym, gdyż i w tym przypadku zdarzenie to mogłoby również wystąpić lub nie. Jakkolwiek w obu opisanych przypadkach mamy do czynienia ze zdarzeniami losowymi, jednak zdajemy sobie sprawę, że zdarzenia te różnią się między sobą.

Różnica ta polega na tym, że „łatwiej” wyciągnąć kulę białą w pierwszym przypadku niż w drugim. Oczywiście słowo „łatwiej” należy rozumieć w ten sposób, że gdybyśmy wielokrotnie wyciągali z urny po jednej kuli i wkładali ją każdorazowo z powrotem, to w pierwszym przypadku udałoby się nam znacznie częściej wyciągnąć kulę białą niż w drugim. Potocznie mówimy, że w pierwszym przypadku szansa wyciągnięcia kuli białej jest większa, w drugim zaś mniejsza. Aby uniknąć takich niedokładnych określeń jak „większa szansa”, „mniejsza szansa”, wprowadzimy pojęcie miary możliwości zajścia zdarzenia losowego. Taką miarę nazywać będziemy *prawdopodobieństwem*.

Podamy klasyczną definicję prawdopodobieństwa, której autorem jest Laplace⁽¹⁾.

OKREŚLENIE 2. Jeśli zdarzenie E rozkłada się na n wykluczających się wzajemnie i jednocześnie możliwych zdarzeń elementarnych, spośród których m sprzyja zajściu interesującego nas zdarzenia A , to *prawdopodobieństwem zdarzenia A* nazywa się ułamek, w którego liczniku znajduje się liczba zdarzeń sprzyjających zajściu zdarzenia A , w mianowniku zaś – liczba wszystkich możliwych zdarzeń.

Oznaczając prawdopodobieństwo zdarzenia A symbolem $P(A)$ możemy więc napisać

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

Ponieważ w opisany wyżej doświadczeniu mieliśmy w urnie 10 kul, spośród których 5 było koloru białego, przeto

$$P(A) = \frac{5}{10} = \frac{1}{2}.$$

⁽¹⁾ Definicję tę podał Laplace w 1812 r. w pracy *Théorie analytique des probabilités*.

Posługując się definicją klasyczną znajdzmy, czemu równa się prawdopodobieństwo wyrzucenia orła przy jednym rzucie monetą. Zakładając, że moneta posiada kształt idealnego krążka o jednakowej grubości oraz że wykonana jest z jednolitego materiału, mamy prawo uważać, że rzucając w dowolny sposób monetę możemy otrzymać jedynie dwa wyniki, a mianowicie orła albo reszkę (pomijamy sytuację, gdy moneta stanie pionowo na swym brzegu). Zajściu interesującego nas zdarzenia sprzyja jeden przypadek. W takim razie prawdopodobieństwo wyrzucenia orła przy jednorazowym rzucie monetą równa się $\frac{1}{2}$.

W definicji prawdopodobieństwa była mowa o zdarzeniu E . Zdarzenie to określa pewne ciało borełowskie zdarzeń \mathcal{B} . Każdemu zdarzeniu należącemu do ciała zdarzeń \mathcal{B} odpowiada jakieś prawdopodobieństwo. W takim razie prawdopodobieństwo możemy rozpatrywać jako funkcję tych zdarzeń. Funkcja ta jest określona na ciele zdarzeń \mathcal{B} . Zbadajmy własności tej funkcji. Ujęto je w postaci następujących twierdzeń.

TWIERDZENIE 1. *Funkcja $P(A)$ może przybierać jedynie wartości nieujemne, zawarte w przedziale $\langle 0, 1 \rangle$, czyli*

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

Istotnie, $P(A)=0$, gdy nie ma zdarzeń sprzyjających zajściu zdarzenia A . Wtedy bowiem

$$P(A) = \frac{0}{n} = 0.$$

Gdy ilość zdarzeń sprzyjających zajściu zdarzenia A równa się ilości wszystkich możliwych zdarzeń, to $P(A)=1$, gdyż wtedy

$$P(A) = \frac{m}{n} = \frac{n}{n} = 1.$$

Z tego wynika, że jeśli $0 < m < n$, to $0 < P(A) < 1$. Oczywiście, gdy nie ma zdarzeń sprzyjających zajściu zdarzenia A , to zdarzenie A jest zdarzeniem niemożliwym, czyli $A = V$. Stąd

$$(2) \quad P(V) = 0.$$

Jeśli $m = n$, czyli jeśli $A = U$, to

$$(3) \quad P(U) = 1.$$

TWIERDZENIE 2. *Jeśli zdarzenie A rozkłada się na dwa zdarzenia A_1 i A_2 (patrz 2.2.2, określenie 10), przy czym zdarzeniu A_1 sprzyja m_1 zdarzeń, a zdarzeniu A_2 sprzyja m_2 zdarzeń, to*

$$(4) \quad P(A) = P(A_1) + P(A_2).$$

Dowód tego twierdzenia jest bardzo prosty. Ponieważ $m = m_1 + m_2$, przeto

$$P(A) = \frac{m}{n} = \frac{m_1 + m_2}{n} = \frac{m_1}{n} + \frac{m_2}{n} = P(A_1) + P(A_2).$$

Z twierdzeń 1 i 2 wynika, że prawdopodobieństwo sumy zdarzeń przeciwnych równa się 1. Mamy bowiem $A + \bar{A} = U$. Stąd na mocy własności (3) i (4)

$$(5) \quad P(A) + P(\bar{A}) = P(U) = 1.$$

Jeśli oznaczymy

$$(6) \quad P(A) = p \quad \text{oraz} \quad P(\bar{A}) = q,$$

to otrzymamy następującą ważną zależność między p i q :

$$(7) \quad p = 1 - q.$$

Wzorem (7) będziemy często posługiwać się w przyszłości, przeto czytelnik powinien dobrze go sobie zapamiętać.

PRZYKŁAD 1. Rzucamy kością do gry. Zdarzenie A polega na wyrzuceniu liczby oczek podzielnej przez 3. Znaleźć $P(\bar{A})$.

Wyrzucenie liczby oczek podzielnej przez 3 jest równoważne wyrzuceniu trzech oczek albo sześciu oczek. Ponieważ prawdopodobieństwo wyrzucenia trzech oczek równa się prawdopodobieństwu wyrzucenia sześciu oczek i równa się $\frac{1}{2}$, przeto posługując się wzorem (4) znajdujemy, że prawdopodobieństwo sumy tych dwóch zdarzeń równa się sumie prawdopodobieństw tych zdarzeń, czyli równa się $\frac{2}{6} = \frac{1}{3}$. Obliczyliśmy, że $P(A) = p = \frac{1}{3}$. Prawdopodobieństwo zdarzenia przeciwnego, tzn. zdarzenia polegającego na wyrzuceniu liczby oczek niepodzielnej przez trzy, znajdziemy łatwo posługując się wzorem (7). Mamy bowiem

$$P(\bar{A}) = q = 1 - p = 1 - \frac{1}{3} = \frac{2}{3}.$$

Rachunek prawdopodobieństwa znajduje liczne zastosowania w badaniach przyrodniczych, mających bezpośrednie znaczenie praktyczne.

PRZYKŁAD 2. Przykład ten dotyczy zastosowania rachunku prawdopodobieństwa do szacowania liczby ryb oraz frakcji (udziału procentowego) poszczególnych gatunków ryb w zamkniętym zbiorniku wodnym (np. w stawie lub jeziorze). Postępuje się w sposób następujący. Dokonuje się połów ryb, ryby wyłowione poddaje się znakowaniu i wpuszcza się z powrotem do wody. Po upływie pewnego czasu (dostatecznego dla „wymieszania się” ryb znakowanych z pozostałymi rybami) dokonuje się ponownie połów i oblicza się frakcję ryb znakowanych w ogólnej liczbie ryb wyłowionych. Ta frakcja jest oszacowaniem⁽¹⁾ prawdopodobieństwa złowienia ryby znakowanej. Przypuśćmy, że wśród złowionych 948 ryb znaleziono 6 ryb znakowanych. Oznacza to, że

$$P(A) = \frac{6}{948} = \frac{1}{158}.$$

Jeżeli wiadomo, że do jeziora wpuszczono $m = 500$ ryb znakowanych, to liczba n wszystkich ryb w stawie może być oszacowana w sposób następujący:

$$\frac{m}{n} = P(A) = \frac{1}{158},$$

stąd

$$n = 500 \cdot 158 = 79000.$$

⁽¹⁾ Problematyczne szacowanie parametrów statystycznych poświęcony jest § 6.4. Patrz również [35], § 1.6.

Twierdzenie 2 nosi nazwę *twierdzenia o prawdopodobieństwie sumy zdarzeń*. Można je sformułować w odniesieniu do dowolnej liczby zdarzeń. Jeśli bowiem jakieś zdarzenie A rozkłada się na zdarzenia A_1, A_2, \dots, A_r , to prawdopodobieństwo sumy zdarzeń równe się sumie prawdopodobieństw tych zdarzeń.

TWIERDZENIE 3. Jeżeli zdarzenie A pociąga za sobą zdarzenie B , czyli jeśli $A \subset B$, to

$$(8) \quad P(A) \leq P(B).$$

Gdy $A \subset B$, to zbiór zdarzeń sprzyjających zajęciu zdarzenia A zawiera się w zbiorze zdarzeń sprzyjających zajęciu zdarzenia B . Zbiór zdarzeń sprzyjających zajęciu zdarzenia A może więc być najwyżej tak liczny jak zbiór zdarzeń sprzyjających zajęciu zdarzenia B . Liczniejszy być nie może, gdyż część nie może być większa od całości. Dowodzi to, że wzór (8) jest słuszny.

PRZYKŁAD 3. Przypuśćmy, że zdarzenie A polega na wyciągnięciu asa z talii kart liczącej 52 karty. Natomiast zdarzenie B polega na wyciągnięciu figury z tej talii kart. Oczywiście zajęcie zdarzenia A pociąga za sobą zajęcie zdarzenia B , gdyż jeśli został wyciągnięty as, to została tym samym wyciągnięta figura. Zbiór zdarzeń sprzyjających zajęciu zdarzenia A zawiera się w zbiorze zdarzeń sprzyjających zajęciu zdarzenia B , ponieważ każdy as jest figurą. Stąd wynika, że ilość asów nie może być większa od ilości figur. Istotnie, w talii kart znajduje się 16 figur, a cztery spośród nich są asami.

Prawdopodobieństwo wyciągnięcia asa równa się $\frac{4}{52} = \frac{1}{13}$, natomiast prawdopodobieństwo wyciągnięcia figury równa się $\frac{16}{52} = \frac{4}{13}$. Widzimy więc, że zgodnie z twierdzeniem 3 jest $P(A) < P(B)$.

2.3.2. Wady klasycznej definicji prawdopodobieństwa

Sformułowana przez Laplace'a klasyczna definicja prawdopodobieństwa jest prosta i nie budzi intuicyjnych sprzeciwów. Wydawać by się mogło, że za jej pomocą moglibyśmy rozwiązać z mniejszą lub większą trudnością dowolne zagadnienie probabilistyczne. Okazuje się jednak, że tak nie jest. Definicja ta posiada trzy poważne wady.

Pierwsza z nich polega na tym, że definicja ta jest *tautologią*. Oznacza to, że w definicji użyte jest słowo definiowane. Objasniając bowiem znaczenie słowa „prawdopodobieństwo” definicja klasyczna posługuje się pojęciem jednakowo możliwych zdarzeń. Zdarzenia jednakowo możliwe to nic innego jak zdarzenia jednakowo prawdopodobne, czyli zdarzenia o jednakowym prawdopodobieństwie ich wystąpienia. Jeśli nie wiadomo, co to jest prawdopodobieństwo, to nie wiadomo również, co to są zdarzenia jednakowo prawdopodobne lub jednakowo możliwe. I na odwrót: jeśli nie wiadomo, co to są zdarzenia jednakowo możliwe, to nie wiadomo, co to jest prawdopodobieństwo.

Druga wada klasycznej definicji prawdopodobieństwa polega na tym, że definicja ta wymaga, aby zbiór zdarzeń sprzyjających zajęciu danego zdarzenia A i zbiór wszystkich możliwych zdarzeń zawierały skończoną ilość elementów. W przeciwnym bowiem przypadku klasyczna definicja prawdopodobieństwa traci sens. Wyobraźmy sobie na przykład, że na prostej l dane są cztery punkty A, C, D, B . Punkty te wyznaczają pewien zbiór odcinków. Wybierzmy dwa spośród nich, a mianowicie odcinki AB i CD . Odcinek CD leży wewnętrz odcinka AB . Za pomocą klasycznej definicji nie potrafimy dać odpowiedzi na następujące pytanie: jeśli wiadomo, że wewnętrz odcinka AB wybrano pewien punkt,

to jakie jest prawdopodobieństwo, że punkt ten został wybrany wewnątrz odcinka CD ? Jak wiadomo, ilość punktów należących do pewnego odcinka nie może być wyrażona liczbą skończoną, natomiast klasyczna definicja wymaga obliczenia ułamka, w którego liczniku i mianowniku występują skończone liczby m i n .

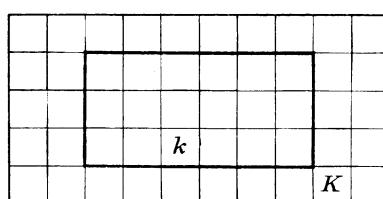
Trzecią wadą klasycznej definicji prawdopodobieństwa jest to, że żąda ona znajomości zbioru zdarzeń sprzyjających zajęciu danego zdarzenia A i zbioru wszystkich możliwych zdarzeń, związanych z realizacją doświadczenia, w wyniku którego może wystąpić zdarzenie A . Chcąc bowiem obliczyć $P(A)$ na podstawie definicji klasycznej, należy znać licznosć obu tych zbiorów. Warunek ten może być na ogół spełniony w odniesieniu do gier losowych, rzadko jednak spełnia się w praktyce przy badaniu zjawisk przyrodniczych lub społecznych.

Wymienione wady klasycznej definicji prawdopodobieństwa spowodowały, że definicja ta została poddana krytyce, która doprowadziła do innych sformułowań definicji prawdopodobieństwa.

2.3.3. Geometryczna definicja prawdopodobieństwa

Klasyczna definicja nie daje nam odpowiedzi, jakie jest prawdopodobieństwo, że punkt należący do odcinka AB należy również do CD , jeśli wiadomo, że CD zawiera się w AB . Przypuśćmy, że długość odcinka $AB=x$, natomiast długość odcinka $CD=y$. Intuicja podsufa nam proste rozwiązanie postawionego zagadnienia: szukane prawdopodobieństwo równa się y/x , tzn. jest stosunkiem długości odcinka CD do długości odcinka AB .

Gdy $x=y$, czyli gdy oba odcinki pokrywają się, prawdopodobieństwo to równa się jedności. Im mniejsza jest długość odcinka CD w stosunku do długości odcinka AB , tym mniejsze jest szukane prawdopodobieństwo. Intuicyjne rozwiązanie jest poprawne. Istotnie, liczba y/x jest szukanym prawdopodobieństwem, jak się o tym przekonamy.



Rys. 1

Rozpatrzmy jeszcze jeden przykład. Na rysunku 1 są dwa prostokąty. Mniejszy z nich leży całkowicie wewnątrz większego. Oznaczmy mniejszy prostokąt literą k , większy zaś – literą K . Należy znaleźć prawdopodobieństwo P , że wybrany na chybił trafił punkt należący do K będzie również należał do k . Powierzchnia obu prostokątów pokryta jest siatką kwadratową. Oznaczmy literą n ilość kwadratów pokrywających prostokąt K , natomiast ilość kwadratów pokrywających prostokąt k – literą m . Jeśli wybrany dowolnie punkt leży wewnątrz K , tzn. że leży on wewnątrz jednego spośród n kwadratów, pokry-

wających K . Jeżeli A oznacza zdarzenie, polegające na tym, że wybrany dowolnie punkt należący do K należy również do k , to zgodnie z definicją klasyczną zajścia zdarzenia A sprzyja m zdarzeń, podczas gdy ogólna ilość wszystkich możliwych zdarzeń równa się n . Na rysunku $n=50$, natomiast $m=18$. Wobec tego $P(A)=m/n=18/50=9/25$. Liczby m i n są miarami powierzchni obu prostokątów.

Podamy obecnie tzw. *geometryczną definicję prawdopodobieństwa*.

Jeśli Q i q są to dwa zbiorы w przestrzeni r -wymiarowej oraz jeśli $q \subset Q$, to prawdopodobieństwo tego, że dowolny punkt należący do Q będzie również należał do q , równa się stosunkowi miary zbioru q do miary zbioru Q .

Za pomocą geometrycznego prawdopodobieństwa szacuje się zasoby drewna lub runa leśnego pewnego ustalonego obszaru lasu, zachwaszczenie upraw rolnych, zasoby złóż rud mineralnych itd.

PRZYKŁAD. W fabryce lakieru przeprowadzono badanie jakości dwóch lakierów kryjących A i B , przeznaczonych głównie do ochrony cienkich blach żelaznych przed korozją. Badaniu poddano dwie próbne partie blach, z których każda liczyła po 20 arkuszy. Arkusze pierwszej partii pokryto lakierem A , natomiast arkusze drugiej partii — lakierem B . Na pokryte powłoką lakieru arkusze blachy skierowano strumień ziaren piasku kwarcowego i jednocześnie poddano blachę działaniu pary wodnej i zmianom temperatury. Po zakończeniu doświadczenia pozostawiono blachę w nasyconych parą wilgotnych komorach przez jednakowy dla obu partii okres czasu, a następnie zbadano frakcję powierzchni skorodowanej blach każdej partii. Uznano, że lakier B jest lakierem lepszym, gdyż wskaźnik skorodowania (prawdopodobieństwo geometryczne) dla lakieru A wyniósł 0,17, natomiast dla lakieru B — tylko 0,02.

2.3.4. Statystyczna, czyli częstościowa definicja prawdopodobieństwa

Nietrudno dostrzec pewne podobieństwo między klasyczną i geometryczną definicją prawdopodobieństwa. Definicja geometryczna jest co prawda wolna od dwóch pierwszych wad definicji klasycznej, nie jest jednak wolna od wady trzeciej. Dla obliczenia prawdopodobieństwa w oparciu o definicję geometryczną należy bowiem znać miary zbiorów Q i q , co w praktyce na ogół nie jest możliwe. Wobec tego statystycy podali propozycję innej definicji, zwanej statystyczną, czyli częstościową. Zanim podamy treść tej definicji, rozpatrzmy parę przykładów.

W urnie znajduje się 5 kul białych i 10 czerwonych. Jeśli skład urny jest znany, to w oparciu o klasyczną definicję prawdopodobieństwa, bez przeprowadzania jakichkolwiek doświadczeń, możemy powiedzieć, że jeśli A oznacza wyciągnięcie z urny kuli białej, to $P(A)=\frac{5}{15}=\frac{1}{3}$.

Zastanówmy się, czy możliwe jest określenie $P(A)$, jeśli skład urny nie jest znany. Aby na to odpowiedzieć, przeprowadźmy najpierw następujące doświadczenie. Będziemy z naszej urny wielokrotnie wyciągali jedną kulę, oglądali jej barwę, notowali wynik oględzin i wkładali kulę z powrotem do urny. Jeśli po każdorazowym wyciągnięciu i włożeniu kuli do urny wstrząśniemy urnę kilkakrotnie, to będziemy mogli uważać, że każdej kuli zapewniliśmy jednakową szansę wyciągnięcia. Oznaczmy symbolem A zdarzenie polegające na wyciągnięciu kuli barwy białej. Stosunek ilości wyciągniętych kul białych do ogólnej ilości ciągnięć będzie częstością zdarzenia A . Częstość ta w zależności od wyniku

eksperymentu będzie przybierała różne wartości. Okazuje się jednak, że w miarę powiększania ilości doświadczeń wahania częstości będą się stawały na ogół coraz mniejsze, oscylując wokół pewnej stałej liczby p . Tą nieznaną liczbą p jest stosunek kul białych do wszystkich kul, znajdujących się w urnie.

Powtarzając wielokrotnie nasze doświadczenie stwarzamy warunki, w których prawidłowość ma możliwość utorowania sobie drogi przez chaos przypadkowości. Rezultat pojedynczego ciągnienia zależy od przypadku: możemy wyciągnąć kulę białą lub czerwoną. W miarę jednak powiększania ilości doświadczeń wpływ przypadku maleje, a coraz wyraźniej zarysuje się prawidłowość wywołana stałym składem urny. Jeśli bowiem każda kula ma jednakową szansę wyciągnięcia, to przy wielokrotnym powtarzaniu doświadczenia każda kula zostanie wyciągnięta mniej więcej jednakową ilość razy, a tym samym stosunek ilości wyciągniętych kul białych do ogólnej ilości ciągnięć będzie zbliżony do składu urny, to znaczy do stosunku ilości kul białych do ilości wszystkich kul w urnie. Widzimy więc, że prawdopodobieństwo zdarzenia A można w pewien sposób określić na drodze eksperimentalnej.

Opisany tok postępowania doprowadził nas do tzw. *statystycznej definicji prawdopodobieństwa*, którą można by sformułować mniej więcej w sposób następujący:

Jeżeli przy wielokrotnej realizacji doświadczeń, w wyniku których może wystąpić zdarzenie A , częstość tego zdarzenia przejawia wyraźną prawidłowość, oscylując wokół pewnej nieznanej liczby p , i jeśli wahania częstości przejawiają tendencję malejącą w miarę wzrostu liczby doświadczeń, to liczba p nazywa się *prawdopodobieństwem zdarzenia A*.

Jak widzimy, statystyczne określenie prawdopodobieństwa, jakkolwiek mało scisłe, lecz za to bardzo sugestywne intuicyjnie, nic nie mówi o ogólnej ilości zdarzeń możliwych, o ilości zdarzeń sprzyjających, o zdarzeniach jednakowo możliwych. Te walory statystycznej definicji prawdopodobieństwa spowodowały, że została ona szeroko rozpowszechniona, szczególnie wśród przedstawicieli nauk eksperimentalnych. Prawdopodobieństwo rozumiane w sensie statystycznym okazało się użyteczną kategorią naukową. Dowiodły tego ciekawe wyniki badań naukowych uzyskane w oparciu o rachunek prawdopodobieństwa w takich dziedzinach wiedzy, jak demografia, biologia, agrotechnika i zootechnika, astronomia, fizyka jądrowa itd. Nowsze badania wykazały, że rachunek prawdopodobieństwa może również oddać cenne usługi naukom ekonomicznym.

W statystycznej definicji prawdopodobieństwa mówi się, że prawdopodobieństwo jest to pewna nieznana nam liczba p , wokół której waha się empirycznie zaobserwowana częstość. Oczywiście liczba ta opisując jakąś stronę realnej rzeczywistości wyraża obiektywną treść. Dowodzi to, że prawdopodobieństwo istnieje niezależnie od doświadczenia. Niemniej statystyczna definicja prawdopodobieństwa przeznacza eksperimentowi bardzo doniosłą rolę, gdyż tylko za pośrednictwem eksperymentu możemy uzyskać aproksymację prawdopodobieństwa i tylko eksperiment może dać w każdym konkretnym przypadku odpowiedź na pytanie, czy prawdopodobieństwo danego zdarzenia ma realny sens, czy też nie. Obserwując np. częstość urodzeń chłopców rzuca się w oczy wyraźna prawidłowość: częstość ta waha się wokół pewnej liczby, która jest nieco wyższa od liczby 0,5. Wobec tego można mówić o prawdopodobieństwie urodzenia chłopca i można na drodze eksperimentalnej uzyskać aproksymację wartości tego prawdopodobieństwa.

Gdy mowa o statystycznej definicji prawdopodobieństwa, warto wspomnieć o określaniu prawdopodobieństwa podanym przez R. Misesa. Jeżeli symbolem m/n oznaczymyczęstość jakiegoś zdarzenia A , to według Misesa prawdopodobieństwo

$$P(A) = \lim_{n \rightarrow \infty} \frac{m}{n}.$$

Określenie to, jakkolwiek rozprzestrzenione dosyć szeroko wśród przyrodników, jest nie do przyjęcia, ponieważ posiada dwie zasadnicze wady. Pierwsza z nich polega na tym, że w definicji Misesa określa się prawdopodobieństwo jako granicę, do której dąży częstość, gdy ilość doświadczeń rośnie do nieskończoności. Ponieważ ilość doświadczeń w praktyce jest zawsze wielkością ograniczoną, przeto definicja Misesa pozbawiona jest realnego sensu.

A oto druga wada tej definicji. Założmy rzeczą absurdalną, że możemy powiększać nieograniczenie liczbę doświadczeń. Okazuje się, że nawet przy takim założeniu definicja jest niesłuszna, gdyż mówi się w niej o granicy częstości, tzn. o granicy zmiennej empirycznej. Jakie wartości przybiera ta zmienność, do jakiej dąży granica oraz czy w ogóle dąży do jakiejś granicy, o tym decyduje wynik każdego eksperymentu z dążącą do nieskończoności serii doświadczeń. Wiemy zaś, że wynik każdego doświadczenia jest zdarzeniem losowym. Wobec tego zdążanie zmiennej m/n do jakiejś granicy jest także zdarzeniem losowym, mającym swoje prawdopodobieństwo. Widać z tego, że definicja Misesa, podobnie jak definicja klasyczna, jest również tautologią, niewidoczną na pierwszy rzut oka⁽¹⁾.

2.3.5. Współczesna definicja prawdopodobieństwa. Aksjomatyka rachunku prawdopodobieństwa

Żadna z omówionych definicji prawdopodobieństwa nie może być podstawą, na której można by zbudować mocną, harmonijną i wolną od sprzeczności konstrukcję naukową.

Szybki rozwój sił wytwórczych oraz osiągnięcia twórczej myśli naukowej doby współczesnej stworzyły nowe tereny zastosowań twierdzeń rachunku prawdopodobieństwa. Zadania, które należało rozwiązać, stawały się coraz bardziej skomplikowane. Dotychczasowy aparat badawczy już nie wystarczał. Należało pokonać trudności związane ze zdefiniowaniem prawdopodobieństwa i zająć się stworzeniem, w oparciu o istniejący dorobek myśli probabilistycznej, samodzielnej gałęzi wiedzy i odrębnej dyscypliny naukowej. Zadanie to zostało zainicjowane przez Bernsteina, a ostatecznie rozwiązane przez Kołmogorowa. Ci dwaj

⁽¹⁾ Błędność koncepcji Misesa polega nie na tym, iż korzysta on w swej definicji z pojęcia granicy, nadając tym samym prawdopodobieństwu abstrakcyjną, matematyczną formę, lecz na tym, że posługuje się pojęciem granicy w sposób niewłaściwy. Określenie Misesa pozbawione jest sensu, gdyż w praktyce nie możemy zwiększać nieograniczenie liczby doświadczeń, a więc nigdy nie możemy uczynić zadość warunkom, o których mówi się w definicji Misesa.

wybitni matematycy radzieccy uważani są powszechnie za twórców nowoczesnej szkoły rachunku prawdopodobieństwa.

Według koncepcji Kołmogorowa rachunek prawdopodobieństwa jest gałęzią matematyki. Punktem wyjścia rachunku prawdopodobieństwa, podobnie jak w innych dyscyplinach matematycznych, powinien być system pewników, które przyjmuje się bez dowodu i z których w drodze poprawnego rozumowania dedukcyjnego wyprowadza się i dowodzi nowych twierdzeń. U podstaw rachunku prawdopodobieństwa w jego nowoczesnej postaci leżą twierdzenia teorii mnogości i teorii miary.

Oto zaproponowany przez Kołmogorowa system pewników, stanowiących fundament rachunku prawdopodobieństwa.

PEWNIK I. *Każdemu zdarzeniu A, wchodzącemu w skład borelowskiego ciała zdarzeń, przyporządkowana jest pewna liczba $P(A)$. Liczba ta czyni zadość warunkowi $0 \leq P(A) \leq 1$ i nazywa się prawdopodobieństwem zdarzenia A.*

Pewnik I jest zarazem definicją prawdopodobieństwa. Widzimy, że definicja wspólna ma charakter formalno-logiczny, gdyż niczego nie zakłada o zdarzeniach ani o sposobie przyporządkowania tym zdarzeniom określonych prawdopodobieństw. Takie ujęcie zagadnienia posiada cenne zalety, gdyż pozwala łatwo powiązać teorię z praktyką. Oto np. jednym ze sposobów takiego przyporządkowania poszczególnym zdarzeniom wchodząącym w skład borelowskiego ciała zdarzeń odpowiednich prawdopodobieństw jest przyjęcie, że te prawdopodobieństwa równają się zaobserwowanym na drodze eksperymentalnej częstościom tych zdarzeń. Łatwość powiązania częstości z prawdopodobieństwami posiada duże znaczenie praktyczne. Warto w tym miejscu przytoczyć ważne słowa, zawarte w artykule Chinczyna pt. *Metoda funkcji dowolnych i walka z idealizmem w dziedzinie teorii prawdopodobieństwa*, opublikowanym w numerze 3 (4) „Przekładów”, zeszyty „Myśli Filozoficznej”. Chinczyn pisze:

„Wszyscy zgadzamy się co do tego, że prawdopodobieństwo zdarzenia powinno być bliskie częstości jego pojawiения się przy dużej liczbie doświadczeń. Jeżeli na przykład rachunek teoretyczny daje na prawdopodobieństwo jakiegoś zdarzenia wartość $\frac{1}{3}$, a przy powtarzaniu doświadczeń – zdarzenie to będzie uporczywie zachodzić zawsze w jednej czwartej wszystkich przypadków, to jednomyslnie zadecydujemy, że rachunek przeprowadzony został błędnie albo opierał się na fałszywych założeniach, albo też popełniono myłkę w obliczeniach. Dzisiaj żaden kurs wykładów rachunku prawdopodobieństwa nie może się obejść bez wyraźnego wskazania – już na samym wstępie – konieczności tej częstościowej interpretacji prawdopodobieństwa. Wszyscy jesteśmy w stanie konkretnie uchwycić istotny sens dowolnego twierdzenia i dowolnego zagadnienia teorii prawdopodobieństwa tylko wtedy, gdy na miejsce prawdopodobieństw wszystkich zdarzeń postawimy odpowiednie częstości”.

PEWNIK II. *Prawdopodobieństwo zdarzenia pewnego równa się jedności:*

$$P(U)=1.$$

Czytelnik zechce porównać treść pewnika I i pewnika II z treścią twierdzenia 1 (patrz 2.3.1); analogia jest oczywista.

PEWNIK III. *Prawdopodobieństwo sumy skończonej lub przeliczalnej ilości parami wylączających się zdarzeń A_1, A_2, \dots równa się sumie prawdopodobieństw poszczególnych zdarzeń:*

$$P(A_1 + A_2 + \dots) = P(A_1) + P(A_2) + \dots$$

Przytoczone trzy pewniki noszą nazwę *systemu pewników rachunku prawdopodobieństwa* lub inaczej *aksjomatyki rachunku prawdopodobieństwa*.

Pytania kontrolne i zadania

1. Jak brzmi klasyczna definicja prawdopodobieństwa?
2. Z talii kart, liczącej 52 karty, wyciągamy jedną kartę. Znaleźć prawdopodobieństwo wyciągnięcia:
a) karty czerwonej, b) figury, c) asa, d) nieasa, e) asa czarnego, f) asa pik.
3. Rzucamy monetą. Znaleźć prawdopodobieństwo wyrzucenia: a) orła, b) reszki, c) orła albo reszki.
4. Rzucamy kością do gry. Znaleźć prawdopodobieństwo wyrzucenia: a) sześciu oczek, b) parzystej liczby oczek, c) podzielnej przez 3 liczby oczek, d) podzielnej przez 4 liczby oczek, e) nieparzystej liczby oczek.
5. Wymienić i omówić wady klasycznej definicji prawdopodobieństwa.
6. Podać geometryczną definicję prawdopodobieństwa.
7. Jak brzmi statystyczna definicja prawdopodobieństwa?
8. Wymienić pewniki rachunku prawdopodobieństwa.

2.4. PODSTAWOWE TWIERDZENIA RACHUNKU PRAWDOPODOBIĘSTWA

2.4.1. Wnioski z aksjomatyki prawdopodobieństwa

Z pewników podanych w poprzednim paragrafie łatwo wyciągnąć kilka ważnych wniosków.

Wniosek 1. *Prawdopodobieństwo zdarzenia niemożliwego równa się zeru.*

Istotnie, ponieważ możemy napisać, że

$$U = U + V,$$

przeto na mocy pewnika III

$$P(U) = P(U) + P(V).$$

Ponieważ jednak pewnik II głosi, że $P(U) = 1$, więc

$$1 = 1 + P(V),$$

czyli

$$(1) \quad P(V) = 0.$$

Wniosek 2. *Suma prawdopodobieństw zdarzeń przeciwnych równa się jedności.*

Korzystając bowiem z określenia zdarzeń przeciwnych możemy napisać, że

$$A + \bar{A} = U.$$

Stąd

$$(2) \quad P(A) + P(\bar{A}) = P(U) = 1.$$

To z kolei prowadzi do znanego już wzoru:

$$P(A) = 1 - P(\bar{A}).$$

Wniosek 3. Jeżeli zdarzenie A pociąga za sobą zdarzenie B , to prawdopodobieństwo zdarzenia A jest nie większe od prawdopodobieństwa zdarzenia B .

Rzeczywiście, zamiast zwrotu „zdarzenie A pociąga za sobą zdarzenie B ” moglibyśmy bowiem użyć powiedzenia „ A zawiera się w B ”. Z tego wynika bezpośrednio, że

$$B = A + C,$$

gdzie $C = B - A$.

Oczywiście, jeśli $P(C) \neq 0$, to

$$P(B) = P(A) + P(C) > P(A).$$

Znak nierówności zamienia się na znak równości, gdy $P(C) = 0$.

2.4.2. Zdarzenia niezależne

W rachunku prawdopodobieństwa i statystyce matematycznej duże znaczenie ma pojęcie zdarzeń niezależnych.

OKREŚLENIE 1. Dwa zdarzenia A i B nazywają się *zdarzeniami niezależnymi*, jeśli zajście jednego z tych zdarzeń nie ma wpływu na prawdopodobieństwo zajścia drugiego (poźniej podamy ścisłą definicję zdarzeń niezależnych).

Rozpatrzmy parę przykładów zdarzeń niezależnych. Rzucamy dwa razy monetą; wyrzucenie orła w pierwszym rzucie nie ma wpływu na prawdopodobieństwo wyrzucenia orła w drugim rzucie. Te dwa zdarzenia są od siebie niezależne.

Ciekawe, że na przekór pozorom, nie jest to dla wszystkich intuicyjnie oczywiste. Niektórzy rozumują w sposób następujący: jeśli w poprzednim rzucie monetą został wyrzucony orzeł, to przy założeniu, że wyrzucenie orła i wyrzucenie reszki jest zdarzeniem jednakowo prawdopodobnym, należy uważać, że w następnym rzucie wyrzucenie reszki jest zdarzeniem bardziej prawdopodobnym niż wyrzucenie orła, gdyż orzeł już raz wypadł, a reszka nie.

Rozumowanie to jest błędne. W opisany przykładzie już z samego charakteru zjawiska wynika, że rezultaty obu rzutów są od siebie niezależne. Intuicja słusznie co prawda podpowiada, że prawdopodobieństwo wyrzucenia w obu rzutach monetą dwóch orłów jest mniejsze, niż prawdopodobieństwo wyrzucenia jednego orła i jednej reszki. Stąd jednak nie wynika bynajmniej, że wyrzucenie w pierwszym rzucie orła zwiększa prawdopodobieństwo wyrzucenia w drugim rzucie reszki. Z tych rozważań wypływa ważna wskazówka praktyczna: rozstrzygnięcie, czy zdarzenia są od siebie zależne, czy też nie, jest na ogół zadaniem trudnym, ponieważ nawet w tak prostym przykładzie, jak rzut dwoma monetami, nie jest to zupełnie oczywiste.

Oto inne przykłady zdarzeń niezależnych. Na stole stoją dwie urny zawierające kule białe i czerwone. Oczywiście wyciągnięcie kuli białej z pierwszej urny nie ma wpływu na prawdopodobieństwo wyciągnięcia kuli czerwonej z drugiej urny. Zdarzeniami niezależnymi będą również wyniki dwukrotnego rzutu kością do gry lub wyniki jednorazowego rzutu dwoma kościemi.

We wszystkich przytoczonych przykładach była mowa o dwóch zdarzeniach. Zobaczmy później, że pojęcie niezależności można rozciągnąć na dowolną liczbę zdarzeń.

O tym, czy zdarzenia są od siebie zależne czy też nie, decyduje między innymi sposób losowania. Przekona nas o tym następujący przykład. W urnie znajduje się 5 kul białych i 10 czerwonych. Doświadczenie polega na tym, że z urny ciągnimy dwukrotnie jedną kulę. Jeśli po wyciągnięciu kuli i obejrzeniu jej barwy włożymy ją z powrotem do urny, to taki sposób losowania nosi nazwę *losowania ze zwracaniem*, jeśli zaś wyciągnięta kula nie powraca do urny, to mówimy, że korzystamy ze schematu *losowania bez zwracania*⁽¹⁾.

Oznaczmy symbolem A zdarzenie polegające na wyciągnięciu kuli białej w pierwszym ciągnieniu, a symbolem B – zdarzenie polegające na wyciągnięciu kuli białej w drugim ciągnieniu. Przekonamy się, że zdarzenia A i B są zdarzeniami niezależnymi, gdy posługujemy się schematem losowania ze zwracaniem, natomiast zdarzenia te są zdarzeniami zależnymi, gdy losowanie jest bez zwracania. Istotnie, gdy po wyciągnięciu i obejrzeniu barwy kuli wkładamy ją z powrotem do urny, to

$$P(A) = P(B) = \frac{5}{15} = \frac{1}{3}.$$

Inaczej przedstawia się sprawa, gdy korzystamy ze schematu losowania bez zwracania. Wtedy bowiem prawdopodobieństwo wyciągnięcia kuli białej w pierwszym ciągnieniu równa się $\frac{5}{15} = \frac{1}{3}$, natomiast prawdopodobieństwo wyciągnięcia kuli białej w drugim ciągnieniu zależy od tego, jaka kula została wyciągnięta w pierwszym ciągnieniu; jeśli w pierwszym ciągnieniu została wyciągnięta kula biała, to $P(B) = \frac{4}{14} = \frac{2}{7}$, a jeśli kula czerwona, to $P(B) = \frac{5}{14}$.

OKREŚLENIE 2. Jeśli prawdopodobieństwo zdarzenia B zależy od dodatkowych warunków, to prawdopodobieństwo zdarzenia B nazywać będziemy *prawdopodobieństwem warunkowym* lub *względnym*.

Prawdopodobieństwo warunkowe zdarzenia B przy założeniu, że zaszło zdarzenie A , oznaczać będziemy symbolem $P(B|A)$. Korzystając z określenia 2 możemy sformułować warunek niezależności zdarzeń.

OKREŚLENIE 3. Dwa zdarzenia A i B są od siebie *niezależne*, jeśli

$$(1) \quad P(A) = P(A|B)$$

lub, co na jedno wychodzi, jeśli

$$(2) \quad P(B) = P(B|A).$$

Niech na przykład A oznacza wyrzucenie na kości do gry liczby oczek mniejszej od 2, a B – wyrzucenie liczby oczek mniejszej od 3. W takim razie mamy

(1) Patrz [35], rozdz. II.

$$P(A) = \frac{1}{6},$$

$$P(A|B) = \frac{1}{2},$$

skąd

$$P(A) \neq P(A|B),$$

a więc zdarzenia A i B są od siebie zależne.

W podobny sposób łatwo sprawdzić, że przy losowaniu ze zwracaniem zdarzenia A i B są niezależne.

W celu dokładnego wyjaśnienia, co oznacza prawdopodobieństwo warunkowe i na czym polega badanie niezależności zdarzeń, podamy jeszcze jeden przykład. Z talii kart ciągnimy jedną kartę. Należy zbadać, czy wyciągnięcie figury i wyciągnięcie karty czarnej są to zdarzenia zależne, czy też nie. Oznaczmy literą A wyciągnięcie figury, natomiast literą B – wyciągnięcie karty czarnej. W takim razie

$$P(A) = \frac{16}{52} = \frac{4}{13},$$

gdyż kart jest 52, a figur 16. Obliczamy $P(A|B)$, czyli prawdopodobieństwo wyciągnięcia figury przy założeniu, że wyciągnięto kartę czarną. Kart czarnych jest 26, a wśród nich znajduje się 8 figur; stąd

$$P(A|B) = \frac{8}{26} = \frac{4}{13}.$$

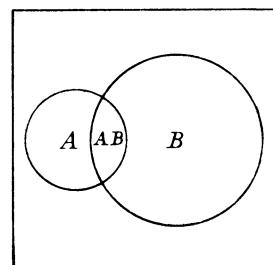
Ponieważ $P(A) = P(A|B)$, przeto zdarzenia A i B są od siebie niezależne (czytelnik sprawdzi z łatwością, że $P(B) = P(B|A)$).

2.4.3. Prawdopodobieństwo iloczynu zdarzeń

Na rysunku 1 przedstawiony jest kwadrat, wewnątrz którego znajdują się dwa okręgi przecinające się ze sobą. Przypuśćmy, że w obrębie kwadratu wybrano na chybił trafili jakiś punkt. Oznaczmy literą A zdarzenie polegające na tym, że ten punkt trafi do lewego okręgu, natomiast literą B – zdarzenie polegające na tym, że punkt znajdzie się wewnątrz prawego okręgu. Przypuśćmy, że powierzchnia lewego koła równa jest $k > 0$, powierzchnia prawnego koła jest równa $r > 0$, a powierzchnia wspólna obu kół wynosi s . W takim razie, jeśli powierzchnia kwadratu równa się n , to

$$P(A) = \frac{k}{n}, \quad P(B) = \frac{r}{n}, \quad P(AB) = \frac{s}{n},$$

$$P(A|B) = \frac{s}{r}, \quad P(B|A) = \frac{s}{k}.$$



Rys. 1

Ze wzoru na prawdopodobieństwo warunkowe łatwo wyprowadzić wzór na prawdopodobieństwo iloczynu zdarzeń A i B . Ponieważ

$$P(A|B) = \frac{s}{r},$$

przeto dzieląc licznik i mianownik prawej strony tej równości przez n otrzymamy

$$P(A|B) = \frac{s}{r} = \frac{\frac{s}{n}}{\frac{r}{n}} = \frac{P(AB)}{P(B)}.$$

Z założenia wynika, że $P(B) \neq 0$; stąd

$$(1) \quad P(AB) = P(B)P(A|B).$$

Przekształcając wzór na $P(B|A)$, otrzymamy

$$(2) \quad P(AB) = P(A)P(B|A).$$

Wzory (1) i (2) noszą nazwę *wzorów na prawdopodobieństwo iloczynu dwóch dowolnych zdarzeń*.

Jeśli zdarzenia A i B są zdarzeniami niezależnymi, to (patrz 2.4.2, określenie 3):

$$P(A|B) = P(A) \quad \text{i} \quad P(B|A) = P(B).$$

Wobec tego wzory (1) i (2) przybiorą prostszą postać, a mianowicie

$$(3) \quad P(AB) = P(A)P(B).$$

Zastosowanie wzoru (2) zilustruje następujący przykład. Z talii kart wyciągamy kolejno dwie karty, bez wkładania ich z powrotem do talii. Należy znaleźć prawdopodobieństwo, że obie karty będą asami. Oznaczmy literą A zdarzenie polegające na tym, że pierwsza karta będzie asem, natomiast literą B – zdarzenie polegające na tym, że druga karta będzie asem. W takim razie mamy

$$P(A) = \frac{4}{52} = \frac{1}{13},$$

$$P(B|A) = \frac{3}{51} = \frac{1}{17}.$$

Stąd

$$P(AB) = P(A)P(B|A) = \frac{1}{13} \cdot \frac{1}{17} = \frac{1}{221}.$$

Zadanie to moglibyśmy także rozwiązać inaczej. Zastanówmy się, na ile sposobów można wyciągnąć dwie karty spośród 52 kart. Aby odpowiedzieć na to pytanie, należy obliczyć C_{52}^2 . Będzie to ogólna ilość wszystkich możliwych zdarzeń elementarnych. Aby znaleźć ilość zdarzeń sprzyjających, należy obliczyć C_4^2 , gdyż tyloma sposobami można wyciągnąć dwa asy spośród czterech asów. Wobec tego

$$P(AB) = \frac{C_4^2}{C_{52}^2} = \frac{\frac{4!}{2!(4-2)!}}{\frac{52!}{2!(52-2)!}} = \frac{1}{221}.$$

Wzór na prawdopodobieństwo iloczynu można rozciągnąć na dowolną liczbę zdarzeń:

$$(4) . \quad P(E_1 E_2 \dots E_n) = P(E_1) P(E_2 | E_1) P(E_3 | E_1 E_2) \dots P(E_n | E_1 \dots E_{n-1}).$$

Wzór (4) jest rzadko stosowany w praktyce, podajemy go więc bez wyprowadzenia⁽¹⁾.

Posługując się tym wzorem rozwiążemy następujące zadanie: Z talii kart ciągnimy bez zwracania trzy karty. Znaleźć prawdopodobieństwo wyciągnięcia asa, króla i damy.

Oznaczmy symbolem E_1 wyciągnięcie asa, E_2 – wyciągnięcie króla, E_3 – wyciągnięcie damy. Wobec tego

$$P(E_1) = \frac{4}{52}, \quad P(E_2 | E_1) = \frac{4}{51}, \quad P(E_3 | E_1 E_2) = \frac{4}{50}.$$

W takim razie

$$P(E_1 E_2 E_3) = \frac{4 \cdot 4 \cdot 4}{52 \cdot 51 \cdot 50} = \frac{8}{16575}.$$

Wzór (3) nosi nazwę *wzoru na prawdopodobieństwo iloczynu dwóch zdarzeń niezależnych*. Ze wzoru tego wynika, że jeśli zdarzenia A i B są od siebie niezależne, to prawdopodobieństwo iloczynu tych zdarzeń równa się iloczynowi ich prawdopodobieństw. Można wykazać, że twierdzenie odwrotne jest także prawdziwe, a więc relacja (3) jest koniecznym i dostatecznym warunkiem niezależności dwóch zdarzeń. To kryterium może być łatwo uogólnione na przypadek dowolnej skończonej liczby zdarzeń.

OKREŚLENIE 1. Zdarzenia E_1, E_2, \dots, E_n są od siebie *niezależne*, jeżeli relacja

$$P(E_{k_1} E_{k_2} \dots E_{k_s}) = P(E_{k_1}) P(E_{k_2}) \dots P(E_{k_s})$$

zachodzi dla każdego podcięgu wskaźników k_1, k_2, \dots, k_s , $1 < s \leq n$, ciągu liczb naturalnych $1, 2, \dots, n$.

Rozważmy przypadek trzech zdarzeń E_1, E_2, E_3 . Czytelnik zechce zwrócić uwagę, że równość

$$P(E_1 E_2 E_3) = P(E_1) P(E_2) P(E_3)$$

nie pociąga jeszcze za sobą niezależności zdarzeń E_1, E_2, E_3 . Objaśni to następujący przykład.

PRZYKŁAD 1. Z urny, zawierającej 24 kartki ponumerowane liczbami 1, 2, ..., 24, losuje się jedną kartkę. Oznaczmy przez E_1 wylosowanie liczby nie większej od 12, przez E_2 – wylosowanie jednej z liczb 1, 2, 3, 4, 5, 7, a przez E_3 – wylosowanie liczby podzielnej przez 3. W takim razie:

$$P(E_1) = \frac{12}{24} = \frac{1}{2},$$

$$P(E_2) = \frac{6}{24} = \frac{1}{4},$$

$$P(E_3) = \frac{8}{24} = \frac{1}{3},$$

$$P(E_1 E_2 E_3) = \frac{1}{24},$$

$$P(E_1) P(E_2) P(E_3) = \frac{1}{2} \cdot \frac{1}{4} \cdot \frac{1}{3} = \frac{1}{24}.$$

⁽¹⁾ Wyprowadzenie znajdziesz czytelnik w książce [8] na str. 30.

Tak więc

$$P(E_1 E_2 E_3) = P(E_1) P(E_2) P(E_3).$$

Łatwo jednak zauważyc, że zdarzenia E_1, E_2, E_3 nie są niezależne, gdyż

$$P(E_1 E_2) = \frac{6}{24} = \frac{1}{4},$$

natomiast

$$P(E_1) P(E_2) = \frac{1}{8}.$$

I podobnie

$$P(E_2 E_3) = \frac{1}{24},$$

natomiast

$$P(E_2) P(E_3) = \frac{1}{12}.$$

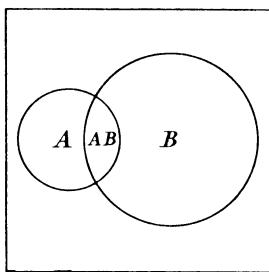
Relacja niezależności spełnia się jedynie dla zdarzeń E_1, E_3 . Mamy bowiem

$$P(E_1 E_3) = P(E_1) P(E_3) = \frac{4}{24} = \frac{1}{6}.$$

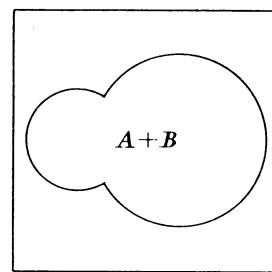
2.4.4. Prawdopodobieństwo sumy zdarzeń

Pewnik III głosi, że prawdopodobieństwo sumy zdarzeń wyłączających się równa się sumie prawdopodobieństw tych zdarzeń. Dotychczas nie wiemy jednak, czemu równa się prawdopodobieństwo sumy dowolnych zdarzeń. Spróbujmy rozwiązać to zagadnienie.

Na rysunku 1 przedstawiony jest kwadrat, wewnątrz którego leżą dwa przecinające się ze sobą okręgi. Oznaczmy literą A zdarzenie polegające na tym, że punkt P , obrany w dowolny sposób wewnątrz kwadratu, znajdzie się wewnątrz lewego okręgu oraz literą B – zdarzenie polegające na tym, że punkt ten znajdzie się wewnątrz prawego okręgu. Interesuje nas zdarzenie $A + B$, tzn. że punkt znajdzie się albo wewnątrz okręgu lewego, albo wewnątrz prawego. Z rysunku widać, że zdarzenie równoważne sumie zdarzeń A i B zajdzie



Rys. 1



Rys. 2

wtedy, gdy punkt P znajdzie się wewnątrz obszaru R , ograniczonego krzywą zamkniętą, utworzoną przez oba okręgi. Obszar ten przedstawiony jest na rysunku 2. Przypuśćmy, że powierzchnia lewego koła równa się k , powierzchnia prawego koła – r , a powierzchnia wspólna obu kół – s . Mając te dane można łatwo znaleźć powierzchnię obszaru R . Oznaczmy szukaną powierzchnię literą g . Z rysunku widać, że

$$(1) \quad g = k + r - s.$$

Stąd już tylko jeden krok do znalezienia wzoru na prawdopodobieństwo sumy dowolnych zdarzeń. Podzielmy obie strony równania (1) przez n , gdzie n oznacza powierzchnię kwadratu; otrzymamy

$$\frac{g}{n} = \frac{k}{n} + \frac{r}{n} - \frac{s}{n},$$

ale

$$\frac{g}{n} = P(A+B), \quad \frac{k}{n} = P(A), \quad \frac{r}{n} = P(B), \quad \frac{s}{n} = P(AB).$$

Wobec tego

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

PRZYKŁAD 1. Z talii liczącej 52 karty wyciągnięto jedną kartę. Znaleźć prawdopodobieństwo, że będzie to figura lub karta czerwona.

Oznaczmy symbolem A wyciągnięcie figury, a symbolem B — wyciągnięcie karty czerwonej. W takim razie

$$P(A) = \frac{16}{52} = \frac{4}{13}, \quad P(B) = \frac{26}{52} = \frac{1}{2}, \quad P(AB) = \frac{8}{52} = \frac{2}{13}.$$

Wobec tego

$$P(A+B) = \frac{8+13-4}{26} = \frac{17}{26}.$$

2.4.5. Wzór na prawdopodobieństwo całkowite i wzór Bayesa

Dane jest zdarzenie A i wzajemnie wyłączające się zdarzenia E_1, E_2, \dots, E_n . Przyjmujemy, że zdarzenie A może zajść tylko łącznie z jednym ze zdarzeń E_i ($i=1, 2, \dots, n$). W takim razie

$$A = AE_1 + AE_2 + \dots + AE_n.$$

Ponieważ zgodnie z założeniem E_1, E_2, \dots, E_n wyłączają się wzajemnie, przeto i zdarzenia AE_1, AE_2, \dots, AE_n są również zdarzeniami wyłączającymi się. Stąd

$$P(A) = P(AE_1) + P(AE_2) + \dots + P(AE_n) = \sum_{i=1}^n P(AE_i).$$

Ze wzoru na prawdopodobieństwo iloczynu dwóch dowolnych zdarzeń wynika jednak, że

$$P(AE_i) = P(E_i)P(A|E_i).$$

Ostatecznie otrzymujemy, że

$$(1) \quad P(A) = \sum_{i=1}^n P(E_i)P(A|E_i).$$

Jest to tzw. wzór na prawdopodobieństwo całkowite.

Zastosowanie tego wzoru wyjaśnia następujący przykład: Mamy cztery urny, w których znajdują się kule białe i czarne. W pierwszej urnie mamy 999 kul białych i jedną czarną, natomiast w każdej z pozostałych trzech urn znajduje się jedna kula biała i jedna czarna.

Wkładamy na chybił trafił rękę do jednej z urn i wyciągamy kulę. Należy znaleźć prawdopodobieństwo, że wyciągnięta kula będzie koloru czarnego.

Aby rozwiązać postawione zadanie, oznaczmy symbolem A wyciągnięcie czarnej kuli, E_1 – wyciągnięcie kuli z urny, w której stosunek kul czarnych do kul białych równa się $1 : 999$, E_2 – wyciągnięcie kuli z urny, w której stosunek kul czarnych do kul białych równa się $1 : 1$.

Ponieważ mamy cztery urny, więc

$$P(E_1) = \frac{1}{4}, \quad P(E_2) = \frac{3}{4}.$$

Prawdopodobieństwo wyciągnięcia kuli czarnej przy założeniu, że włożyliśmy rękę do pierwszej urny, wynosi $\frac{1}{1000}$. Jest to prawdopodobieństwo warunkowe zdarzenia A przy założeniu, że zaszło zdarzenie E_1 . Znaleźliśmy więc

$$P(A|E_1) = \frac{1}{1000}.$$

Analogicznie znajdujemy

$$P(A|E_2) = \frac{1}{2}.$$

Mamy już wszystkie dane potrzebne do obliczenia $P(A)$. Podstawiając je do wzoru (1) otrzymamy

$$P(A) = \frac{1}{4} \cdot \frac{1}{1000} + \frac{3}{4} \cdot \frac{1}{2} = 0,37525.$$

Nietrudno spostrzec, że prawdopodobieństwo wyciągnięcia kuli czarnej byłoby znacznie mniejsze, gdybyśmy wszystkie kule zsypali do jednej urny. Wtedy $P(A)$ równałoby się bowiem

$$\frac{4}{1006} \approx 0,003976.$$

Na marginesie naszego przykładu zanotujemy więc ważne spostrzeżenie: wielkość prawdopodobieństwa zależy od sposobu losowania. Prawdopodobieństwo interesującego nas zdarzenia A może być inne, gdy losowania dokonujemy z populacji⁽¹⁾ nie podzielonej na części (które odtąd nazywać będziemy *warstwami*), a inne, gdy populacja przed losowaniem podzielona została na warstwy, a losowanie odbywa się w ten sposób, że najpierw losuje się warstwę, następnie przeprowadza się losowanie wewnątrz warstwy⁽²⁾.

Wzór na prawdopodobieństwo całkowite służy do obliczania prawdopodobieństwa zdarzenia A przy założeniu, że zdarzenie to może zajść jedynie łącznie z jednym z wyłącznie wzajemnie wyłączających się zdarzeń E_1, E_2, \dots, E_n .

Obecnie wyprowadzimy wzór, który pozwala obliczyć prawdopodobieństwo, że zaszło jedno z wzajemnie wyłącznie wzajemnie wyłączających się zdarzeń E_1, E_2, \dots, E_n , jeśli wiadomo, że zaszło zdarzenie A . Innymi słowy, wyprowadzimy wzór, który podaje, czemu się równa $P(E_i|A)$. Posługując się twierdzeniem o prawdopodobieństwie iloczynu dwóch dowolnych zdarzeń

⁽¹⁾ Przez *populację* w statystyce rozumie się zbiór jednostek podlegających badaniu. Pełny zbiór interesujących nas jednostek nazywa się *populacją generalną*. Jeśli z populacji generalnej wylosujemy lub w jakiś sposób wybierzemy pewną ilość jednostek, to zbiór tych jednostek nazywa się *populacją próbą* lub krótko *próbką* (patrz 6.1). Wspominaliśmy o tym już w przykładzie 4 z 1.1.3.

⁽²⁾ Patrz [35], § 2.5.

możemy napisać, że

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

Stąd

$$(2) \quad P(E_i|A) = \frac{P(E_i) P(A|E_i)}{P(A)}, \quad P(A) \neq 0.$$

Ponieważ jednak na mocy wzoru (1)

$$P(A) = P(E_1) P(A|E_1) + P(E_2) P(A|E_2) + \dots + P(E_n) P(A|E_n),$$

więc

$$(3) \quad P(E_i|A) = \frac{P(E_i) P(A|E_i)}{P(E_1) P(A|E_1) + P(E_2) P(A|E_2) + \dots + P(E_n) P(A|E_n)}.$$

Wzór (3) nosi nazwę *wzoru Bayesa*. Sens praktyczny tego wzoru, posiadającego liczne zastosowania, wyjaśnia dwa przykłady.

PRZYKŁAD 1. Mamy cztery urny, w których znajdują się kule białe i czarne. Skład poszczególnych urn jest taki sam, jak w przykładzie ilustrującym zastosowanie wzoru na prawdopodobieństwo całkowite. Z jednej z urn została wyciągnięta kula czarna. Należy znaleźć prawdopodobieństwo, że kula ta została wyciągnięta z urny, w której stosunek kul czarnych do kul białych równa się 1 : 1000.

Poprzednio obliczyliśmy, że

$$P(E_1) = \frac{1}{4}, \quad P(A|E_1) = \frac{1}{1000} = 0,001, \quad P(A) = 0,37525 \approx 0,375.$$

Wobec tego

$$P(E_1|A) = \frac{0,25 \cdot 0,001}{0,375} \approx 0,00067.$$

PRZYKŁAD 2. Sklep spożywczy otrzymuje spirytus od trzech dostawców, których oznaczymy literami X, Y, Z. Z praktyki wiadomo, że dostawy X zawierają średnio 2% butelek z uszkodzonymi pierścieniami nakrętek, dostawy Y – 6%, a dostawy Z – 1%.

Od odbiorcy wpłynęła reklamacja, w której klient skarży się, że kupiona przez niego butelka o naruszonym pierścieniu zamiast spirytusu zawierała wodę. Dochodzenie ustaliło, że w wymienionym w reklamacji dniu nabycia butelki w magazynie sklepu znajdowała się mniej więcej jednakowa ilość towaru od każdego z dostawców. Chcemy obliczyć, jakie jest prawdopodobieństwo, że reklamowana butelka pochodzi od dostawcy Y.

Wprowadzimy następujące symbole: A będzie oznaczać wylosowanie butelki z naruszonym pierścieniem, E₁ – wylosowanie butelki od dostawcy X, E₂ – wylosowanie butelki od dostawcy Y, E₃ – wylosowanie butelki od dostawcy Z.

Mamy:

$$P(E_1) = \frac{1}{3}, \quad P(A|E_1) = 0,02,$$

$$P(E_2) = \frac{1}{3}, \quad P(A|E_2) = 0,06,$$

$$P(E_3) = \frac{1}{3}, \quad P(A|E_3) = 0,01.$$

Obliczamy P(A):

$$P(A) = \frac{1}{3}(0,02 + 0,06 + 0,01) = 0,03.$$

Wobec tego

$$P(E_2|A) = \frac{\frac{1}{3} \cdot 0,06}{0,03} = \frac{2}{3}.$$

2.4.6. O konieczności ścisłego formułowania zagadnień probabilistycznych

Przy rozwiązywaniu zagadnień z rachunku prawdopodobieństwa należy zawsze z niezwykłą starannością i uwagą zdefiniować:

- 1° warunki, w jakich realizuje się doświadczenie,
- 2° zdarzenie, które stanowi przedmiot naszych zainteresowań,
- 3° sposób przyporządkowania poszczególnym zdarzeniom odpowiadających im prawdopodobieństw.

W przeciwnym przypadku grozi nam otrzymanie błędnych wyników. Zilustrujemy to na przykładach.

PRZYKŁAD 1. Z talii liczącej 52 karty wyciągnięto 3 karty. Należy znaleźć prawdopodobieństwo, że będą to same asy.

Zadanie to zredagowane jest wadliwie, gdyż nie określono wyraźnie warunków, w jakich realizuje się doświadczenie. W zadaniu nie powiedziano, czy kartę po wyciągnięciu wkłada się z powrotem do talii, czy też nie.

Jeśli losowanie jest bezzwrotne, to szukane prawdopodobieństwo równa się

$$\frac{4}{52} \cdot \frac{3}{51} \cdot \frac{2}{50} = \frac{1}{5525},$$

jeśli natomiast losowanie jest zwrotne, to prawdopodobieństwo to wyniesie

$$\frac{4}{52} \cdot \frac{4}{52} \cdot \frac{4}{52} = \frac{1}{2197}.$$

Jak tego należało oczekiwać, otrzymaliśmy różne wyniki.

PRZYKŁAD 2. Oto treść dwóch zadań, które pozornie niczym się nie różnią między sobą:

a) Z urny zawierającej 5 kul białych i 5 kul czarnych wyciągamy kolejno dwie kule nie wkładając ich z powrotem do urny. Znaleźć prawdopodobieństwo, że będą to kule różnobarwne.

b) Z urny zawierającej 5 kul białych i 5 kul czarnych wyciągamy kolejno dwie kule nie wkładając ich z powrotem do urny. Znaleźć prawdopodobieństwo, że wyciągnemy kulę białą i czarną.

Czytając uważnie tekst obu zadań, spostrzeżemy niewątpliwie, że w zadaniu a) interesuje nas zdarzenie, polegające na tym, że pierwsza kula będzie koloru białego, a druga czarnego lub – na odwrót – pierwsza będzie koloru czarnego, a druga białego. Szukane prawdopodobieństwo wynosi

$$2 \cdot \frac{5}{10} \cdot \frac{5}{9} = \frac{5}{9}.$$

W zadaniu b) natomiast interesuje nas zdarzenie polegające na tym, że pierwsza kula będzie koloru białego, druga zaś koloru czarnego. Prawdopodobieństwo tego zdarzenia równa się

$$\frac{5}{10} \cdot \frac{5}{9} = \frac{5}{18}.$$

Przykład ten wykazuje, że najmniejsza nieostrożność w redagowaniu definicji interesującego nas zdarzenia może łatwo doprowadzić do błędnych rezultatów.

Jeśli jednak należy bardzo starannie redagować określenia zdarzeń, to rzecz oczywista, również bardzo uważnie należy czytać te określenia. Przykład 3 podaje typowy błąd, jaki często trafia się osobom, nie stosującym się do tego wymogu.

PRZYKŁAD 3. Rzucamy kości i monetę. Należy znaleźć prawdopodobieństwo wyrzucenia albo orła, albo reszki, albo trzech oczek.

Niektórzy rozwiązuje to zadanie w następujący sposób: Niech A oznacza wyrzucenie orła, B – wyrzucenie reszki, a C – wyrzucenie trzech oczek. W takim razie

$$P(A) = \frac{1}{2}, \quad P(B) = \frac{1}{2}, \quad P(C) = \frac{1}{6}.$$

Wobec tego

$$P(A + B + C) = \frac{1}{2} + \frac{1}{2} + \frac{1}{6} = \frac{7}{6}.$$

Otrzymana liczba jest większa od jedności. Oznacza to, że w obliczeniach popełniono błąd. Łatwo wskazać, gdzie się on kryje. Przy rozwiązywaniu zadania zastosowano wzór na prawdopodobieństwo sumy zdarzeń wyłączających się, a przecież wystarczy tylko przeczytać uważnie tekst zadania, aby przekonać się, że zdarzenia A i C oraz B i C nie są zdarzeniami wyłączającymi się.

Oto jak powinno wyglądać prawidłowe rozwiązanie zadania:

$$P(A + B + C) = P(U + C) = P(U) = 1.$$

Poprawne zdefiniowanie interesującego nas zdarzenia i warunków, w jakich realizuje się to zdarzenie, nie wystarcza do jednoznacznego rozwiązania zagadnień probabilistycznych. Aby uzyskać taką jednoznaczność, należy podać sposób, w jaki prawdopodobieństwo, będące jak wiadomo funkcją pewnego ciała zdarzeń \mathcal{B} , zostało określone na tym ciele zdarzeń. Sprawę tę wyjaśni przykład 4, do rozpatrzenia którego właśnie przystępujemy.

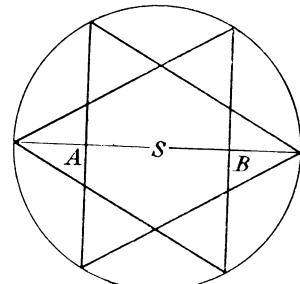
PRZYKŁAD 4⁽¹⁾. W okrąg wpisany jest trójkąt równoboczny. Należy znaleźć prawdopodobieństwo, że wybrana na chybił trafili cięciwa okręgu okaże się dłuższa od boku trójkąta.

Rozwiązanie 1. Obieramy w okręgu średnicę S (rys. 1). Na średnicach tej leżą środki cięciw prostopadłych do średnicy. Ze względu na symetrię okręgu warunkom zadania czynią zadość te cięciwy, których środki leżą na odcinku AB . Jak wiadomo, wysokość trójkąta równobocznego równa się $\frac{\sqrt{3}}{2}R$, łatwo więc obliczyć, że długość odcinka

$$AB = 2(2R - \frac{\sqrt{3}}{2}R) = R.$$

Stąd szukane prawdopodobieństwo wynosi $\frac{1}{2}$.

Rozwiązanie 2. W poprzednim rozwiążaniu rozpatrywaliśmy zbiór cięciw okręgu, których środki leżały na średnicach okręgu. W niniejszym rozwiążaniu zajmiemy się zbio-



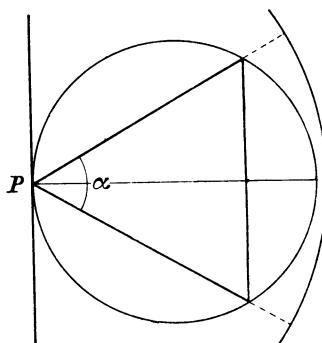
Rys. 1

⁽¹⁾ Patrz [11], str. 33 - 34.

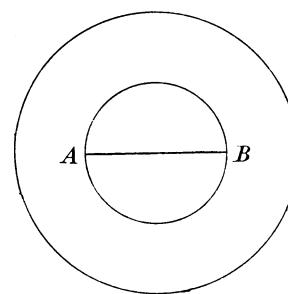
rem cięciw, które leżą na promieniach półokręgu zakreślonego promieniem większym od S w ten sposób, że średnica tego półokręgu jest styczna do okręgu w punkcie P (patrz rys. 2). Oczywiście, ze względu na symetrię okręgu, warunkom zadania czynią zadość cięciwy leżące wewnątrz kąta α . Ponieważ kąt ten równa się 60° , przeto szukane prawdopodobieństwo wynosi

$$\frac{60^\circ}{180^\circ} = \frac{1}{3}.$$

Rozwiązanie 3. Przyjrzyjmy się jeszcze raz rysunkowi 1. Wszystkie cięciwy, których środki leżą wewnątrz odcinka AB , są dłuższe od boku trójkąta wpisanego w okrąg. Zakreśl-



Rys. 2



Rys. 3

my ze środka okręgu nowy okrąg o średnicy AB . Wszystkie cięciwy o środkach leżących wewnątrz okręgu, którego długość średnicy wynosi AB , spełniają warunki zadania (patrz rys. 3). Ponieważ pole większego koła równe jest R^2 , a pole mniejszego koła $\frac{1}{4}R^2$, przeto interesujące nas prawdopodobieństwo równe jest $\frac{1}{4}$.

Widzimy, że otrzymaliśmy trzy różne rozwiązania naszego zadania. Zadanie to znane jest w rachunku prawdopodobieństwa pod nazwą *paradoksu Bertranda*. Które z podanych wyżej trzech rozwiązań jest poprawne? Oczywiście wszystkie one są poprawne. Mamy tu właściwie do czynienia z trzema różnymi zadaniami. W każdym z tych zadań w inny sposób przyporządkowano prawdopodobieństwa interesującym nas zdarzeniom i dlatego właśnie otrzymaliśmy różne wyniki.

Prawdopodobieństwo jest funkcją elementów zbioru zdarzeń. Aksjomatyka rachunku prawdopodobieństwa nie określa jednoznacznie tej funkcji. Sposób określenia funkcji zależy od rozwiązywanego zagadnienia. Wynika on z warunków, w których przebiega doświadczenie, z okoliczności, które mają wpływ na realizację zdarzenia, a tym samym i na sposób przyporządkowania temu zdarzeniu odpowiadającego mu prawdopodobieństwa⁽¹⁾.

Poprawne zdefiniowanie warunków, w jakich realizuje się doświadczenie, zdarzeń, których prawdopodobieństwa szukamy i funkcji wiążącej prawdopodobieństwa z odpo-

⁽¹⁾ Szersze omówienie tych zagadnień znajdzie czytelnik w pracach [5] i [25].

wiązającymi im zdarzeniami nastręcza w praktyce wiele trudności. Duże usługi oddają w takich przypadkach modele zdarzeń losowych, które w uproszczonej postaci, lecz z zachowaniem wszystkich istotnych cech prezentują mechanizm zdarzeń losowych. Tymi modelami są gry losowe: kości, karty, rzucanie monetą, wyciąganie kul z urny, gra na loterii i inne.

Nie dlatego wykłady rachunku prawdopodobieństwa ilustrowane są przykładami, w których jest mowa o grach losowych, że trudno podać inne, ciekawsze przykłady ważnych zagadnień naukowych, lecz dlatego, że gry losowe są modelami zdarzeń losowych. Jeśli model posiada wszystkie ważne cechy oryginału, to badając model poznajemy oryginał. Posługiwanie się grami losowymi należy do metodyki badań probabilistycznych, gdyż ułatwia studiowanie rachunku prawdopodobieństwa, rozwija wyobraźnię, pomaga w rozwiązywaniu zagadnień probabilistycznych, ćwiczy intuicję i ułatwia kontrolę wniosków otrzymanych w drodze rozważań dedukcyjnych (umożliwia bowiem sprawdzenie słuszności tych wniosków bądź za pomocą eksperymentu myślowego, bądź za pośrednictwem bezpośredniego doświadczenia). Należy o tym dobrze pamiętać w czasie studiowania twierdzeń rachunku prawdopodobieństwa i przerabiania przykładów ilustrujących te twierdzenia. Bagatelizowanie zadań i przykładów, w których jest mowa o grach losowych, utrudni znacznie i bez tego niełatwwe zadanie opanowania podstaw teorii prawdopodobieństwa. Warto w tym miejscu dodać, że korzystanie z modeli zdarzeń losowych znajdujemy zarówno w pracach klasycznych, jak i współczesnych.

Pytania kontrolne i zadania

1. Czemu równa się prawdopodobieństwo zdarzenia niemożliwego?
2. Czemu równa się suma prawdopodobieństw zdarzeń przeciwnych?
3. Podać określenie i przykłady różnicy zdarzeń.
4. Podać określenie i przykłady zdarzeń niezależnych.
5. Opisać, na czym polega schemat losowania zwrotnego i bezzwrotnego.
6. W którym z tych dwóch schematów zdarzenia są zależne, a w którym niezależne?
7. Podać warunek niezależności zdarzeń.
8. Rzucamy kością do gry. Zbadać, czy wyrzucenie parzystej liczby oczek i wyrzucenie podzielnej przez trzy liczby oczek są to zdarzenia zależne, czy też nie.
9. Z talii kart ciągnimy jedną kartę. Zbadać, czy wyciągnięcie asa i wyciągnięcie pika są to zdarzenia od siebie zależne, czy nie.
10. Podać określenie i przykłady prawdopodobieństw warunkowych.
11. Jak brzmi twierdzenie o prawdopodobieństwie iloczynu dwóch dowolnych zdarzeń? Podać brzmienie tego twierdzenia, gdy zdarzenia są niezależne.
12. Rzucamy jeden raz trzema kościemi do gry. Znaleźć prawdopodobieństwo wyrzucenia trzech szóstek.
13. Z urny, zawierającej 5 kul białych i 5 kul czerwonych, wyciągamy kolejno dwie kule bez wkładania ich z powrotem do urny. Znaleźć prawdopodobieństwo, że pierwsza kula będzie czerwona, a druga biała.
14. Z urny o tym samym składzie co w zadaniu 13 wyciągamy kolejno cztery kule, posługując się losowaniem bezzwrotnym. Znaleźć prawdopodobieństwo, że wśród wyciągniętych kul znajdą się trzy czerwone i jedna biała.
15. Jak brzmi twierdzenie o prawdopodobieństwie sumy dwóch dowolnych zdarzeń? Podać brzmienie tego twierdzenia, gdy zdarzenia wyłączają się wzajemnie.

16. Rzucamy dwa razy monetą. Znaleźć prawdopodobieństwo, że przynajmniej raz wyrzucimy orła.

17. Rzucono dwie kości do gry. Jakie jest prawdopodobieństwo zdarzenia polegającego na tym, że wyrzucona suma oczek równa się 8, jeśli wiadomo, że wyrzucono parzystą liczbę oczek?

18. Z talii kart, liczącej 52 karty, wyjęto dwie karty bez wkładania kart do talii. Znaleźć prawdopodobieństwo, że

- a) druga karta będzie asem,
- b) druga karta będzie asem, jeśli wiadomo, że pierwsza była asem.

19. Pewna osoba ma losy dwóch loterii fantowych. Prawdopodobieństwo wygrania na pierwszej loterii równa się 0,1, a prawdopodobieństwo wygrania na drugiej loterii równa się 0,3. Znaleźć prawdopodobieństwo:

- a) wygrania na pierwszej i drugiej loterii,
- b) wygrania albo na pierwszej, albo na drugiej loterii,
- c) niewygrania na żadnej loterii.

20. W jednej urnie znajdują się jedna kula biała i trzy czarne, w drugiej natomiast znajdują się trzy kule białe i sześć czarnych. Sięgamy na chybił trafili do jednej z urn i wyciągamy kulę. Jakie jest prawdopodobieństwo, że wyciągnięta kula będzie koloru białego?

21. Z talii kart, liczącej 52 karty, wyciągamy jedną kartę i nie oglądając jej wkładamy do drugiej talii, liczącej też 52 karty. Następnie po przetasowaniu drugiej talii ciągnemy z niej jedną kartę. Jakie jest prawdopodobieństwo, że wyciągnięta karta będzie asem?

22. Podać wzór na prawdopodobieństwo całkowite.

23. Podać wzór Bayesa.

24. Wyjaśnić, czym różni się ujęcie zagadnienia we wzorze na prawdopodobieństwo całkowite i we wzorze Bayesa.

25. Co to są modele zdarzeń losowych i na czym polegają korzyści posługiwania się nimi?

3.1. ZMIENNE LOSOWE

3.1.1. Pojęcia ogólne

Przystępujemy do zdefiniowania jednego z podstawowych pojęć rachunku prawdopodobieństwa, a mianowicie pojęcia zmiennej losowej. Zanim podamy formalną definicję zmiennej losowej, przytoczymy tu najpierw rozpowszechnione w literaturze przedmiotu następujące intuicyjno-poglądowe określenie tego pojęcia: zmienną losową nazywa się taka wielkość, która w wyniku doświadczenia przyjmuje określoną wartość, znaną po zrealizowaniu doświadczenia, a nie dającą się przewidzieć przed realizacją doświadczenia. Inaczej mówiąc, zmienna losowa jest to taka zmienna, która w wyniku doświadczenia przybiera jedną i tylko jedną wartość ze zbioru tych wszystkich wartości, jakie ta zmiana może przyjąć.

Zmienne losowe oznaczamy na ogólnie wielkimi końcowymi literami alfabetu: X , Y , ... Wartości, jakie te zmienne przybierają, nazywać będziemy *realizacjami zmiennych losowych* lub krótko *realizacjami* i oznaczać odpowiednio małymi literami x , y , ... Małymi literami będziemy też oznaczać wielkości zmienne, ale nie losowe.

Rozpatrzmy parę przykładów zmiennych losowych.

PRZYKŁAD 1. Rzucamy jeden raz monetą. W wyniku realizacji doświadczenia możemy otrzymać dwa zdarzenia: E_1 – wyrzucenie orła, E_2 – wyrzucenie reszki. Przyporządkujemy zdarzeniu E_1 liczbę 0, a zdarzeniu E_2 liczbę 1. Liczby 0 i 1 są realizacjami zmiennej losowej X , określonej na zbiorze zdarzeń E_1 i E_2 .

PRZYKŁAD 2. Z talii kart ciągnimy jedną kartę. Jeśli talia liczy 52 karty, to zbiór zdarzeń możliwych (przestrzeń zdarzeń) zawiera 52 zdarzenia. Przypuśćmy, że zdarzenia należące do tego zbioru zostały ponumerowane liczbami naturalnymi od 1 do 52. W takim razie zmienna losowa X może przybierać 52 wartości: $x_1 = 1, x_2 = 2, \dots, x_{52} = 52$.

Z wartościami zmiennej losowej związane są określone prawdopodobieństwa, mówi się więc niekiedy, że zmienna losowa jest to taka zmienna, która przybiera różne wartości z różnym prawdopodobieństwem. Odpowiednikiem zmiennych losowych są w statystyce cechy statystyczne.

Rozpatruje się dwa rodzaje zmiennych losowych:

1. zmienne skokowe, czyli dyskretne,
2. zmienne ciągłe.

OKREŚLENIE 1. *Zmiennymi losowymi skokowymi lub dyskretnymi* nazywamy takie zmienne losowe, które mają skończony lub przeliczalny zbiór wartości.

Z tego wynika, że zmienne losowe skokowe mogą przybierać tylko niektóre wartości liczbowe (najczęściej wartości liczb naturalnych). Zmienną losową typu skokowego jest np. dobowa liczba zgonów, urodzeń, małżeństw w Polsce, liczba koni w poszczególnych gospodarstwach na terenie Dolnego Śląska albo wydajność pracy robotnika mierzona w sztukach wyrobów na godzinę.

OKREŚLENIE 2. *Zmiennymi losowymi ciągłymi* nazywamy takie zmienne losowe, które mogą przybierać dowolne wartości liczbowe z pewnego przedziału liczbowego (przy czym w szczególności może to być przedział nieskończony)⁽¹⁾.

Zbiór wartości, jakie mogą przybierać zmienne losowe ciągłe, jest więc mocy continuum. Zmienną losową typu ciągłego jest np. wzrost, waga, wiek poszczególnych osób, grubość arkusza blachy, wytrzymałość belki stalowej na zginanie, opór przewodnika elektrycznego itd.

Po tych uwagach wstępnych podamy obecnie formalne określenie zmiennej losowej:

OKREŚLENIE 3. *Zmienną losową X* nazywa się funkcję $X=X(e)$, określoną na zbiorze zdarzeń elementarnych E , taką że dla każdej liczby rzeczywistej x zbiór A zdarzeń elementarnych $e \in E$, dla których $X(e) < x$, spełnia warunek $A \in \mathcal{B}$.

Wynika stąd, że

$$P(A) = P(-\infty < X < x).$$

Oznacza to, że każdej wartości x może być przyporządkowane odpowiednie prawdopodobieństwo, czyli że prawdopodobieństwo to jest funkcją x . Funkcja ta nazywa się *rozkładem prawdopodobieństwa zmiennej losowej X* .

Spośród zmiennych losowych, z którymi często spotykamy się w praktyce, należy specjalnie wymienić dwie grupy zmiennych:

1. wyniki pomiarów,
2. wartości cech jednostek statystycznych, wylosowanych z populacji generalnej.

Działalności produkcyjnej człowieka towarzyszy na każdym kroku czynność mierzenia, a wyniki pomiarów są zmiennymi losowymi. Podobnie na każdym kroku zmierzni jesteśmy wypowiadać sądy nie na podstawie pełnego zbioru informacji, a jedynie na podstawie części tego zbioru. Słuszność takich sądów jest zdarzeniem losowym. Wnioskowanie o całości na podstawie części wynika ze specjalnych okoliczności, uniemożliwiających lub utrudniających zbadanie tej całości. Wypowiadaniem sądów o populacji generalnej na podstawie znajomości stosunków panujących w próbce wylosowanej z tej populacji zajmuje się *metoda reprezentacyjna* [35], stanowiąca dział statystyki matematycznej. Metoda reprezentacyjna ma szereg ważnych zastosowań w przemyśle, w handlu, rolnictwie i innych działach wytwórczości. Metodą reprezentacyjną posługuje się statystyczna kontrola jakości produkcji, statystyczny odbiór towarów. Za pomocą tej metody dokonuje się szacowania plonów, bada się budżety rodzinne, przeprowadza się analizę rynku itp. (patrz 6.2). Podstawowym aparatem badawczym metody reprezentacyjnej jest rachunek prawdopodobieństwa, który podaje ujęte w postaci twierdzeń formalne prawidła działań na zmiennych losowych.

(¹) Później poznamy bardziej ścisłe określenie zmiennej losowej ciągłej (patrz 3.4).

3.2. ROZKŁAD I DYSTRYBUANTA ZMIENNEJ LOSOWEJ SKOKOWEJ

Przypuśćmy, że zmienna losowa X jest zmienną typu skokowego, która może przybierać wartości x_1, x_2, \dots odpowiednio z prawdopodobieństwem p_1, p_2, \dots . Każdej realizacji zmiennej losowej X przyporządkowane jest więc pewne prawdopodobieństwo. Te prawdopodobieństwa można traktować jako funkcję określoną na zbiorze wartości, jakie może przybierać zmienna losowa X .

OKREŚLENIE 1. *Rozkładem skokowej zmiennej losowej X nazywa się prawdopodobieństwo tego, że zmienna X przybierze wartość x_i ($i = 1, 2, \dots$).*

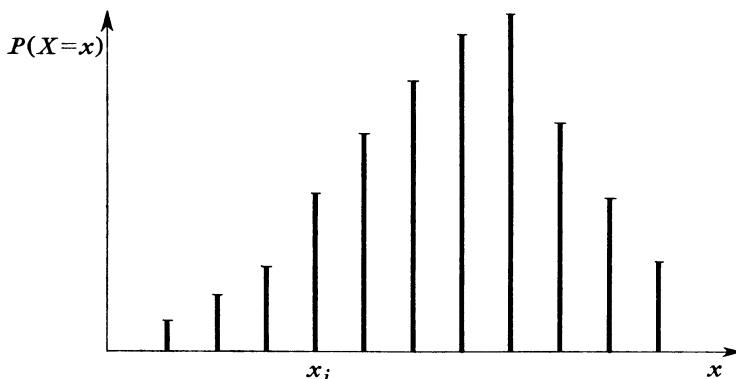
Zapisujemy to krótko w sposób następujący:

$$P(X=x_i) = p_i.$$

Oczywiście

$$\sum_i^{\infty} p_i = 1.$$

Niech x będzie pewną liczbą rzeczywistą, a X zmienną losową typu skokowego. Prawdopodobieństwo tego, że zmienna losowa X przybierze wartość mniejszą od x , jest funkcją x . Oznaczmy ją symbolem $F(x)$.



Rys. 1

OKREŚLENIE 2. Funkcja $F(x) = P(X < x)$ nazywa się *dystrybuantą zmiennej losowej X* . Z określenia dystrybuanty wynika, że

$$0 \leq F(x) \leq 1.$$

Łatwo wykazać związek istniejący między rozkładem i dystrybuantą. Przypuśćmy, że wartości zmiennej losowej X zostały uszeregowane w porządku rosnącym, tzn.

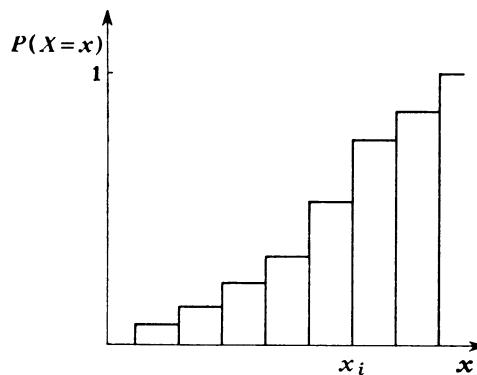
$$x_1 \leq x_2 \leq \dots \leq x_i \leq \dots \leq x_{n-1} \leq x_n.$$

Niech $x_i < x \leq x_{i+1}$, w takim razie

$$F(x) = P(X < x) = p_1 + p_2 + \dots + p_i.$$

W celu lepszego wyjaśnienia pojęcia rozkładu i dystrybuanty posłużymy się wykresami.

Na rysunku 1 przedstawiono rozkład prawdopodobieństw. Wartości zmiennej losowej odłożone są na osi x . Prawdopodobieństwa odpowiadające poszczególnym wartośćom zmiennej przedstawione są za pomocą odcinków prostopadłych do osi odciętych. Wykres przedstawiony na rysunku 1 przypomina swym wyglądem dobrze znany ze statystyki histogram częstości (patrz 6.1). Nic w tym dziwnego. Analogia między rozkładem prawdopodobieństw a rozkładem częstości szeregu rozdzielczego jest bardzo duża. Wyjaśnia to statystyczna definicja prawdopodobieństwa.



Rys. 2

Na rysunku 2 przedstawiony jest wykres dystrybuanty. Widzimy, że wartości dystrybuanty otrzymuje się przez kumulowanie wartości rozkładu. Odpowiednikiem dystrybuanty w statystyce jest częstość skumulowana.

Uwaga. Skala osi rzędnych na obu wykresach jest różna.

Wysokość ostatniego odcinka wykresu dystrybuanty równa się łącznej wysokości wszystkich odcinków wykresu funkcji rozkładu.

Jeżeli $x \geq x_n$, to

$$P(X \leq x) = p_1 + p_2 + \dots + p_n = 1.$$

Wynika stąd, że wysokość ostatniego odcinka równa się jedności.

Rozkład i dystrybuanta są to dwie najważniejsze charakterystyki zmiennej losowej. Oczywiście, gdy znamy rozkład, zawsze można znaleźć dystrybuantę, i podobnie, gdy znamy dystrybuantę – możemy odtworzyć rozkład. Jeżeli znamy któryś z tych dwóch charakterystyk zmiennej losowej, to mamy pełną informację „o sposobie zachowania się” tej zmiennej; wiemy bowiem, jakie wartości ona przybiera i jakie prawdopodobieństwo odpowiada tym wartośćom.

W zastosowaniach praktycznych wykres rozkładu zmiennej losowej zastępuje histogram częstości badanej cechy statystycznej.

Podamy kilka przykładów ilustrujących znaczenie, jakie w zastosowaniach praktycznych odgrywa badanie rozkładu zmiennych losowych (cech statystycznych).

PRZYKŁAD 1. Sklepy obuwnicze mają obowiązek opracowywania co pewien okres czasu (np. raz w roku) planu zaopatrzenia, w oparciu o który sporządza się zamówienia na poszczególne rodzaje (asortymenty) obuwia, realizowane później bądź przez krajowe wytwórnie obuwia, bądź też za pośrednictwem handlu zagranicznego. Przy sporządzaniu asortymentowego planu zaopatrzenia bierze się pod uwagę wiele czynników, np. modeł, gusty i przyzwyczajenia klientów, rozkład dochodów ludności, sezon roku, strukturę ludności według płci, wieku, a także – co jest szczególnie ważne – rozkład dwóch cech, od których w pierwszym rzędzie zależy dobranie wygodnego obuwia, a mianowicie długość i szerokość stopy. Aby poznać ten rozkład, przeprowadza się stosowne badania antropometryczne i w oparciu o uzyskany tą drogą materiał statystyczny otrzymuje się potrzebne informacje o rozkładzie badanych cech statystycznych, co z kolei umożliwia prawidłowe sporządzenie planu zaopatrzenia sieci sklepów detalicznych w obuwiu.

Podobny przykład można byłoby również podać z zakresu handlu artykułami konfekcyjnymi, a przede wszystkim odzieżą.

PRZYKŁAD 2. Jednym z podstawowych surowców używanych w budownictwie jest tzw. pospółka, czyli mieszanina piasku, żwiru oraz drobnych kamyczków. Aby ta mieszanina mogła uzyskać miano surowca przemysłowego, jej skład, tzn. struktura ziaren według ich granulacji, musi odpowiadać określonym normom. Rodzi się tu natychmiast pytanie, jak można określić normę dla mieszaniny ogromnej ilości drobnych ziarenek piasku, żwiru i małych kamyczków, jeśli mieszanina ta nie jest na ogół produktem otrzymanym sztucznie przez zmieszanie piasku, żwiru i kamieni, lecz jest otrzymywana w sposób naturalny przez eksploatację złóż pospolitej zalegających dna rzek. A jednak normę taką można ustalić i można również sprawdzić, czy materiał wydobyty z określonego miejsca złoża odpowiada normie. Czyni się to zwykle w ten sposób, że za pomocą сит sortuje się frakcję próbnej wydobytej partii pospolitej i sporządza się wykres dystrybuanty rozkładu granulacji ziaren pospolitej. Jeśli wykreślona dystrybuanta przebiega między dwoma liniami określającymi gorne i dolne dopuszczalne położenie wykresu dystrybuanty, to pospółka nadaje się do celów eksploatacji przemysłowej.

3.3. NIEKTÓRE ROZKŁADY ZMIENNEJ LOSOWEJ SKOKOWEJ

3.3.1. Rozkład zero-jedynkowy

Dana jest zmienna losowa X , która może przybierać jedynie dwie wartości: $x_1 = 1$ i $x_2 = 0$.

Jeśli

$$P(X = x_1) = p,$$

to

$$P(X = x_2) = 1 - p = q,$$

gdzie

$$P(X = x_1) + P(X = x_2) = 1.$$

Widzimy więc, że zmienna losowa X przybiera wartość 1 z prawdopodobieństwem p , a wartość 0 z prawdopodobieństwem q .

Taki rozkład zmiennej losowej nazywa się *rozkładem zero-jedynkowym*.

Przypuśćmy, że interesuje nas jakieś zdarzenie A . Zajście zdarzenia A będziemy nazywali *sukcesem*, a niezajście tego zdarzenia – *niepowodzeniem*. Jeżeli prawdopodobieństwo sukcesu równa się p , to prawdopodobieństwo niepowodzenia wynosi $1 - p$, czyli q .

Z rozkładem zero-jedynkowym mamy np. do czynienia przy jednorazowym rzucie monetą. Oznaczmy wyrzucenie orła liczbą 1, natomiast wyrzucenie reszki 0. W takim razie zmienna losowa X przybiera wartość 1 z prawdopodobieństwem $p = \frac{1}{2}$ oraz wartość 0 z prawdopodobieństwem $q = \frac{1}{2}$.

A oto inny przykład. Rzucamy jeden raz kością do gry. Interesuje nas wyrzucenie sześciu oczek. Prawdopodobieństwo sukcesu wynosi $\frac{1}{6}$, natomiast prawdopodobieństwo niepowodzenia $\frac{5}{6}$. Oznaczając sukces liczbą 1, a niepowodzenie liczbą 0 otrzymamy rozkład zero-jedynkowy.

Zmienna losowa X przybiera wartość 1 z prawdopodobieństwem $\frac{1}{6}$, 0 z prawdopodobieństwem $\frac{5}{6}$.

3.3.2. Rozkład dwumianowy

Wyprowadzenie wzoru rozkładu dwumianowego poprzedzimy przykładem. Rzucamy trzy razy monetą. Niech zmienną losową X będzie liczba wyrzuconych orłów. Zmienna losowa w naszym przykładzie może przybierać wartości 0, 1, 2, 3. Znajdziemy prawdopodobieństwa odpowiadające poszczególnym wartościom tej zmiennej losowej. W tym celu wypiszemy wszystkie możliwe sytuacje, które mogą wystąpić przy trzech rzutach monetą.

Tablica 1

Sytuacja	Liczba orłów X	Prawdopodobieństwo
$O O O$	3	$\frac{1}{8}$
$O O R \}$ $O R O \}$ $R O O \}$	2	$\frac{3}{8}$
$R R O \}$ $R O R \}$ $O R R \}$	1	$\frac{3}{8}$
$R R R$	0	$\frac{1}{8}$

Otrzymaliśmy 8 sytuacji. Jedna z nich sprzyja wyrzuceniu trzech orłów, trzy sprzyjają wyrzuceniu dwóch orłów i trzy – wyrzuceniu jednego orła, jedna sytuacja sprzyja nie-wyrzuceniu orła w ogóle (zdarzenie to jest równoważne wyrzuceniu trzech reszek).

Dzieląc liczbę sytuacji sprzyjających wystąpieniu poszczególnych wartości zmiennej losowej przez liczbę wszystkich możliwych sytuacji znajdziemy interesujące nas prawdopodobieństwo. Znaleziony rozkład nazywa się *rozkładem dwumianowym*.

Zajmiemy się obecnie wyprowadzeniem wzoru na rozkład i dystrybuantę w rozkładzie dwumianowym.

Przypuśćmy, że interesuje nas zajście zdarzenia A . Niech prawdopodobieństwo wystąpienia A przy jednorazowej realizacji doświadczenia równa się p . Przeprowadzamy n nieza-

leżnych doświadczeń. Chcemy znaleźć rozkład zmiennej losowej X , która jest liczbą sukcesów przy n -krotnej realizacji doświadczenia. Zmienna losowa X może przybierać wartości $0, 1, 2, \dots, n$.

Zbadamy, czemu równa się prawdopodobieństwo, że zmienna losowa X przybiera wartość k . Interesujące nas zdarzenie jest równoważne zdarzeniu, że wśród n wyników doświadczeń będzie k sukcesów i $n-k$ niepowodzeń. Zapiszmy to zdarzenie w sposób następujący:

$$\underbrace{AA\dots A}_{k \text{ razy}} \cdot \underbrace{\bar{A}\bar{A}\dots\bar{A}}_{n-k \text{ razy}}.$$

W takim razie prawdopodobieństwo tego zdarzenia równa się

$$\underbrace{pp\dots p}_{k \text{ razy}} \cdot \underbrace{qq\dots q}_{n-k \text{ razy}} = p^k q^{n-k}.$$

Otrzymany wynik byłby rozwiązaniem interesującego nas zagadnienia, gdybyśmy szukali prawdopodobieństwa wystąpienia k sukcesów w pewnym określonym porządku. Tak jednak nie jest. Ponumerujmy poszczególne doświadczenia liczbami od 1 do n . Numer tych doświadczeń, w których pojawiło się zdarzenie A , nazwijmy numerami wyróżnionymi. Mamy odpowiedzieć na pytanie, ile k -elementowych zbiorów takich liczb wyróżnionych można utworzyć ze zbioru zawierającego n liczb.

Odpowiedź jest prosta. Daje ją wzór na ilość kombinacji z n po k elementów, czyli C_n^k (patrz 1.1.4, wzór (1)). Wobec tego

$$(1) \quad P(X=k) = C_n^k p^k q^{n-k}.$$

Jest to wzór rozkładu dwumianowego.

A oto wzór dystrybuanty w tym rozkładzie

$$(2) \quad P(X < x) = \sum_{k < x} C_n^k p^k q^{n-k},$$

przy czym sumowanie rozciąga się na wszystkie wartości zmiennej losowej, które są mniejsze od x . Prawa strona wzoru (1) jest składnikiem rozwinięcia dwumianu Newtona (patrz 1.2, wzór (6)):

$$(p+q)^n.$$

Tłumaczy to pochodzenie nazwy „rozkład dwumianowy”.

Łatwo udowodnić, że

$$\sum_{k=0}^n P(X=k) = 1.$$

Mamy bowiem

$$\sum_{k=0}^n P(X=k) = \sum_{k=0}^n C_n^k p^k q^{n-k} = (p+q)^n = 1,$$

gdyż $p+q=1$.

W celu poglądowego objaśnienia sposobu korzystania ze wzoru (1) zajmiemy się obecnie obliczaniem prawdopodobieństw odpowiadających poszczególnym wartościom zmiennej losowej, którą jest liczba białych kul wyciągniętych z urny, jeśli ciągnimy kolejno 5 kul i każdorazowo, po obejrzeniu barwy, wkładamy kule z powrotem do urny oraz jeśli frakcja białych kul w urnie wynosi $\frac{1}{3}$.

Wprowadzimy oznaczenie literowe: $n=5$, $p=\frac{1}{3}$, $q=\frac{2}{3}$. Zmienna losowa X może przybierać wartości 0, 1, 2, 3, 4, 5.

Wartości zmiennej losowej i obliczone za pomocą wzoru dwumianowego odpowiadające im prawdopodobieństwa podaje tablica 2.

Tablica 2

k	$C_n^k p^k q^{n-k} = P(X=k)$	$\sum_{i=0}^k C_n^i p^i q^{n-i} = P(X \leq k)$
1	2	3
0	$C_5^0 \left(\frac{1}{3}\right)^0 \left(\frac{2}{3}\right)^5 = \frac{32}{243}$	$\sum_{i=0}^0 C_5^i \left(\frac{1}{3}\right)^i \left(\frac{2}{3}\right)^{5-i} = \frac{32}{243}$
1	$C_5^1 \left(\frac{1}{3}\right)^1 \left(\frac{2}{3}\right)^4 = \frac{80}{243}$	$\sum_{i=0}^1 C_5^i \left(\frac{1}{3}\right)^i \left(\frac{2}{3}\right)^{5-i} = \frac{112}{243}$
2	$C_5^2 \left(\frac{1}{3}\right)^2 \left(\frac{2}{3}\right)^3 = \frac{80}{243}$	$\sum_{i=0}^2 C_5^i \left(\frac{1}{3}\right)^i \left(\frac{2}{3}\right)^{5-i} = \frac{192}{243}$
3	$C_5^3 \left(\frac{1}{3}\right)^3 \left(\frac{2}{3}\right)^2 = \frac{40}{243}$	$\sum_{i=0}^3 C_5^i \left(\frac{1}{3}\right)^i \left(\frac{2}{3}\right)^{5-i} = \frac{232}{243}$
4	$C_5^4 \left(\frac{1}{3}\right)^4 \left(\frac{2}{3}\right)^1 = \frac{10}{243}$	$\sum_{i=0}^4 C_5^i \left(\frac{1}{3}\right)^i \left(\frac{2}{3}\right)^{5-i} = \frac{242}{243}$
5	$C_5^5 \left(\frac{1}{3}\right)^5 \left(\frac{2}{3}\right)^0 = \frac{1}{243}$	$\sum_{i=0}^5 C_5^i \left(\frac{1}{3}\right)^i \left(\frac{2}{3}\right)^{5-i} = \frac{243}{243} = 1$
Razem		1

Jak to wynika z nagłówków, kolumna 2 tablicy podaje wartości rozkładu, natomiast kolumna 3 zawiera wartości dystrybuanty. Wykres rozkładu przedstawiony jest na rysunku 1, a wykres dystrybuanty – na rysunku 2. Osie odciętych na obu wykresach mają jednakowe skale, osie rzędnych mają różne skale.

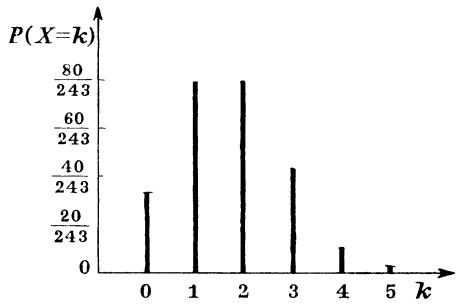
Przedstawiony na rysunku 1 rozkład jest asymetryczny. Tłumaczy się to tym, że $p \neq q$. Gdyby $p=q=\frac{1}{2}$, to wykres byłby symetryczny.

Rozkład dwumianowy jest jednym z najważniejszych rozkładów teoretycznych. Oto główne cechy tego rozkładu:

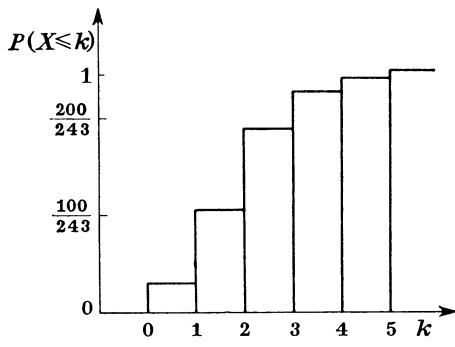
1. Rozkład dwumianowy jest rozkładem sumy n nieszczelnych zmiennych losowych o rozkładzie zero-jedynkowym.
2. Zmienna losowa w rozkładzie dwumianowym jest zmienną skokową, przybierającą wartości liczb całkowitych nieujemnych.

3. Jeśli $p = q$, to rozkład jest symetryczny, natomiast jeśli $p \neq q$, to rozkład jest asymetryczny.

W statystyce mamy do czynienia z rozkładem dwumianowym przy losowaniu zwracającym elementów z populacji ograniczonej lub przy losowaniu bezzwracającym z populacji nieograniczonej, jeśli wynik pojedynczego losowania jest zmienną losową o rozkładzie zero-jedynkowym.



Rys. 1



Rys. 2

Zmienna losowa posiada rozkład dwumianowy, gdy czyni ona zadość warunkom doświadczenia, zwanego *schematem Bernoulliego*.

OKREŚLENIE 1. Jeśli w wyniku realizacji pewnego doświadczenia może pojawić się zdarzenie A , które nazywać będziemy sukcesem, to liczba sukcesów, jaka może wystąpić przy n -krotnej realizacji doświadczeń, jest zmienną losową czyniącą zadość *schematowi Bernoulliego*, gdy wyniki poszczególnych doświadczeń są zdarzeniami niezależnymi oraz gdy $P(A)$ jest stałe i równa się p .

3.3.3. Rozkład hipergeometryczny

Jak wiadomo (patrz 2.4.2), wyniki losowania w losowaniu ze zwracaniem są zdarzeniami niezależnymi, natomiast przy losowaniu bez zwracania – zdarzeniami zależnymi.

Jeśli z urny zawierającej pewną ilość kul białych i czarnych ciągnimy n razy kulę, wkładając ją każdorazowo po obejrzeniu barwy z powrotem do urny, to przez cały czas losowania skład urny nie ulega zmianie. Aby znaleźć prawdopodobieństwo wyciągnięcia k kul białych w n kolejnych ciągnieniach, korzystamy ze wzoru dwumianowego.

Inaczej się sprawia przedstawia, gdy po wyciągnięciu kuli nie zwracamy jej z powrotem do urny. Po każdym losowaniu skład urny zmienia się. Chcąc obliczyć prawdopodobieństwo wyciągnięcia k kul białych w n losowaniach nie możemy korzystać ze wzoru dwumianowego, gdyż warunki schematu Bernoulliego nie są spełnione (wyniki losowań są zdarzeniami zależnymi).

Rozpatrzmy to zagadnienie w postaci ogólnej. Z populacji generalnej liczącej N elementów pobrano próbki. Niech próbka liczy n elementów, gdzie $n \leq N$. Przy pobieraniu elementów do próbki korzystano ze schematu losowania bez zwracania, to znaczy wyloso-

wane elementy nie powracały do populacji. Wiadomo, że w populacji generalnej R elementów ma cechę A (gdzie $R \leq N$), a $N - R$ elementów ma cechę B . Należy znaleźć prawdopodobieństwo, że w próbce znajdzie się k elementów mających cechę A (oczywiście k nie może być większe od mniejszej z dwóch liczb n i R).

Przy rozwiązywaniu postawionego zadania skorzystamy z klasycznej definicji prawdopodobieństwa i wzorów kombinatoryki⁽¹⁾. Ogólna ilość sposobów wyciągnięcia n elementów ze zbioru liczącego N elementów równa się C_N^n . W populacji znajduje się R elementów mających cechę A . Nas interesuje wyciągnięcie k takich elementów. Wiemy, że ze zbioru liczącego R elementów można utworzyć tyle różnych podzbiorów k -elementowych, ile wynosi C_R^k . Ponieważ próbka liczy n elementów, więc jeśli k z nich ma mieć cechę A , to $n-k$ musi mieć cechę B . W populacji generalnej znajduje się $N-R$ elementów mających cechę B . W takim razie liczba sposobów wylosowania $n-k$ elementów ze zbioru liczącego $N-R$ elementów równa się C_{N-R}^{n-k} .

Oznaczmy symbolem $P(X=k)$ prawdopodobieństwo tego, że wśród n elementów próbki, pobranych w drodze losowania bez zwracania z populacji generalnej liczącej N elementów, znajdzie się k elementów mających cechę A , jeśli wiadomo, że populacja generalna posiada R takich elementów. W takim razie

$$(1) \quad P(X=k) = \frac{C_R^k C_{N-R}^{n-k}}{C_N^n}.$$

W mianowniku prawej strony tego wzoru znajduje się ilość wszystkich możliwych zdarzeń przy losowaniu n elementów z populacji liczącej N elementów. W liczniku podana jest liczba zdarzeń sprzyjających wyciągnięciu k i tylko k elementów mających cechę A .

Rozkład prawdopodobieństw, określony wzorem (1), nazywa się *rozkładem hipergeometrycznym*.

Jeśli licznik i mianownik wzoru (1) podzielimy przez $\frac{R!(N-R)!}{n!(N-n)!}$, to po łatwych przekształceniach otrzymamy

$$(2) \quad P(X=k) = \frac{C_n^k C_{N-n}^{R-k}}{C_N^R}.$$

PRZYKŁAD 1. Korzystając ze wzoru (1) obliczyć rozkład prawdopodobieństw wygranej w popularnej grze liczbowej Toto-Lotek. Jak wiadomo, gra liczbową polega na trafnym skreśleniu sześciu spośród 49 liczb 1, 2, ..., 49 umieszczonych na karcie do gry.

Oznaczmy symbolami P_0, P_1, \dots, P_6 prawdopodobieństwa trafnego skreślenia odpowiednio: żadnej liczby, jednej, dwóch, ... i wreszcie sześciu liczb.

Mamy

$$P_0 = \frac{C_6^0 C_{43}^6}{C_{49}^6} = 0,43596498,$$

$$P_1 = \frac{C_6^1 C_{43}^5}{C_{49}^6} = 0,41301946,$$

(1) Z zadaniami tego typu spotkaliśmy się już w 2.4.3.

$$P_2 = \frac{C_6^2 C_{43}^4}{C_{49}^6} = 0,13237803 ,$$

$$P_3 = \frac{C_6^3 C_{43}^3}{C_{49}^6} = 0,01765040 ,$$

$$P_4 = \frac{C_6^4 C_{43}^2}{C_{49}^6} = 0,00096860 ,$$

$$P_5 = \frac{C_6^5 C_{43}^1}{C_{49}^6} = 0,00001845 ,$$

$$P_6 = \frac{C_6^6 C_{43}^0}{C_{49}^6} = 0,00000008 .$$

Jak wynika z obliczeń, grając w Toto-Lotka 100 razy, musimy się spodziewać, że nic nie wygramy w 98 przypadkach; wygranej mamy prawo oczekwać tylko w dwóch przypadkach.

Dla należytego poznania rozkładu hipergeometrycznego rozpatrzmy jeszcze następujący przykład. Z talii liczącej 52 karty ciągnimy bez zwracania 10 kart. Mamy obliczyć prawdopodobieństwo, że wśród wylosowanych kart będą 0, 1, 2, 3, 4 asy.

Posługując się wzorem (1) lub (2) znajdujemy, że

$$P(X=0) = 0,41344 ,$$

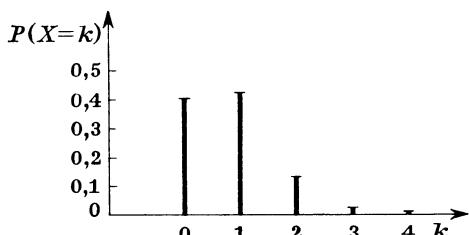
$$P(X=1) = 0,42405 ,$$

$$P(X=2) = 0,14312 ,$$

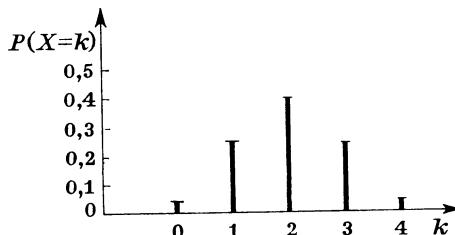
$$P(X=3) = 0,01862 ,$$

$$P(X=4) = 0,00078 .$$

Obliczone na podstawie warunków zadania wartości rozkładu hipergeometrycznego przedstawione są graficznie na rysunku 1. Widzimy, że otrzymany rozkład jest asyme-



Rys. 1



Rys. 2

tryczny. Asymetria ta maleje, gdy liczебność próbki rośnie do $N/2$, i rośnie, gdy liczебność próbki zbliża się do liczебności populacji.

Oto rozkład prawdopodobieństw wyciągnięcia k asów ($k=0, \dots, 4$) z talii zawierającej 52 karty, jeśli próbka liczy 26 kart:

$$P(X=0)=0,05522,$$

$$P(X=1)=0,24970,$$

$$P(X=2)=0,39016,$$

$$P(X=3)=0,24970,$$

$$P(X=4)=0,05522.$$

Rozkład ten, przedstawiony na rysunku 2, jest symetryczny.

Porównując wzrokowo wykresy rozkładu dwumianowego i hipergeometrycznego dostrzegamy duże podobieństwo między tymi rozkładami.

Wzór (1) został wyprowadzony dla przypadku, gdy populacja generalna zawiera dwie grupy elementów: elementy o cesze A i elementy o cesze \bar{A} , którą oznaczyliśmy symbolem B .

Oczywiście wzór ten można z łatwością rozszerzyć na populacje o większej ilości grup elementów. Wyjaśnimy to na przykładzie. W urnie znajduje się 5 kul białych, 2 kule czerwone i 3 kule zielone. Z urny tej losujemy bez zwracania 4 kule. Należy obliczyć prawdopodobieństwo, że wyciągniemy 2 kule białe, 1 czerwoną i 1 zieloną.

Oznaczmy szukane prawdopodobieństwo symbolem $P(2, 1, 1)$. W takim razie

$$P(2, 1, 1) = \frac{C_5^2 C_2^1 C_3^1}{C_{10}^4} = \frac{2}{7}.$$

Na zakończenie uwagi o rozkładzie hipergeometrycznym warto jeszcze raz podkreślić, że z rozkładem tym mamy do czynienia przy losowaniu bez zwracania, to znaczy gdy wyniki losowania są zdarzeniami zależnymi. Gdy natomiast korzystamy z losowania ze zwracaniem, czyli gdy wyniki losowania są zdarzeniami niezależnymi, to mamy do czynienia z rozkładem dwumianowym.

3.3.4. Rozkład Poissona

Dana jest zmienna losowa X o rozkładzie dwumianowym, tzn.

$$P(X=k) = C_n^k p^k q^{n-k}.$$

Załóżmy, że przy $n \rightarrow \infty$ p zmienia się w ten sposób, że $np = m$, gdzie $m > 0$ jest pewną stałą. W takim razie $p = m/n$. Ponieważ

$$P(X=k) = C_n^k p^k (1-p)^{n-k} = \frac{n(n-1)\dots(n-k+1)}{k!} p^k (1-p)^{n-k}$$

(patrz 1.1.1, upraszczanie silni), przeto podstawiając zamiast p ułamek m/n otrzymamy

$$\begin{aligned} P(X=k) &= \frac{n(n-1)\dots(n-k+1)}{k!} \left(\frac{m}{n}\right)^k \left(1 - \frac{m}{n}\right)^{n-k} = \\ &= \frac{n}{n} \cdot \frac{n-1}{n} \cdot \dots \cdot \frac{n-k+1}{n} \left(1 - \frac{m}{n}\right)^{-k} \frac{m^k}{k!} \left(1 - \frac{m}{n}\right)^n. \end{aligned}$$

Gdy $n \rightarrow \infty$, to

$$(1) \quad \lim_{n \rightarrow \infty} P(X=k) = \frac{m^k}{k!} \lim_{n \rightarrow \infty} \left(1 - \frac{m}{n}\right)^n,$$

gdyż

$$\lim_{n \rightarrow \infty} \frac{n}{n} \cdot \frac{n-1}{n} \cdot \dots \cdot \frac{n-k+1}{n} \left(1 - \frac{m}{n}\right)^{-k} = 1.$$

Podstawiając $-n/m = z$, $n = -mz$ otrzymujemy

$$\lim_{n \rightarrow \infty} \left(1 - \frac{m}{n}\right)^n = \lim_{z \rightarrow -\infty} \left(1 + \frac{1}{z}\right)^{-mz} = \lim_{z \rightarrow -\infty} \left[\left(1 + \frac{1}{z}\right)^z\right]^{-m}.$$

Z analizy wiadomo (patrz [19], str. 45), że

$$\lim_{z \rightarrow -\infty} \left(1 + \frac{1}{z}\right)^z = e.$$

Stąd

$$(2) \quad \lim_{z \rightarrow -\infty} \left[\left(1 + \frac{1}{z}\right)^z\right]^{-m} = e^{-m}.$$

Ostatecznie, na mocy wzorów (1) i (2),

$$(3) \quad P(X=k) = \frac{m^k}{k!} e^{-m}.$$

Rozkład określony wzorem (3) nazywa się *rozkładem Poissona*. Rozkład ten daje dobre przybliżenie wartości rozkładu dwumianowego, gdy n jest dostatecznie duże, a p – małe.

W tablicy 1 (por. [4], str. 104) podano przykładowo wartości rozkładu dwumianowego i rozkładu Poissona dla $k=0, 1, \dots, 11$ przy założeniu, że $p=0,02$, a $n=100$.

Tablica 1

k	Rozkład dwumianowy $P(X=k) = C_n^k p^k (1-p)^{n-k}$	Rozkład Poissona $P(X=k) = \frac{(np)^k}{k!} e^{-np}$
1	2	3
0	0,13262	0,13533
1	0,27065	0,27067
2	0,27342	0,27067
3	0,18228	0,18044
4	0,09021	0,09022
5	0,03535	0,03609
6	0,01142	0,01203
7	0,00313	0,00343
8	0,00073	0,00085
9	0,00015	0,00019
10	0,00003	0,000038
11	0,00001	0,000007

Porównując liczby kolumn drugiej i trzeciej widzimy, że rozkład Poissona daje bardzo dobre przybliżenie rozkładu dwumianowego. To przybliżenie jest tym lepsze, im p jest mniejsze, a n większe. Właśnie ze względu na to, że rozkład Poissona stosuje się wtedy, gdy p jest małe, znany jest w literaturze także pod nazwą *rozkładu rzadkich zdarzeń* lub *prawa małych liczb*. Rozkład Poissona daje na ogół dostatecznie dobre przybliżenie rozkładu dwumianowego, gdy $p < 0,2$ i $n > 20$.

Rozkład Poissona ma bardzo szerokie zastosowanie w statystycznym odbiorze towarów i w statystycznej kontroli jakości produkcji. Ma on również inne ważne zastosowania. W wielu dziedzinach nauki spotyka się zjawiska o tym rozkładzie. Szczególnie licznych przykładów takich zjawisk dostarcza demografia, biologia, fizyka, mechanika, astronomia.

Dla uniknięcia żmudnych rachunków związanych z korzystaniem ze wzoru na rozkład Poissona opracowane zostały specjalne tablice. Tablice te podane są w końcu niniejszej książki.

Przystąpimy obecnie do rozpatrzenia kilku przykładów zastosowań rozkładu Poissona, zaczerpniętych z praktyki statystycznego odbioru towarów.

PRZYKŁAD 1. Jeżeli wiadomo, że wadliwość towaru (tzn. przeciętny procent braków) wynosi 2%, to jakie jest prawdopodobieństwo, że w partii towaru liczącej 100 sztuk znajdzie się nie więcej niż 3 sztuki złe?

W zadaniu tym $p = 0,02$, $n = 100$, $np = 2$, $k = 3$. Należy obliczyć

$$P(X < 4) = P(X=0) + P(X=1) + P(X=2) + P(X=3).$$

Obliczenie tego prawdopodobieństwa za pomocą wzoru (3) nastręczałoby wiele kłopotu. Znacznie wygodniej jest posłużyć się tablicą V, która podaje wartości

$$P(X < k) = \sum_{r=0}^k \frac{(np)^r e^{-np}}{r!}.$$

Dla $np = 2$ i $k = 3$ znajdujemy w tej tablicy liczbę 0,857. Jest to właśnie szukane prawdopodobieństwo.

PRZYKŁAD 2. Wadliwość towaru wynosi 2%. Ile dobrych sztuk towaru należy dodać do partii towaru liczącej 100 sztuk, aby z prawdopodobieństwem nie mniejszym niż 0,95 uniknąć reklamacji (to znaczy, aby z prawdopodobieństwem nie mniejszym niż 0,95 partia zawierała 100 sztuk dobrych)?

W zadaniu tym znamy $p = 0,02$, $n = 100$, $P(X < k) = 0,95$, $np = 2$. Szukamy k .

W tablicy V w wierszu zaopatrzonym nagłówkiem 2 szukamy liczby większej od 0,95. Najmniejszą spośród liczb większych od 0,95 jest liczba 0,983. Liczba ta znajduje się w kolumnie, w której nagłówkiem jest liczba 5. Wobec tego dla uniknięcia reklamacji z prawdopodobieństwem większym niż 0,95 należy do partii towaru dodać 5 sztuk dobrych.

PRZYKŁAD 3. Wadliwość towaru jest nieznana. Czy można z prawdopodobieństwem większym niż 0,90 twierdzić, że wadliwość ta jest mniejsza niż 0,02, jeśli w partii liczącej 100 sztuk znalazły się 7 sztuk złych?

Przy założeniu, że $p = 0,02$, mamy

$$P(X > 0) = 0,865 < 0,90.$$

Oznacza to, że gdybyśmy nawet w naszej partii towaru nie znaleźli ani jednego braku, to i tak nie moglibyśmy twierdzić z prawdopodobieństwem większym niż 0,90, że wadliwość towaru jest mniejsza niż 0,02. Ponieważ w partii towaru znalazły się 7 sztuk złych, a przy wadliwości 0,02 i wielkości partii wynoszącej 100 sztuk prawdopodobieństwo tego, że liczba braków będzie większa niż 6, wynosi 0,005, więc uznając hipotezę, że $p < 0,02$, za prawdziwą, musielibyśmy przyjąć, że zaszło zdarzenie mało prawdopodobne, polegające na trafieniu do partii towaru o wadliwości mniejszej niż 0,02 aż 7 sztuk złych. Stąd płynie wniosek, że można sądzić, iż wadliwość towaru jest większa niż 0,02.

Trzy przytoczone wyżej przykłady są fragmentaryczną ilustracją zagadnień z zakresu statystycznego odbioru towarów i sposobu rozwiązywania tych zagadnień za pomocą rozkładu Poissona.

3.4. DYSTRYBUANTA ZMIENNEJ LOSOWEJ CIĄGŁEJ GĘSTOŚĆ PRAWDOPODOBIEŃSTWA

Niech X będzie zmienną losową typu ciągłego. Analogicznie do określenia 2 w § 3.2 *dystrybuantą* tej zmiennej nazywać będziemy funkcję $F(x)$, gdzie

$$(1) \quad F(x) = P(X < x).$$

$F(x)$ jest prawdopodobieństwem, więc

$$(2) \quad 0 \leq F(x) \leq 1.$$

Przypuśćmy, że x_1 i x_2 są to dwie dowolne liczby rzeczywiste. Założymy, że $x_2 > x_1$. Znajdziemy $P(x_1 \leq X < x_2)$. Ponieważ zdarzenie polegające na tym, że $X < x_2$, rozkłada się (patrz 2.2.2, określenie 10) na dwa zdarzenia

$$X < x_1 \quad \text{oraz} \quad x_1 \leq X < x_2,$$

przeto

$$P(X < x_2) = P(X < x_1) + P(x_1 \leq X < x_2).$$

Stąd

$$(3) \quad P(x_1 \leq X < x_2) = P(X < x_2) - P(X < x_1).$$

Posługując się symbolem dystrybuanty równość powyższą można zapisać inaczej:

$$(4) \quad P(x_1 \leq X < x_2) = F(x_2) - F(x_1).$$

Wzór (4) pozwoli nam udowodnić ważne

TWIERDZENIE 1. *Jeśli X jest zmienną losową ciągłą, to*

$$P(X = x_0) = 0,$$

gdzie x_0 jest dowolną stałą.

Dowód. Na mocy wzoru (4)

$$P(x_0 \leq X < x_0 + \Delta x) = F(x_0 + \Delta x) - F(x_0).$$

Gdy $\Delta x \rightarrow 0^+$, to ponieważ dystrybuanta jest funkcją przynajmniej prawostronnie ciągłą (patrz [8], str. 43), więc różnica

$$F(x_0 + \Delta x) - F(x_0)$$

także dąży do zera. Stąd, na mocy oczywistej nierówności

$$0 \leq P(X = x_0) \leq P(x_0 \leq X < x_0 + \Delta x),$$

otrzymujemy

$$P(X=x_0)=0,$$

czego należało dowieść.

Pamiętamy, że gdy była mowa o zmiennych losowych skokowych, rozpatrywaliśmy prawdopodobieństwo $P(X=x_i)=p_i$ zwane rozkładem zmiennej losowej skokowej. Prawdopodobieństwo to traci sens w odniesieniu do zmiennych losowych ciągłych, gdyż jak wynika z twierdzenia 1,

$$P(X=x)=0.$$

Analogonem rozkładu zmiennej losowej skokowej jest tzw. *gęstość zmiennej losowej ciągłej*. Przejdziemy obecnie do zdefiniowania tego ważnego pojęcia.

Z definicji dystrybuanty wynika, że

$$P(x_1 \leq X < x_2) = F(x_2) - F(x_1).$$

Gdy $\Delta x = x_2 - x_1 \rightarrow 0$, to

$$P(x_1 \leq X < x_2) = dF(x) + r(x),$$

gdzie $r(x)$ jest nieskończenie małą rzędą wyższego niż Δx . Z określenia różniczki funkcji wynika jednak, że

$$dF(x) = F'(x) dx.$$

We wzorze tym symbolem $F'(x)$ oznaczono pochodną dystrybuanty. Mamy więc

$$(5) \quad F'(x) = \lim_{\Delta x \rightarrow 0} \frac{F(x_2) - F(x_1)}{\Delta x} = \lim_{\Delta x \rightarrow 0} \frac{P(x_1 \leq X < x_2)}{\Delta x}.$$

OKREŚLENIE 1. Jeśli dystrybuanta $F(x)$ ma pochodną w punkcie x , to pochodna ta nazywa się *gęstością prawdopodobieństwa* zmiennej losowej X w punkcie x .

Oznaczając gęstość prawdopodobieństwa symbolem $f(x)$ mamy

$$(6) \quad f(x) = F'(x) \quad \text{oraz} \quad F(x) = \int_{-\infty}^x f(t) dt.$$

Z określenia 1 wynika, że

$$(7) \quad f(x) \geq 0,$$

$$(8) \quad P(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} f(t) dt,$$

$$(9) \quad \int_{-\infty}^{\infty} f(x) dx = 1.$$

OKREŚLENIE 2. Zmienna losowa X jest *zmienną ciągłą w danym przedziale*, jeśli w tym przedziale gęstość $f(x)$ istnieje i jest funkcją ciągłą względem x w całym przedziale z wyjątkiem co najwyżej skończonej ilości punktów.

3.5. NIEKTÓRE ROZKŁADY ZMIENNEJ LOSOWEJ CIĄGLEJ

3.5.1. Rozkład prostokątny

Rozkładem prostokątnym nazywamy rozkład, którego gęstość określa następująca relacja:

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

We wzorach tych a i b są to dowolne stałe, przy czym $a < b$.

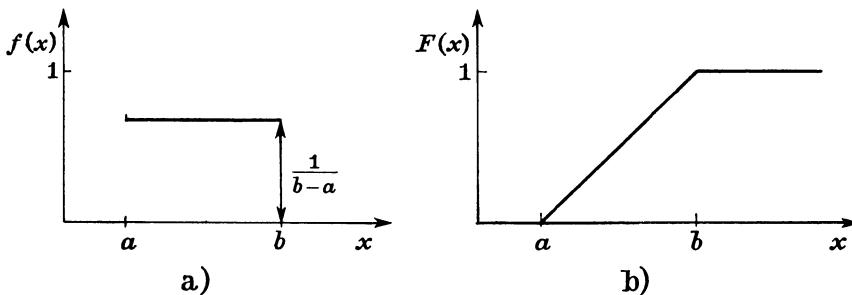
Aby znaleźć dystrybuantę rozkładu prostokątnego, wykorzystamy wzór (6) podany w poprzednim paragrafie. Ze wzoru tego wynika, że dla znalezienia dystrybuanty należy scałkować funkcję gęstości w przedziale od $-\infty$ do x , czyli

$$F(x) = \int_{-\infty}^x f(x) dx.$$

Ze względu na wzory (1) otrzymujemy

$$(2) \quad F(x) = \begin{cases} 0 & \text{dla } x < a, \\ \frac{x-a}{b-a} & \text{dla } a \leq x \leq b, \\ 1 & \text{dla } x > b. \end{cases}$$

Czytelnik sprawdzi bez trudu, że $P(a \leq X \leq b) = 1$.



Rys. 1

Wykresy funkcji gęstości i dystrybuanty przedstawione są odpowiednio na rysunkach 1a i 1b. Wykres gęstości rozkładu prostokątnego usprawiedliwia nazwę tego rozkładu.

Ze względu na przebieg krzywej gęstości w rozkładzie prostokątnym, rozkład ten znany jest także pod nazwą *rozkładu jednostajnego* lub *rozkładu jednakowych prawdopodobieństw*.

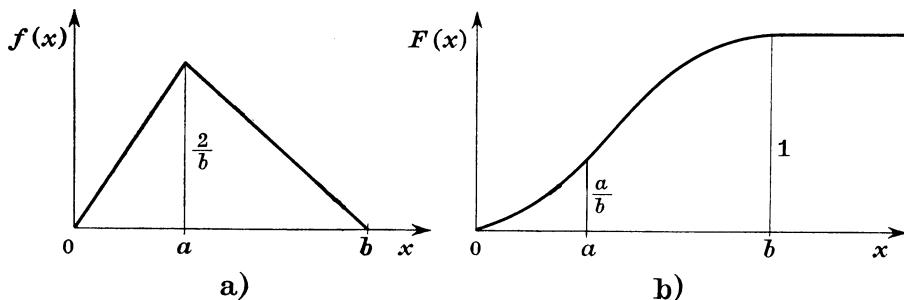
3.5.2. Rozkład trójkątny

W *rozkładzie trójkątnym* gęstość jest określona wzorami

$$(1) \quad f(x) = \begin{cases} 0 & \text{dla } x < 0, \\ \frac{2x}{ab} & \text{dla } 0 \leq x \leq a, \\ \frac{2x}{ab-b^2} + \frac{2}{b-a} & \text{dla } a \leq x \leq b, \\ 0 & \text{dla } x > b. \end{cases}$$

Na mocy wzorów (1) i wzoru (6) z § 3.4 mamy

$$(2) \quad F(x) = \begin{cases} 0 & \text{dla } x < 0, \\ \frac{x^2}{ab} & \text{dla } 0 \leq x \leq a, \\ \frac{x^2 - a^2}{ab - b^2} + \frac{2(x-a)}{b-a} + \frac{a}{b} & \text{dla } a \leq x \leq b, \\ 1 & \text{dla } x > b. \end{cases}$$



Rys. 1

Wzory powyższe staną się zupełnie oczywiste, gdy czytelnik wyprowadzając je oprze się na wykresie gęstości oraz na wykresie dystrybuanty, które przedstawione są na rysunkach 1a i 1b.

3.5.3. Rozkład normalny

W rozkładzie normalnym gęstość wyraża się wzorem

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

gdzie m i σ są to pewne stałe. We wzorze tym oznaczono

$$\exp \left[-\frac{(x-m)^2}{2\sigma^2} \right] = e^{-\frac{(x-m)^2}{2\sigma^2}}.$$

Wprowadzając nową zmienną

$$(2) \quad t = \frac{x-m}{\sigma},$$

która nazywać będziemy *zmienną standaryzowaną*, otrzymamy

$$(3) \quad f(t) = \frac{1}{\sigma \sqrt{2\pi}} e^{-t^2/2}.$$

Oznaczając

$$(4) \quad \varphi(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$$

otrzymujemy zależność między funkcją gęstości zwykłej zmiennej i funkcją gęstości zmiennej standaryzowanej. Zależność ta wyraża się wzorem

$$(5) \quad f(t) = \frac{1}{\sigma} \varphi(t).$$

Wykażemy, że

$$\int_{-\infty}^{\infty} f(x) dx = 1.$$

W tym celu do wzoru (1) wprowadźmy zmienną standaryzowaną

$$t = \frac{x-m}{\sigma}.$$

Stąd

$$dx = \sigma dt.$$

Wobec tego

$$\frac{1}{\sigma \sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(x-m)^2/2\sigma^2} dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-t^2/2} dt.$$

Całka, występująca po prawej stronie tej równości, przez podstawienie $z = t/\sqrt{2}$ sprowadza

się do całki

$$\frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} e^{-z^2} dz = 1 ,$$

wiadomo bowiem (patrz 1.3.1, wzór (6)), że

$$(6) \quad \int_{-\infty}^{\infty} e^{-z^2} dz = \sqrt{\pi} ,$$

gdzie jest to znana nam całka Eulera-Poissona.

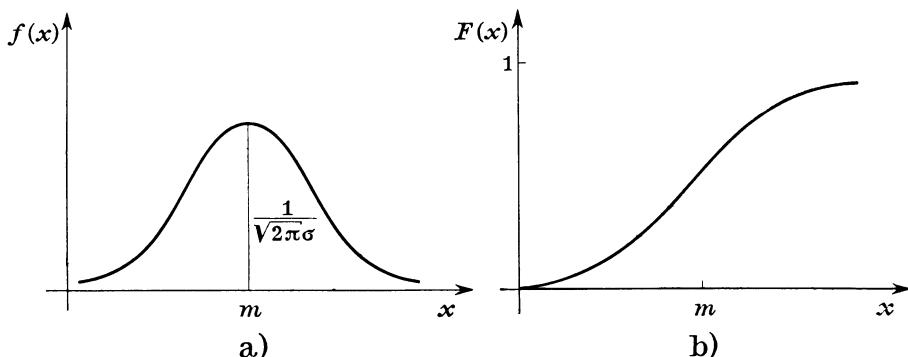
Ponieważ całka

$$\int_0^t e^{-z^2} dz$$

nie może być wyrażona za pomocą funkcji elementarnych, przeto wartość tej całki oblicza się korzystając z rozwinięcia w szereg funkcji

$$(7) \quad e^{-z^2} = 1 - z^2 + \frac{z^4}{2} - \frac{z^6}{3} + \dots + (-1)^{n-1} \cdot \frac{z^{2(n-1)}}{n-1} \pm \dots$$

Bezpośrednie obliczanie za pomocą wzoru (7) wartości gęstości i dystrybuanty rozkładu normalnego wymaga żmudnych rachunków. Dla uniknięcia konieczności wykonywania tych rachunków zostały opracowane tablice gęstości i dystrybuanty standaryzowanej zmiennej losowej o rozkładzie normalnym.



Rys. 1

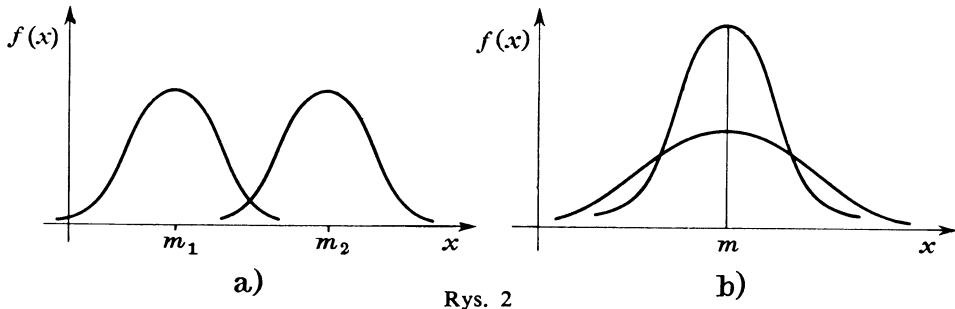
W końcu niniejszej książki podane są takie tablice; za ich pomocą znajduje się bez trudu wartość funkcji gęstości

$$\varphi(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$$

lub dystrybuanty

$$\Phi(t) = \frac{1}{2} + \frac{1}{\sqrt{2\pi}} \int_0^t e^{-v^2/2} dv.$$

Rozkład normalny należy do najważniejszych rozkładów teoretycznych. W praktyce na każdym kroku napotykamy zmienne losowe posiadające rozkład bardzo zbliżony do rozkładu normalnego. Klasycznych przykładów zastosowań rozkładu normalnego dostarczają rysunki 1a i 1b.



Rys. 2

cza teoria błędów obserwacji. Jako potocze przykłady wymienimy wzrost i wagę ludzi. Rozkład wartości wzrostu i wagi ludzi jest bardzo zbliżony do rozkładu normalnego. Dobre wyobrażenie o kształcie krzywej gęstości rozkładu normalnego możemy otrzymać np. przy mieleniu kawy na młynku. Znajdująca się w szufladce młynka zmietana kawa przypomina swym kształtem dzwon. Każdy pionowy przekrój tego dzwonu jest bardzo podobny do kształtu powierzchni pod krzywą gęstości rozkładu normalnego. Charakterystyczny wygląd krzywej gęstości i dystrybuanty rozkładu normalnego demonstrują rysunki 1a i 1b.

Rozkład normalny jest rozkładem symetrycznym względem prostej $x=m$. Obliczając pochodną funkcji

$$f'(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-m)^2}{2\sigma^2}\right)$$

i przyrównując ją do zera sprawdzimy łatwo, że gęstość osiąga maksimum w punkcie $x=m$. Wartość maksymalna funkcji gęstości wynosi

$$f(m) = \frac{1}{\sigma\sqrt{2\pi}}.$$

Na lewo i na prawo od swego maksimum krzywa gęstości rozkładu normalnego stale opada, zbliżając się asymptotycznie do osi odciętych. Punkty przegięcia krzywej gęstości rozkładu normalnego mają odcięte $m \pm \sigma$.

Liczby m i σ są parametrami rozkładu normalnego. Liczba m określa położenie osi symetrii rozkładu, natomiast od parametru σ zależy wartość maksymalna funkcji gęstości. Na rysunku 2a przedstawiono dwa wykresy rozkładu normalnego o tym samym para-

metrze σ , lecz różniące się między sobą parametrem m . Na rysunku 2b przedstawione są dwa wykresy rozkładu normalnego mające wspólny parametr m i różne wartości parametru σ .

Na zakończenie uwagi o rozkładzie normalnym przytoczymy jeden przykład liczbowy. Istnieje twierdzenie (poznamy je później), z którego wynika, że przy dużych wartościach n rozkład normalny daje dobre przybliżenie rozkładu dwumianowego. Dla danych wartości p , q i n parametry rozkładu normalnego wyrażają się wzorami

$$(8) \quad m = np ,$$

$$(9) \quad \sigma = \sqrt{npq} .$$

PRZYKŁAD 1. Rzucamy 10000 razy monetą. Przyjmując, że prawdopodobieństwo wyrzucenia orła równa się $\frac{1}{2}$, obliczyć prawdopodobieństwo tego, że liczba wyrzuconych orłów zawiera się w granicach od 5050 do 5100.

W zadaniu tym zmienną losową X , posiadającą rozkład zbliżony do rozkładu normalnego, jest liczba wyrzuconych orłów. Rozwiążanie zadania bez pomocy tablic nastreżyczyłoby wiele kłopotu. Jak wiemy, funkcję gęstości rozkładu normalnego określają dwa parametry m i σ . Zarówno rozkład ilości wyrzuconych orłów, o które chodzi w naszym zadaniu, jak i rozkład stabilizowanej zmiennej standaryzowanej t jest rozkładem normalnym, przy czym różnica między rozkładem normalnym liczby wyrzuconych orłów i rozkładem normalnym zmiennej t polega na tym, że te dwa rozkłady mają różne wartości parametrów m i σ .

Posługując się wzorami (8) i (9) obliczmy wartości tych parametrów w rozkładzie normalnym zmiennej losowej X . Otrzymujemy

$$m = 5000 , \quad \sigma = 50 .$$

Wobec tego zmienna losowa X ma rozkład normalny o parametrach $m = 5000$ i $\sigma = 50$. Zapisujemy to krótko: zmienna losowa X ma rozkład $N(5000, 50)$. W celu sprowadzenia tego rozkładu do rozkładu $N(0, 1)$, tzn. do rozkładu normalnego o parametrach $m = 0$ i $\sigma = 1$, należy zmienną losową X zestandaryzować, posługując się wzorem (2). Po wykonaniu standaryzacji można już korzystać z tablic rozkładu normalnego.

W naszym zadaniu mamy obliczyć

$$P(x_1 \leq X \leq x_2) = P(5050 \leq X \leq 5100) .$$

Wykonujemy standaryzację, polegającą na wprowadzeniu zmiennej t :

$$t_1 = \frac{5050 - 5000}{50} = 1 , \quad t_2 = \frac{5100 - 5000}{50} = 2 .$$

Mamy więc obliczyć

$$(10) \quad P(t_1 \leq T \leq t_2) = \frac{1}{\sqrt{2\pi}} \int_{t_1}^{t_2} e^{-v^2/2} dv .$$

Ale dla $t_1 \geq 0$ i $t_2 \geq 0$

$$(11) \quad \frac{1}{\sqrt{2\pi}} \int_{t_1}^{t_2} e^{-v^2/2} dv = \frac{1}{\sqrt{2\pi}} \left[\int_0^{t_2} e^{-v^2/2} dv - \int_0^{t_1} e^{-v^2/2} dv \right].$$

Oznaczmy

$$\Phi(t_1) = \frac{1}{\sqrt{2\pi}} \int_0^{t_1} e^{-v^2/2} dv$$

oraz

$$\Phi(t_2) = \frac{1}{\sqrt{2\pi}} \int_0^{t_2} e^{-v^2/2} dv.$$

Wobec tego

$$P(t_1 \leq T \leq t_2) = \Phi(t_2) - \Phi(t_1).$$

Wartości $\Phi(t_1)$ i $\Phi(t_2)$ odczytujemy z tablic zamieszczonych w końcu niniejszej książki.

W naszym przypadku mamy

$$\Phi(2) = 0,4773, \quad \Phi(1) = 0,3413.$$

Stąd szukane prawdopodobieństwo

$$P(1 \leq T \leq 2) = 0,1360.$$

Oczywiście tyle samo równa się też

$$P(x_1 \leq X \leq x_2) = P(5050 \leq X \leq 5100) = 0,1360.$$

Dalsze przykłady ilustrujące sposób korzystania z tablic rozkładu normalnego znajdzie czytelnik w części III, § 6.5.

3.5.4. Rozkład gamma

► Gęstość rozkładu gamma⁽¹⁾ określają wzory

$$(1) \quad f(x) = \begin{cases} \frac{x^a e^{-x/b}}{\Gamma(a+1) b^{a+1}} & \text{dla } x > 0, \\ 0 & \text{dla } x \leq 0, \end{cases}$$

gdzie $a > 0$, $b > 0$.

Jak widać, $f(x) \geq 0$ i jest funkcją ciągłą.

⁽¹⁾ Przed przystąpieniem do studiowania rozkładu gamma należy gruntownie zapoznać się z § 1.3.

Wykażemy, że

$$\int_0^\infty f(x) dx = 1 .$$

Rozpatrzmy w tym celu

$$(2) \quad I = \int_0^\infty \frac{x^a e^{-x/b}}{b^{a+1}} dx .$$

Podstawiając $y = \frac{x}{b}$, $x = by$, $dx = b dy$ do wzoru (2) otrzymamy

$$(3) \quad I = \int_0^\infty \frac{(by)^a e^{-y}}{b^{a+1}} b dy = \int_0^\infty y^a e^{-y} dy = \Gamma(a+1) .$$

Natomiast

$$\int_0^\infty f(x) dx = \frac{I}{\Gamma(a+1)} = 1 .$$

Dystrybuanta rozkładu gamma jest postaci

$$(4) \quad F(x) = \begin{cases} \int_0^x f(t) dt & \text{dla } x > 0, \\ 0 & \text{dla } x \leq 0 . \end{cases}$$

Wartości $F(x)$ znajdujemy całkując przez części całkę po prawej stronie równania

$$(5) \quad 1 - F(x) = \int_x^\infty \frac{t^a e^{-t/b}}{\Gamma(a+1) b^{a+1}} dt .$$

Mamy

$$1 - F(x) = \frac{1}{\Gamma(a+1) b^{a+1}} \int_x^\infty t^a e^{-t/b} dt .$$

Podstawiamy

$$\frac{t}{b} = z, \quad t = bz, \quad dt = bdz$$

i otrzymujemy

$$1 - F(x) = \frac{1}{\Gamma(a+1) b^{a+1}} \int_{x/b}^\infty (bz)^a e^{-z} b dz = \frac{1}{\Gamma(a+1)} \int_{x/b}^\infty z^a e^{-z} dz .$$

Jeśli a jest liczbą całkowitą nieujemną, to $\Gamma(a+1) = a!$. Natomiast

$$\begin{aligned} \int_{x/b}^{\infty} z^a e^{-z} dz &= - \int_{x/b}^{\infty} z^a d(e^{-z}) = \\ &= -[z^a e^{-z}]_{x/b}^{\infty} + a \int_{x/b}^{\infty} e^{-z} z^{a-1} dz = e^{-x/b} \left(\frac{x}{b} \right)^a + a \int_{x/b}^{\infty} e^{-z} z^{a-1} dz . \end{aligned}$$

Otrzymaliśmy wzór rekurencyjny, który po rozwinięciu względem a może być napisany w sposób następujący:

$$\int_{x/b}^{\infty} z^a e^{-z} dz = e^{-x/b} \left[\frac{a!}{a!} \left(\frac{x}{b} \right)^a + \frac{a!}{(a-1)!} \left(\frac{x}{b} \right)^{a-1} + \dots + \frac{a!}{1} \cdot \frac{x}{b} + \frac{a!}{0!} \right].$$

Stąd

$$1 - F(x) = e^{-x/b} \left[1 + \frac{x}{b} + \frac{1}{2!} \left(\frac{x}{b} \right)^2 + \dots + \frac{1}{a!} \left(\frac{x}{b} \right)^a \right],$$

czyli ostatecznie

$$(6) \quad F(x) = 1 - e^{-x/b} \left[1 + \frac{x}{b} + \frac{1}{2!} \left(\frac{x}{b} \right)^2 + \dots + \frac{1}{a!} \left(\frac{x}{b} \right)^a \right].$$

Wzorem tym warto się posługiwać, gdy a jest nieduże oraz gdy a jest liczbą całkowitą dodatnią lub gdy jest wielokrotnością $\frac{1}{2}$. Gdy a jest wielokrotnością $\frac{1}{2}$, wartość $a!$ obliczamy korzystając ze wzoru rekurencyjnego

$$a! = a\Gamma(a)$$

oraz z równości

$$\Gamma(\frac{1}{2}) = \sqrt{\pi}$$

(patrz 1.3.1, wzór (5)).

Wartości gęstości i dystrybuanty rozkładu gamma są stablicowane. Ponieważ tablice te tworzą gruby foliąt i znajdują się jedynie w bibliotekach instytucji naukowych, więc gdy występuje konieczność obliczenia wartości dystrybuanty rozkładu gamma, korzysta się na ogół ze wzoru (6).

Rozkład gamma znajduje liczne zastosowania w różnych dziedzinach statystyki matematycznej. Szczególnie ważną rolę rozkład ten odgrywa w statystycznej kontroli jakości produkcji i statystycznym odbiorze towarów.

3.5.5. Rozkład beta

► Funkcja gęstości w rozkładzie *beta* wyraża się wzorami

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

gdzie $a > 0$ i $b > 0$. Funkcja $f(x)$ jest nieujemna wewnątrz całego przedziału $\langle 0, 1 \rangle$, wartości zerowe osiąga na końcach tego przedziału, jeśli $a > 1$ i $b > 1$.

Jeśli $a-1>0$ i $b-1>0$, to $f(x)$ osiąga maksimum w punkcie

$$x_0 = \frac{a-1}{a+b-2} ;$$

podobnie jeśli $a-1<0$ i $b-1<0$, to $f(x)$ osiąga minimum w punkcie

$$x_0 = \frac{a-1}{a+b-2} .$$

Łatwo wykazać, że

$$(2) \quad \int_0^1 f(x) dx = 1 .$$

Istotnie,

$$\int_0^1 f(x) dx = \frac{1}{B(a, b)} \int_0^1 x^{a-1} (1-x)^{b-1} dx = \frac{B(a, b)}{B(a, b)} = 1$$

(patrz 1.3.2, wzór (1)).

Dystrybuanta rozkładu beta wyraża się wzorami

$$(3) \quad F(x) = \begin{cases} \int_0^x f(t) dt & \text{dla } 0 \leq x \leq 1, \\ 0 & \text{dla } x < 0, \\ 1 & \text{dla } x > 1. \end{cases}$$

Miedzy funkcjami gamma i beta zachodzi prosty związek (patrz 1.3.2, wzór (2))

$$B(a, b) = \frac{\Gamma(a) \Gamma(b)}{\Gamma(a+b)} ,$$

który pozwala korzystać przy obliczaniu wartości funkcji beta z tablic funkcji gamma.

Można udowodnić, że dystrybuanta rozkładu beta spełnia następującą tożsamość:

$$(4) \quad F(1-x) = 1 - F(x) .$$

Najważniejszą dziedziną zastosowań rozkładu beta jest statystyczna kontrola jakości produkcji i statystyczny odbiór towarów. 

Pytania kontrolne i zadania

1. Podać określenie zmiennej losowej.
2. Podać przykłady zmiennych losowych.
3. Podać określenia i przykłady zmiennych losowych skokowych i ciągłych.
4. Co to jest rozkład prawdopodobieństw?
5. Podać definicję funkcji rozkładu i dystrybuanty zmiennej losowej typu skokowego. Wyjaśnić analogię między funkcją rozkładu i częstością oraz dystrybuantą i częstością skumulowaną.
6. Podać określenie rozkładu zero-jedynkowego. Wymienić kilka przykładów zmiennych losowych o tym rozkładzie.

7. Opisać, na czym polega schemat Bernoulliego.
8. Podać wzór na funkcję rozkładu i dystrybuantę w rozkładzie dwumianowym. Wymienić własności tego rozkładu. Podać przykłady zmiennych losowych o tym rozkładzie.
9. Rzucamy 6 razy monetą. Zakładając, że prawdopodobieństwo wyrzucenia orła $p=1/2$, obliczyć prawdopodobieństwo:
- wyrzucenia orła jeden raz,
 - wyrzucenia orła dwa lub trzy razy,
 - wyrzucenia przynajmniej dwa razy orła,
 - wyrzucenia najwyżej pięć razy orła,
 - niewyrzucenia ani razu reszki.
10. Rzucamy 5 razy kością do gry. Zakładając, że kość ma kształt idealnego sześciangu, obliczyć prawdopodobieństwo:
- wyrzucenia dwa razy sześciu oczek,
 - wyrzucenia raz sześciu oczek i dwa razy pięciu oczek,
 - wyrzucenia przynajmniej raz sześciu oczek,
 - niewyrzucenia ani razu dwóch oczek,
 - wyrzucenia jednego oczka albo dwóch oczek, albo ..., albo sześciu oczek.
11. Podać wzór na funkcję rozkładu i dystrybuantę w rozkładzie hipergeometrycznym.
12. Wyjaśnić pogłówno (na przykładzie), kiedy korzystamy z rozkładu dwumianowego, a kiedy z rozkładu hipergeometrycznego.
13. W urnie znajdują się dwie kule białe i cztery kule czarne. Z urny tej wyciągamy kolejno trzy kule nie wkładając ich z powrotem do urny. Obliczyć prawdopodobieństwo wyciągnięcia:
- trzech kul czarnych,
 - jednej kuli białej i dwóch kul czarnych,
 - jednej kuli czarnej i dwóch kul białych,
 - dodając obliczone prawdopodobieństwa zbadać, czy zdarzenia wymienione w a), b), c) stanowią przestrzeń zdarzeń.
14. Podać wzór na funkcję rozkładu i dystrybuantę w rozkładzie Poissona.
15. Dla jakiego p i n rozkład Poissona daje dostatecznie dobre przybliżenie rozkładu dwumianowego?
16. W fabryce żarówek ustalono, że przeciętny procent braków wynosi 2%. Posługując się tablicami rozkładu Poissona obliczyć prawdopodobieństwo, że w partii towaru liczącej 100 żarówek
- znajdą się dwie żarówki złe (dwa braki),
 - znajdą się ponad dwa braki,
 - nie będzie braków,
 - będzie nie więcej niż trzy braki.
17. Stwierdzono, że w czasie transportu towaru powstaje przeciętnie 3% braków. Przedsiębiorstwo ekspediujące chce dodawać do każdej partii gratis dwie sztuki towaru. Jakie duże muszą być partie towaru, aby z prawdopodobieństwem nie mniejszym niż 0,99 uniknąć reklamacji, tzn. aby z prawdopodobieństwem nie mniejszym niż 0,99 liczba braków w partii była nie większa niż dwa?
18. Podać definicję gęstości prawdopodobieństwa.
19. Zmienna losowa X jest zmienną ciągłą. Czemu równa się $P(X=x)$?
20. Na czym polega związek między gęstością prawdopodobieństwa i dystrybuantą?
21. Podać wzór na gęstość prawdopodobieństwa i dystrybuantę w rozkładzie normalnym.
22. Na czym polega i do czego służy standaryzowanie zmiennej?
23. Podać własności rozkładu normalnego.
24. Zmienna losowa X posiada rozkład normalny o parametrach $m=12$, $\sigma=2$. Posługując się tablicami rozkładu normalnego obliczyć:
- $P(X<15)$,
 - $P(X<7)$,
 - $P(8 < X \leq 16)$,
 - $P(8 < X \leq 13)$,
 - jeśli wiadomo, że $P(X < x) = 0,9773$, to czemu równa się x ?

3.6. ZMIENNE LOSOWE DWUWYMIAROWE

3.6.1. Sformułowanie zagadnienia

W punkcie 3.1.1 podaliśmy określenie zmiennej losowej jednowymiarowej. Okreście-
nie to można łatwo uogólnić na przypadek zmiennych losowych dwuwymiarowych.

OKREŚLENIE 1. *Zmienną losową dwuwymiarową (X, Y) nazywa się parę funkcji $X=X(e)$, $Y=Y(e)$ określonych na zbiorze zdarzeń elementarnych E , takich że dla każdej pary liczb rzeczywistych x, y zbiór A zdarzeń elementarnych $e \in E$, dla których $X(e) < x$, $Y(e) < y$, spełnia warunek $A \in \mathcal{B}$.*

Wynika stąd, że

$$P(A) = P(X < x, Y < y).$$

Prawdopodobieństwo to nazywa się *rozkładem zmiennej losowej (X, Y)*. Realizacje zmiennej losowej (X, Y) będziemy oznaczać parą małych liter opatrzonych indeksami, a mianowicie (x_1, y_1) , (x_2, y_2) , ... Dwuwymiarowa zmienna losowa bywa często interpretowana geometrycznie. Niech (x_i, y_i) oznacza realizację zmiennej losowej (X, Y). Realizacji tej odpowiada pewien punkt na płaszczyźnie. Punkt ten, jak wszelkie inne punkty odpowiadające realizacjom zmiennej losowej (X, Y), będziemy nazywać *punktem eksperymentalnym*. Położenie punktu na płaszczyźnie określają dwie składowe X, Y . Jeżeli składowe są zmiennymi losowymi, to mówimy, że mamy do czynienia z *dwuwymiarową zmienną losową (X, Y) lub z wektorem losowym $\mathbf{X}=(X, Y)$* .

Odpowiednikiem zmiennych losowych w statystyce są cechy statystyczne. Populacja rozpatrywana ze względu na dwie cechy nazywa się *populacją dwuwymiarową*. Dwie cechy, ze względu na które bada się populację, są odpowiednikiem dwuwymiarowej zmiennej losowej, natomiast poszczególne obserwacje statystyczne, wyrażające wartości każdej z tych cech u poszczególnych jednostek statystycznych, należących do danej populacji, są odpowiednikami realizacji dwuwymiarowej zmiennej losowej. Przykładem dwuwymiarowej populacji może być np. pewna kategoria robotników w danym zakładzie produkcyjnym rozpatrywana ze względu na staż pracy zawodowej i wysokość zarobków.

Analogicznie jak w przypadku zmiennych losowych jednowymiarowych, wśród zmiennych losowych dwuwymiarowych można wyróżnić zmienne losowe skokowe i zmienne losowe ciągłe.

3.6.2. Dwuwymiarowa zmienna losowa skokowa

OKREŚLENIE 1. Dwuwymiarowa zmienna losowa (X, Y) jest zmienną *skokową*, jeśli składowe X i Y mają tylko skończony lub przeliczalny zbiór wartości.

OKREŚLENIE 2. *Rozkładem skokowej zmiennej dwuwymiarowej (X, Y) nazywa się prawdo-
podobieństwo zdarzenia ($X=x_i, Y=y_j$) $i=1, 2, \dots, j=1, 2, \dots$*

Na oznaczenie funkcji rozkładu dwuwymiarowej zmiennej losowej skokowej używamy następującego zapisu:

$$P(X=x_i, Y=y_j) = p_{ij}.$$

Wartości, jakie przybiera zmienna losowa skokowa i prawdopodobieństwa odpowiadające poszczególnym wartościom tej zmiennej są zestawione w tablicy 1, zwanej *tablicą dwudzielną* lub *korelacyjną*.

Tablica 1

$X \backslash Y$	y_1	y_2	...	y_J	...	y_s	\sum
x_1	p_{11}	p_{12}	...	p_{1J}	...	p_{1s}	p_1
x_2	p_{21}	p_{22}	...	p_{2J}	...	p_{2s}	p_2
...
x_t	p_{t1}	p_{t2}	...	p_{tJ}	...	p_{ts}	p_t
...
x_r	p_{r1}	p_{r2}	...	p_{rJ}	...	p_{rs}	p_r
\sum	q_1	q_2	...	q_J	...	q_s	1

Jeśli zmienne X i Y mogą przybierać przeliczalny zbiór wartości, to

$$\sum_i \sum_j p_{ij} = 1.$$

Jeśli natomiast $i=1, 2, \dots, r$, a $j=1, 2, \dots, s$, to

$$(1) \quad \sum_{i=1}^r \sum_{j=1}^s p_{ij} = 1.$$

Sumę (1) otrzymuje się dodając do siebie wszystkie prawdopodobieństwa zawarte w tablicy 1. Można to uczynić dwoma sposobami: sumując najpierw wzduż wierszy, a następnie dodając sumy ostatniej kolumny lub – na odwrót – sumując wzduż kolumn, a następnie dodając sumy ostatniego wiersza. Wynika stąd, że suma ostatniej kolumny równa się sumie ostatniego wiersza i równa się jedności, czyli

$$(2) \quad \sum_{i=1}^r p_i = 1,$$

$$(3) \quad \sum_{j=1}^s q_j = 1.$$

Z równości (2) i (3) wynika, że prawdopodobieństwa zawarte w ostatniej kolumnie i w ostatnim wierszu tablicy 1 tworzą rozkłady. Rozkłady te noszą nazwę *rozkładów brzegowych* skokowej zmiennej losowej (X, Y).

We wzorze (2)

$$(4) \quad p_i = \sum_{j=1}^s p_{ij},$$

natomiaszt we wzorze (3)

$$(5) \quad q_j = \sum_{i=1}^r p_{ij}.$$

Napiszmy sumę, stojącą po prawej stronie wzoru (4), w rozwiniętej postaci

$$(6) \quad p_i = p_{i1} + p_{i2} + \dots + p_{is}.$$

Po podzieleniu obu stron równości (6) przez p_i otrzymamy

$$(7) \quad \frac{p_{i1}}{p_i} + \frac{p_{i2}}{p_i} + \dots + \frac{p_{is}}{p_i} = 1.$$

Ponieważ suma (7) równa się jedności, a jej składniki są nieujemne, przeto otrzymaliśmy rozkład prawdopodobieństw. Jest to tak zwany *warunkowy rozkład prawdopodobieństw* zmiennej Y pod warunkiem, że zmienna $X=x_i$ ⁽¹⁾.

Oznaczmy

$$(8) \quad p_{j|i} = \frac{p_{ij}}{p_i}, \quad p_{i|j} = \frac{p_{ij}}{q_j},$$

gdzie $p_{j|i}$ jest to prawdopodobieństwo warunkowe y_j przy założeniu, że $X=x_i$, a $p_{i|j}$ jest to prawdopodobieństwo warunkowe x_i przy założeniu, że $Y=y_j$. Wobec tego wzór (7) można napisać krótko w postaci następującej:

$$(9) \quad \sum_{j=1}^s p_{j|i} = 1.$$

Podobnie warunkowy rozkład prawdopodobieństw zmiennej X pod warunkiem, że zmienna $Y=y_j$, wyraża się wzorem

$$(10) \quad \sum_{i=1}^r p_{i|j} = 1.$$

Na mocy wzoru (8)

$$(11) \quad p_{ij} = p_i p_{j|i} = q_j p_{i|j}.$$

Ze wzoru (11) wynika, że *rozkład łączny dwuwymiarowej zmiennej losowej równa się iloczynowi bezwarunkowego rozkładu jednej ze zmiennych przez warunkowy rozkład drugiej zmiennej*. Terminem *rozkład łączny* nazwaliśmy p_{ij} . Termin ten dobrze podkreśla fakt, że p_{ij} dotyczy zmiennej dwuwymiarowej, podczas gdy p_i , q_j , $p_{i|j}$, $p_{j|i}$ dotyczą zmiennych jednowymiarowych.

Jeżeli dla wszystkich i, j

$$(12) \quad p_{j|i} = q_j$$

⁽¹⁾ Pojęcie prawdopodobieństwa warunkowego jest czytelnikowi znane z 2.4.2. Była tam mowa o prawdopodobieństwie warunkowym zdarzeń, teraz natomiast zajmujemy się prawdopodobieństwem warunkowym dotyczącym zmiennych losowych.

lub, co na jedno wychodzi,

$$(13) \quad p_{i|j} = p_i,$$

to zmienne losowe X i Y są *niezależne*.

Wzór (11) przybiera w tym przypadku prostszą postać, a mianowicie

$$(14) \quad p_{ij} = p_i \cdot q_j.$$

Równość (14) wyraża konieczny i dostateczny warunek niezależności dwóch zmiennych losowych. Warunek ten posiada analogiczną postać do warunku niezależności dwóch zdarzeń losowych (patrz 2.4.2).

Na mocy wzoru (11)

$$(15) \quad \sum_i \sum_j p_{ij} = \sum_i p_i \sum_j p_{j|i} = \sum_j q_j \sum_i p_{i|j}, \quad 1 \leq i \leq r, \quad 1 \leq j \leq s.$$

Gdy zmienne losowe są niezależne, to

$$(16) \quad \sum_i \sum_j p_{ij} = \sum_i p_i \sum_j q_j.$$

OKREŚLENIE 3. Funkcja $F(x, y) = P(X < x, Y < y)$ nazywa się *dystrybuantą dwuwymiarową zmiennej losowej (X, Y)* .

Gdy dwuwymiarowa zmienna losowa jest typu skokowego, to z określenia dystrybuanty wynika, że

$$(17) \quad F(x, y) = \sum_{x_i < x} \sum_{y_j < y} P(X = x_i, Y = y_j).$$

oraz

$$(18) \quad F(+\infty, +\infty) = \sum_{x_i < \infty} \sum_{y_j < \infty} P(X = x_i, Y = y_j) = 1.$$

Dystrybuanta brzegowa zmiennej X wyraża się wzorem

$$(19) \quad F(x, +\infty) = \sum_{x_i < x} \sum_{y_j < \infty} P(X = x_i, Y = y_j).$$

3.6.3. Dwuwymiarowa zmienna losowa ciągła

OKREŚLENIE 1. *Gęstością* $f(x, y)$ dwuwymiarowej zmiennej losowej (X, Y) nazywa się pochodna mieszana dystrybuanty w punkcie (x, y) , tzn.

$$f(x, y) = \frac{\partial^2 F}{\partial x \partial y}.$$

OKREŚLENIE 2. Jeśli dystrybuanta $F(x, y)$ jest funkcją ciągłą oraz jeśli gęstość $f(x, y)$ jest również funkcją ciągłą z wyjątkiem co najwyżej zbioru punktów należących do skończonej liczby krzywych, to dwuwymiarowa zmienna losowa (X, Y) jest zmienną *ciągłą*.

Na mocy określenia 1 mamy⁽¹⁾

$$(2) \quad F(x, y) = \int_{-\infty}^x \int_{-\infty}^y f(u, v) du dv$$

oraz

$$(3) \quad F(+\infty, +\infty) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy = 1,$$

$$(4) \quad F(-\infty, y) = F(x, -\infty) = 0.$$

Dystrybuanta brzegowa zmiennej losowej X wyraża się wzorem

$$(5) \quad F(x, \infty) = \int_{-\infty}^x \int_{-\infty}^{\infty} f(u, v) du dv = \int_{-\infty}^x f_1(u) du.$$

We wzorze tym

$$(6) \quad f_1(x) = \int_{-\infty}^{\infty} f(x, y) dy$$

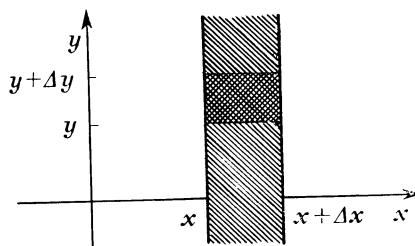
oznacza gęstość brzegową zmiennej X .

Dzięki symetrii wzorów, dystrybuanta i gęstość brzegowa zmiennej Y mają analogiczną postać.

Omawiając dwuwymiarową zmienną losową skokową podaliśmy definicję prawdopodobieństwa warunkowego (patrz 3.6.2, wzór (8)). Gdy zmienna (X, Y) jest zmienną ciągłą, przez prawdopodobieństwo warunkowe $P(y \leq Y < y + \Delta y | x \leq X < x + \Delta x)$ będziemy rozumieli wyrażenie

$$(7) \quad P(y \leq Y < y + \Delta y | x \leq X < x + \Delta x) = \frac{P(y \leq Y < y + \Delta y, x \leq X < x + \Delta x)}{P(x \leq X < x + \Delta x)},$$

przy czym zakłada się, że $P(x \leq X < x + \Delta x) > 0$.



Rys. 1

⁽¹⁾ Czytelnik mający braki w wiadomościach z zakresu rachunku całkowego może mimo tego przystąpić do studlowania własności dwuwymiarowej zmiennej losowej typu ciągłego, jeśli przyswoi sobie dobrze definicje, wzory i pojęcia omówione w 3.6.2. Pamiętać bowiem należy, że całkowanie oznacza sumowanie nieskończonego szeregu zbieżnego. Przy czytaniu znaku \int można więc w wyobrażni zastąpić znakiem \sum .

Porównując wzór (7) ze wzorem (8), 3.6.2, łatwo dostrzec podobieństwo między nimi. Jak wiadomo, dla zmiennych losowych ciągłych mamy

$$P(X=x) = P(Y=y) = P(X=x, Y=y) = 0.$$

Prawdopodobieństwo warunkowe (7) jest to prawdopodobieństwo tego, że losowo wybrany punkt (X, Y) znajdzie się w prostokącie $y \leq Y < y + \Delta y, x \leq X < x + \Delta x$, jeśli wiadomo, że punkt ten leży wewnątrz obszaru $-\infty < Y < \infty, x \leq X < x + \Delta x$ (rys. 1). Z określenia dystrybuanty zmiennej ciągłej wynika, że

$$(8) \quad P(y \leq Y < y + \Delta y | x \leq X < x + \Delta x) = \frac{\int_x^{x+\Delta x} \int_{y}^{y+\Delta y} f(u, v) du dv}{\int_x^{x+\Delta x} \int_{-\infty}^{\infty} f(u, v) du dv}.$$

Oczywiście

$$(9) \quad P(-\infty < Y < \infty | x \leq X < x + \Delta x) = 1,$$

mamy bowiem

$$(10) \quad P(-\infty < Y < \infty | x \leq X < x + \Delta x) = \frac{\int_x^{x+\Delta x} \int_{-\infty}^{\infty} f(u, v) du dv}{\int_x^{x+\Delta x} \int_{-\infty}^{\infty} f(u, v) du dv} = 1.$$

Dystrybuanta w rozkładzie warunkowym wyraża się wzorem

$$(11) \quad P(Y < y | x \leq X < x + \Delta x) = \frac{\int_x^{x+\Delta x} \int_{-\infty}^y f(u, v) du dv}{\int_x^{x+\Delta x} \int_{-\infty}^{\infty} f(u, v) du dv},$$

natomiast gęstość $f(y|x)$ – wzorem

$$(12) \quad f(y|x) = \frac{f(x, y)}{f_1(x)}.$$

Przypominamy, że

$$f_1(x) = \int_{-\infty}^{\infty} f(x, y) dy.$$

Formułując definicję prawdopodobieństwa warunkowego (patrz wzór (7)) zajmowaliśmy się zmienną Y przy założeniu, że zmienna X spełnia nierówność $x \leq X < x + \Delta x$. Łatwo wyprowadzić wzory symetryczne do wzorów (7) - (12) dla zmiennej X , przy założeniu, że zmienna Y czyni zadość nierówności $y \leq Y < y + \Delta y$.

Na zakończenie rozważań o zmiennej losowej dwuwymiarowej zajmiemy się zbadaniem koniecznego i dostatecznego warunku niezależności zmiennych losowych X i Y , jeśli zmienne te są ciągłe.

OKREŚLENIE 3. Zmienne losowe X i Y są od siebie *niezależne*, jeżeli dla każdej pary liczb x_1, x_2 oraz y_1, y_2

$$(13) \quad P(x_1 \leq X < x_2, y_1 \leq Y < y_2) = P(x_1 \leq X < x_2) P(y_1 \leq Y < y_2).$$

W szczególności zmienne losowe X i Y są niezależne, jeżeli

$$(14) \quad P(X < x, Y < y) = P(X < x) P(Y < y).$$

Z określenia dystrybuanty wynika, że równość (14) można napisać inaczej, a mianowicie

$$(15) \quad F(x, y) = F_1(x) F_2(y),$$

gdzie

$$F_1(x) = F(x, \infty), \quad F_2(y) = F(\infty, y).$$

Jest to konieczny i dostateczny warunek niezależności zmiennych losowych (dowód znajdziesz czytelnik w pracy [8], str. 64 - 65).

4.1. PARAMETRY OPISOWE

OKREŚLENIE 1. Liczba charakteryzująca w pewien sposób zbiór wartości, jakie może przybierać zmienność losowa, nazywa się *parametrem opisowym* tej zmiennej lub krótko *parametrem*.

Dzięki parametrom opisowym możemy za pomocą kilku liczb uzyskać dostatecznie dobre wyobrażenie o rozkładzie zmiennej losowej. Zilustrujemy to na przykładzie. Przypuszcmy, że interesują nas zarobki robotników w pewnej fabryce. Najpełniejszą informację w tej sprawie uzyskamy studując listy płacy i zapoznając się z zarobkiem każdego robotnika. Jeżeli liczba zatrudnionych w fabryce jest duża, to studiowanie list płacy zajmie nam wiele czasu, a przyniesie mało korzyści, gdyż i tak nie zapamiętamy przecież wszystkich pozycji występujących w listach płacy. Rozsądniej uczynimy, jeżeli zamiast interesować się zarobkami wszystkich robotników, zbadamy np., jaka wysokość zarobków powtarza się w listach płacy najczęściej oraz ile wynosi zarobek najwyższy i najniższy.

Wysokość zarobków, która w listach płacy występuje najczęściej, może być uznana za płacę typową dla robotników danej fabryki, za pewną wielkość przeciętną, dającą wyobrażenie o wysokości zarobków wszystkich robotników w tej fabryce, natomiast płaca najwyższa i płaca najniższa wyznaczają przedział, w którym są zawarte wszystkie odchylenia od tej przeciętnej.

Zarobki poszczególnych robotników możemy uważać za realizacje zmiennej losowej. W takim razie płaca najczęstsza oraz płaca najwyższa i płaca najniższa są to parametry opisowe tej zmiennej. Określenie 1 zredagowane jest bardzo ogólnie. Nie preczuje ono bliżej, w jaki sposób parametr opisowy ma charakteryzować zbiór wartości zmiennej losowej. Wynika stąd, że liczba parametrów opisowych może być dowolnie wielka. W praktyce korzysta się jednak zaledwie z kilku parametrów, najbardziej wygodnych w użyciu.

Największe znaczenie praktyczne mają dwie grupy parametrów. Do pierwszej grupy zaliczamy parametry reprezentujące przeciętną wielkość zmiennej losowej, do drugiej grupy natomiast – parametry dające wyobrażenie o tym, jak bardzo poszczególne wartości zmiennej losowej odchylają się od tej przeciętnej wielkości.

Najważniejszym parametrem, należącym do pierwszej grupy, jest tak zwana *wartość przeciętna*, znana także pod nazwą *wartości oczekiwanej* lub *nadziei matematycznej*.

W drugiej grupie na pierwszym miejscu wymienić należy *wariancję*, *odchylenie standarde* i *odchylenie przeciętne*.

4.2. WARTOŚĆ PRZECIETNA

4.2.1. Określenie wartości przeciętnej. Przykłady

Niech X będzie zmienną losową typu skokowego, przybierającą wartości x_1, x_2, \dots, x_n odpowiednio z prawdopodobieństwami p_1, p_2, \dots, p_n .

OKREŚLENIE 1. *Wartość przeciętna zmiennej losowej skokowej X jest to suma iloczynów poszczególnych wartości tej zmiennej i odpowiadających tym wartościom prawdopodobieństw:*

$$(1) \quad E(X) = \sum_{i=1}^n x_i p_i,$$

gdzie $E(X)$ oznacza wartość przeciętną.

Jeżeli X jest zmienną losową skokową, która może przybierać przeliczalną ilość wartości x_1, x_2, \dots odpowiednio z prawdopodobieństwami p_1, p_2, \dots , to

$$(2) \quad E(X) = \sum_{i=1}^{\infty} x_i p_i$$

przy założeniu, że nieskończony szereg, którego sumą jest wartość przeciętna $E(X)$, jest bezwzględnie zbieżny⁽¹⁾.

PRZYKŁAD 1. Rzucamy monetę. Oznaczmy wyrzucenie orła liczbą 1, natomiast wyrzucenie reszki – liczbą 0. Niech $P(1)=P(0)=\frac{1}{2}$. W takim razie

$$E(X) = 1 \cdot \frac{1}{2} + 0 \cdot \frac{1}{2} = \frac{1}{2}.$$

PRZYKŁAD 2. Zmienna losowa X przybiera wartości $x_1=10, x_2=100, x_3=1000$ odpowiednio z prawdopodobieństwami $p_1=0,5, p_2=0,4, p_3=0,1$. Obliczyć wartość przeciętną zmiennej losowej X .

Obliczenie wygodnie ująć w następującą tabelkę:

x_i	p_i	$x_i p_i$
10	0,5	5
100	0,4	40
1000	0,1	100
	1	145

stąd

$$E(X) = 145.$$

PRZYKŁAD 3. Zmienna losowa X o rozkładzie zero-jedynkowym przybiera wartość 1 z prawdopodobieństwem p oraz 0 z prawdopodobieństwem q . W takim razie

$$(3) \quad E(X) = 1 \cdot p + 0 \cdot q = p.$$

⁽¹⁾ Określenie szeregu bezwzględnie zbieżnego znajdzie czytelnik w pracy [19], str. 176.

PRZYKŁAD 4. Zmienna losowa ma rozkład dwumianowy o parametrach p, n , to znaczy

$$P(X=k) = C_n^k p^k q^{n-k}$$

(patrz 3.3.2, wzór (1)). W takim razie

$$(4) \quad E(X) = \sum_{k=0}^n k C_n^k p^k q^{n-k}.$$

Wartość sumy stojącej po prawej stronie równości (4) obliczamy łatwo różniczkując dwumian Newtona względem p . Mamy bowiem

$$\begin{aligned} (p+q)^n &= \sum_{k=0}^n C_n^k p^k q^{n-k}, \\ [(p+q)^n]' &= n(p+q)^{n-1} = \sum_{k=0}^n k C_n^k p^{k-1} q^{n-k}. \end{aligned}$$

Po pomnożeniu powyższej równości przez p otrzymamy

$$(5) \quad np(p+q)^{n-1} = \sum_{k=0}^n k C_n^k p^k q^{n-k}.$$

Ponieważ prawe strony wzorów (4) i (5) są jednakowe, więc

$$E(X) = np(p+q)^{n-1}.$$

Ale

$$p+q=1,$$

skąd ostatecznie

$$(6) \quad E(X) = np.$$

► **PRZYKŁAD 5.** Zmienna losowa X ma rozkład hipergeometryczny (patrz 3.3.3, wzór (1))

$$P(X=k) = \frac{C_R^k C_{N-R}^{n-k}}{C_N^n}.$$

Ponieważ

$$\sum_{k=0}^n P(X=k) = \sum_{k=0}^n \frac{C_R^k C_{N-R}^{n-k}}{C_N^n} = 1,$$

więc

$$(7) \quad \sum_{k=0}^n C_R^k C_{N-R}^{n-k} = C_N^n.$$

Wartość przeciętna w rozkładzie hipergeometrycznym wyraża się wzorem

$$(8) \quad E(X) = \sum_{k=0}^n k P(X=k) = \sum_{k=0}^n k \frac{C_R^k C_{N-R}^{n-k}}{C_N^n}.$$

Zauważmy jednak, że dla $k \geq 1$

$$\begin{aligned} kC_R^k &= \frac{kR!}{k!(R-k)!} = \frac{R!}{(k-1)!(R-k)!} = \\ &= \frac{R(R-1)!}{(k-1)![R-1-(k-1)]!} = RC_{R-1}^{k-1}. \end{aligned}$$

Stąd

$$(9) \quad E(X) = \sum_{k=1}^n R \frac{C_{R-1}^{k-1} C_{N-R}^{n-k}}{C_N^n} = \frac{R}{C_N^n} \sum_{k=1}^n C_{R-1}^{k-1} C_{N-R}^{n-k}.$$

Podstawiając $k-1=m$ otrzymujemy

$$E(X) = \frac{R}{C_N^n} \sum_{m=0}^{n-1} C_{R-1}^m C_{N-R}^{n-m-1}.$$

Podstawiając z kolei $R-1=S$ mamy

$$E(X) = \frac{R}{C_N^n} \sum_{m=0}^{n-1} C_S^m C_{(N-1)-S}^{(n-1)-m}.$$

Na mocy wzoru (7)

$$\sum_{m=0}^{n-1} C_S^m C_{(N-1)-S}^{(n-1)-m} = C_{N-1}^{n-1}.$$

W takim razie

$$E(X) = \frac{RC_{N-1}^{n-1}}{C_N^n} = n \frac{R}{N}.$$

Ale $R/N=p$, gdyż R oznacza liczbę elementów mających cechę A w populacji liczącej N elementów. Stąd ostatecznie

$$(10) \quad E(X) = np. \quad \blacktriangleleft$$

PRZYKŁAD 6. Zmienna losowa X ma rozkład Poissona (patrz 3.3.4, wzór (3))

$$P(X=k) = \frac{(np)^k}{k!} e^{-np}.$$

Wobec tego

$$E(X) = \sum_{k=0}^{\infty} k \frac{(np)^k}{k!} e^{-np} = np e^{-np} \sum_{k=1}^{\infty} \frac{(np)^{k-1}}{(k-1)!}.$$

Ponieważ

$$e^{-np} = \sum_{m=0}^{\infty} \frac{(np)^m}{m!},$$

więc

$$(11) \quad E(X) = np e^{-np} e^{np} = np.$$

Wszystkie te przykłady obliczania wartości przeciętnej dotyczyły zmiennej losowej typu skokowego. Obecnie przystępujemy do rozpatrzenia wartości przeciętnej zmiennej losowej ciągłej.

Niech X będzie zmienną losową typu ciąglego, zaś $f(x)$ niech będzie gęstością prawdopodobieństwa tej zmiennej.

OKREŚLENIE 2. Wartością przeciętną zmiennej losowej ciągkiej X nazywa się całka

$$(12) \quad E(X) = \int_{-\infty}^{\infty} x f(x) dx.$$

Ciąka ta istnieje, jeśli istnieje całka

$$\int_{-\infty}^{\infty} |x| f(x) dx.$$

Niech $F(x)$ oznacza dystrybuantę zmiennej losowej X . W takim razie

$$dF(x) = f(x) dx.$$

Wobec tego

$$E(X) = \int_{-\infty}^{\infty} x dF(x).$$

PRZYKŁAD 7. Zmienna losowa X ma rozkład prostokątny. Funkcja gęstości w tym rozkładzie dana jest wzorami (patrz 3.5.1, wzór (1))

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

Wartość przeciętna w rozkładzie prostokątnym równa się

$$(13) \quad E(X) = \frac{1}{b-a} \int_a^b x dx = \frac{1}{b-a} \frac{b^2 - a^2}{2} = \frac{a+b}{2}.$$

PRZYKŁAD 8. Zmienna losowa X ma rozkład normalny. Jak wiadomo (patrz 3.5.3, wzór (1)), funkcja gęstości rozkładu normalnego wyraża się wzorem

$$f(x) = \frac{1}{\sigma \sqrt{2\pi}} \exp\left(-\frac{(x-m)^2}{2\sigma^2}\right).$$

W takim razie

$$E(X) = \frac{1}{\sigma \sqrt{2\pi}} \int_{-\infty}^{\infty} x \exp\left(-\frac{(x-m)^2}{2\sigma^2}\right) dx.$$

Podstawiamy (patrz 3.5.3, wzór (2))

$$t = \frac{x-m}{\sigma}, \quad x = t\sigma + m, \quad dx = \sigma dt.$$

Wobec tego

$$E(X) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (t\sigma + m) e^{-t^2/2} dt = \frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{\infty} te^{-t^2/2} dt + \frac{m}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-t^2/2} dt .$$

Ale

$$\frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{\infty} te^{-t^2/2} dt = \frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-t^2/2} d\left(\frac{t^2}{2}\right) = \frac{\sigma}{\sqrt{2\pi}} \left[-e^{-t^2/2}\right]_{-\infty}^{\infty} = 0 ,$$

natomajst

$$\int_{-\infty}^{\infty} e^{-t^2/2} dt = \sqrt{2\pi} ,$$

stąd

$$(14) \quad E(X) = m .$$

4.2.2. Twierdzenia o wartości przeciętnej

TWIERDZENIE 1. *Wartość przeciętna stałej równa się tej stałej:*

$$(1) \quad E(C) = C .$$

Dowód. Stałą można traktować jako zmienną losową, która przybiera tylko jedną wartość z prawdopodobieństwem równym jedności, stąd

$$E(C) = C \cdot 1 = C .$$

TWIERDZENIE 2. *Wartość przeciętna sumy dwóch zmiennych losowych X i Y równa się sumie wartości przeciętnych tych zmiennych:*

$$(2) \quad E(X + Y) = E(X) + E(Y) .$$

Dowód. Rozpatrzymy dwa przypadki.

A. Niech X i Y będą zmiennymi losowymi skokowymi. Stosownie do przyjętego zapisu będziemy oznaczać

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

$$\sum_j p_{ij} = p_i , \quad \sum_i p_{ij} = q_j$$

(patrz 3.6.2, wzory (4) i (5)). W takim razie mamy

$$E(X + Y) = \sum_{i=j=1}^{\infty} (x_i + y_j) p_{ij} = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} (x_i + y_j) p_{ij} .$$

Stąd

$$\begin{aligned} E(X+Y) &= \sum_{i=1}^{\infty} x_i \sum_{j=1}^{\infty} p_{ij} + \sum_{j=1}^{\infty} y_j \sum_{i=1}^{\infty} p_{ij} = \\ &= \sum_{i=1}^{\infty} x_i p_i + \sum_{j=1}^{\infty} y_j q_j = E(X) + E(Y). \end{aligned}$$

B. Niech X i Y będą zmiennymi losowymi ciągłymi. Mamy

$$\begin{aligned} E(X+Y) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x+y) f(x, y) dx dy = \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f(x, y) dx dy + \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f(x, y) dx dy = \\ &= \int_{-\infty}^{\infty} x \left[\int_{-\infty}^{\infty} f(x, y) dy \right] dx + \int_{-\infty}^{\infty} y \left[\int_{-\infty}^{\infty} f(x, y) dx \right] dy. \end{aligned}$$

W powyższych wzorach $f(x, y)$ oznacza dwuwymiarową gęstość zmiennych X i Y . Jak wiadomo (patrz 3.6.3, wzór (6)):

$$f_1(x) = \int_{-\infty}^{\infty} f(x, y) dy$$

oznacza brzegową gęstość zmiennej X , natomiast

$$f_2(y) = \int_{-\infty}^{\infty} f(x, y) dx$$

oznacza brzegową gęstość zmiennej Y . Wobec tego

$$E(X+Y) = \int_{-\infty}^{\infty} x f_1(x) dx + \int_{-\infty}^{\infty} y f_2(y) dy = E(X) + E(Y).$$

TWIERDZENIE 3. *Wartość przeciętna sumy dowolnej skończonej liczby zmiennych losowych równa się sumie wartości przeciętnych tych zmiennych.*

Dowód tego twierdzenia, będącego uogólnieniem twierdzenia 2, znajdzie czytelnik w podręczniku [8].

TWIERDZENIE 4. *Wartość przeciętna iloczynu dwóch niezależnych zmiennych losowych równa się iloczynowi wartości przeciętnych tych zmiennych:*

$$(3) \quad E(XY) = E(X) E(Y).$$

Dowód. Rozpatrzymy dwa przypadki.

A. Niech X i Y będą zmiennymi losowymi skokowymi. W takim razie

$$\begin{aligned} E(XY) &= \sum_{i=j=1}^{\infty} x_i y_j p_i q_j = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} x_i y_j p_i q_j = \\ &= \sum_{i=1}^{\infty} x_i p_i \sum_{j=1}^{\infty} y_j q_j = E(X) E(Y). \end{aligned}$$

B. Niech X i Y będą zmiennymi losowymi ciągłyimi. Wobec tego

$$\begin{aligned} E(XY) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xyf(x, y) dx dy = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xyf_1(x)f_2(y) dx dy = \\ &= \int_{-\infty}^{\infty} xf_1(x) dx \int_{-\infty}^{\infty} yf_2(y) dy = E(X)E(Y). \end{aligned}$$

Wniosek. Stała można wynieść przed znak wartości przeciętnej.

Istotnie, na mocy twierdzenia 4

$$E(CX) = E(C)E(X),$$

ponieważ jednak $E(C) = C$, więc

$$(4) \quad E(CX) = CE(X).$$

Oto kilka przykładów zastosowań poznanych twierdzeń o wartości przeciętnej.

PRZYKŁAD 1. Dana jest zmienna losowa $Z = X - E(X)$. Obliczyć wartość przeciętną zmiennej Z .

Mamy

$$E(Z) = E[X - E(X)] = E(X) - E(X) = 0.$$

PRZYKŁAD 2. Zmienna losowa

$$U = X_1 + X_2 + \dots + X_n$$

jest sumą zmiennych losowych, mających jednakową wartość przeciętną m . Znaleźć wartość przeciętną zmiennej losowej U .

Mamy

$$E(U) = E(X_1) + E(X_2) + \dots + E(X_n) = nm.$$

PRZYKŁAD 3. Zmienna losowa $U = X_1 + X_2 + \dots + X_n$ jest sumą n niezależnych zmiennych losowych o rozkładzie zero-jedynkowym mających jednakową wartość przeciętną p . W takim razie

$$E(U) = np$$

jest wartością przeciętną w rozkładzie dwumianowym. Wynik ten otrzymaliśmy już innym sposobem (patrz 4.2.1, wzór (6)).

PRZYKŁAD 4. Na stronie 74 obliczone zostały wartości prawdopodobieństw wygranych w grze liczbowej Toto-Lotek. Prawdopodobieństwa te nie dają jednak możliwości obliczenia nadziei matematycznej wygranej, gdyż wysokość wygranych nie jest stała, lecz zależy od liczby grających i od liczby kuponów podlegających premiowaniu w każdej grupie wygranych; regulamin gry podaje jednak, że z ogólnej sumy pieniędzy uzyskanej ze sprzedaży kuponów połowę przeznacza się na cele społeczne i pokrycie kosztów Państwowego Przedsiębiorstwa Toto-Lotek, połowę zaś wypłaca się w postaci premii uczestnikom gry. Korzystając z tej informacji można łatwo obliczyć nadzieję matematyczną wygranej, gdyż wiedząc, że cena kuponu wynosi 2 zł, otrzymujemy natychmiast, że nadzieję matematyczną wygranej wynosi 1 zł.

4.3. WARIANCJA I ODCHYLENIE STANDARDOWE

4.3.1. Określenia i przykłady

Wariancja i odchylenie standardowe są to parametry opisowe drugiego rodzaju, służące do charakterystyki rozproszenia wartości zmiennej losowej.

Niech X będzie zmienną losową, a $E(X)$ wartością przeciętną tej zmiennej. Różnicę

$$Z = X - E(X)$$

nazywać będziemy *odchyleniem wartości zmiennej losowej od wartości przeciętnej* tej zmiennej.

OKREŚLENIE 1. *Wariancją zmiennej losowej X nazywa się wartość przeciętna kwadratu odchylenia wartości tej zmiennej od wartości przeciętnej $E(X)$:*

$$(1) \quad V(X) = E(Z^2) = E[X - E(X)]^2,$$

gdzie $V(X)$ oznacza wariancję zmiennej losowej X .

Jeśli X jest zmienną losową typu skokowego, mogącą przybierać skońzoną ilość wartości x_1, x_2, \dots, x_n odpowiednio z prawdopodobieństwami p_1, p_2, \dots, p_n , to

$$(2) \quad V(X) = \sum_{i=1}^n [x_i - E(X)]^2 p_i.$$

Jeżeli natomiast X jest skokową zmienną losową, która może przybierać przeliczalny zbiór wartości, to

$$(3) \quad V(X) = \sum_{i=1}^{\infty} [x_i - E(X)]^2 p_i$$

przy założeniu, że szereg stojący po prawej stronie równości (3) jest zbieżny.

Gdy X jest zmienną losową ciągłą, a $f(x)$ jest gęstością prawdopodobieństwa tej zmiennej, to

$$(4) \quad V(X) = \int_{-\infty}^{\infty} [x - E(X)]^2 f(x) dx.$$

Udowodnimy ważne twierdzenie, z którego niejednokrotnie będziemy korzystać przy obliczaniu wariancji.

TWIERDZENIE 1. *Zachodzi związek*

$$(5) \quad V(X) = E(X^2) - E^2(X).$$

Twierdzenie to udowodnimy zakładając, że zmienna losowa X jest zmienną skokową.

Dowód. Mamy

$$\begin{aligned} V(X) &= \sum_i [x_i - E(X)]^2 p_i = \\ &= \sum_i [x_i^2 - 2x_i E(X) + E^2(X)] p_i = \\ &= \sum_i x_i^2 p_i - 2E(X) \sum_i x_i p_i + E^2(X) \sum_i p_i. \end{aligned}$$

Ale

$$\sum_i x_i^2 p_i = E(X^2), \quad \sum_i x_i p_i = E(X)$$

oraz

$$\sum_i p_i = 1.$$

Wobec tego

$$V(X) = E(X^2) - 2E(X)E(X) + E^2(X) = E(X^2) - E^2(X).$$

Wariancja, jako suma iloczynów o nieujemnych czynnikach, jest wielkością nieujemną. Istotnie,

$$V(X) = \sum_i [x_i - E(X)]^2 p_i.$$

Iloczyny, stojące pod znakiem sumy po prawej stronie tej równości, są nieujemne. W takim razie

$$(6) \quad V(X) = E(X^2) - E^2(X) \geq 0.$$

Obok wariancji, jako miary rozproszenia używa się także pierwiastka z wariancji.

OKREŚLENIE 2. Pierwiastek kwadratowy z wariancji nazywa się *odchyleniem standardowym*.

Odchylenie standardowe oznacza się na ogół małą grecką literą σ .

PRZYKŁAD 1. Zmienna losowa X przybiera wartości $x_1 = 10$, $x_2 = 100$, $x_3 = 1000$ odpowiednio z prawdopodobieństwami $p_1 = 0,5$, $p_2 = 0,4$, $p_3 = 0,1$. Obliczyć wariancję i odchylenie standardowe zmiennej X .

W przykładzie 2 z 4.2.1 obliczyliśmy wartość przeciętną $E(X) = 145$. Ponieważ zgodnie ze wzorem (5) wariancja zmiennej X równa się

$$V(X) = E(X^2) - E^2(X),$$

więc pozostaje obliczyć $E(X^2)$.

Oto pomocnicza tabelka, z której korzystamy przy obliczaniu wariancji, jeżeli wartość przeciętna $E(X)$ jest znana:

x_i	x_i^2	p_i	$x_i^2 p_i$
10	100	0,5	50
100	10000	0,4	4000
1000	1000000	0,1	100000
			104050

Zatem

$$V(X) = 104050 - 145^2 = 104050 - 21025 = 83025.$$

Odchylenie standardowe

$$\sigma_x = \sqrt{V(X)} = \sqrt{83025} \approx 288.$$

PRZYKŁAD 2. Zmienna losowa X ma rozkład zero-jedynkowy. Obliczyć wariancję tej zmiennej.

Jak wiadomo z przykładu 3 (patrz 4.2.1), w rozkładzie zero-jedynkowym $E(X)=p$. Zbadamy, czemu równa się $E(X^2)$ w tym rozkładzie. Mamy

$$(7) \quad E(X^2)=1^2 \cdot p + 0^2 \cdot q = p,$$

skąd

$$(8) \quad V(X)=E(X^2)-E^2(X)=p-p^2=p(1-p)=pq.$$

PRZYKŁAD 3. Zmienna losowa X ma rozkład dwumianowy

$$P(X=k)=C_n^k p^k q^{n-k}.$$

Obliczyć wariancję zmiennej X .

W celu obliczenia $V(X)$ zróżniczujemy dwukrotnie względem p obie strony równania

$$(p+q)^n = \sum_{k=0}^n C_n^k p^k q^{n-k}.$$

Otrzymujemy

$$n(n-1)(p+q)^{n-2} = \sum_{k=2}^n C_n^k k(k-1)p^{k-2}q^{n-k}.$$

Mnożąc obie strony tej równości przez p^2 i uwzględniając, że $p+q=1$, otrzymamy dalej

$$n(n-1)p^2 = \sum_k C_n^k k^2 p^k q^{n-k} - \sum_k C_n^k k p^k q^{n-k}.$$

Ale

$$\sum_k C_n^k k^2 p^k q^{n-k} = E(X^2),$$

zaś

$$\sum_k C_n^k k p^k q^{n-k} = E(X) = np$$

(patrz 4.2.1, wzór (6)), skąd

$$E(X^2) = n^2 p^2 - np^2 + np.$$

Wobec tego

$$(9) \quad V(X) = n^2 p^2 - np^2 + np - n^2 p^2 = np - np^2 = np(1-p) = npq.$$

► **PRZYKŁAD 4.** Zmienna losowa ma rozkład hipergeometryczny dany wzorem

$$P(X=k) = \frac{C_R^k C_{N-R}^{n-k}}{C_N^n}.$$

Znaleźć wariancję zmiennej losowej o tym rozkładzie.

Przy obliczaniu $V(X)$ pomocne nam będzie następujące wyrażenie:

$$E(X^2) - E(X) = E(X^2 - X) = E[X(X-1)].$$

Na mocy określenia wartości przeciętnej

$$E[X(X-1)] = \frac{1}{C_N^n} \sum_{k=2}^n k(k-1) C_R^k C_{N-R}^{n-k}.$$

Ponieważ dla $k \geq 2$

$$\begin{aligned} k(k-1) C_R^k &= \frac{k(k-1)R!}{k!(R-k)!} = \frac{R!}{(k-2)![R-2-(k-2)]!} = \\ &= \frac{R(R-1)(R-2)!}{(k-2)![R-2-(k-2)]!} = R(R-1) C_{R-2}^{k-2}, \end{aligned}$$

wobec tego

$$E(X^2) - E(X) = \frac{R(R-1)}{C_N^n} \sum_{k=2}^n C_{R-2}^{k-2} C_{N-R}^{n-k}.$$

Podstawiając $m = k-2$ otrzymujemy

$$E(X^2) - E(X) = \frac{R(R-1)}{C_N^n} \sum_{m=0}^{n-2} C_{R-2}^m C_{N-R}^{(n-2)-m}.$$

Podstawiając $S = R-2$ mamy dalej

$$E(X^2) - E(X) = \frac{R(R-1)}{C_N^n} \sum_{m=0}^{n-2} C_S^m C_{(N-2)-S}^{(n-2)-m}.$$

Na mocy znanego wzoru

$$\sum_{k=0}^n C_R^k C_{N-R}^{n-k} = C_N^n$$

(patrz 4.2.1, wzór (7)) mamy

$$\sum_{m=0}^{n-2} C_S^m C_{(N-2)-S}^{(n-2)-m} = C_{N-2}^{n-2}.$$

Wobec tego

$$E(X^2) - E(X) = \frac{R(R-1)}{C_N^n} C_{N-2}^{n-2} = \frac{R(R-1)n(n-1)}{N(N-1)},$$

a stąd

$$E(X^2) = \frac{R(R-1)n(n-1)}{N(N-1)} + \frac{nR}{N},$$

gdyż

$$E(X) = \frac{nR}{N}$$

(patrz 4.2.1, wzór (10)). W takim razie

$$\begin{aligned}
 (10) \quad V(X) &= E(X^2) - E^2(X) = \\
 &= \frac{R(R-1) n(n-1)}{N(N-1)} + \frac{nR}{N} - \frac{n^2 R^2}{N^2} = \\
 &= \frac{nR}{N^2(N-1)} [N(R-1)(n-1) + N(N-1) - (N-1)nR] = \\
 &= \frac{nR}{N^2(N-1)} [NRn - NR - Nn + N + N^2 - N - NRn + nR] = \\
 &= \frac{nR}{N^2(N-1)} [(N^2 - Nn) + (nR - NR)] = \\
 &= \frac{nR}{N^2(N-1)} [N(N-n) - R(N-n)] = \\
 &= \frac{nR}{N^2(N-1)} (N-R)(N-n).
 \end{aligned}$$

Podstawiając $p=R/N$, $q=(N-R)/N$, otrzymujemy

$$(11) \quad V(X) = npq \frac{1-n/N}{1-1/N}.$$



PRZYKŁAD 5. Zmienna losowa ma rozkład Poissona. Obliczyć wariancję tej zmiennej.

W przykładzie 6 (z 4.2.1) wykazaliśmy, że wartość przeciętna zmiennej losowej o rozkładzie Poissona równa się np . W celu obliczenia $V(X)$ zajmiemy się znalezieniem $E(X^2)$.

Mamy

$$E(X^2) = e^{-np} \sum_{k=0}^{\infty} k^2 \frac{(np)^k}{k!} = e^{-np} \sum_{k=1}^{\infty} k \frac{(np)^k}{(k-1)!}.$$

Jak wiadomo (patrz 4.2.1, przykład 6),

$$e^{np} = \sum_{m=0}^{\infty} \frac{(np)^m}{m!} = \sum_{k=1}^{\infty} \frac{(np)^{k-1}}{(k-1)!},$$

czyli

$$npe^{np} = \sum_{k=1}^{\infty} \frac{(np)^k}{(k-1)!}.$$

Podstawiając w powyższej równości $np=r$ i różniczkując względem r otrzymujemy

$$re^r + e^r = \sum_{k=1}^{\infty} \frac{kr^{k-1}}{(k-1)!}.$$

Pomnożmy obie strony tej równości przez r

$$r^2 e^r + re^r = \sum_{k=1}^{\infty} \frac{kr^k}{(k-1)!}.$$

Korzystając z tego wyniku widzimy więc, że

$$E(X^2) = e^{np} \frac{r^2 + r}{e^{np}} = r^2 + r = (np)^2 + np .$$

Stąd ostatecznie

$$V(X) = E(X^2) - E^2(X) = (np)^2 + np - (np)^2 = np .$$

PRZYKŁAD 6. Zmienna losowa X ma rozkład normalny, którego funkcja gęstości wyraża się wzorem

$$f(x) = \frac{1}{\sigma \sqrt{2\pi}} \exp\left(-\frac{(x-m)^2}{2\sigma^2}\right) .$$

Znaleźć wariancję zmiennej X .

W przykładzie 8 (z 4.2.1, wzór (14)) wykazaliśmy, że $E(X) = m$. Aby można było zastosować wzór $V(X) = E(X^2) - E^2(X)$, musimy zbadać, czemu równa się $E(X^2)$. Mamy

$$E(X^2) = \frac{1}{\sigma \sqrt{2\pi}} \int_{-\infty}^{\infty} x^2 \exp\left(-\frac{(x-m)^2}{2\sigma^2}\right) dx .$$

Wprowadzając zmienną standaryzowaną

$$t = \frac{x-m}{\sigma} , \quad x = t\sigma + m , \quad dx = \sigma dt ,$$

mamy

$$\begin{aligned} E(X^2) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (t\sigma + m)^2 e^{-t^2/2} dt = \\ &= \frac{1}{\sqrt{2\pi}} \left[\sigma^2 \int_{-\infty}^{\infty} t^2 e^{-t^2/2} dt + 2\sigma m \int_{-\infty}^{\infty} t e^{-t^2/2} dt + m^2 \int_{-\infty}^{\infty} e^{-t^2/2} dt \right] = \\ &= \frac{1}{\sqrt{2\pi}} [\sigma^2 I_1 + 2\sigma m I_2 + m^2 I_3] , \end{aligned}$$

gdzie

$$I_1 = \int_{-\infty}^{\infty} t^2 e^{-t^2/2} dt , \quad I_2 = \int_{-\infty}^{\infty} t e^{-t^2/2} dt , \quad I_3 = \int_{-\infty}^{\infty} e^{-t^2/2} dt .$$

Zajmijmy się całką

$$I_1 = \int_{-\infty}^{\infty} t^2 e^{-t^2/2} dt .$$

Całkując przez części otrzymujemy

$$I_1 = \int_{-\infty}^{\infty} t(t e^{-t^2/2}) dt = - \int_{-\infty}^{\infty} t d(e^{-t^2/2}) = -t e^{-t^2/2} \Big|_{-\infty}^{\infty} + \int_{-\infty}^{\infty} e^{-t^2/2} dt = \sqrt{2\pi} ,$$

gdyż stosując wzór de L'Hospitala łatwo przekonać się, że

$$te^{-t^2/2} \Big|_{-\infty}^{\infty} = 0 .$$

W punkcie 4.2.1 wykazaliśmy, że

$$I_2 = 0 , \quad I_3 = \sqrt{2\pi} ,$$

wobec tego

$$E(X^2) = \frac{1}{\sqrt{2\pi}} (\sigma^2 \sqrt{2\pi} + m^2 \sqrt{2\pi}) = \sigma^2 + m^2 .$$

W takim razie

$$V(X) = E(X^2) - E^2(X) = \sigma^2 + m^2 - m^2 = \sigma^2 .$$

PRZYKŁAD 7. Liczne skargi klientów sklepów mięsnych i spożywczych na powolną pracę sprzedawców, powodującą formowanie się długich kolejek, skłoniły kierownictwo tych sklepów do bliższego zbadania tej sprawy. Jak się okazało, oprócz dużych strat czasu, które można było złożyć na karb wybredności i niezdecydowania klientów, bardzo poważną pozycję w przeciętnym czasie obsługi klienta zajmowało ważenie towaru. Sprzedawcy kontrolowani przez wiele par oczu klientów starali się ważyć rzetelnie, ale jednocześnie bojąc się manka usiłowali zredukować „nadwyżki” przysługującą klientom do minimum. Przy ważeniu szynki czy sera powodowało to odcinanie cienkich plasterków i wielokrotnie dokładanie ich bądź odejmowanie od ważonej porcji, przy ważeniu owoców, pomidorów czy ogórków – usuwanie okazów bardziej dorodnych i zastępowanie ich mniejszymi (lub odwrotnie) połączone z mozołnym wyszukiwaniem owocu czy też warzywa o ciężarze nadającym się do najlepszego wyrównania wagi; przy ważeniu mąki lub cukru – ostrożne zsypywanie z szufelki okruchów ważonej substancji do woreczka stojącego na wadze, w obawie aby tych ziarenek nie wpadło ani za dużo, ani za mało.

Aby usprawnić ważenie, kierownictwo przeprowadziło stosowny instruktaż, poinformowało sprzedawców, że dopuszczalna nadwaga wynosi 1% wagi sprzedanego towaru, a dla zniechęcenia sprzedawców do pogoni za osiągalną z tego tytułu superatą wprowadziło system wyróżnień, nagród i premii pieniężnych, uzależniony od liczby załatwionych klientów w okresie szczytowego natężenia ruchu w sklepach. Jednym z kryteriów premiowania sprzedawców była wprawa i wrodzona rzeczność ważenia, którą określało się w drodze doświadczalnej. Jeden z eksperymentów polegał na tym, że każdy ze sprzedawców miał odważyć 100 kilogramowych woreczków napełnionych piaskiem, wykonując wszystkie związane z tym czynności techniczne. Podstawa premiowania byłaby: 1) czas wykonania eksperymentu, 2) wielkość odchylenia przeciętnej wagi woreczka piasku od wymaganej wagi 1 kg, 3) wariancja wagi woreczków. Sprzedawcy, u których wartość przynajmniej jednego z tych parametrów przekraczała dopuszczalny pułap, zostali przesunięci do innej pracy. Podjęte przez kierownictwo kroki zaradcze przyczyniły się do znacznego zwiększenia przepustowości sklepów, zmniejszenia kolejek, zwiększenia wydajności pracy sprzedawców, a jako dodatkową korzyść dały możliwość obiektywnego oceniania przydatności sprzedawców do wykonywanego zawodu.

4.3.2. Twierdzenia o wariancji

TWIERDZENIE 1. *Wariancja stałej równa się zeru:*

(1)

$$V(C) = 0 .$$

Dowód. Ponieważ

$$V(X) = E(X^2) - E^2(X) ,$$

więc

$$V(C) = E(C^2) - E^2(C) = C^2 - C^2 = 0.$$

TWIERDZENIE 2. *Wariancja iloczynu stałej C przez zmienną losową X równa się iloczynowi kwadratu tej stałej przez wariancję zmiennej losowej X :*

$$(2) \quad V(CX) = C^2 V(X).$$

Dowód. Mamy

$$V(CX) = E[(CX - E(CX))^2] = C^2 E(X^2) - C^2 E^2(X) = C^2 V(X).$$

TWIERDZENIE 3. *Wariancja sumy dwóch niezależnych zmiennych losowych równa się sumie wariancji tych zmiennych:*

$$(3) \quad V(X + Y) = V(X) + V(Y).$$

Dowód. Mamy

$$V(X + Y) = E(X + Y)^2 - E^2(X + Y).$$

Ale

$$\begin{aligned} E(X + Y)^2 &= E(X^2 + 2XY + Y^2) = \\ &= E(X^2) + 2E(XY) + E(Y^2) = \\ &= E(X^2) + 2E(X)E(Y) + E(Y^2), \end{aligned}$$

założyliśmy bowiem, że zmienne X i Y są niezależne. Natomiast

$$\begin{aligned} E^2(X + Y) &= E(X + Y)E(X + Y) = [E(X) + E(Y)]^2 = \\ &= E^2(X) + E^2(Y) + 2E(X)E(Y). \end{aligned}$$

Wobec tego

$$\begin{aligned} V(X + Y) &= E(X^2) + E(Y^2) + 2E(X)E(Y) - E^2(X) - E^2(Y) - 2E(X)E(Y) = \\ &= E(X^2) - E^2(X) + E(Y^2) - E^2(Y) = V(X) + V(Y). \end{aligned}$$

TWIERDZENIE 4. *Wariancja sumy n niezależnych zmiennych losowych X_1, X_2, \dots, X_n równa się sumie wariancji tych zmiennych:*

$$(4) \quad V\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n V(X_i).$$

Dowód tego twierdzenia przebiega analogicznie do dowodu twierdzenia 3.

TWIERDZENIE 5. *Wariancja różnicy dwóch niezależnych zmiennych losowych równa się sumie wariancji tych zmiennych*

$$(5) \quad V(X - Y) = V(X) + V(Y).$$

Dowód. Mamy

$$\begin{aligned}
 V(X - Y) &= E(X - Y)^2 - E^2(X - Y) = \\
 &= E(X^2) - 2E(X)E(Y) + E(Y^2) - E^2(X) + 2E(X)E(Y) - E^2(Y) = \\
 &= V(X) + V(Y).
 \end{aligned}$$

Oto parę przykładów na zastosowanie udowodnionych twierdzeń.

PRZYKŁAD 1. Obliczyć wariancję różnicy zmiennej losowej i jej wartości przeciętnej.

Mamy

$$V[X - E(X)] = V(X) + V[E(X)] = V(X).$$

PRZYKŁAD 2. Obliczyć wariancję sumy n niezależnych zmiennych losowych, posiadających ten sam rozkład zero-jedynkowy o parametrze p . Mamy

$$V(X_1 + X_2 + \dots + X_n) = V(X_1) + V(X_2) + \dots + V(X_n) = nV(X).$$

Ale $V(X) = pq$ (patrz 4.3.1, wzór (8)), wobec tego

$$V(X_1 + X_2 + \dots + X_n) = npq.$$

Wynik ten innym sposobem otrzymaliśmy w przykładzie 3 (z 4.3.1). Jest to wariancja zmiennej losowej o rozkładzie dwumianowym.

PRZYKŁAD 3. Obliczyć wariancję i odchylenie standardowe średniej arytmetycznej n niezależnych zmiennych losowych o jednakowym rozkładzie. Mamy

$$\begin{aligned}
 (6) \quad V(\bar{X}) &= V\left(\frac{X_1 + X_2 + \dots + X_n}{n}\right) = \frac{1}{n^2} [V(X_1) + V(X_2) + \dots + V(X_n)] = \\
 &= \frac{1}{n^2} \cdot nV(X) = \frac{V(X)}{n},
 \end{aligned}$$

skąd

$$(7) \quad \sigma(\bar{X}) = \sqrt{\frac{V(X)}{n}} = \frac{\sigma(X)}{\sqrt{n}}.$$

4.4. ODCHYLENIE PRZECIĘTNE

OKREŚLENIE 1. *Odchyleniem przeciętnym* nazywamy wartość przeciętną bezwzględnego odchylenia wartości zmiennej losowej od wartości przeciętnej $E(X)$.

Oznaczając odchylenie przeciętne literą d , otrzymamy

$$(1) \quad d = \sum_{-\infty < x_i < \infty} |x_i - E(X)| p_i,$$

gdy zmienna losowa jest skokowa, lub

$$(2) \quad d = \int_{-\infty}^{\infty} |x - E(X)| f(x) dx,$$

gdy zmienna losowa jest ciągła.

Odchylenie przeciętne, podobnie jak wariancja lub odchylenie standardowe, jest miarą rozproszenia, a tym samym należy do parametrów opisowych drugiego rodzaju.

Rachunki, związane z obliczeniem odchylenia przeciętnego, są na ogół znacznie prostsze od rachunków potrzebnych do znalezienia odchylenia standardowego. Pomimo tej zalety odchylenie przeciętne stosuje się w rachunku prawdopodobieństwa i statystyce matematycznej rzadziej niż odchylenie standardowe. Składa się na to kilka przyczyn, których nie będziemy tu szerzej omawiać. Wskażemy tylko na to, że działania matematyczne na liczbach bezwzględnych nastręczają wiele trudności. I tak na przykład, klasycznymi metodami analitycznymi nie można rozwiązać zagadnień związanych ze znajdowaniem ekstremum funkcji wartości bezwzględnych. Jak wiadomo, funkcja $y=|x|$ nie ma pochodnej w punkcie $x=0$.

Pomimo niedogodności natury teoretycznej związanych z odchyleniem przeciętnym, parametr ten stosuje się jednak w praktyce, gdyż jest intuicyjnie bardziej zrozumiały i łatwiej się go oblicza od odchylenia standardowego.

4.5. MOMENTY

Omówione poprzednio parametry opisowe: wartość przeciętna, wariancja i odchylenie przeciętne należą do wspólnej grupy parametrów, zwanych *momentami*.

Momenty dzielą się na

- 1) momenty absolutne,
- 2) momenty względne

oraz na

- 1) momenty zwykłe,
- 2) momenty centralne.

OKREŚLENIE 1. *Momentem absolutnym rzędu k* nazywa się wartość przeciętną zmiennej losowej $|X-C|^k$, gdzie C oznacza dowolną liczbę rzeczywistą, zwaną *punktem odniesienia*, natomiast k jest liczbą naturalną.

OKREŚLENIE 2. *Momentem względnym rzędu k* lub krótko – *momentem rzędu k* nazywa się wartość przeciętna zmiennej losowej $(X-C)^k$.

OKREŚLENIE 3. Momenty, których punkt odniesienia $C=0$, nazywają się *momentami zwykłymi*. Momenty te oznacza się zwykle symbolem m_k , to znaczy

$$(1) \quad m_k = E(X^k).$$

Zgodnie z tym określeniem

$$(2) \quad m_k = \sum_{-\infty < x_i < \infty} x_i^k p_i,$$

gdy zmienna losowa X jest zmienną skokową, lub

$$(3) \quad m_k = \int_{-\infty}^{\infty} x^k f(x) dx,$$

gdy zmienna jest ciągła.

Omówiona w § 4.2 wartość przeciętna $E(X)$ jest momentem zwykłym pierwszego rzędu. Możemy więc napisać

$$(4) \quad m_1 = E(X^1) = E(X).$$

OKREŚLENIE 4. Momenty, których punkt odniesienia $C = E(X)$, nazywają się *momentami centralnymi*. Momenty centralne oznacza się na ogół symbolem μ_k , to znaczy

$$(5) \quad \mu_k = E[X - E(X)]^k = E(X - m_1)^k.$$

Przykładem momentu centralnego absolutnego jest odchylenie przeciętne. Odchylenie przeciętne jest momentem centralnym pierwszego rzędu. Wariancja natomiast jest przykładem momentu centralnego drugiego rzędu. Istotnie,

$$(6) \quad \mu_2 = E[X - E(X)]^2 = E(X - m_1)^2 = V(X).$$

Oczywiście wariancja jest momentem względnym.

Można udowodnić, że jeśli istnieje moment absolutny rzędu k , to istnieje również moment względny tego samego rzędu. Wynika to z twierdzenia, że każdy szereg bezwzględnie zbieżny jest zbieżny [19].

Między momentami centralnymi i momentami zwykłymi zachodzi ścisły związek. Znając wszystkie momenty m_1, m_2, \dots, m_k możemy moment centralny μ_k wyrazić za pomocą momentów zwykłych. Oto zależności między momentami centralnymi i zwykłymi dla $k = 0, 1, 2$ i 3 :

$$(7) \quad \begin{aligned} \mu_0 &= E(X - m_1)^0 = m_0 = 1, \\ \mu_1 &= E(X - m_1)^1 = m_1 - m_1 = 0, \\ \mu_2 &= E(X - m_1)^2 = m_2 - m_1^2 = V(X), \\ \mu_3 &= E(X - m_1)^3 = m_3 - 3m_2 m_1 + 2m_1^3. \end{aligned}$$

Czytelnik sprawdzi bez trudu, że zależności te można przedstawić za pomocą następującego ogólnego wzoru ([17], str. 50):

$$(8) \quad \mu_k = \sum_{j=0}^k C_k^j m_{k-j} (-m_1)^j.$$

Momenty zostały wprowadzone przez Czebyszewa (patrz § 2.1). Duże znaczenie momentów w statystyce polega na tym, że są one wygodnymi parametrami opisowymi populacji statystycznych. Momenty wyższych rzędów stosuje się do charakterystyki asymetrii i ekscesu rozkładów statystycznych.

Na zakończenie obliczymy dla przykładu centralny moment czwartego rzędu w rozkładzie normalnym.

Mamy

$$E[X - E(X)]^4 = \frac{1}{\sigma \sqrt{2\pi}} \int_{-\infty}^{\infty} (x - m)^4 \exp\left(-\frac{(x-m)^2}{2\sigma^2}\right) dx.$$

Podstawmy

$$\frac{x-m}{\sigma} = t, \quad x - m = t\sigma, \quad dx = \sigma dt,$$

w takim razie

$$E[X - E(X)]^4 = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} t^4 \sigma^5 e^{-t^2/2} dt = \frac{\sigma^4}{\sqrt{2\pi}} \int_{-\infty}^{\infty} t^4 e^{-t^2/2} dt = \sigma^4 \cdot I.$$

Obliczamy

$$\begin{aligned} I &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} t^4 e^{-t^2/2} dt = -\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} t^3 de^{-t^2/2} = -\frac{1}{\sqrt{2\pi}} \left[t^3 e^{-t^2/2} \right]_{-\infty}^{\infty} - \int_{-\infty}^{\infty} e^{-t^2/2} dt^3 = \\ &= \frac{3}{\sqrt{2\pi}} \int_{-\infty}^{\infty} t^2 e^{-t^2/2} dt = -\frac{3}{\sqrt{2\pi}} \int_{-\infty}^{\infty} t de^{-t^2/2} = \frac{-3}{\sqrt{2\pi}} \left[te^{-t^2/2} \right]_{-\infty}^{\infty} - \int_{-\infty}^{\infty} e^{-t^2/2} dt = 3, \end{aligned}$$

gdzie

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-t^2/2} dt = 1.$$

Wobec tego

$$E[X - E(X)]^4 = 3\sigma^4.$$

Postępując analogicznie, czytelnik sprawdzi z łatwością, że moment centralny trzeciego rzędu równa się零. Znajdowanie poszczególnych momentów danego rozkładu teoretycznego nastręcza czasami pewne trudności. Cennym narzędziem, ułatwiającym znalezienie momentów, jest tak zwana funkcja charakterystyczna.

4.6. FUNKCJE CHARAKTERYSTYCZNE

OKREŚLENIE 1. *Funkcją charakterystyczną* zmiennej losowej X nazywa się funkcja

$$(1) \quad \varphi(t) = E(e^{itX}).$$

We wzorze tym t oznacza dowolną liczbę rzeczywistą, natomiast $i = \sqrt{-1}$. Zgodnie z tym określeniem

$$(2) \quad \varphi(t) = \sum_k p_k e^{itx_k},$$

gdy X jest zmienną losową skokową, lub

$$(3) \quad \varphi(t) = \int_{-\infty}^{\infty} e^{itx} dF(x),$$

gdy X jest zmienną losową ciągłą.

Ponieważ (patrz 1.4, wzór (6))

$$(4) \quad e^{itX} = \cos tX + i \sin tX,$$

więc moduł e^{itX} równa się

$$(5) \quad |e^{itX}| = \sqrt{\cos^2 tX + \sin^2 tX} = 1.$$

Na mocy wzoru (4) mamy również

$$(6) \quad \varphi(t) = E(e^{itX}) = E(\cos tX + i \sin tX) = E(\cos tX) + iE(\sin tX)$$

oraz

$$(7) \quad \varphi(-t) = E(\cos tX) - iE(\sin tX).$$

Stąd

$$(8) \quad \varphi(-t) = \overline{\varphi(t)},$$

gdzie $\overline{\varphi(t)}$ jest funkcją sprzężoną z funkcją $\varphi(t)$.

Oczywiście

$$(9) \quad \varphi(0) = E(e^0) = 1.$$

Udowodnimy następujące

TWIERDZENIE 1. *Funkcja charakterystyczna sumy dwóch niezależnych zmiennych losowych X i Y równa się iloczynowi funkcji charakterystycznych tych zmiennych.*

Dowód. W dowodzie wykorzystamy następujący

LEMAT. *Jeżeli U_1 i U_2 są to niezależne zmienne losowe, zaś*

$$V_1 = g_1(U_1) \quad \text{oraz} \quad V_2 = g_2(U_2),$$

to V_1 i V_2 są również niezależnymi zmiennymi losowymi, jeżeli tylko V_1 jest jedno-jednoznaczną funkcją U_1 , a V_2 jest jedno-jednoznaczną funkcją U_2 .

Dowód lematu znaleźć można w pracy [8], str. 65 - 66.

Mamy

$$E[e^{it(X+Y)}] = E(e^{itX} \cdot e^{itY}).$$

Na mocy lematu

$$(10) \quad E(e^{itX} \cdot e^{itY}) = E(e^{itX}) \cdot E(e^{itY}).$$

Metodą indukcji matematycznej można łatwo uogólnić twierdzenie 1 na dowolną skończoną liczbę niezależnych zmiennych losowych X_1, X_2, \dots, X_n .

Funkcja charakterystyczna jest wartościowym narzędziem teoretycznym rachunku prawdopodobieństwa. Funkcja ta oddaje np. cenne usługi przy znajdowaniu momentów. Zauważmy bowiem, że n -ta pochodna funkcji charakterystycznej wyraża się wzorem

$$(11) \quad \varphi^{(n)}(t) = i^n \int_{-\infty}^{\infty} x^n e^{itx} dF(x).$$

Ale

$$\int_{-\infty}^{\infty} x^n dF(x) = m_n,$$

wobec tego, jeśli pochodna ta istnieje, to

$$(12) \quad \varphi^{(n)}(0) = i^n \int_{-\infty}^{\infty} x^n dF(x) = i^n E(X^n) = i^n m_n,$$

skąd

$$(13) \quad m_n = \frac{1}{i^n} \varphi^{(n)}(0).$$

Za pomocą funkcji charakterystycznej łatwo obliczyć wartość przeciętną i wariancję.
Oznaczmy

$$(14) \quad g(t) = \ln \varphi(t).$$

Stąd

$$(15) \quad g'(t) = \frac{\varphi'(t)}{\varphi(t)},$$

$$(16) \quad g''(t) = \frac{\varphi''(t) \varphi(t) - [\varphi'(t)]^2}{\varphi^2(t)}.$$

Na mocy wzoru (15) mamy więc

$$(17) \quad g'(0) = \varphi'(0) = iE(X) = im_1$$

oraz na mocy wzoru (16)

$$(18) \quad g''(0) = \varphi''(0) - [\varphi'(0)]^2 = i^2 [E(X^2) - E^2(X)] = -V(X) = -\mu_2.$$

Stąd

$$(19) \quad E(X) = \frac{g'(0)}{i},$$

$$(20) \quad V(X) = -g''(0).$$

PRZYKŁAD 1. Za pomocą funkcji charakterystycznej znaleźć wartość przeciętną i wariancję w rozkładzie dwumianowym.

W celu znalezienia funkcji charakterystycznej zmiennej X o rozkładzie dwumianowym zauważmy, że zmienna X jest sumą niezależnych zmiennych losowych o rozkładzie zero-jedynkowym.

Funkcja charakterystyczna zmiennej losowej o rozkładzie zero-jedynkowym wyraża się wzorem

$$(21) \quad \varphi(t) = E(e^{itX}) = pe^{it} + q.$$

Na mocy twierdzenia 1 funkcja charakterystyczna zmiennej losowej o rozkładzie dwumianowym ma postać

$$(22) \quad \varphi(t) = (pe^{it} + q)^n.$$

W takim razie

$$\varphi'(t) = n(pe^{it} + q)^{n-1} i p e^{it}.$$

Wobec tego

$$E(X) = \frac{1}{i} \varphi'(0) = np.$$

Znaleźliśmy wartość przeciętną, którą otrzymaliśmy już poprzednio (patrz 4.2.1 i 4.2.2).

W celu znalezienia wariancji obliczmy

$$\begin{aligned} \varphi''(t) &= n(n-1)(pe^{it} + q)^{n-2} i p e^{it} i p e^{it} + n(pe^{it} + q)^{n-1} i^2 p e^{it} = \\ &= -[n(n-1)(pe^{it} + q)^{n-2} p^2 e^{2it} + n(pe^{it} + q)^{n-1} p e^{it}]. \end{aligned}$$

Na mocy wzoru (18)

$$\begin{aligned} g''(0) &= -[n(n-1)p^2 + np] - (inp)^2 = \\ &= -[n^2 p^2 - np^2 + np] + n^2 p^2 = np^2 - np, \end{aligned}$$

skąd

$$V(X) = -g''(0) = np - np^2 = np(1-p) = npq.$$

Otrzymaliśmy znaną nam już (patrz 4.3.1 i 4.3.2) wartość wariancji w rozkładzie dwumianowym.

PRZYKŁAD 2. Znaleźć funkcję charakterystyczną w rozkładzie Poissona.

Mamy

$$(23) \quad \varphi(t) = E(e^{itX}) = \sum_{k=0}^{\infty} e^{itk} \frac{(np)^k}{k!} e^{-np} = e^{-np} \sum_{k=0}^{\infty} \frac{(np e^{it})^k}{k!} = e^{-np} e^{np e^{it}} = e^{np(e^{it}-1)}.$$

PRZYKŁAD 3. Znaleźć funkcję charakterystyczną w rozkładzie prostokątnym.

Mamy

$$(24) \quad \varphi(t) = E(e^{itX}) = \frac{1}{b-a} \int_a^b e^{itx} dx = \frac{e^{itx}}{(b-a)it} \Big|_a^b = \frac{e^{itb} - e^{ita}}{(b-a)it}.$$

► **PRZYKŁAD 4.** Znaleźć funkcję charakterystyczną w rozkładzie normalnym standaryzowanym.

Mamy

$$(25) \quad \varphi(t) = E(e^{itX}) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{itx} e^{-x^2/2} dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(x-it)^2/2} e^{-t^2/2} dx = e^{-t^2/2},$$

gdyż podstawieniem $x - it = u$ sprowadzamy całkę

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(x-it)^2/2} dx$$

do całki

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-u^2/2} du,$$

której wartość, jak wiadomo, równa się jedności. \blacktriangleleft

Uwaga. Podane wyżej przykłady ilustrowały sposób znajdowania funkcji charakterystycznej dla danego rozkładu prawdopodobieństw. Okazuje się, że możliwa jest również operacja odwrotna, polegająca na znalezieniu funkcji rozkładu prawdopodobieństw w oparciu o daną funkcję charakterystyczną. Istnieje twierdzenie, udowodnione przez Lévy'ego (twierdzenie to i dowód znajdzie czytelnik w [8], str. 128 - 130), które głosi, że znając funkcję charakterystyczną jakiegoś rozkładu można znaleźć dystrybuantę i gęstość w tym rozkładzie.

Gdy X jest zmienną losową skokową, przejście od funkcji charakterystycznej do funkcji rozkładu odbywa się za pomocą wzoru

$$(26) \quad p_k = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-itk} \varphi(t) dt.$$

Natomiast gdy X jest zmienną losową typu ciągłego, to

$$(27) \quad F'(x) = f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} \varphi(t) dt.$$

► **PRZYKŁAD 5.** Dana jest funkcja charakterystyczna $\varphi(t) = pe^{it} + q$. Znaleźć rozkład prawdopodobieństw zmiennej losowej X , jeśli zmienna ta jest typu skokowego.

Na podstawie wzoru (26) mamy

$$p_k = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-itk} (pe^{it} + q) dt = \frac{p}{2\pi} \int_{-\pi}^{\pi} e^{it(1-k)} dt + \frac{q}{2\pi} \int_{-\pi}^{\pi} e^{-itk} dt = \frac{p}{2\pi} I_1 + \frac{q}{2\pi} I_2.$$

Ale

$$\begin{aligned} I_1 &= \int_{-\pi}^{\pi} e^{it(1-k)} dt = \left[\frac{e^{(1-k)t}}{(1-k)i} \right]_{-\pi}^{\pi} = \\ &= \frac{e^{(1-k)\pi i} - e^{-(1-k)\pi i}}{(1-k)i} = \frac{2}{1-k} \sin((1-k)\pi) \quad \text{dla } k \neq 1 \end{aligned}$$

(patrz 1.4, wzór (9)), a

$$\begin{aligned} I_2 &= \int_{-\pi}^{\pi} e^{-itk} dt = \left[\frac{e^{-itk}}{-ki} \right]_{-\pi}^{\pi} = \\ &= \frac{e^{-k\pi i} - e^{k\pi i}}{-ki} = \frac{e^{k\pi i} - e^{-k\pi i}}{ki} = \frac{2}{k} \sin k\pi \quad \text{dla } k \neq 0, \end{aligned}$$

wobec tego

$$p_k = \frac{p \sin(1-k)\pi}{(1-k)\pi} + \frac{q \sin k\pi}{k\pi} = 0 \quad \text{dla } k \neq 1 \quad \text{i } k \neq 0.$$

Natomiast dla $k=0$

$$\begin{aligned} p_0 &= \frac{1}{2\pi} \int_{-\pi}^{\pi} (pe^{it} + q) dt = \frac{p}{2\pi} \int_{-\pi}^{\pi} e^{it} dt + \frac{q}{2\pi} \int_{-\pi}^{\pi} dt = \frac{p}{2\pi} \left[\frac{e^{it}}{i} \right]_{-\pi}^{\pi} + \frac{q}{2\pi} [t]_{-\pi}^{\pi} = \\ &= \frac{p}{2\pi} \cdot \frac{e^{i\pi} - e^{-i\pi}}{i} + \frac{q}{2\pi} 2\pi = \frac{p}{2\pi} 2 \sin \pi + q = q, \end{aligned}$$

a dla $k=1$

$$p_1 = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-it} (pe^{it} + q) dt = \frac{1}{2\pi} \int_{-\pi}^{\pi} (qe^{-it} + p) dt = p.$$

Stąd ostatecznie

$$p_0 = q, \quad p_1 = p.$$

Są to prawdopodobieństwa rozkładu zero-jedynkowego.

PRZYKŁAD 6. Dana jest funkcja charakterystyczna $\varphi(t) = e^{-t^2/2}$. Znaleźć gęstość rozkładu zmiennej losowej X wiedząc, że zmienna ta jest ciągła.

Na podstawie wzoru (27) mamy

$$\begin{aligned} f(x) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} e^{-t^2/2} dt = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-(t+ix)^2/2} e^{(ix)^2/2} dt = \\ &= e^{-x^2/2} \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-(t+ix)^2/2} dt = e^{-x^2/2} \frac{\sqrt{2\pi}}{2\pi} = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}. \end{aligned}$$

Znaleziona gęstość prawdopodobieństwa jest gęstością zmiennej losowej standaryzowanej o rozkładzie normalnym (porównaj wzór (4) z 3.5.3).

PRZYKŁAD 7. Znaleźć rozkład sumy dowolnej ilości niezależnych zmiennych losowych, z których każda ma rozkład normalny.

Dana jest zmienna losowa

$$(28) \quad X = X_1 + X_2 + \dots + X_n.$$

Występujące po prawej stronie znaku równości niezależne zmienne losowe mają rozkład normalny o parametrach m_k , σ_k ($k=1, 2, \dots, n$). Na mocy twierdzenia o wartości przeciętnej i wariancji sumy niezależnych zmiennych losowych możemy napisać, że

$$(29) \quad \begin{aligned} m &= m_1 + m_2 + \dots + m_n, \\ \sigma^2 &= \sigma_1^2 + \sigma_2^2 + \dots + \sigma_n^2. \end{aligned}$$

Na podstawie twierdzenia 1 z § 4.6, z którego wynika, że funkcja charakterystyczna sumy niezależnych zmiennych losowych równa się iloczynowi funkcji charakterystycznych tych zmiennych, możemy również napisać⁽¹⁾, że

$$(30) \quad E(e^{itX}) = \prod_{k=1}^n \exp(m_k it - \frac{1}{2}\sigma_k^2 t^2),$$

gdyż funkcja charakterystyczna rozkładu normalnego ma postać

$$(31) \quad e^{mit - \frac{1}{2}\sigma^2 t^2}.$$

Istotnie, ponieważ $\varphi(t) = E(e^{itX})$, więc jeśli $X = aZ + b$, to

$$(32) \quad E[e^{it(aZ+b)}] = E[e^{itaZ+itb}] = E[e^{itb} \cdot e^{itaZ}] = e^{itb} E(e^{itaZ}) = e^{itb} \varphi(at).$$

W przykładzie 4, § 4.6, wykazaliśmy, że funkcja charakterystyczna standaryzowanej zmiennej losowej o rozkładzie normalnym ma postać $e^{-t^2/2}$. Ponieważ jednak $t = (x-m)/\sigma$, a stąd $x = t\sigma + m$, wobec tego na mocy wzoru (32)

$$E(e^{itX}) = E[e^{it(t\sigma+m)}] = e^{itm} \varphi(\sigma r) = e^{itm} \cdot e^{-\sigma^2 r^2/2} = e^{itm - \sigma^2 r^2/2}.$$

W ten sposób wykazaliśmy słuszność wzoru (31). Literę r wprowadziliśmy zamiast parametru funkcji charakterystycznej t , gdyż literę t zarezerwowaliśmy uprzednio na oznaczenie zmiennej standaryzowanej. Oczywiście zastąpienie litery t literą r ma na celu jedynie uniknięcie nieporozumień, które mogłyby powstać w związku z różnymi znaczeniami litery t we wzorach

$$\varphi(t) = E(e^{itX}), \quad t = \frac{x-m}{\sigma}.$$

Wzór (30) na mocy (29) można również przedstawić w prostszej postaci

$$\prod_{k=1}^n \exp(m_k it - \frac{1}{2}\sigma_k^2 t^2) = \exp(mit - \frac{1}{2}\sigma^2 t^2),$$

gdyż mnożenie potęg następuje się dodawaniem ich wykładników. Porównanie strony prawej tej równości ze wzorem (31) pozwala na stwierdzenie, że otrzymaliśmy funkcję

⁽¹⁾ Symbol \prod we wzorze (30) jest znakiem iloczynu, przy czym mnożenie rozciąga się na czynniki ponumerowane liczbami naturalnymi od 1 do n .

charakterystyczną rozkładu normalnego. Wynika stąd, że zmienna losowa (28) ma rozkład normalny.

Zastosowanie funkcji charakterystycznych nie ogranicza się do zagadnienia znajdowania momentów rozkładu. Z funkcji charakterystycznej korzysta się przy dowodzeniu wielu twierdzeń rachunku prawdopodobieństwa, między innymi grupy ważnych twierdzeń znanych pod nazwą twierdzeń granicznych (będą one przedmiotem naszych rozważań w następnym rozdziale).

Na zakończenie uwag o funkcjach charakterystycznych przytoczymy bez dowodu dwa ważne twierdzenia o funkcjach charakterystycznych, tzw. *proste i odwrotne twierdzenie Lévy'ego* (patrz str. 122, uwaga).

TWIERDZENIE 2 (proste). Jeżeli ciąg

$$F_1(x), F_2(x), \dots, F_n(x), \dots$$

dystrybuant zmiennych losowych

$$X_1, X_2, \dots, X_n, \dots$$

jest zbieżny do dystrybuanty $F(x)$, to również ciąg

$$\varphi_1(t), \varphi_2(t), \dots, \varphi_n(t), \dots$$

funkcji charakterystycznych zmiennych losowych $X_1, X_2, \dots, X_n, \dots$ jest zbieżny do funkcji charakterystycznej $\varphi(t)$ zmiennej losowej X .

TWIERDZENIE 3 (odwrotne). Jeżeli ciąg

$$\varphi_1(t), \varphi_2(t), \dots, \varphi_n(t), \dots$$

funkcji charakterystycznych zmiennych losowych

$$X_1, X_2, \dots, X_n, \dots$$

jest zbieżny do funkcji ciąglej $\varphi(t)$, to ciąg

$$F_1(x), F_2(x), \dots, F_n(x), \dots$$

dystrybuant zmiennych losowych $X_1, X_2, \dots, X_n, \dots$ jest zbieżny do dystrybuanty $F(X)$ pewnej zmiennej losowej X , przy czym na mocy twierdzenia prostego $\varphi(t)$ jest funkcją charakterystyczną tej zmiennej losowej.

4.7. MÓMENTY ZMIENNEJ LOSOWEJ DWUWYMIAROWEJ

4.7.1. Regresja pierwszego i drugiego rodzaju⁽¹⁾

OKREŚLENIE 1. Momentem wzajemnym rzędu $l+k$ dwuwymiarowej zmiennej losowej (X, Y) nazywamy wartość oczekiwana zmiennej losowej

$$[(X - C)^l (Y - D)^k].$$

⁽¹⁾ Opracowano na podstawie [15].

Liczby l i k mogą przybierać dowolne wartości ze zbioru liczb całkowitych nieujemnych. Liczby C i D , które nazywać będziemy *współrzędnymi punktu odniesienia*, są dowolnymi liczbami rzeczywistymi.

OKREŚLENIE 2. Momenty, których współrzędne punktu odniesienia $C=D=0$, nazywają się *momentami zwykłymi*. Momenty te oznacza się na ogół symbolem m_{lk} .

Zgodnie z tym określeniem

$$(1) \quad m_{lk} = E(X^l Y^k) = \lim_{\substack{a \rightarrow -\infty \\ b \rightarrow +\infty}} \lim_{\substack{c \rightarrow -\infty \\ d \rightarrow +\infty}} \sum_{a < x_l < b} \sum_{c < y_j < d} x_i^l y_j^k p_{ij} = \\ = \sum_i \sum_j x_i^l y_j^k p_{ij},$$

gdy zmienna (X, Y) jest skokowa, oraz

$$(2) \quad m_{lk} = E(X^l Y^k) = \lim_{\substack{a \rightarrow -\infty \\ b \rightarrow +\infty}} \lim_{\substack{c \rightarrow -\infty \\ d \rightarrow +\infty}} \int_a^b \int_c^d x^l y^k f(x, y) dx dy = \\ = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x^l y^k f(x, y) dx dy,$$

gdy zmienna (X, Y) jest ciągła.

Najczęściej korzysta się z momentów pierwszego i drugiego rzędu. Momentami pierwszego rzędu są wartości oczekiwane zmiennych losowych X i Y :

$$(3) \quad m_{10} = E(X^1 Y^0) = E(X)$$

oraz

$$(4) \quad m_{01} = E(X^0 Y^1) = E(Y).$$

Moment drugiego rzędu określony wzorem

$$(5) \quad m_{11} = E(XY)$$

nazywa się *momentem mieszanym*.

Pozostałe dwa momenty drugiego rzędu wyrażają się wzorami

$$(6) \quad m_{20} = E(X^2 Y^0) = E(X^2),$$

$$(7) \quad m_{02} = E(X^0 Y^2) = E(Y^2).$$

OKREŚLENIE 3. Momenty, których współrzędne punktu odniesienia $C=E(X)$ oraz $D=E(Y)$, nazywamy *momentami centralnymi*.

Momenty centralne oznacza się zwykle symbolem μ_{lk} . Mamy więc

$$(8) \quad \mu_{lk} = E[(X - m_{10})^l (Y - m_{01})^k].$$

Oczywiście

$$(9) \quad \mu_{10} = E[(X - m_{10})^1 (Y - m_{01})^0] = 0$$

oraz

$$(10) \quad \mu_{01} = E[(X - m_{10})^0 (Y - m_{01})^1] = 0.$$

Duże znaczenie w dalszych rozważaniach będą miały trzy momenty centralne drugiego rzędu.

Momenty

$$(11) \quad \mu_{20} = E[(X - m_{10})^2] = V(X)$$

i

$$(12) \quad \mu_{02} = E[(Y - m_{01})^2] = V(Y)$$

są to wariancje zmiennej losowej X i zmiennej losowej Y . Mieszany moment centralny drugiego rzędu

$$(13) \quad \mu_{11} = E[(X - m_{10})(Y - m_{01})]$$

znany jest pod nazwą *kowariancji*. Kowariancję oznacza się często symbolem $C(X, Y)$.

Momenty centralne dwuwymiarowej zmiennej losowej dają się wyrazić za pomocą momentów zwykłych – i na odwrót. Łatwo wykazać na przykład, że

$$(14) \quad \mu_{11} = m_{11} - m_{10} m_{01}.$$

Rzeczywiście,

$$\begin{aligned} \mu_{11} &= E[(X - m_{10})(Y - m_{01})] = \\ &= E(XY) - m_{10} E(Y) - m_{01} E(X) + m_{10} m_{01} = \\ &= m_{11} - m_{10} m_{01} - m_{01} m_{10} + m_{10} m_{01} = m_{11} - m_{10} m_{01}. \end{aligned}$$

Podobnie dowodzi się, że

$$(15) \quad \mu_{20} = m_{20} - m_{10}^2$$

oraz

$$(16) \quad \mu_{02} = m_{02} - m_{01}^2.$$

Dla kowariancji można udowodnić następujące ważne twierdzenie:

TWIERDZENIE 1. Jeżeli zmienne losowe X i Y są niezależne, to kowariancja $C(X, Y)$ tych zmiennych jest równa zeru.

Dowód ⁽¹⁾. Mamy

$$C(X, Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - m_{10})(y - m_{01}) f(x, y) dx dy.$$

⁽¹⁾ Dowód przeprowadzono dla zmiennych ciągłych. Przebiega on podobnie, gdy zmienne są skokowe.

Gdy zmienne losowe są niezależne, to

$$f(x, y) = f_1(x)f_2(y).$$

Wobec tego

$$\begin{aligned} C(X, Y) &= \int_{-\infty}^{\infty} (x - m_{10}) f_1(x) dx \int_{-\infty}^{\infty} (y - m_{01}) f_2(y) dy = \\ &= E(X - m_{10}) E(Y - m_{01}) = \mu_{10} \mu_{01}. \end{aligned}$$

Na mocy wzorów (9) i (10) otrzymujemy więc

$$C(X, Y) = 0.$$

Twierdzenie odwrotne jest nieprawdziwe.

Obok momentów zdefiniowanych dotychczas istnieje jeszcze jedna grupa momentów. Są to tzw. *momenty warunkowe*. Terminem tym określa się momenty jednej ze zmiennych X, Y przy założeniu, że pozostała zmienna przybrała pewną określoną wartość.

W dalszych rozważaniach będziemy korzystać z dwóch momentów warunkowych, a mianowicie z *warunkowej wartości oczekiwanej i warunkowej wariancji*. Jeżeli zmienna (X, Y) jest ciągła, parametry te wyrażają się odpowiednio wzorami

$$(17) \quad E(Y|X=x) = m_{01}(x) = \frac{\int_{-\infty}^{\infty} y f(x, y) dy}{\int_{-\infty}^{\infty} f(x, y) dy} = \int_{-\infty}^{\infty} y f(y|x) dy,$$

$$\begin{aligned} (18) \quad V(Y|X=x) &= \frac{\int_{-\infty}^{\infty} [y - m_{01}(x)]^2 f(x, y) dy}{\int_{-\infty}^{\infty} f(x, y) dy} = \\ &= \int_{-\infty}^{\infty} [y - m_{01}(x)]^2 f(y|x) dy. \end{aligned}$$

Są to momenty warunkowe zmiennej Y . Analogiczną parę wzorów można napisać dla zmiennej X .

Gdy zmienne X i Y są niezależne, to

$$m_{01}(x) = \frac{\int_{-\infty}^{\infty} y f_1(x)f_2(y) dy}{\int_{-\infty}^{\infty} f_1(x)f_2(y) dy} = \int_{-\infty}^{\infty} y f_2(y) dy = \mu_{01}.$$

W dwuwymiarowym rozkładzie zmiennej (X, Y) warunkowa wartość oczekiwana $E(Y|X=x)$ jest jakąś funkcją zmiennej X . Możemy więc napisać

$$(19) \quad E(Y|X=x) = g_1(x).$$

Zastępując w powyższym wzorze $E(Y|X=x)$ prostszym symbolem y otrzymamy

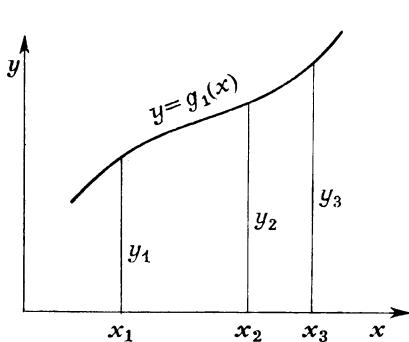
$$(20) \quad y = g_1(x).$$

Równanie (20) jest znane w statystyce matematycznej pod nazwą *równania regresji pierwszego rodzaju* zmiennej Y względem zmiennej X .

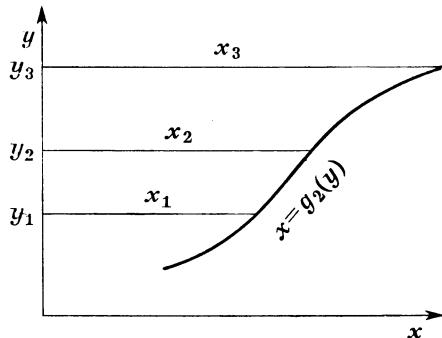
Przez zamianę liter x i y we wzorach (19) i (20) otrzymuje się równanie regresji zmiennej losowej X względem zmiennej Y

$$(21) \quad x = g_2(y).$$

Jeżeli zmienna (X, Y) jest ciągła, to obrazami geometrycznymi funkcji (20) i (21) są pewne linie. Linie te noszą nazwę *linii regresji pierwszego rodzaju*.



Rys. 1



Rys. 2

Na rysunku 1 przedstawiona jest linia regresji Y względem X . Rzędne tej krzywej przedstawiają wartości oczekiwane zmiennej Y przy założeniu, że zmienna $X=x$. Jeżeli równanie tej krzywej jest znane, to każdej wartości zmiennej X umiemy przyporządkować wartość oczekiwana zmiennej Y .

Rysunek 2 wyobraża linię regresji X względem Y .

TWIERDZENIE 2. *Linie regresji pierwszego rodzaju mają tę własność, że wartość oczekiwana kwadratu odchylenia wartości zmiennej Y od tej linii jest minimum.*

Dowód. Mamy wykazać, że

$$E[Y - g(x)]^2 = \min.$$

Ale

$$\begin{aligned} E[Y - g(x)]^2 &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} [y - g(x)]^2 f(x, y) dx dy = \\ &= \int_{-\infty}^{\infty} f_1(x) dx \int_{-\infty}^{\infty} [y - g(x)]^2 f(y|x) dy. \end{aligned}$$

Jak wiadomo, $E(Y-u)^2$ osiąga minimum, gdy $u=E(Y)$. Wtedy

$$E(Y-u)^2 = V(Y).$$

Stąd, aby wartość $E[Y - g(x)]^2$ była minimalna, trzeba, aby

$$g(x) = E(Y|X=x),$$

wtedy bowiem

$$\int_{-\infty}^{\infty} [y - E(Y|X=x)]^2 f(y|x) dy$$

równa się $V(Y|X=x)$, to znaczy osiąga wartość minimalną. Wobec tego

$$\begin{aligned} E[Y - g(x)]^2 &= \int_{-\infty}^{\infty} f_1(x) dx \int_{-\infty}^{\infty} [y - m_{01}(x)]^2 f(y|x) dy = \\ &= V(Y|X=x) = \min. \end{aligned}$$

W praktyce zdarza się stosunkowo rzadko, aby znana była postać funkcji $g(x)$. Na ogół postępuje się więc w ten sposób, że z populacji dwuwymiarowej pobiera się próbki i sporządza punktowy wykres rozrzutu. Punkty na wykresie tworzą mniej lub bardziej wyraźną smugę. Ta smuga punktów stanowi informację, w oparciu o którą wysuwa się hipotezę, że funkcja $g(x)$ należy do określonej klasy funkcji (np. do klasy funkcji liniowych, wykładniczych, potęgowych lub do klasy wielomianów).

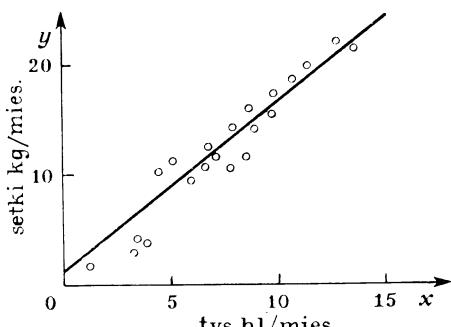
Rysunek 3 przedstawia punktowy wykres rozrzutu, sporządzony w oparciu o materiał liczbowy, zebrany w związku z badaniem zależności między zużyciem chmielu a produkcją brzeczki. Materiał liczbowy pochodzi z Browaru Piastowskiego we Wrocławiu.

Smuga punktów na wykresie zarysowana jest tak wyraźnie, że można śmiało wysunąć hipotezę, iż $g(x)$ należy do zbioru funkcji liniowych.

W celu wyznaczenia parametrów funkcji $g(x)$ minimizuje się wyrażenie $E[Y - g(x)]^2$.

Jeżeli $g(x) \equiv \alpha x$, to dla znalezienia wartości parametru α należy obliczyć wartość minimalną wyrażenia $S \equiv E[Y - \alpha x]^2$. W tym celu oblicza się pochodną $\partial S / \partial \alpha$ i przyrównuje ją do zera. Z otrzymanego w ten sposób równania wyznacza się α . Tak wyznaczona linia nazywa się *linią regresji drugiego rodzaju*. Jeżeli hipoteza dotycząca klasy funkcji, do której należy $g(x)$, jest prawdziwa, to linia regresji drugiego rodzaju pokrywa się z linią regresji pierwszego rodzaju. W zastosowaniach interesują nas zawsze linie regresji pierwszego rodzaju. Ponieważ na ogół nie znamy równań tych linii, przeto linie regresji pierwszego rodzaju zastępujemy łatwiejszymi do wyznaczenia liniami regresji drugiego rodzaju.

Postępując w taki sposób rzadko jesteśmy wolni od obaw, czy linia regresji wyznaczona została właściwie, gdyż informacja, jakiej dostarcza punktowy wykres rozrzutu, jest skąpa, a więc hipoteza dotycząca klasy funkcji zawierającej $g(x)$ może łatwo okazać się fałszywa.



Rys. 3

Zdarza się niekiedy, że poza wykresem rozrzutu dysponujemy innymi jeszcze informacjami, dającymi podstawę do wysunięcia hipotezy o klasie funkcji, w skład której wchodzi $g(x)$. I tak np. znamy czasami równania asymptot linii regresji lub wiemy, że linia ta przechodzi przez początek układu, albo że nie przecina dodatniej części osi X i ujemnej części osi Y . Informacje takie są bardzo cenne. Pochodzą one zawsze ze źródeł pozastatystycznych. Jak wiadomo, jednym z warunków efektywnego badania statystycznego jest dokładna znajomość przedmiotu badania i tej dziedziny wiedzy, która się nim zajmuje. Oznacza to na przykład, że badając metodami statystycznymi skuteczność penicyliny w zwalczaniu gruźlicy trzeba być ftyzjologiem, natomiast badając wpływ ceny masła na wielkość spożycia tłuszczów jadalnych trzeba być ekonomistą. Sama znajomość statystyki jest niewystarczająca. Tylko połączenie informacji pozastatystycznych i statystycznych prowadzi do owocnych wyników badania.

Zasada ta obowiązuje w pełni przy wyznaczaniu linii regresji.

OKREŚLENIE 4. Jeżeli równanie regresji pierwszego rodzaju wyraża się wzorem

$$(22) \quad y = \alpha_{21} x + \beta_{20},$$

to mówimy, że regresja Y względem X jest *liniowa*.

Wzór (22) jest równaniem regresji Y względem X . Wielkości α_{21} i β_{20} są to pewne stałe, zwane *parametrami regresji*. Indeksy stojące obok parametrów służą do odróżniania parametrów regresji Y względem X od parametrów regresji X względem Y . Pierwszy indeks wskazuje zmienną zależną w równaniu regresji, drugi natomiast – zmienną niezależną.

Liniowe równanie regresji X względem Y wyraża się wzorem

$$(23) \quad x = \alpha_{12} y + \beta_{10}.$$

Czasami będziemy pisali równanie (22) z pominięciem indeksów, to znaczy

$$(24) \quad y = \alpha x + \beta;$$

należy wtedy rozumieć, że rozważania dotyczą obu linii regresji, tak regresji Y względem X , jak i regresji X względem Y .

Jeżeli wiadomo, że w rozkładzie dwuwymiarowej zmiennej losowej (X, Y) linie regresji są prostymi, to dla wyznaczenia wartości parametrów α i β należy zminimizować wyrażenie

$$E [Y - \alpha X - \beta]^2 = \int_R [y - \alpha x - \beta]^2 dP,$$

gdzie R^2 oznacza dwuwymiarową przestrzeń całkowania, a dP – różniczkę dwuwymiarowego rozkładu prawdopodobieństwa.

Obliczymy pochodne cząstkowe wyrażenia w nawiasach po lewej stronie względem α i β .

Mamy

$$\frac{\partial}{\partial \alpha} E(Y - \alpha X - \beta)^2 = -2E[(Y - \alpha X - \beta)X]$$

oraz

$$\frac{\partial}{\partial \beta} E(Y - \alpha X - \beta)^2 = -2E(Y - \alpha X - \beta).$$

Przyrównując obliczone pochodne do zera, otrzymujemy tzw. *układ równań normalnych*

$$(25) \quad \begin{aligned} E(XY - \alpha X^2 - \beta X) &= 0, \\ E(Y - \alpha X - \beta) &= 0. \end{aligned}$$

Układ ten po zastąpieniu wartości oczekiwanych odpowiednimi momentami można zapisać następująco:

$$(26) \quad \begin{aligned} m_{11} - \alpha m_{20} - \beta m_{10} &= 0, \\ m_{01} - \alpha m_{10} - \beta &= 0. \end{aligned}$$

Z rozwiązania układu równań (26) otrzymujemy

$$(27) \quad \beta = \beta_{20} = m_{01} - \alpha_{21} m_{10},$$

$$(28) \quad \alpha = \alpha_{21} = \frac{m_{11} - m_{01} m_{10}}{m_{20} - m_{10}^2}.$$

Wzór (28) można napisać inaczej, a mianowicie

$$(29) \quad \alpha = \alpha_{21} = \frac{\mu_{11}}{\mu_{20}}.$$

Podobnie otrzymuje się

$$(30) \quad \beta_{10} = m_{10} - \alpha_{12} m_{01}$$

oraz

$$(31) \quad \alpha_{12} = \frac{\mu_{11}}{\mu_{02}}.$$

Parametry α_{21} i α_{12} nazywają się *współczynnikami regresji*.

Wstawiając wyrażenie (27) do wzoru (22) i (30) do (23) można przedstawić równania regresji następująco:

$$(32) \quad y = \alpha_{21}(x - m_{10}) + m_{01},$$

$$(33) \quad x = \alpha_{12}(y - m_{01}) + m_{10}.$$

Z powyższych równań wynika, że obie proste regresji przechodzą przez punkt o współrzędnych (m_{10}, m_{01}) . Punkt ten nazywać będziemy *środkiem ciężkości populacji*.

Znajomość równań regresji pozwala wyrazić zależność między zmiennymi losowymi za pomocą funkcji opisującej liczbowo związek, jaki łączy ze sobą te zmienne. Wprowadzenie wzoru funkcyjnego jest bardzo wygodne, gdyż pozwala każdej wartości zmiennej losowej występującej w roli argumentu przyporządkować stosowną

wartość pozostałą zmiennej, występującej w roli funkcji. Znaczenie równań linii regresji polega właśnie na tym, że dają one możliwość oceny poszczególnych wartości jednej zmiennej w oparciu o wartości, jakie przybiera druga zmienna. Ocena ta może być gorsza lub lepsza, bardziej dokładna lub mniej dokładna. Używając przeto pojęcia „ocena”, należy w ślad za tym wprowadzić pojęcie „dokładności oceny” i skonstruować miarę tej dokładności.

Oczywiście ocena będzie tym lepsza, a jej dokładność tym większa, im suma bezwzględnych wartości błędów, jakie popełnimy zastępując rzeczywiste wartości zmiennej losowej wartościami otrzymanymi z równania linii regresji, okaże się mniejsza. Zdanie to posiada interpretację geometryczną. Ocena będzie tym lepsza, im bardziej na wykresie rozrzutu punkty skupiają się wokół linii regresji lub, co na jedno wychodzi, im rozproszenie punktów wokół linii jest mniejsze.

Nazwijmy *i-tą resztką regresji Y względem X* lub krótko *resztką* wielkość zdefiniowaną wzorem

$$z_i = y_i - g_1(x_i),$$

która jest realizacją zmiennej losowej $Z = Y - g_1(x)$ zwanej *błędem losowym* lub *błędem resztkowym regresji Y względem X*.

Jako miary rozproszenia punktów wokół linii regresji używa się zwykle tzw. *wariancji resztkowej* $V(Z)$ określonej wzorem

$$(34) \quad V(Z) = E(Z^2) = E(Y - y)^2.$$

Jeśli linia regresji jest linią prostą, to

$$\begin{aligned} V(Z) &= E[(Y - (\alpha_{21}X + \beta_{20}))^2] = E[Y - m_{01} + \alpha_{21}m_{10} + \beta_{20} - \alpha_{21}X - \beta_{20}]^2 = \\ &= E[(Y - m_{01}) - \alpha_{21}(X - m_{10})]^2 = E(W - \alpha_{21}U)^2, \end{aligned}$$

gdzie $W = Y - m_{01}$ oraz $U = X - m_{10}$. I dalej

$$\begin{aligned} V(Z) &= E(W^2 - 2\alpha_{21}WU + \alpha_{21}^2U^2) = E(W^2) - 2\alpha_{21}E(WU) + \alpha_{21}^2E(U^2) = \\ &= \mu_{02} - 2\alpha_{21}\mu_{11} + \alpha_{21}^2\mu_{20}; \end{aligned}$$

stąd

$$(35) \quad V(Z) = \mu_{02} - 2\alpha_{21}\mu_{11} + \alpha_{21} \frac{\mu_{11}}{\mu_{20}} \mu_{20} = \mu_{02} - \alpha_{21}\mu_{11}.$$

Posługując się symbolami wariancji i kowariancji można przedstawić wzór (35) w postaci

$$(36) \quad V(Z) = V(Y) - \alpha_{21}C(X, Y) = V(Y) - \alpha_{21}^2V(X).$$

Analogicznie można wykazać, że jeżeli $Z = X - g_2(y)$, to

$$(37) \quad V(Z) = \mu_{20} - \alpha_{12}\mu_{11}.$$

Dla odróżnienia błędu losowego regresji Y względem X od błędu losowego regresji X względem Y wprowadzimy oznaczenia

$$Z_{21} = Y - g_1(x), \quad Z_{12} = X - g_2(y).$$

Pierwiastek kwadratowy z wariancji resztowej nazywać będziemy *standardowym błędem oceny*⁽¹⁾ lub *średnim błędem resztowym* i oznaczać go symbolem σ_{21} i σ_{12} odpowiednio dla regresji Y względem X i regresji X względem Y . W takim razie

$$(38) \quad \sigma_{21} = \sqrt{\mu_{02} - \alpha_{21} \mu_{11}}$$

oraz

$$(39) \quad \sigma_{12} = \sqrt{\mu_{20} - \alpha_{12} \mu_{11}}.$$

4.7.2. Korelacja. Stosunek koreacyjny i współczynnik korelacji

Zastosowanie wariancji resztowej nie ogranicza się do mierzenia wyłącznie rozrzutu punktów dookoła linii regresji. Zauważmy bowiem, że im mniejszy jest ten rozrzułt, tym ściślejsza jest więź między zmiennymi losowymi X i Y . Gdy wszystkie punkty leżą na linii regresji, rozrzutu nie ma wcale i $V(Z_{21})=0$. W takim przypadku można już mówić nie o zależności między zmiennymi losowymi X i Y , lecz o zwykłym związku funkcyjnym. Wobec tego wielkość $V(Z_{21})$ można zastosować do mierzenia siły zależności między dwiema zmiennymi losowymi. Wariancja resztowa rzeczywiście ma takie zastosowanie, jakkolwiek nie w postaci określonej definicją.

Miara zależności między zmiennymi losowymi powinna bowiem spełniać następujące warunki:

1. powinna być wielkością niemianowaną;
2. powinna być wielkością unormowaną, przybierającą wartości należące do pewnego skończonego przedziału liczbowego;
3. powinna przybierać wartości rosnące, gdy zależność przybiera na sile, a malejące, gdy zależność słabnie;
4. nie powinna zależeć od tego, czy mierzy się zależność X od Y , czy Y od X .

Żadnego z tych warunków $V(Z_{21})$ nie spełnia. Za pomocą kilku prostych operacji matematycznych można jednak skonstruować wielkość o żądanych własnościach. Do spełnienia warunków 1 i 2 wystarczy podzielić $V(Z_{21})$ przez $V(Y)$ (lub $V(Z_{12})$ przez $V(X)$). Istotnie, ponieważ $V(Z_{21})$ i $V(Y)$ mają jednakowe miano, przeto

$$\frac{V(Z_{21})}{V(Y)}$$

jest wielkością niemianowaną.

Ponieważ

$$0 \leq \frac{V(Z_{21})}{V(Y)} \leq 1,$$

więc wielkość ta jest unormowana w przedziale $\langle 0, 1 \rangle$.

(1) *Standard error of estimation.*

Warunek 3 zostanie spełniony, jeśli zamiast $\frac{V(Z_{21})}{V(Y)}$ i $\frac{V(Z_{12})}{V(X)}$ wprowadzimy wielkości

$$(1) \quad \eta_{21}^2 = 1 - \frac{V(Z_{21})}{V(Y)},$$

$$(2) \quad \eta_{12}^2 = 1 - \frac{V(Z_{12})}{V(X)}.$$

Wielkości η_{21} i η_{12} zdefiniowane wzorami (1) i (2) znane są pod nazwą *stosunków koreacyjnych*⁽¹⁾. Oczywiście

$$(3) \quad 0 \leq \eta_{21}^2 \leq 1.$$

Stosunek koreacyjny η_{21} przybiera wartość równą jedności wtedy i tylko wtedy, gdy $V(Z_{21})=0$, tzn. gdy zależność łącząca zmienne losowe X i Y jest zależnością funkcyjną.

Gdy

$$\eta_{21} \neq 0,$$

to o zmiennych losowych mówi się, że są ze sobą *skorelowane*. Gdy

$$\eta_{21} = 0,$$

mówimy, że zmienne są *nieskorelowane*. Wszystkie uwagi wypowiadane o stosunku koreacyjnym η_{21} dotyczą również η_{12} .

Brak więzi koreacyjnej między zmiennymi losowymi nie oznacza bynajmniej, że zmienne te są niezależne. Jak wiadomo, zmienne losowe X i Y są niezależne, jeżeli

$$(4) \quad F(x, y) = F_1(x)F_2(y),$$

natomiast zmienne te są nieskorelowane, gdy

$$(5) \quad \eta_{21} = 0 \quad \text{lub} \quad \eta_{12} = 0.$$

Relacje (4) i (5) nie są równoważne.

Dla uczynienia zadość warunkowi 4, przyjmiemy założenie, że obie linie regresji pierwszego rodzaju są liniami prostymi. W takim razie

$$V(Z_{21}) = V(Y) - \alpha_{21} C(X, Y)$$

i analogicznie

$$V(Z_{12}) = V(X) - \alpha_{12} C(X, Y).$$

Stąd

$$\eta_{21}^2 = 1 - \frac{V(Y) - \alpha_{21} C(X, Y)}{V(Y)} = \alpha_{21} \frac{C(X, Y)}{V(Y)} = \alpha_{21} \frac{\mu_{11}}{\mu_{02}} = \alpha_{21} \alpha_{12}.$$

(¹) Miarę tę wprowadził K. Pearson.

Podobnie

$$\eta_{12}^2 = 1 - \frac{V(X) - \alpha_{12} C(X, Y)}{V(X)} = \alpha_{12} \frac{C(X, Y)}{V(X)} = \alpha_{12} \frac{\mu_{11}}{\mu_{20}} = \alpha_{12} \alpha_{21}.$$

Zatem gdy linie regresji są liniami prostymi, to

$$\eta_{21} = \eta_{12}.$$

Wielkość

$$(6) \quad \rho = \pm \sqrt{\alpha_{12} \alpha_{21}}$$

nazywamy *współczynnikiem korelacji*. Ponieważ współczynnik korelacji jest szczególnym przypadkiem stosunku korelacyjnego, więc na mocy wzoru (3) mamy również

$$\rho^2 \leq 1, \quad \text{czyli} \quad -1 \leq \rho \leq 1.$$

Gdy $\rho > 0$, mówimy, że między zmiennymi losowymi X i Y zachodzi *korelacja dodatnia*. W przypadku korelacji dodatniej wzrostowi wartości jednej zmiennej odpowiada wzrost wartości oczekiwanych drugiej zmiennej. Zauważmy bowiem, że

$$\rho = \pm \sqrt{\alpha_{12} \alpha_{21}} = \frac{\mu_{11}}{\pm \sqrt{\mu_{20} \mu_{02}}}.$$

Jak widać, znak współczynnika korelacji zależy od znaku μ_{11} .

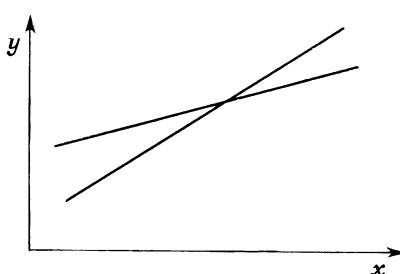
Z drugiej strony

$$\alpha_{12} = \frac{\mu_{11}}{\mu_{02}}, \quad \alpha_{21} = \frac{\mu_{11}}{\mu_{20}}.$$

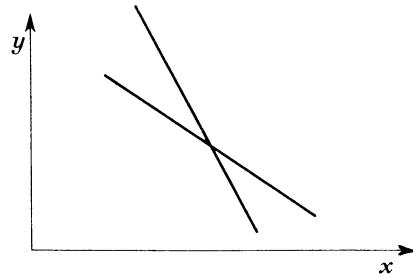
Ponieważ $\mu_{20} \geq 0$ i $\mu_{02} \geq 0$, więc znaki α_{12} i α_{21} zależą również od znaku μ_{11} . Stąd gdy $\rho > 0$, to $\alpha_{12} > 0$ i $\alpha_{21} > 0$. Ale α_{12} i α_{21} są parametrami kierunkowymi prostych regresji. Wobec tego, gdy $\rho > 0$, to prosta regresji $y = \alpha_{21}x + \beta_{20}$ tworzy z osią odciętych kąt ostry, czyli y jest funkcją rosnącą x .

Gdy $\rho < 0$, mamy do czynienia z *korelacją ujemną* i wtedy wzrostowi wartości jednej zmiennej towarzyszy zmniejszanie się warunkowych wartości oczekiwanych drugiej zmiennej.

Położenie linii regresji w przypadku korelacji dodatniej przedstawione jest na rysunku 1, natomiast położenie tych linii w przypadku korelacji ujemnej wyobrażone jest na rysunku 2.



Rys. 1



Rys. 2

Współczynnik korelacji równa się +1 lub -1 wtedy i tylko wtedy, gdy wszystkie punkty (x, y) leżą na linii prostej.

TWIERDZENIE 1. *Gdy $\rho^2=1$, obie linie regresji pokrywają się.*

Dowód. Założymy, że $\rho=+1$ (dowód przebiega analogicznie, jeśli założymy, że $\rho=-1$). W takim razie z definicji współczynnika korelacji wynika, że

$$\rho = \alpha_{12} \alpha_{21} = 1.$$

Parametr α_{12} jest tangensem kąta nachylenia prostej regresji $x = \alpha_{12}y + \beta_{10}$ do osi rzędnych. Współczynnik kątowy tej prostej, wzięty w odniesieniu do osi odciętych, równa się $1/\alpha_{12}$. Współczynnik kątowy prostej regresji $y = \alpha_{21}x + \beta_{20}$ względem osi odciętych jest równy α_{21} . Tangens kąta ψ zawartego między obu prostymi wynosi zatem

$$\frac{\frac{1}{\alpha_{12}} - \alpha_{21}}{1 + \frac{\alpha_{21}}{\alpha_{12}}} = \frac{1 - \alpha_{12} \alpha_{21}}{\alpha_{12} + \alpha_{21}} = 0.$$

Ponieważ tangens kąta równa się zeru, gdy kąt równa się zeru, przeto otrzymany wynik wskazuje, że proste pokrywają się, co należało okazać.

Gdy współczynnik korelacji równa się zeru, zmienne losowe X i Y są ze sobą nieskorelowane. Można łatwo wykazać (dowód przebiega tak samo jak dowód twierdzenia 1), że gdy $\rho=0$, to kąt ψ między prostymi regresji równa się $\pi/2$. Zauważmy dalej, że współczynnik korelacji równa się zeru wtedy i tylko wtedy, gdy $C(X, Y)=0$. Ponieważ jednak

$$\alpha_{12} = \frac{C(X, Y)}{V(Y)}, \quad \alpha_{21} = \frac{C(X, Y)}{V(X)},$$

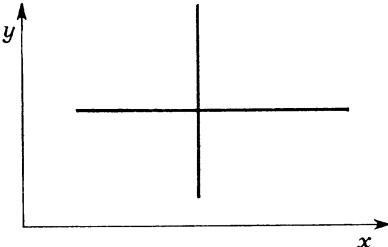
przeto gdy $\rho=0$, to $\alpha_{12}=0$ i $\alpha_{21}=0$. Wobec tego, gdy zmienne X i Y nie są skorelowane, to proste regresji przecinają się pod kątem $\psi=\pi/2$, przy czym prosta regresji Y względem X jest równoległa do osi odciętych, zaś prosta regresji X względem Y jest równoległa do osi rzędnych (rys. 3).

Uprednio udowodnione zostało twierdzenie (patrz 4.7.1), że jeśli zmienne losowe X i Y są niezależne, to kowariancja $C(X, Y)$ równa się zeru. Obecnie twierdzenie to wypowiadamy w nieco innej formie.

TWIERDZENIE 2. *Jeżeli zmienne X i Y są niezależne, to są również nieskorelowane.*

Twierdzenie odwrotne nie jest prawdziwe. Duże znaczenie teoretyczne i praktyczne ma twierdzenie przeciwstawne do twierdzenia 2.

TWIERDZENIE 3. *Jeżeli zmienne losowe X i Y są skorelowane, to są również zależne.*



Rys. 3

Twierdzenie 3 nie wymaga dowodu, gdyż jest twierdzeniem przeciwnym do twierdzenia 2, które jest prawdziwe, a jak wiadomo, dwa twierdzenia przeciwnostawne mogą być albo oba prawdziwe, albo oba fałszywe.

Na zakończenie uwag o współczynniku korelacji udowodnimy ważne

TWIERDZENIE 4. *Dla dowolnych zmiennych losowych X i Y istnieje zawsze przekształcenie liniowe, sprowadzające te zmienne do postaci, w której współczynnik korelacji pomiędzy zmiennymi równa się zeru.*

Dowód. Korzystając ze znanych wzorów na przesunięcie i obrót układu współrzędnych mamy

$$X' = (X - m_{10}) \cos \psi + (Y - m_{01}) \sin \psi,$$

$$Y' = -(X - m_{10}) \sin \psi + (Y - m_{01}) \cos \psi.$$

Aby zmienne X' i Y' były nieskorelowane, potrzeba i wystarcza, aby

$$C(X', Y') = E(X' Y') = 0.$$

Zgodnie ze wzorem (13) z 4.7.1 mamy

$$E(X' Y') = E\{[(X - m_{10}) \cos \psi + (Y - m_{01}) \sin \psi] [-(X - m_{10}) \sin \psi + (Y - m_{01}) \cos \psi]\}.$$

Po wymnożeniu i wprowadzeniu znaku wartości oczekiwanej do nawiasu mamy

$$E(X' Y') = E[(X - m_{10})(Y - m_{01})] \cos 2\psi - \frac{1}{2}[E(X - m_{10})^2 - E(Y - m_{01})^2] \sin 2\psi.$$

Dzielimy obie strony powyższej równości przez $\cos 2\psi$:

$$\begin{aligned} \frac{E(X' Y')}{\cos 2\psi} &= E[(X - m_{10})(Y - m_{01})] - \frac{1}{2}[E(X - m_{10})^2 - E(Y - m_{01})^2] \operatorname{tg} 2\psi = \\ &= \mu_{11} - \frac{1}{2}(\mu_{20} - \mu_{02}) \operatorname{tg} 2\psi. \end{aligned}$$

Jeżeli obierzemy taki kąt obrotu ψ , że

$$(7) \quad \operatorname{tg} 2\psi = \frac{2\mu_{11}}{\mu_{20} - \mu_{02}},$$

to otrzymamy $E(X' Y') = 0$, a stąd $\rho(X', Y') = 0$, co należało okazać.

Na marginesie powyższego twierdzenia zauważmy, że gdybyśmy ze wzoru (7) wyznaczyli kąt ψ i przyjęli

$$\lambda = \operatorname{tg} \psi,$$

to prosta o równaniu

$$(8) \quad y - m_{01} = \lambda(x - m_{10})$$

ma tę własność, że suma kwadratów odległości punktów (x, y) od tej prostej jest minimum. Prosta (8) znana jest pod nazwą *prostej regresji ortogonalnej*. Ze wzoru (8) wynika, że ta prosta przechodzi przez środek ciężkości. W drodze elementarnych przekształceń otrzymuje się, że współczynnik kątowy prostej (8) jest równy

$$(9) \quad \lambda = \frac{1}{2\mu_{11}} \left(\mu_{20} - \mu_{02} \pm \sqrt{(\mu_{20} - \mu_{02})^2 + 4\mu_{11}^2} \right).$$

4.7.3. Współczynnik zależności

Współczynnik korelacji nie jest jedyną znaną miarą zależności między zmiennymi losowymi. Miar takich jest więcej. Dzieli się je na dwie grupy. Do pierwszej należą tzw. *miary parametryczne*, tzn. takie miary, które zależą od wartości, jakie przybierają wzajemnie od siebie zmienna losowa. Do drugiej grupy natomiast należą tzw. *miary nieparametryczne*, czyli takie miary, które zależą nie od wartości przybieranych przez zmienną losową, lecz od odpowiadających tym wartościom prawdopodobieństw lub od gęstości rozkładu tych zmiennych.

Współczynnik korelacji należy do grupy miar parametrycznych. Należą do niej również takie miary, jak stosunek korelacyjny η , Spearmana współczynnik korelacji rang R , czy też Kendalla współczynnik korelacji rang τ (patrz [17]). Do drugiej grupy należą takie miary, jak Yula współczynnik skojarzenia Q , Pearsona współczynnik skojarzenia Q , Yula współczynnik powiązania Y , Pearsona współczynnik średniej kwadratowej wielodzielności C , czy też Czuprowa współczynnik T . Szczegółowe informacje na temat każdej z tych miar znajdzie czytelnik w [17].

Najważniejszą miarą w grupie miar parametrycznych jest oczywiście współczynnik korelacji. Współczynnik ten jest niekiedy stosowany niechętnie, a to ze względu na jego trzy poważne wady:

1º Brak skorelowania między badanymi zmiennymi losowymi nie oznacza, że te zmienne są niezależne.

2º Współczynnik korelacji może być stosowany tylko w przypadku regresji liniowej. (Wynika stąd, że współczynnik korelacji należy stosować z największą ostrożnością w przypadku dyskretnych zmiennych losowych, które rzadko wykazują taką zależność, która może być opisana za pomocą regresji liniowej).

3º Współczynnik korelacji nie może być stosowany w przypadku niemierzalnych zmiennych losowych.

Dużą niedogodnością związaną ze stosowaniem współczynnika korelacji jest fakt, że dokładny rozkład współczynnika korelacji z próbki jest znany tylko w przypadku, gdy dwuwymiarowa zmienna losowa ma rozkład normalny (patrz 4.7.4). Wadę tę jednak ma większość znanych miar zależności zmiennych losowych.

Najważniejszą miarą w grupie miar nieparametrycznych jest Pearsona współczynnik średniej kwadratowej wielodzielności C , oparty na parametrze χ^2 (patrz 6.5.3). Współczynnik ten ma następujące wady:

1º Gdy zbiór danych liczbowych, stanowiących podstawę do obliczania wartości współczynnika, jest skończony, to współczynnik ten nie może nigdy osiągnąć wartości równej jedności. Wada ta jest szczególnie niemiła, gdy liczność zbioru danych jest mała.

2º Współczynnik jest zdefiniowany dla przypadku, gdy badane zmienne losowe mogą przybierać tylko skończony zbiór wartości.

Umiejętność właściwego mierzenia zależności między zmiennymi losowymi ma dość znaczenie w badaniach naukowych wszelkiego typu. Przecież głównym celem nauki jest właśnie poznawanie związków i zależności łączących ze sobą poszczególne zjawiska. Umiejętność wykrywania i mierzenia zależności między zjawiskami i procesami losowymi

ma szczególne znaczenie w naukach społecznych, a wśród nich przede wszystkim w ekonomii. Badania eksperymentalne takich zależności są na ogół trudne, kłopotliwe i kosztowne ze względu na bardzo ograniczone możliwości przeprowadzania sztucznych, laboratoryjnych eksperymentów w takich dziedzinach, w których obiektem badania są sami ludzie, ich sprawy i działalność. Nic dziwnego, że właśnie w tych dziedzinach szczególna rola przy badaniu i mierzeniu zależności między zjawiskami przypada metodom probabilistycznym i statystycznym.

Wymagania, jakie powinna spełniać dobrze zdefiniowana miara zależności między zmiennymi losowymi, są bardzo rygorystyczne. Oto one:

1º Miara zależności winna przyjmować wartość 0 wtedy i tylko wtedy, gdy zmienne losowe są niezależne, a wartość 1 wtedy i tylko wtedy, gdy zależność między badanymi zmiennymi jest absolutna⁽¹⁾.

2º Miara zależności winna spełniać swoje zadanie nie tylko w przypadku regresji liniowej, ale i krzywoliniowej.

3º Miara ma być tak zdefiniowana, aby można było znaleźć jej rozkład w próbce, i to możliwie bez żadnych założeń co do rozkładu zmiennej losowej (X, Y). Ponieważ warunek ten jest niesłychanie trudny do spełnienia, przeto można kontentować się znajomością granicznego rozkładu tej miary w próbce, a przynajmniej jej rozkładu warunkowego.

4º Miara ma zachować swą przydatność tak w przypadku zmiennych losowych ciągły, jak i dyskretnych. W szczególności miara winna spełniać swoje zadanie, gdy zmienne losowe X, Y są zmiennymi zero-jedynkowymi, tzn. gdy zwykle stosuje się jakieś specjalne miary zależności (np. współczynnik skojarzenia Yula lub Pearsona). Miara ma zachować również swą przydatność i w tym przypadku, gdy jedna lub obie zmienne są niemierzalne.

5º Miara ma być tak zdefiniowana, aby umożliwiała poznanie funkcyjnej zależności między tą miarą a współczynnikiem korelacji, przynajmniej w przypadku gdy rozkład zmiennej losowej (X, Y) jest normalny.

6º Miara powinna mieć również dwie dodatkowe zalety o charakterze pragmatycznym:

a) koncepcja i postać analityczna miary muszą być wystarczająco proste, umożliwiające jej szeroką popularyzację i masowe zastosowanie w praktyce;

b) obliczenia związane z tą miarą muszą być również proste.

Jak widać, warunki⁽²⁾, którym musi odpowiadać dobrze zdefiniowana miara zależności, nie są łatwe do spełnienia. Tłumaczy to fakt istnienia wielu różnych miar przydatnych do różnych celów i stosowanych w specyficznych sytuacjach oraz mających swoje indywidualne zalety i wady. Próby zdefiniowania miary „idealnej”, a przynajmniej takiej, która spełniałaby możliwie jak najwięcej wymienionych wyżej warunków, bynajmniej nie

(1) Przypominamy, że zmienne losowe X, Y są niezależne wtedy i tylko wtedy, gdy $F(x, y) = F_1(x) \cdot F_2(y)$ (patrz 3.6.3, określenie 3). Wyjaśniamy również, że o zależności absolutnej między ciągłymi zmiennymi losowymi X, Y będziemy mówić wtedy i tylko wtedy, gdy obszar zmienności zmiennej losowej (X, Y) jest miary zero. Warunek ten dla zmiennych dyskretnych ma bardziej zawiąz postać.

(2) Warunki te zostały sformułowane przez H. Steinhausa na jednym z prowadzonych przez niego seminariów poświęconych metodom badania współzależności cech antropometrycznych.

ustały. Przedstawimy tu jedną z takich prób, a mianowicie zaproponowany przez autora tzw. *współczynnik zależności* δ ⁽¹⁾

Rozważmy dwuwymiarową zmienną losową ciągłą (X, Y) . Warunek dostateczny i konieczny na to, aby zmienne X i Y były niezależne, może być, jak wiadomo, wyrażony za pomocą tożsamości:

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

gdzie $f(x, y)$ oznacza gęstość zmiennej losowej (X, Y) , a $f_1(x)$ oraz $f_2(y)$ są gęstościami brzegowymi tej zmiennej.

OKREŚLENIE 1. *Współczynnikiem zależności stochastycznej* między zmiennymi losowymi X i Y lub krótko *współczynnikiem zależności* nazywać będziemy wielkość

$$(1) \quad \delta = \sqrt{1 - \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \min[f(x, y), f_1(x) \cdot f_2(y)] dx dy}.$$

Łatwo sprawdzić, że

$$0 \leq \delta \leq 1.$$

Zajmiemy się obecnie adaptacją definicji współczynnika zależności, zredagowanej uprzednio tylko dla przypadku ciągłego, również i do przypadku dyskretnego. Jak wiadomo, dwie dyskretne zmienne losowe X, Y są niezależne wtedy i tylko wtedy, gdy

$$P(X=x, Y=y) = P(X=x) \cdot P(Y=y),$$

przy czym x może przybierać jedną z wartości: x_1, x_2, \dots, x_r , stanowiących zbiór możliwych realizacji zmiennej losowej X , y zaś może przybierać jedną z wartości: y_1, y_2, \dots, y_s , tworzących zbiór możliwych realizacji zmiennej losowej Y .

W celu uproszczenia symboliki wprowadzimy oznaczenia analogiczne do oznaczeń stosowanych w 3.6.2, a mianowicie:

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

oraz

$$P(X=x_i) = p_i, \quad P(Y=y_j) = q_j.$$

W dalszym ciągu potrzebne nam będą dwie macierze, a mianowicie macierz

$$(2) \quad \mathbf{P}_1 = \begin{bmatrix} p_{11} & p_{12} & \dots & p_{1s} \\ p_{21} & p_{22} & \dots & p_{2s} \\ \dots & \dots & \dots & \dots \\ p_{r1} & p_{r2} & \dots & p_{rs} \end{bmatrix}$$

oraz macierz

$$(3) \quad \mathbf{P}_2 = \begin{bmatrix} p_1 q_1 & p_1 q_2 & \dots & p_1 q_s \\ p_2 q_1 & p_2 q_2 & \dots & p_2 q_s \\ \dots & \dots & \dots & \dots \\ p_r q_1 & p_r q_2 & \dots & p_r q_s \end{bmatrix}.$$

⁽¹⁾ Opracowano w oparciu o artykuł autora: *On the Measurement of Stochastic Dependence*, Zastosowania matematyki X (1969).

Przypuśćmy na chwilę, że definicja współczynnika zależności została zredagowana za pomocą relacji

$$(4) \quad \delta^2 = 1 - \sum_{i,j} \min(p_{ij}, p_i q_j),$$

a więc podobnie, jak to miało miejsce w przypadku ciągłym. Oznaczmy przez M zbiór par (i,j) , dla których $p_{ij} > p_i q_j$ oraz przez K – zbiór par (i,j) , dla których $p_{ij} \leq p_i q_j$. Łatwo sprawdzić, że

$$(5) \quad 1 - \sum_{(i,j) \in M} p_i q_j = \sum_{(i,j) \in K} p_i q_j.$$

Stąd

$$\begin{aligned} (6) \quad \delta^2 &= 1 - \sum_{i,j} \min(p_{ij}, p_i q_j) = \\ &= 1 - \sum_{(i,j) \in M} p_i q_j - \sum_{(i,j) \in K} p_{ij} = \\ &= \sum_{(i,j) \in K} p_i q_j - \sum_{(i,j) \in K} p_{ij} = \\ &= \sum_{(i,j) \in M} p_{ij} - \sum_{(i,j) \in M} p_i q_j. \end{aligned}$$

Nie zmniejszając ogólności dalszych rozważań możemy założyć, że $s \geq r$, czyli liczba kolumn macierzy \mathbf{P}_1 jest nie mniejsza od liczby wierszy tej macierzy oraz że każdy wiersz i każda kolumna macierzy zawiera przynajmniej jeden element dodatni (w przeciwnym wypadku macierz zredukowałaby się do mniejszych rozmiarów, lecz znowu nie miałaby ani pustych wierszy, ani kolumn).

Oznaczmy literą k liczbę par (i,j) w K oraz literą m liczbę par (i,j) w M . Między k i m zachodzą następujące związki:

$$k + m = rs, \quad 0 < k \leq rs, \quad 0 \leq m < rs.$$

Jeżeli $k = rs$, to

$$p_{ij} = p_i q_j \quad \text{dla wszystkich } i, j.$$

W takim razie

$$\sum_{(i,j) \in M} p_{ij} - \sum_{(i,j) \in M} p_i q_j = 0.$$

Można jednak również wykazać (dowód pomijamy), że

$$\sum_{(i,j) \in M} p_{ij} - \sum_{(i,j) \in M} p_i q_j \leq 1 - \frac{1}{r}.$$

Z tego wynika, że gdybyśmy zaakceptowali definicję współczynnika zależności dla przypadku dyskretnego taką, jaka została zaproponowana w relacji (4), to współczynnik zależności zamiast pożądanej nierówności $0 \leq \delta \leq 1$ spełniałby nierówność $0 \leq \delta \leq 1 - 1/r$. Oznacza to, że wprowadzona na użytek tymczasowy definicja współczynnika zależności

(4) musi ulec modyfikacji, przybierając ostatecznie kształt następujący:

$$(7) \quad \delta = \sqrt{\frac{1 - \sum_{i,j} \min(p_{ij}, p_i q_j)}{1 - \frac{1}{\min(r, s)}}}.$$

Taka jest postać współczynnika zależności δ w przypadku, gdy obie zmienne losowe są dyskretne. Teraz już współczynnik zależności spełnia nierówność

$$0 \leq \delta \leq 1.$$

Zwracamy uwagę czytelnika, że r oznacza liczbę wartości, jakie może przybierać dyskretna zmienna losowa X , natomiast s jest liczbą wartości, jakie może przybierać dyskretna zmienna losowa Y . Tak więc liczby r, s są dane. Założyliśmy uprzednio, że $r \leq s$. Przypuśćmy, że w jakimś konkretnym przypadku liczba r jest bardzo duża. Wtedy oczywiście ułamek $1/\min(r, s) = 1/r$ niewiele różni się od zera, tak że można przyjąć, iż zachodzi przybliżona równość

$$(8) \quad \delta^2 \approx 1 - \sum_{i,j} \min(p_{ij}, p_i q_j).$$

Jest to przypadek analogiczny do przypadku ciągłego, gdyż wzór (8) ma konstrukcję podobną do konstrukcji wzoru (1).

Nic w tym zresztą dziwnego. Gdy obie dyskretne zmienne losowe X, Y mogą przybierać bardzo wiele wartości, to stają się one podobne do zmiennych ciągłych. Dla podkreślenia tej własności zmiennych można je określić mianem zmiennych *quasi-ciągły*.

Z drugiej strony, gdy zmienne losowe X, Y są ciągłe, można je przekształcić w *quasi-dyskretne* za pomocą łączenia ich wartości w klasy, podobnie jak to się czyni przy budowaniu szeregu rozdzielczego. Otrzyma się wtedy dwudzielną tablicę o r wierszach i s kolumnach. Oczywiście traktując zmienne ciągłe jako quasi-dyskretne musimy w sposób arbitralny ustalić wartości liczb r i s , gdyż tym razem liczby te nie są znane. Jeżeli estymujemy wartość parametru δ w oparciu o dane liczbowe z próbki (będzie o tym mowa w 6.7), to przy ustalaniu wartości liczb r i s winniśmy dążyć do tego, aby liczby te były jak największe, a jednocześnie starać się przestrzegać warunku, aby liczność poszczególnych wierszy i kolumn tablicy dwudzielnej nie była w miarę możliwości mniejsza od 5. Liczba 5 jest oczywiście progiem umownym. Jak wiadomo, próg taki przyjmuje się zwykle przy budowaniu szeregów rozdzielczych, dając do tego, aby częstotliwości odpowiadające poszczególnym klasom dawały wiarygodne oceny odpowiednich prawdopodobieństw rozkładu analizowanych zmiennych losowych.

Zwracamy uwagę, że wartości liczb r i s ustala się niezależnie, jednak gdy nic temu nie stoi na przeszkodzie, przyjmuje się, że liczby r i s są równe lub prawie równe.

Wielka przydatność praktyczna współczynnika zależności polega na tym, że będąc nieparametryczną miarą zależności może być on stosowany, gdy obie zmienne są mierzalne, gdy obie są niemierzalne, gdy jedna jest mierzalna, a druga niemierzalna, gdy zmienne są ciągłe lub gdy są one dyskretne, gdy mogą przybierać wiele lub tylko dwie wartości. Jak z tego wynika, miara ta jest bardzo uniwersalna. Jak zobaczymy później, związane

z nią obliczenia są proste i mało uciążliwe. Obok tych zalet praktycznych współczynnik zależności ma cenne zalety teoretyczne:

1° Dwie zmienne losowe X, Y są niezależne wtedy i tylko wtedy, gdy $\delta^2 = 0$.

2° Jeżeli X i Y są ciągłymi zmiennymi losowymi, to $\delta^2 = 1$ wtedy i tylko wtedy, gdy cała masa prawdopodobieństwa łącznego rozkładu tych zmiennych jest rozpostarta na obszarze miary zero.

3° Jeżeli X i Y są dyskretnymi zmiennymi losowymi oraz jeżeli $r=s=2$, to $\delta^2 = 0$ wtedy i tylko wtedy, gdy $p_{ij} = p_i q_j$ dla $i, j = 1, 2$ oraz $\delta^2 = 1$ wtedy i tylko wtedy, gdy sumy wszystkich wierszy i kolumn dwudzielnej tablicy są sobie równe i jeżeli każda kolumna (wiersz) zawiera jeden i tylko jeden element. Innymi słowy, $\delta^2 = 1$ wtedy i tylko wtedy, gdy niezerowe elementy macierzy P_1 są sobie równe i leżą na jednej z przekątnych macierzy.

Przykłady zastosowań współczynnika zależności oraz sposób obliczania tego współczynnika zostaną podane w 6.7.5.

4.7.4. Dwuwymiarowy rozkład normalny

W rozdziale 3 przedstawiono różne rozkłady jednowymiarowej zmiennej losowej najczęściej spotykane w praktyce oraz podano własności tych rozkładów. W sposób naturalny rodzi się pytanie, czy można uogólnić te rozkłady na przypadek dwuwymiarowy. Okazuje się, że takie uogólnienie prowadzi do tak skomplikowanych wzorów, że ciekawe wyniki można otrzymać tylko w dwóch przypadkach, dotyczących zmiennej losowej ciągłej, a mianowicie w przypadku rozkładu jednostajnego i rozkładu normalnego. Uogólnienie rozkładu jednostajnego jest tak proste, że nie będziemy mu poświęcać uwagi. Zajmiemy się natomiast starannym zbadaniem właściwości rozkładu normalnego w przypadku dwuwymiarowej zmiennej losowej (X, Y).

Rozkładem normalnym dwuwymiarowej zmiennej losowej (X, Y) nazywa się rozkład, którego gęstość wyraża się wzorem

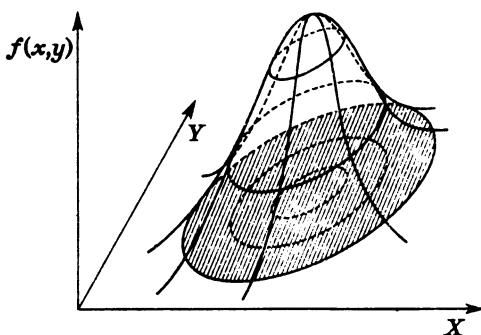
$$(1) \quad f(x, y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp \left\{ -\frac{1}{2(1-\rho^2)} \left[\frac{(x-m_1)^2}{\sigma_1^2} - \right. \right. \\ \left. \left. - \frac{2\rho(x-m_1)(y-m_2)}{\sigma_1\sigma_2} + \frac{(y-m_2)^2}{\sigma_2^2} \right] \right\},$$

gdzie

$$\begin{aligned} m_1 &= E(X), & m_2 &= E(Y), \\ \sigma_1 &= [E(X-m_1)^2]^{\frac{1}{2}}, & \sigma_2 &= [E(Y-m_2)^2]^{\frac{1}{2}}, \\ \rho &= (\alpha_{12} \alpha_{21})^{\frac{1}{2}}. \end{aligned}$$

Wyjaśniamy, że dla uproszczenia oznaczeń naruszyliśmy nieco konwencję notacji wprowadzoną w 4.7.1 i na oznaczenie wartości przeciętnych $E(X)$ i $E(Y)$ zastosowaliśmy zamiast m_{10} i m_{01} prostsze symbole m_1 i m_2 .

Przypomnijmy, że w przypadku zmiennej losowej jednowymiarowej rozkład normalny był zdeterminowany dwoma parametrami, a mianowicie m i σ . Można było więc oczekiwać, że w przypadku zmiennej losowej dwuwymiarowej w rozkładzie wystąpią cztery parametry, gdy tymczasem, jak widać, jest ich pięć. Tym piątym parametrem jest współczynnik korelacji ρ . Może to stanowić ilustrację, jak bardzo komplikują się wzory rozkładów przy przechodzeniu do coraz to wyższych wymiarów zmiennych losowych. Antycypując dalsze rozważania podamy, że liczba parametrów rozkładu normalnego k -wymiarowej zmiennej losowej wynosi $\frac{1}{2}k(k+3)$.



Rys. 1

Nietrudno zauważyć, że przyrównując do stałej c^2 wyrażenie występujące w nawiasach kwadratowych we wzorze (1) otrzymamy równanie elipsy, zwanej *elipsą jednakowej gęstości prawdopodobieństwa* w rozkładzie normalnym. Rodzina takich elips odpowiadających różnym wartościom parametru c nazywa się *widmem rozkładu normalnego*. Charakterystyczny wygląd dwuwymiarowego rozkładu normalnego i jego widma przedstawia rysunek 1. Jak widać, jest to rozkład jednomodalny. Jego osi symetrii przechodzi przez punkt o współrzędnych m_1, m_2 . Wartość ekstremum zależy od trzech parametrów: σ_1, σ_2, ρ i wynosi

$$(2) \quad \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}.$$

Korzystając ze wzoru (12) z 3.6.3 można wykazać, że po wykonaniu odpowiednich przekształceń otrzymamy następujące wyrażenia na rozkłady warunkowe:

$$(3) \quad f(y|x) = \frac{1}{\sqrt{2\pi}\sigma_2\sqrt{1-\rho^2}} \exp \left\{ -\frac{1}{2(1-\rho^2)\sigma_2^2} \left[y - m_2 - \rho \frac{\sigma_2}{\sigma_1} (x - m_1) \right]^2 \right\},$$

$$(4) \quad f(x|y) = \frac{1}{\sqrt{2\pi}\sigma_1\sqrt{1-\rho^2}} \exp \left\{ -\frac{1}{2(1-\rho^2)\sigma_1^2} \left[x - m_1 - \rho \frac{\sigma_1}{\sigma_2} (y - m_2) \right]^2 \right\},$$

Przyjrzyjmy się uważnie wyrażeniu (3). Dostrzeżemy niewątpliwie, że przedstawia ono gęstość jednowymiarowego rozkładu normalnego o parametrach:

$$(5) \quad E(Y|X=x) = m_2 + \rho \frac{\sigma_2}{\sigma_1} (x - m_1)$$

oraz

$$(6) \quad V(Y|X=x) = \sigma_2^2 (1 - \rho^2).$$

Zauważmy jednak, że

$$\rho \frac{\sigma_2}{\sigma_1} = \alpha_{21}$$

natomast

$$m_2 - \rho \frac{\sigma_2}{\sigma_1} m_1 = \beta_{20}.$$

Stąd wyrażenie (5) może być przedstawione w postaci

$$y = \alpha_{21} x + \beta_{20},$$

czyli

$$E(Y|X=x) = \alpha_{21} x + \beta_{20}.$$

Oznacza to, że regresja I rodzaju Y względem X w rozkładzie normalnym jest funkcją liniową zmiennej x . Drogą zwykłej zamiany symboli łatwo wykazać, że analogicznie regresja liniowa I rodzaju X względem Y jest funkcją liniową zmiennej y . Łącząc oba te twierdzenia otrzymujemy ważny wniosek, że linie regresji w dwuwymiarowym rozkładzie normalnym są liniami prostymi. Wniosek ten ma prostą interpretację geometryczną. Obierzmy dowolną stałą c i napiszmy równanie elipsy jednakowej gęstości prawdopodobieństwa w rozkładzie normalnym zawierające tę stałą. Otrzymamy

$$\frac{(x - m_1)^2}{\sigma_1^2} - 2\rho \frac{(x - m_1)(y - m_2)}{\sigma_1 \sigma_2} + \frac{(y - m_2)^2}{\sigma_2^2} = c^2.$$

Poprowadźmy dwie pary stycznych do tej elipsy, pierwszą parę równolegle do osi rzędnych, drugą zaś równolegle do osi odciętych. Z kolei połączmy odcinkami punkty styczności pierwszej pary stycznych oraz punkty styczności drugiej pary stycznych. Można wykazać, że odcinki te są średnicami elipsy, tzn. przechodzą przez jej środek. Otóż prosta przechodząca przez pierwszą parę punktów styczności lub, co na jedno wychodzi, pokrywająca się z pierwszym z dwóch odcinków jest prostą regresji Y względem X . Druga prosta jest prostą regresji X względem Y (patrz rys. 2 i 3).

Po omówieniu konsekwencji wynikających ze wzoru (5) powróćmy do wzoru (6). Zauważmy, że prawa strona tego wzoru nie zależy od x . Oznacza to, że warunkowa wariancja $V(Y|X=x)$ jest stała w rozkładzie normalnym. Przypomnijmy, że w 4.7.1 zdefiniowana została wariancja resztkowa (wzór (35)):

$$V(Z) = \mu_{02} - \alpha_{21} \mu_{11}.$$

Ponieważ

$$\mu_{02} = \sigma_2^2 \quad \text{oraz} \quad \alpha_{21} \mu_{11} = \sigma_2^2 \rho^2,$$

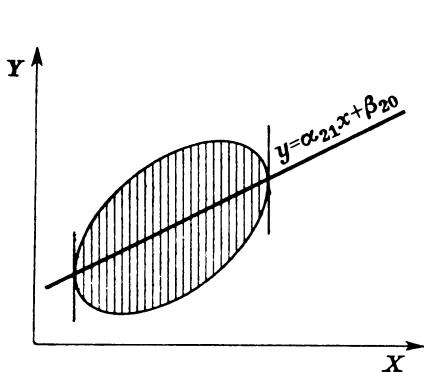
przeto

$$V(Z) = \sigma_2^2(1 - \rho^2)$$

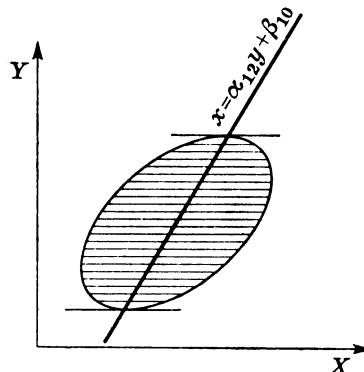
oraz

$$(7) \quad \sigma_{21} = \sigma_2 \sqrt{1 - \rho^2},$$

gdzie σ_{21} oznacza standardowy błąd oceny regresji Y względem X .



Rys. 2



Rys. 3

W wyniku podobnych rozważań można wykazać, że

$$(8) \quad \sigma_{12} = \sigma_1 \sqrt{1 - \rho^2}.$$

Postać wzoru na gęstość dwuwymiarowego rozkładu normalnego ulega znacznemu uproszczeniu, jeżeli zmienne X i Y zastąpimy wielkościami

$$\frac{X - m_1}{\sigma_1} \quad \text{i} \quad \frac{Y - m_2}{\sigma_2}.$$

Wtedy nowe zmienne, na oznaczenie których zachowamy te same litery X i Y , spełniają dwa warunki

$$E(X) = E(Y) = 0$$

oraz

$$V(X) = V(Y) = 1.$$

Wzór (1) przybierze teraz postać

$$(9) \quad f(x, y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)}(x^2 - 2\rho xy + y^2)\right].$$

Jak widać, tym razem gęstość dwuwymiarowego rozkładu normalnego zależy już tylko od jednego parametru ρ .

Zgodnie z twierdzeniem 4 z 4.7.2 jeżeli zmienne X i Y zastąpimy wielkościami

$$X \cos \psi + Y \sin \psi$$

oraz

$$-X \sin \psi + Y \cos \psi$$

i jeżeli obierzemy taki kąt obrotu, że

$$\operatorname{tg} 2\psi = \frac{2\mu_{11}}{\mu_{20} - \mu_{02}},$$

to korelacja między tymi wielkościami będzie równa zero. Zauważmy jednak, że ponieważ przyjęliśmy, iż $V(X) = V(Y) = 1$, więc również $\mu_{20} = \mu_{02} = 1$. Oznacza to, że mianownik ułamka stojącego po prawej stronie znaku równości przekształca się w zero. Wynika z tego, że $2\psi = \frac{1}{2}\pi$, czyli $\psi = \frac{1}{4}\pi$. Tak więc jeżeli (aby nie komplikować symboliki) zachowamy na oznaczenie tych wielkości stosowane poprzednio litery X i Y , to $\rho(X, Y) = 0$. Wyrażenie (9) przybierze więc postać

$$(10) \quad f(x, y) = \frac{1}{2\pi} \exp \left[-\frac{1}{2}(x^2 + y^2) \right].$$

Teraz już gęstość dwuwymiarowego rozkładu normalnego nie zależy od żadnego parametru. Krzywe jednakowej gęstości prawdopodobieństwa nie są już elipsami, lecz okręgami o równaniu

$$x^2 + y^2 = c^2.$$

Nasuwa się tu pytanie, czy skoro postać (10) funkcji gęstości dwuwymiarowego rozkładu normalnego nie zależy od żadnego parametru, nie można byłoby stablicować wartości tej funkcji. Łatwo jednak dostrzec, że nie ma potrzeby opracowywania specjalnych tablic, gdyż można wykorzystać do obliczenia wartości tej funkcji istniejące tablice funkcji gęstości jednowymiarowego rozkładu normalnego.

Rzeczywiście,

$$\frac{1}{2\pi} \exp \left[-\frac{1}{2}(x^2 + y^2) \right] = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}x^2) \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}y^2).$$

Stąd

$$(11) \quad f(x, y) = f(x) \cdot f(y).$$

Na mocy (11) wzór na dystrybuantę dwuwymiarowego rozkładu normalnego o parametrach $m_1 = m_2 = 0$, $\sigma_1 = \sigma_2 = 1$, $\rho = 0$ przybierze postać

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

Jak wynika z relacji (11), gdy zmienne losowe X , Y są nieskorelowane, to są one także niezależne. Twierdzenie odwrotne jest także prawdziwe. Można więc udowodnić następujące ważne

TWIERDZENIE 1. Jeżeli zmienne losowe X i Y mają rozkłady normalne i są nieskorelowane, to są również niezależne i – na odwrót – jeżeli są niezależne, to są również nieskorelowane.

Jeżeli więc wyjściowe zmienne losowe X i Y zestandardyzujemy, tak że ich wartości oczekiwane będą równe zeru, natomiast wariancje będą równe jedności, a następnie dokonamy obrotu układu współrzędnych o 45° , to w wyniku takiego przekształcenia otrzymamy zmienne losowe o rozkładach $N(0, 1)$, a przy tym nieskorelowane, a więc i niezależne. Czytelnik zechce zapamiętać ważną wskazówkę praktyczną, którą warto kierować się przy tym przekształceniu: należy najpierw dokonać standaryzacji zmiennych X i Y za pomocą wzorów

$$X' = \frac{X - m_1}{\sigma_1}, \quad Y' = \frac{Y - m_2}{\sigma_2}.$$

Dzięki temu wzory na obrót przybiorą postać:

$$X'' = \frac{\sqrt{2}}{2}(X' + Y'), \quad Y'' = \frac{\sqrt{2}}{2}(-X' + Y').$$

Po wykonaniu tych podstawień wracamy do pierwotnych oznaczeń zastępując X'' i Y'' przez X i Y .

4.7.5. Związki między współczynnikiem korelacji i współczynnikiem zależności w rozkładzie normalnym

Rozważmy zmienną losową (X, Y) o rozkładzie normalnym i niech gęstość tego rozkładu ma postać

$$f(x, y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left[-\frac{x^2 - 2\rho xy + y^2}{2(1-\rho^2)}\right],$$

przy czym ρ oznacza współczynnik korelacji. Pozostałe parametry: $E(X) = E(Y) = 0$ oraz $V(X) = V(Y) = 1$.

Zgodnie z definicją współczynnika zależności w przypadku zmiennych losowych ciągłych (definicja 1 w 4.7.3) parametr ten w rozważanym przez nas rozkładzie normalnym wyraża się wzorem

$$(1) \quad \delta^2 = \frac{1}{2\pi} \iint_G \left\{ \frac{1}{\sqrt{1-\rho^2}} \exp\left[-\frac{x^2 - 2\rho xy + y^2}{2(1-\rho^2)}\right] - \exp\left[-\frac{x^2 + y^2}{2}\right] \right\} dx dy,$$

gdzie G jest takim obszarem całkowania, że dla każdego punktu $(x, y) \in G$ zachodzi nierówność $f(x, y) > f_1(x)f_2(y)$. Wykażemy, że obszar G leży między dwiema gałęziami hiperboli, przy czym asymptotami tej hiperboli są proste

$$y = \frac{\rho x}{1 + \sqrt{1-\rho^2}} \quad \text{oraz} \quad x = \frac{\rho y}{1 + \sqrt{1-\rho^2}}.$$

► Zauważmy w tym celu, że brzeg obszaru G wyznacza się z warunku

$$\frac{1}{\sqrt{1-\rho^2}} \exp \left[-\frac{x^2 - 2\rho xy + y^2}{2(1-\rho^2)} \right] = \exp \left(-\frac{x^2 + y^2}{2} \right).$$

Po zlogarytmowaniu mamy

$$-\frac{1}{2} \ln(1-\rho^2) - \frac{x^2 - 2\rho xy + y^2}{2(1-\rho^2)} = -\frac{x^2 + y^2}{2}, \quad \rho \neq 1.$$

W wyniku elementarnych przekształceń otrzymamy

$$-\rho x^2 + 2xy - \rho y^2 = (1-\rho^2)C,$$

gdzie

$$C = \frac{1}{\rho} \ln(1-\rho^2).$$

Występująca w tym równaniu forma kwadratowa może być łatwo sprowadzona do postaci kanonicznej

$$\lambda_1 x^2 + \lambda_2 y^2$$

drogą wyznaczenia pierwiastków charakterystycznych λ_1 i λ_2 z równania

$$\begin{vmatrix} -\rho - \lambda & 1 \\ 1 & -\rho - \lambda \end{vmatrix} = 0.$$

Oto te pierwiastki:

$$\lambda_1 = 1 - \rho, \quad \lambda_2 = -1 - \rho.$$

Stąd

$$\frac{x^2}{C(1+\rho)} - \frac{y^2}{C(1-\rho)} = 1.$$

Łatwo zauważyć, że krzywa przedstawiona powyższym równaniem jest hiperbolą. Równania asymptot tej hiperboli mają postać

$$y = \sqrt{\frac{1-\rho}{1+\rho}} x, \quad x = \sqrt{\frac{1+\rho}{1-\rho}} y.$$

Dokonując obrotu układu współrzędnych o kąt $-\pi/4$ równania asymptot przybiorą postać

$$y = \frac{\sqrt{1+\rho} + \sqrt{1-\rho}}{\sqrt{1+\rho} - \sqrt{1-\rho}} x, \quad x = \frac{\sqrt{1+\rho} - \sqrt{1-\rho}}{\sqrt{1+\rho} + \sqrt{1-\rho}} y.$$

Stąd ostatecznie

$$y = \frac{\rho x}{1 + \sqrt{1-\rho^2}}, \quad x = \frac{\rho y}{1 + \sqrt{1-\rho^2}}.$$



Wzór (1) wyraża związek zachodzący między parametrami ρ i δ . A. Smoluk stosując przybliżone metody całkowania opracował tablicę, przedstawiającą zależność między tymi parametrami (tabl. 1) przy założeniu, że zmienna losowa (X, Y) ma rozkład normalny. Tablica ta pozwala wyrazić współczynnik korelacji za pomocą współczynnika zależności i vice versa. Korzyść, jaka płynie z tego, że znając jeden z dwóch parametrów ρ i δ można za pomocą tablicy znaleźć pozostały, nie ogranicza się bynajmniej do tego, że nie musimy przy tym wykonywać żadnych obliczeń. Jeżeli bowiem w konkretnym przypadku obliczymy na podstawie danych liczbowych z próbki współczynnik zależności, a następnie obliczymy na podstawie tych samych danych współczynnik korelacji i korzystając z tablicy znajdziemy tablicową wartość współczynnika zależności oraz jeżeli okaże się, że współczynniki obliczony i oszacowany za pomocą tablic różnią się znacznie między sobą, to jest to ważną przesłanką do podejrzenia, że rozkład, z którego pochodzą dane liczbowe, stanowiące podstawę obliczeń, nie jest normalny.

Tablica 1

ρ	δ^2	δ
0,05	0,0160	0,13
10	0321	18
15	0485	22
20	0654	26
25	0828	29
30	1010	32
35	1201	35
40	1402	37
45	1616	40
50	1846	43
55	2095	46
60	2367	49
65	2669	52
70	3008	55
75	3397	58
80	3856	62
85	4417	66
90	5143	72
92	5522	74

4.7.6. Uwagi o wielowymiarowych zmiennych losowych

Nie ma w zasadzie żadnych istotnych trudności w uogólnieniu rozważań dotyczących zmiennych losowych dwuwymiarowych na przypadek wielowymiarowy. Można więc zrezygnować z systematycznego demonstrowania analogii między przypadkiem 2-wymiarowym i k -wymiarowym poświęcając więcej uwagi zagadnieniom mającym specjalne znaczenie w zastosowaniach praktycznych.

OKREŚLENIE 1. *Zmienną losową k -wymiarową (X_1, X_2, \dots, X_k) nazywa się ciąg funkcji $X_1 = X_1(e), X_2 = X_2(e), \dots, X_k = X_k(e)$ określonych na zbiorze zdarzeń elementarnych E , takich że dla każdego ciągu liczb rzeczywistych x_1, x_2, \dots, x_k zbiór A zdarzeń elementarnych $e \in E$, dla których $X_1(e) < x_1, X_2(e) < x_2, \dots, X_k(e) < x_k$, spełnia warunek $A \in \mathcal{B}$.*

Wynika stąd, że

$$P(A) = P(X_1 < x_1, X_2 < x_2, \dots, X_k < x_k).$$

OKREŚLENIE 2. Funkcja $F(x_1, x_2, \dots, x_k) = P(X_1 < x_1, X_2 < x_2, \dots, X_k < x_k)$ nazywa się *dystrybuantą k -wymiarowej zmiennej losowej (X_1, X_2, \dots, X_k)*.

OKREŚLENIE 3. *Gęstością $f(x_1, x_2, \dots, x_k)$ k -wymiarowej zmiennej losowej X_1, X_2, \dots, X_k nazywamy pochodną mieszaną dystrybuanty w punkcie (x_1, x_2, \dots, x_k) , tzn.*

$$f(x_1, x_2, \dots, x_k) = \frac{\partial^k F}{\partial x_1 \partial x_2 \dots \partial x_k}.$$

Jak widać, uogólnienie podstawowych pojęć wprowadzonych przy omawianiu zmiennych losowych 2-wymiarowych na przypadek zmiennych k -wymiarowych, chociaż nie nastręcza trudności, jednak prowadzi do tego, że notacja staje się uciążliwa i zawiła. Zmusza to do wprowadzenia jakiejś nowej, wygodniejszej konwencji oznaczania tych zmiennych. Konwencją taką jest zapis macierzowy.

Niech

$$\mathbf{X} = (X_1, X_2, \dots, X_k)$$

będzie wektorem, którego składowymi są zmienne losowe X_1, X_2, \dots, X_k . Wektor \mathbf{X} nazywać będziemy *wektorem losowym*. Termin ten jest synonimem pojęcia zmiennej losowej k -wymiarowej. Niech dalej

$$\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ik})$$

oznacza wektor, którego składowymi są dane liczby rzeczywiste

$$x_{i1}, x_{i2}, \dots, x_{ik}.$$

Wyjaśniamy, że indeks i został wprowadzony po to, aby umożliwić odróżnienie i -tej realizacji np. drugiej składowej wektora losowego \mathbf{X} , tzn. realizacji x_{i2} zmiennej losowej X_2 , od liczby x_2 , będącej drugą realizacją jednowymiarowej zmiennej losowej X , lub od zmiennej nielosowej x_2 . Wektorowa interpretacja k -wymiarowej zmiennej losowej pozwala utrzymać pełną analogię zapisu dystrybuanty i gęstości zmiennej k -wymiarowej z zapisem dystrybuanty i gęstości zmiennej jednowymiarowej. Rzeczywiście, notacja

$$F(\mathbf{x}), \quad f(\mathbf{x})$$

oraz notacja

$$F(x), \quad f(x)$$

różnią się tylko tym, że w pierwszym przypadku w roli argumentu funkcji występuje wektor \mathbf{x} , a w drugim liczba x . Jak zobaczymy, analogia ta zostanie zachowana i w przyszłości.

Przypuśćmy, że przeprowadzono serię doświadczeń, w trakcie których obserwowano wartości, jakie przyjmują składowe X_1, X_2, \dots, X_k wektora \mathbf{X} . Jeżeli liczba doświadczeń wynosi n , to wyniki doświadczeń można zapisać w postaci następującej macierzy o rozmiarach $(n \times k)$:

$$\begin{bmatrix} x_{11} & x_{12} & \dots & x_{1k} \\ x_{21} & x_{22} & \dots & x_{2k} \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix}.$$

Jeżeli oznaczymy wiersze tej macierzy symbolami $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$, to każdemu doświadczeniu z serii n doświadczeń będzie przyporządkowany odpowiedni wektor \mathbf{x}_i ($i=1, 2, \dots, n$). Wektory te nazywać będziemy *realizacjami k-wymiarowej zmiennej losowej* (lub *wektora losowego* \mathbf{X}). Ponieważ nazwa ta jest długa, a przez to nieporęczna w częstym użyciu, więc wektory \mathbf{x}_i będą mieli również określać mianem *punktów empirycznych*. Stosownie do tego wektor losowy \mathbf{X} nazywać będziemy *przestrzenią punktów empirycznych*. Na oznaczenie, że punkt \mathbf{x} należy do przestrzeni \mathbf{X} , stosuje się zapis $\mathbf{x} \in \mathbf{X}$. Ponieważ na ogół na określenie \mathbf{X} i \mathbf{x} posługiwać się będziemy nazwami wektora losowy i punkt empiryczny, przeto byłoby dobrze, aby czytelnik od razu przyswoił sobie te pojęcia i zdawał sobie sprawę z ich odmiennej treści.

Zauważmy, że zgodnie z definicją zmiennej losowej k -wymiarowej \mathbf{X} każdemu zdarzeniu $A \in \mathcal{B}$ przyporządkowany jest punkt empiryczny $\mathbf{x} \in \mathbf{X}$ oraz prawdopodobieństwo $P(\mathbf{x}) = P(X_1 < x_1, X_2 < x_2, \dots, X_k < x_k)$. W dążeniu do jak największej zwięzości notacji zapisuje się to (E, \mathcal{B}, P) . Zapis ten oznacza, że na elementach borelowskiego ciała \mathcal{B} , wygenerowanego przez zbiór E zdarzeń elementarnych, określona została miara P zwana prawdopodobieństwem⁽¹⁾.

Pełną charakterystykę wektora losowego \mathbf{X} daje dystrybuanta $F(\mathbf{x})$ lub – w przypadku gdy \mathbf{X} ma wszystkie składowe ciągłe – gęstość $f(\mathbf{x})$.

OKREŚLENIE 4. Zmienne losowe X_1, X_2, \dots, X_k są *niezależne*⁽²⁾, jeżeli

$$F(x_1, x_2, \dots, x_k) = F_1(x_1) F_2(x_2) \dots F_k(x_k),$$

gdzie x_1, x_2, \dots, x_k są składowymi dowolnego wektora $\mathbf{x} \in \mathbf{X}$, natomiast F_1, F_2, \dots, F_k są symbolami dystrybuant brzegowych zmiennych losowych X_1, X_2, \dots, X_k .

Określenie to daje się uogólnić na przeliczalną ilość zmiennych losowych, ponieważ jednak uogólnienie to ma wyłącznie teoretyczne znaczenie, przeto nie będziemy go tu przytaczać.

W zastosowaniach praktycznych przyjmuje się zwykle, że gęstość zmiennej \mathbf{X} ma następującą postać.

$$(1) \quad f(\mathbf{x}) = (2\pi)^{-k/2} |\Gamma_k|^{-\frac{1}{2}} \exp \left[-\frac{1}{2} (\mathbf{x} - \mathbf{m})^T \Gamma_k^{-1} (\mathbf{x} - \mathbf{m}) \right],$$

⁽¹⁾ Trójka (E, \mathcal{B}, P) nazywa się *przestrzenią probabilistyczną*.

⁽²⁾ Używa się także nazwy *stochastycznie niezależne*.

gdzie $(\mathbf{x} - \mathbf{m})^T$ oznacza transponowaną macierz $\mathbf{x} - \mathbf{m}$,

$$\mathbf{m} = \begin{bmatrix} m_1 \\ m_2 \\ \vdots \\ m_k \end{bmatrix}, \quad \boldsymbol{\Gamma}_k = \begin{bmatrix} \gamma_{11} & \gamma_{12} & \dots & \gamma_{1k} \\ \gamma_{21} & \gamma_{22} & \dots & \gamma_{2k} \\ \dots & \dots & \dots & \dots \\ \gamma_{k1} & \gamma_{k2} & \dots & \gamma_{kk} \end{bmatrix},$$

przy czym $\gamma_{ij} = C(X_i, X_j)$. Symbol $|\boldsymbol{\Gamma}_k|$ oznacza wyznacznik macierzy $\boldsymbol{\Gamma}_k$. Tak więc

$$m_i = E(X_i),$$

$$\gamma_{ii} = E(X_i - m_i)^2 = V(X_i),$$

$$\gamma_{ij} = E[(X_i - m_i)(X_j - m_j)] = C(X_i, X_j).$$

Oczywiście

$$\gamma_{ij} = \gamma_{ji}.$$

Rozkład o gęstości danej wzorem (1) jest uogólnieniem rozkładu normalnego na przypadek zmiennej losowej k -wymiarowej.

Warto zwrócić uwagę, że wielowymiarowy rozkład normalny jest określony za pomocą momentów zmiennych losowych jedno- i dwuwymiarowych. Łatwo obliczyć liczbę stałych parametrów występujących w funkcji gęstości rozkładu normalnego. Liczba ta jest równa sumie liczby składowych wektora \mathbf{m} , liczby wyrazów leżących na przekątnej macierzy $\boldsymbol{\Gamma}_k$ oraz, ponieważ macierz $\boldsymbol{\Gamma}_k$ jest symetryczna, liczby elementów leżących nad główną przekątną, czyli

$$k + k + \frac{k(k-1)}{2} = \frac{k(k+3)}{2}.$$

Aby zapisać, że wektor losowy \mathbf{X} ma rozkład normalny, stosuje się oznaczenie $N(\mathbf{m}, \boldsymbol{\Gamma}_k)$.

Analizując własności rozkładu dwuwymiarowego przekonaliśmy się, że

1° warunkowa wartość oczekiwana $E(Y|X=x)$ jest funkcją liniową,

2° warunkowa wariancja $V(Y|X=x)$ jest stała.

Można wykazać, że obie te własności są zachowane w przypadku regresji Y względem X_1, X_2, \dots, X_k . W tym celu umówmy się, że od tej pory będziemy rozpatrywać wektor losowy \mathbf{X} nie o k , lecz o $k+1$ składowych, przy czym pierwszych k składowych oznaczymy symbolami X_1, X_2, \dots, X_k , a ostatnią składową oznaczymy symbolem Y . Można wykazać, stosując postępowanie analogiczne do tego, którym uzasadnialiśmy wzór (5) i (6) w 4.7.4, że jeżeli \mathbf{X} ma rozkład normalny, prawdziwe są następujące relacje:

$$1^\circ \quad E(Y|X_1=x_1, X_2=x_2, \dots, X_k=x_k) = \alpha_1 x_1 + \alpha_2 x_2 + \dots + \alpha_k x_k + \alpha_0,$$

$$2^\circ \quad V(Y|X_1=x_1, X_2=x_2, \dots, X_k=x_k) = \text{const.},$$

przy czym $\alpha_1, \alpha_2, \dots, \alpha_k, \alpha_0$ są to stałe parametry.

Zatem funkcja liniowa

$$(2) \quad y = \alpha_1 x_1 + \alpha_2 x_2 + \dots + \alpha_k x_k + \alpha_0$$

jest regresją I rodzaju zmiennej Y względem zmiennych X_1, X_2, \dots, X_k . Zwracamy uwagę czytelnika, iż w wyrażeniu (2) y jest funkcją zmiennych x_1, x_2, \dots, x_k . Wobec tego zmienną y można nazwać *zmienną zależną* lub *funkcją*, natomiast zmienne x_1, x_2, \dots, x_k (które nie są zmiennymi losowymi) można określić mianem *zmiennych niezależnych* lub *argumentów*. Takie też nazwy będą stosowane w niniejszym tekście. Z drugiej strony jednak zmienne X_1, X_2, \dots, X_k , tzn. zmienne losowe, z reguły nie są oczywiście niezależne (są przecież na ogół skorelowane między sobą, na co wskazują elementy γ_{ij} macierzy Γ_k , niekoniecznie równe zeru dla $i \neq j$). Tak więc o zmiennych X_1, X_2, \dots, X_k nie można mówić „zmienne niezależne” mając na myśli, iż występują one w roli argumentów w funkcji regresji. Popełnia się wtedy podwójny błąd: określa się mianem „niezależnych” zmienne stochastycznie zależne oraz utożsamia się argumenty funkcji regresji x_1, x_2, \dots, x_k , które są zmiennymi w zwykłym sensie, ze zmiennymi X_1, X_2, \dots, X_k , które są zmiennymi losowymi. Dla uniknięcia takich nieporozumień w literaturze anglosaskiej zmienną losową Y określa się mianem *regresora*, natomiast zmienne X_1, X_2, \dots, X_k nazywane bywają *regresandami*. Stosowane są również nazwy *zmienna wyjaśniająca (explaining variable)* i *zmienne wyjaśniające (explaining variables)*. Terminy te nie będą tu stosowane, gdyż po tych wyjaśnieniach nie zachodzi po temu potrzeba.

Funkcja regresji jest znakomitym narzędziem *prognozy statystycznej*. Jeśli chcemy podkreślić, że posługujemy się równaniem regresji dla predykcji wartości zmiennej Y , gdy dane są wartości zmiennych X_1, X_2, \dots, X_k , to można stosować nazwę *predykta* na określenie zmiennej Y oraz nazwę *predykanty* na określenie zmiennych X_1, X_2, \dots, X_k .

Przypominamy, że stałe parametry funkcji regresji wyznaczyliśmy w 4.7.1 metodą najmniejszych kwadratów. Tę samą metodę wykorzystamy obecnie.

Należy więc zminimizować wyrażenie:

$$E(Y - \alpha_1 X_1 - \alpha_2 X_2 - \dots - \alpha_k X_k - \alpha_0)^2$$

względem $\alpha_0, \alpha_1, \alpha_2, \dots, \alpha_k$.

Po zróżniczkowaniu i przyrównaniu do zera pochodnych cząstkowych otrzymujemy układ równań liniowych

$$E(Y) - \alpha_1 E(X_1) - \alpha_2 E(X_2) - \dots - \alpha_k E(X_k) - \alpha_0 = 0,$$

$$E(X_1 Y) - \alpha_1 E(X_1 X_1) - \alpha_2 E(X_1 X_2) - \dots - \alpha_k E(X_1 X_k) - \alpha_0 E(X_1) = 0,$$

$$E(X_k Y) - \alpha_1 E(X_k X_1) - \alpha_2 E(X_k X_2) - \dots - \alpha_k E(X_k X_k) - \alpha_0 E(X_k) = 0.$$

Z pierwszego równania znajdujemy, że

$$(3) \quad \alpha_0 = E(Y) - \sum_{j=1}^k \alpha_j E(X_j).$$

Widać stąd, że rozwiązanie układu uprości się, jeżeli zastąpimy zmienne X_1, X_2, \dots, X_k, Y zmiennymi $X_1 - E(X_1), X_2 - E(X_2), \dots, X_k - E(X_k), Y - E(Y)$. Wtedy $\alpha_0 = 0$, a układ

równań normalnych przybierze postać

$$(4) \quad \begin{aligned} \gamma_{11}\alpha_1 + \gamma_{12}\alpha_2 + \dots + \gamma_{1k}\alpha_k &= \gamma_{10}, \\ \gamma_{21}\alpha_1 + \gamma_{22}\alpha_2 + \dots + \gamma_{2k}\alpha_k &= \gamma_{20}, \\ &\dots \dots \dots \dots \dots \dots \\ \gamma_{k1}\alpha_1 + \gamma_{k2}\alpha_2 + \dots + \gamma_{kk}\alpha_k &= \gamma_{k0}. \end{aligned}$$

Wyjaśniamy, że $\gamma_{ij} = C(X_i, X_j)$ oraz $\gamma_{i0} = C(X_i, Y)$.

Zapis tego układu równań i dalszych związanych z nim działań ulegnie uproszczeniu, jeżeli wprowadzi się notację macierzową. Rozważmy macierz:

$$(5) \quad \left[\begin{array}{cccc|c} \gamma_{11} & \gamma_{12} & \dots & \gamma_{1k} & | & \gamma_{10} \\ \gamma_{21} & \gamma_{22} & \dots & \gamma_{2k} & | & \gamma_{20} \\ \dots & \dots & \dots & \dots & | & \dots \\ \gamma_{k1} & \gamma_{k2} & \dots & \gamma_{kk} & | & \gamma_{k0} \\ \hline \gamma_{01} & \gamma_{02} & \dots & \gamma_{0k} & | & \gamma_{00} \end{array} \right] = \left[\begin{array}{c|c} \Gamma_k & \Gamma_0 \\ \hline \Gamma_0^T & \gamma_{00} \end{array} \right],$$

gdzie macierze Γ_k i Γ_0 mają odpowiednio wymiary $(k \times k)$ oraz $(k \times 1)$, natomiast $\gamma_{00} = C(Y, Y) = V(Y)$.

Obecnie układ równań (4) można zapisać w postaci

$$(6) \quad \Gamma_k \mathbf{A} = \Gamma_0,$$

gdzie \mathbf{A} jest wektorem o wymiarach $(k \times 1)$:

$$\mathbf{A} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_k \end{bmatrix}.$$

Stąd

$$(7) \quad \mathbf{A} = \Gamma_k^{-1} \Gamma_0,$$

przy czym zakłada się, że wyznacznik $|\Gamma_k| \neq 0$.

Niekiedy zdarza się, że w obliczeniach występują zamiast zmiennych X_1, X_2, \dots, X_k czy też zmiennych $X_1 - E(X_1), X_2 - E(X_2), \dots, X_k - E(X_k)$ zmienne standaryzowane:

$$\frac{X_1 - E(X_1)}{\sqrt{V(X_1)}}, \quad \frac{X_2 - E(X_2)}{\sqrt{V(X_2)}}, \quad \dots, \quad \frac{X_k - E(X_k)}{\sqrt{V(X_k)}}, \quad \frac{Y - E(Y)}{\sqrt{V(Y)}}.$$

Można wykazać (patrz np. [1]), że wtedy

$$(8) \quad \mathbf{A} = \mathbf{VR}_k^{-1} \mathbf{R}_0,$$

gdzie

$$(9) \quad \mathbf{R}_k = \begin{bmatrix} 1 & \rho_{12} & \dots & \rho_{1k} \\ \rho_{21} & 1 & \dots & \rho_{2k} \\ \dots & \dots & \dots & \dots \\ \rho_{k1} & \rho_{k2} & \dots & 1 \end{bmatrix}, \quad \mathbf{R}_0 = \begin{bmatrix} \rho_{10} \\ \rho_{20} \\ \vdots \\ \rho_{k0} \end{bmatrix},$$

przy czym

$$(10) \quad \rho_{ij} = \rho(X_i, X_j), \quad \rho_{i0} = \rho(X_i, Y),$$

a

$$(11) \quad \mathbf{V} = \sqrt{V(Y)} \begin{bmatrix} \frac{1}{\sqrt{V(X_1)}} & 0 & \dots & 0 \\ 0 & \frac{1}{\sqrt{V(X_2)}} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \frac{1}{\sqrt{V(X_k)}} \end{bmatrix}.$$

Przypominamy (patrz 4.7.2), że w przypadku zmiennej losowej dwuwymiarowej (X, Y) i regresji Y względem X wariancja resztka

$$V(Z_{21}) = V(Y | X = x) = V(Y) - \alpha_{21} C(X, Y).$$

Znajdziemy obecnie wariancję resztową w przypadku regresji Y względem X_1, X_2, \dots, X_k , tzn. wariancję zmiennej

$$(12) \quad Z = Y - \sum_{j=1}^k \alpha_j X_j.$$

Uwaga. Tym razem zrezygnowaliśmy z umieszczania obok zmiennej Z indeksów wskazujących zmienność występującą w roli zmiennej zależnej i zmiennej grającej rolę zmiennych niezależnych. Nie stworzy to nieporozumienia, gdyż każdorazowo będziemy określać wyraźnie rolę zmiennych. Mamy

$$V(Z) = E(Y - U)^2, \quad U = \sum_{j=1}^k \alpha_j X_j.$$

Przypominamy, że

$$E(X_1) = E(X_2) = \dots = E(X_k) = 0, \quad E(Y) = 0.$$

Tak więc

$$\begin{aligned} V(Z) &= E(Y^2 - 2YU + U^2) = E(Y^2 - 2 \sum_{j=1}^k \alpha_j X_j Y + \sum_{i,j=1}^n \alpha_i \alpha_j X_i X_j) = \\ &= \gamma_{00} - 2\mathbf{A}^T \boldsymbol{\Gamma}_0 + \mathbf{A}^T \boldsymbol{\Gamma}_k \mathbf{A}. \end{aligned}$$

Ponieważ $\Gamma_k \mathbf{A} = \Gamma_0$, przeto $\mathbf{A}^T \Gamma_k \mathbf{A} = \mathbf{A}^T \Gamma_0$. Stąd

$$(13) \quad V(Z) = \gamma_{00} - \mathbf{A}^T \Gamma_0$$

lub inaczej

$$(14) \quad V(Z) = V(Y) - \sum_{j=1}^k \alpha_j C(X_j, Y).$$

Z równości (14) widać wyraźnie, że $V(Z)$ nie zależy od zmiennych x_1, x_2, \dots, x_k , a więc jest wielkością stałą, tak jak to sygnalizowała relacja 2°.

Zauważmy, że w przypadku zmiennej losowej dwuwymiarowej (X, Y) prawdziwe jest następujące

TWIERDZENIE 1. Jeżeli $E(X) = E(Y) = 0$ i $U = \alpha X$, to

$$(15) \quad \rho(Y, U) = \rho(X, Y).$$

Rzeczywiście,

$$\rho^2(Y, U) = \frac{C^2(Y, U)}{V(Y)V(U)} = \frac{\alpha^2 E^2(X, Y)}{\alpha^2 V(X)V(Y)} = \rho^2(X, Y).$$

Nasuwa to pomysł, aby parametr $\rho(Y, U)$ wykorzystać do mierzenia łącznego wpływu, jaki na zmienną Y wywierają wszystkie lub niektóre ze zmiennych X_1, X_2, \dots, X_k .

OKREŚLENIE 5. Parametr $\rho(Y, U)$, gdzie $U = \sum_{j=1}^k \alpha_j X_j$, nazywa się *współczynnikiem korelacji wielowymiarowej* (zwanej także *korelacją wielokrotną* lub *wieloraką*). Współczynnik ten oznaczać będziemy symbolem ρ_0 .

Niekiedy należy zbadać wpływ, jaki na zmienną Y wywiera nie cały zbiór zmiennych X_1, X_2, \dots, X_k , lecz dowolna część tego zbioru. Oznaczmy literą I zbiór indeksów zmiennych X_j . Tak więc $j \in I$, $j = 1, 2, \dots, k$. Oznaczmy dalej symbolem I_s podzbiór zbioru I . Wobec tego $I_s \subset I$, przy czym $s = 1, 2, \dots, 2^k - 1$, gdyż wśród wszystkich podzbiorów, jakie można utworzyć z elementów zbioru I , jest $\binom{k}{1}$ podzbiorów jednoelementowych, $\binom{k}{2}$ podzbiorów dwuelementowych, itd., a jak wiadomo

$$\sum_{j=1}^k \binom{k}{j} = 2^k - 1.$$

Umówmy się, że numerując te podzbiory liczbami $1, 2, \dots, 2^k - 1$ uporządkujemy uprzednio podzbiory w kolejności leksykograficznej:

$$(16) \quad \{1\}, \{2\}, \dots, \{k\}, \{1, 2\}, \{1, 3\}, \dots, \{1, k\}, \dots, \{1, 2, \dots, k\}.$$

Na przykład jeżeli $k = 3$, to $2^3 - 1 = 7$, a więc mamy 7 podzbiorów, które uporządkowane w kolejności leksykograficznej tworzą ciągi

$$\{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}.$$

Tak więc miejsce każdego podzbioru w ciągu jest jednoznacznie określone, a odpowiadający danemu podzbiorowi numer kolejny s może być stosunkowo łatwo znaleziony.

Przypuśćmy, że w ciągu (16) chcemy wyróżnić podzbiór $\{2, 3\}$ i znaleźć korelację między zmienną Y a zmienną będącą kombinacją liniową zmiennych X_2 i X_3 . Ponieważ podzbiór $\{2, 3\}$ stoi na szóstym miejscu, przeto mamy znaleźć $\rho_0^{(6)}$, przy czym górny indeks informuje właśnie, że chodzi o korelację wielowymiarową między wyróżnioną zmienną Y a szóstym podzbiorem zbioru zmiennych występujących w roli argumentów. Zauważmy dalej, że

$$\rho_0^{(6)} = \rho(Y, U^{(6)}) ,$$

gdzie

$$U^{(6)} = \alpha_2^{(6)} X_2 + \alpha_3^{(6)} X_3 .$$

Taka konwencja znakowania uwalnia nas od stosowania wielocyfrowych indeksów, jakie w ślad za Yulem przyjęto zwykle używać w analizie regresji i korelacji zmiennych wielowymiarowych.

Niech $\rho_0^{(s)} = \rho(Y, U^{(s)})$ oraz $\rho_0^{(t)} = \rho(Y, U^{(t)})$, przy czym $I_s, I_t \subset I$.

TWIERDZENIE 2. Jeżeli $I_s \subset I_t \subset I$, to

$$\rho_0^{(s)} \leq \rho_0^{(t)} .$$

Istotnie, niech

$$Z_s = Y - \sum_{j \in I_s} \alpha_j^{(s)} X_j$$

oraz

$$Z_t = Y - \sum_{j \in I_t} \alpha_j^{(t)} X_j .$$

W takim razie jeżeli $I_s \subset I_t$, to

$$V(Z_s) \geq V(Z_t) .$$

Ale

$$\rho_0^{(s)} = \sqrt{1 - \frac{V(Z_s)}{V(Y)}}, \quad \rho_0^{(t)} = \sqrt{1 - \frac{V(Z_t)}{V(Y)}} .$$

Stąd

$$\rho_0^{(s)} \leq \rho_0^{(t)} .$$

TWIERDZENIE 3. Jeżeli $Z = Y - U$, a $U = \sum_{j=1}^k \alpha_j X_j$, to

$$\rho^2(Y, Z) = \frac{V(Z)}{V(Y)} .$$

Zauważmy bowiem, że

$$\begin{aligned} \rho^2(Y, Z) &= \frac{E^2[Y(Y - U)]}{V(Y)V(Z)} = \frac{[E(Y)^2 - E(YU)]^2}{V(Y)V(Z)} = \\ &= \frac{[E(Y)^2 - \sum_j \alpha_j C(X_j, Y)]^2}{V(Y)V(Z)} = \frac{V^2(Z)}{V(Y)V(Z)} . \end{aligned}$$

Stąd

$$(17) \quad \rho^2(Y, Z) = \frac{V(Z)}{V(Y)} = \varphi^2.$$

Wielkość φ nazywa się *współczynnikiem zbieżności*. Między współczynnikiem zbieżności i współczynnikiem korelacji zachodzi prosty związek

$$(18) \quad \varphi^2 = 1 - \rho^2.$$

TWIERDZENIE 4. Jeżeli $U = \sum_{j=1}^k \alpha_j X_j$, to $\rho(X_i, Z) = 0$ dla każdego $i \in I$.

Mamy bowiem

$$\rho(X_i, Z) = \frac{E[X_i(Y - \sum_{j \in I} \alpha_j X_j)]}{V(X_i)V(Z)} = \frac{\gamma_{i0} - \sum_{j \in I} \alpha_j \gamma_{ij}}{V(X_i)V(Z)}.$$

Zauważmy jednak, że ponieważ i -te równanie normalne ma postać

$$\sum_{j=1}^k \alpha_j \gamma_{ij} = \gamma_{i0},$$

więc licznik drugiego ułamka jest równy零.

Twierdzenie 4 ma duże znaczenie teoretyczne i praktyczne. Głosi ono, że jeżeli parametry α_j kombinacji liniowej $\sum_{j=1}^k \alpha_j X_j$ wyznaczone zostały metodą najmniejszych kwadratów, to błąd losowy nie jest skorelowany z żadną ze zmiennych X_1, X_2, \dots, X_k .

Niekiedy w zastosowaniach praktycznych pojawia się problem mierzenia zależności zmiennej Y od ustalonej zmiennej X_i , $i \in I$. Chodzi przy tym nie o bezpośrednią korelację między Y i X_i , bo do jej zmierzenia wystarczyły zwykły współczynnik korelacji $\rho(Y, X_i)$, lecz o korelację między zmienną Y a zmienną X_i po wyeliminowaniu wpływu, jaki na zmienną Y wywierają wszystkie pozostałe zmienne o numerach należących do I .

Oznaczmy parametry regresji zmiennej Y względem zmiennych X_j , $j = 1, 2, \dots, k$ oraz $j \neq i$, symbolami β_j . Oznaczmy dalej parametry regresji zmiennej X_i względem tych samych zmiennych X_j symbolami β'_j . Niech

$$(19) \quad W = Y - \sum_{\substack{j=1 \\ j \neq i}}^k \beta_j X_j,$$

$$(20) \quad W' = X_i - \sum_{\substack{j=1 \\ j \neq i}}^k \beta'_j X_j.$$

OKREŚLENIE 6. *Współczynnikiem korelacji cząstkowej* $\rho_i(W, W')$ nazywa się współczynnik korelacji między zmiennymi losowymi W i W' .

Obliczanie współczynnika korelacji wielowymiarowej ρ_0 oraz współczynników korelacji cząstkowej $\rho_1, \rho_2, \dots, \rho_k$ bezpośrednio z definicji jest niewygodne. W dobie współczesnej obliczenia związane z wyznaczeniem parametrów regresji i korelacji w przypadku

wielowymiarowych zmiennych losowych wykonywane są zwykle za pomocą elektronicznych maszyn cyfrowych. Programy, za pomocą których większość maszyn⁽¹⁾ wykonuje obliczenia współczynników korelacji wielowymiarowej, oparte są o wzór

$$(21) \quad \rho_0 = \sqrt{1 - \Gamma/\gamma_{00} \Gamma_{00}},$$

gdzie Γ jest wyznacznikiem macierzy Γ_k , natomiast Γ_{00} jest algebraicznym dopełnieniem elementu γ_{00} tego wyznacznika (patrz [1]).

Współczynnik korelacji cząstkowej ρ_j , $j = 1, 2, \dots, k$, otrzymuje się ze wzoru

$$(22) \quad \rho_j = \frac{-\Gamma_{j0}}{\sqrt{\Gamma_{00} \Gamma_{jj}}}$$

(patrz tamże).

Współczynnik korelacji cząstkowej ma za zadanie mierzyć wpływ zmiennej X_j na zmienną Y , po wyeliminowaniu wpływu pozostałych zmiennych $X_1, X_2, \dots, X_{j-1}, X_{j+1}, \dots, X_k$ na zmienną Y . Zadanie to współczynnik korelacji cząstkowej wypełnia źle. Po pierwsze, jeżeli zmienne X_1, X_2, \dots, X_k są silnie skorelowane zarówno ze zmienną Y , jak i między sobą, to wszystkie współczynniki korelacji cząstkowej będą przyjmować małe co do modułu wartości, gdyż po wyeliminowaniu wpływu zmiennych $X_1, X_2, \dots, X_{j-1}, X_{j+1}, \dots, X_k$ na zmienną Y wpływ zmiennej X_j (silnie przecież skorelowanej z tymi zmiennymi) na zmienną Y będzie już znikomy, mimo silnego skorelowania X_j z Y .

Po drugie, dla współczynników korelacji cząstkowej powinna być spełniona nierówność

$$\sum_{j=1}^k \rho_j \leq \rho_0 \leq 1,$$

gdyż oczywiście siła łącznego wpływu wywieranego przez zmienne X_1, X_2, \dots, X_k , mierzona za pomocą współczynników korelacji cząstkowej, nie powinna przekroczyć jedności, a tymczasem łatwo można podać przykłady, w których nierówność powyższa jest fałszywa.

Właśnie z tych dwóch względów do mierzenia siły wpływu wywieranego na zmienną Y przez poszczególne zmienne X_1, X_2, \dots, X_k z osobna, jak i wspólnie, w dowolnej kombinacji, jaką można utworzyć ze zbioru k -elementowego, stosować będziemy wielkości H_j i H_0 , do zdefiniowania których właśnie przechodzimy.

OKREŚLENIE 7. Wielkość

$$(23) \quad H_j = \frac{\rho_{j0}^2}{1 + \sum_{i \neq j} |\rho_{ij}|}, \quad i = 1, 2, \dots, k,$$

nazywać będziemy wskaźnikiem pojemności indywidualnej nośnika informacji X_j .

Z określenia 7 wynika bezpośrednio, że

$$0 \leq H_j \leq 1.$$

⁽¹⁾ W tym i polskie maszyny serii „Odra” produkcji Wrocławskich Zakładów Elektronicznych ELWRO.

OKREŚLENIE 8. Wielkość

$$H_0 = \sum_{j=1}^k H_j$$

nazywać będziemy *wskaznikiem pojemności integralnej* zbioru nośników informacji X_1, X_2, \dots, X_k . Z definicji parametru H_0 wynika, że może on przyjmować tylko wartości nieujemne. Nasuwa się przypuszczenie, że parametr ten nie przekracza jedności. Potwierdzenie lub obalenie tej hipotezy środkami formalnymi jest rzeczą trudną. Należy więc odwołać się do metod eksperymentalnych. Wykonane przez autora doświadczenia zdają się wskazywać, że znalezienie kontrprzykładu, który pozwoliłby na obalenie hipotezy, nie jest sprawą łatwą. Zauważmy mianowicie, że jeżeli zmienna X_i jest skorelowana ze zmienną Y , to

$$X_i = \rho_{i0} Y + Z_i$$

przy założeniu, że $E(X_i) = E(Y) = 0$ oraz $V(X_i) = V(Y) = 1$. Analogiczną równość możemy napisać dla zmiennych X_j oraz Y :

$$X_j = \rho_{j0} Y + Z_j.$$

Stąd

$$\rho_{ij} = E(X_i X_j) = \rho_{i0} \rho_{j0} E(Y^2) + \rho_{i0} E(Y Z_j) + \rho_{j0} E(Y Z_i) + E(Z_i Z_j).$$

Ale

$$\rho_{i0} E(Y Z_j) = \rho_{i0} E[(X_j - \rho_{j0} Y) Y] = \rho_{i0} E(X_j Y) - \rho_{i0} \rho_{j0} E(Y^2) = 0.$$

Podobnie

$$\rho_{j0} E(Y Z_i) = 0.$$

Wobec tego

$$\rho_{ij} = \rho_{i0} \rho_{j0} + E(Z_i Z_j).$$

Równość ta pozwala udowodnić następujące

TWIERDZENIE 5. Jeżeli $E(Z_i Z_j) = 0$, to

$$0 \leq H_0 \leq 1.$$

Dowód. Mamy

$$H_0 = \sum_{j=1}^k \frac{\rho_{j0}^2}{1 + \sum_{i \neq j} |\rho_{ij}|}.$$

Zauważmy, że ułamek

$$\begin{aligned} H_j &= \frac{\rho_{j0}^2}{1 + \sum_{i \neq j} |\rho_{ij}|} = \frac{\rho_{j0}^2}{1 + \sum_{i \neq j} |\rho_{i0}| |\rho_{j0}|} = \frac{\rho_{j0}^2}{1 - \rho_{j0}^2 + |\rho_{j0}| \sum_{i=1}^k |\rho_{i0}|} = \\ &= \frac{\rho_{j0}^2}{1 - \rho_{j0}^2 + k |\rho_{j0}| |\rho|} \leq \frac{\rho_{j0}^2}{k |\rho_{j0}| |\rho|}, \end{aligned}$$

gdyż $1 - \rho_{jo}^2 \geq 0$. Wobec tego

$$\frac{\rho_{jo}^2}{1 + \sum_{i \neq j} |\rho_{ij}|} \leq \frac{|\rho_{jo}|}{k|\rho|}.$$

Stąd

$$H_0 \leq \sum_{j=1}^k \frac{|\rho_{jo}|}{k|\rho|} = 1,$$

co należało okazać.

Warto tu jeszcze raz podkreślić, że o ile parametry H_1, H_2, \dots, H_k służą do mierzenia indywidualnego wpływu, jaki na zmienną Y wywierają zmienne X_1, X_2, \dots, X_k , a parametr H_0 ma za zadanie mierzyć łączny wpływ, jaki te zmienne wywierają na Y , o tyle różnica $1 - H_0$ określa *silę wpływu, jaki na zmienną Y wywierają wszystkie zmienne nie objęte badaniem*, których oczywiście ani nie znamy, ani nie jesteśmy w stanie poznać. Im różnica $1 - H_0$ jest mniejsza, tym węższy jest margines naszej ignorancji w procesie opisu zachowania się zmiennej Y .

Pytania kontrolne i zadania

1. Podać określenie parametru opisowego.
2. Na jakie dwie grupy można podzielić parametry opisowe? Podać przykłady parametrów, należących do każdej z tych grup i wyjaśnić, czym różnią się parametry każdej grupy.
3. Podać określenie wartości przeciętnej zmiennej losowej typu skokowego i typu ciągłego.
4. Z partii towaru o wadliwości $w = 10\%$ pobrano w drodze losowania próbę liczącą 1500 sztuk. Obliczyć wartość przeciętną liczby braków w próbce.
5. Zmienna losowa X ma rozkład prostokątny. Gęstość w tym rozkładzie określona jest wzorami

$$f(x) = \begin{cases} 1 & \text{dla } 0 < x < 1, \\ 0 & \text{dla } x \leq 0 \text{ lub } x \geq 1. \end{cases}$$

Obliczyć wartość przeciętną zmiennej losowej X .

6. Czemu równa się wartość przeciętna w rozkładzie dwumianowym, w rozkładzie hipergeometrycznym i w rozkładzie Poissona?
7. Rzucamy dwiema kościemi do gry. Obliczyć wartość przeciętną sumy oczek wyrzuconych w jednym rzucie na obu kościach.
8. Rzucamy kością i monetą. Wyniki rzutu kością niech będą realizacjami zmiennej losowej X , natomiast wyniki rzutu monetą niech będą realizacjami zmiennej losowej Y (umawiamy się, że wyrzucenie reszki traktować będziemy jako wyrzucenie zera, natomiast wyrzucenie orła uważać będziemy jako wyrzucenie jedynki). Obliczyć $E(XY)$.
9. Zmienna losowa X ma rozkład dwumianowy o parametrach p, n . Znaleźć wartość przeciętną zmiennej losowej X/n .
10. Podać określenie wariancji zmiennej losowej typu skokowego i ciągłego.
11. Czemu równa się wariancja w rozkładzie dwumianowym i w rozkładzie hipergeometrycznym?
12. Zmienna losowa X ma rozkład Poissona. Wartość przeciętna $E(X) = 2$. Obliczyć wariancję $V(X)$.
13. Zmienna losowa X ma rozkład dwumianowy o parametrach n, p . Obliczyć wariancję zmiennej losowej X/n .

14. Co nazywamy odchyleniem standardowym?

15. Rzucamy kościami do gry. Obliczyć wartość przeciętną i odchylenie standardowe zmiennej losowej X , którą jest liczba wyrzuconych oczek.

16. Dana jest zmienna losowa, która jest sumą n zmiennych losowych o jednakowym rozkładzie, i zmienna losowa nX , przy czym rozkład zmiennej X jest taki sam jak rozkład zmiennych tworzących wspomnianą sumę. Czy wariancja sumy równa się wariancji nX ?

17. Podać określenie odchylenia przeciętnego.

18. Zmienna losowa X przybiera wartości $x_1 = 10, x_2 = 100, x_3 = 1000$ odpowiednio z prawdopodobieństwami $p_1 = 0,5, p_2 = 0,4, p_3 = 0,1$. Obliczyć odchylenie przeciętne zmiennej losowej X .

19.. Zmienna losowa X ma rozkład normalny $N(0, 1)$. Obliczyć odchylenie przeciętne zmiennej X .

20. Rzucamy 10 razy monetą. Oznaczając wyrzucenie reszki liczbą 0, a wyrzucenie orła liczbą 1, obliczyć odchylenie przeciętne wyników rzutów.

21. Wyjaśnić, co rozumiemy przez momenty absolutne i względne oraz przez momenty zwykłe i centralne. Podać odpowiednie przykłady.

22. Podać określenie funkcji charakterystycznej.

23. Za pomocą funkcji charakterystycznej obliczyć wartość przeciętną, wariancję i odchylenie standardowe w rozkładzie prostokątnym, w rozkładzie Poissona i w rozkładzie normalnym.

24. Podać definicję momentu względnego, momentu zwykłego i momentu centralnego dwuwymiarowej zmiennej losowej skokowej i ciągłej.

25. Wymienić ważniejsze momenty pierwszego i drugiego rzędu zmiennej losowej dwuwymiarowej.

26. Co to są momenty warunkowe? Podać przykłady takich momentów.

27. Podać definicje regresji pierwszego i drugiego rodzaju i wyjaśnić różnicę między tymi pojęciami.

28. Wyjaśnić związek, jaki zachodzi między momentami warunkowymi i regresją pierwszego rodzaju.

29. Jaką własność ekstremalną ma linia regresji pierwszego rodzaju?

30. Podać definicje współczynników regresji liniowej.

31. Na czym polega różnica między regresją Y względem X i regresją X względem Y ?

32. Jak mierzy się dokładność oceny wartości, jakie przybiera zmienna Y , gdy jako narzędzie oceny stosuje się regresję liniową $y = \alpha x + \beta$? Co to jest standardowy błąd oceny?

33. Podać definicje współczynnika korelacji i stosunku korelacyjnego.

34. Wymienić własności współczynnika korelacji.

35. Czy brak korelacji między zmiennymi losowymi jest równoważny niezależności zmiennych losowych?

36. Podać definicję regresji ortogonalnej.

37. Podać definicję współczynnika zależności, w przypadku gdy obie zmienne losowe są ciągle i w przypadku gdy zmienne losowe są dyskretnie.

38. Spróbuj objaśnić istotne różnice i podobieństwa między współczynnikiem korelacji i współczynnikiem zależności.

39. Podać wzór na gęstość dwuwymiarowego rozkładu normalnego. Od ilu parametrów zależy gęstość w tym rozkładzie?

40. Wymienić ważniejsze własności dwuwymiarowego rozkładu normalnego.

41. Warunkowa wartość oczekiwana w dwuwymiarowym rozkładzie normalnym jest liniowa. Jak to można wyrazić inaczej?

42. Warunkowa wariancja jest stała w dwuwymiarowym rozkładzie normalnym, stały więc jest również pierwiastek z tej wariancji. Jak to można wysłowić inaczej? Jaka jest tego geometryczna interpretacja?

43. Co to jest wektor losowy? Podać synonimy tego pojęcia.

44. Co to jest punkt empiryczny?

45. Co to znaczy, że zmienne losowe X_1, X_2, \dots, X_k są niezależne?

46. Wyjaśnić analogię między funkcją gęstości dwu- i wielowymiarowego rozkładu normalnego.

- 47.** Jak oblicza się współczynnik korelacji wielowymiarowej (wielokrotnej, wielorakiej) i współczynnik korelacji cząstkowej? Zinterpretować sens tych dwóch miar korelacji.
- 48.** Podać wzór na współczynnik zbieżności.
- 49.** Jaki związek zachodzi między współczynnikiem zbieżności i współczynnikiem korelacji?
- 50.** Co to jest wskaźnik pojemności indywidualnej nośnika informacji?
- 51.** Jaki związek łączy wskaźniki pojemności indywidualnej ze wskaźnikiem pojemności integralnej?
- 52.** W jaki sposób wykorzystujemy współczynnik pojemności integralnej do optymalnego wyboru zmiennych, które powinny wejść do równania regresji ?

5.1. PRAWO WIELKICH LICZB

5.1.1. Wprowadzenie

Prawdopodobieństwo jest pojęciem matematycznym, posiadającym ściśle określony sens. Oczywiście matematyczna treść pojęcia prawdopodobieństwa jest inna niż jego treść potoczna. Prawdopodobieństwo rozumiane w sensie matematycznym jest pewną funkcją, określoną na zbiorze zdarzeń losowych. W rozumieniu potocznym prawdopodobieństwo jest natomiast miarą możliwości zajścia interesującego nas zdarzenia losowego. Między matematyczną i praktyczną interpretacją terminu „prawdopodobieństwo” nie ma żadnego konfliktu.

Jest rzeczą znaną od dawna, że zdarzenia losowe, rozpatrywane w masie, wykazują pewną prawidłowość w swym występowaniu. Prawidłowość ta sprowadza się do tego, że jedne zdarzenia losowe występują w danych warunkach rzadziej, a inne częściej. Zdarzenia występujące częściej przyjęto nazywać zdarzeniami bardziej prawdopodobnymi, a zdarzenia pojawiające się rzadziej – zdarzeniami mniej prawdopodobnymi. Termin „prawdopodobieństwo” został więc skojarzony z częstością występowania zdarzeń losowych. Częstość występowania zdarzeń może być wyrażona liczbowo. Liczba, wyrażająca częstość, jest liczbą doświadczalną. Prawdopodobieństwo – rozumiane w sensie matematycznym – jest abstrakcyjnym uogólnieniem częstości i tak jak każde pojęcie matematyczne obok treści formalnej ma również obiektywną treść realną. O tym, że jedne zdarzenia zachodzą częściej, a inne rzadziej, wiemy dobrze z codziennego doświadczenia. Stwierdzając, że jedne spośród zdarzeń losowych pojawiają się częściej, a inne rzadziej, zestawiamy ze sobą liczbę realizacji danego zdarzenia losowego i ogólną liczbę doświadczeń (przez doświadczenie rozumiemy realizację warunków, w których dane zdarzenie może się pojawić). Interesując się częstością występowania zdarzeń – interesujemy się nie każdą z tych liczb oddzielnie, lecz ich wzajemnym stosunkiem. Jeśli z praktyki wiadomo, że zdarzenie losowe A zachodzi częściej niż zdarzenie losowe B , to oznacza to, że przy zmianie liczby doświadczeń zmieni się także liczba realizacji zdarzeń losowych A i B , lecz dla danej liczby doświadczeń liczba realizacji zdarzenia A przy wielokrotnym powtarzaniu serii doświadczeń będzie na ogół większa od liczby realizacji zdarzenia B . Powiedzieliśmy, że tak dzieje się „na ogół”. Znaczy to, że od tej zasady mogą się zdarzyć odstępstwa (mamy bowiem do czynienia ze zdarzeniami losowymi), lecz odstępstwa takie zdarzają się rzadko, tym rzadziej, im liczba doświadczeń jest większa. Fakt ten, od dawna znany z doświadczenia, przez długi czas nie posiadał interpretacji naukowej.

PRZYKŁAD 1. Rzucamy n razy monetą i kością do gry. Wyrzucenie orła uważać będącym za zdarzenie równoważne zajściu zdarzenia A , natomiast wyrzucenie szóstki traktować będącym jako zdarzenie równoważne wystąpieniu zdarzenia B . Wiemy z góry, rzucając monetą i kością do gry, że zdarzenie A powinno pojawiać się częściej niż zdarzenie B . Zdajemy sobie jednak również sprawę z tego, że przy wielokrotnym powtarzaniu serii n rzutów mogą wystąpić takie serie, w których prawidłowość ta będzie naruszona. Od czego w danych warunkach (warunki te w naszym przykładzie sprowadzają się do tego, że w trakcie dokonywania rzutów posługujemy się stale tą samą monetą i tą samą kostką do gry) zależy możliwość naruszenia tej prawidłowości? Praktyka uczy, że możliwość ta zależy od liczb rzutów, zmniejszając się, gdy liczba rzutów rośnie.

Znany z doświadczenia fakt, że prawidłowość w występowaniu zjawisk losowych akcentuje się coraz silniej w miarę wzrostu liczby doświadczeń, nie budził nigdy sprzeciwów intuicyjnych czy też logicznych. Naukowe wyjaśnienie tego faktu podane zostało jednak po raz pierwszy dopiero w roku 1713, gdy pośmiertnie opublikowano dzieło J. Bernoulliego *De Arte Coniectandi Tractatus*⁽¹⁾. W dziele tym Bernoulli udowodnił twierdzenie (ze względu na przywiązywaną do niego wagę nazwane przez autora *twierdzeniem złotym*), z którego wynika, że prawdopodobieństwo tego, iż częstość zdarzenia losowego A będzie wykazywać wahania mniejsze co do wartości bezwzględnej od dowolnie małej dodatniej liczby ε , dąży do jedności, gdy liczba doświadczeń n nieskończonie wzrasta. Z twierdzenia tego dalej wynika, że jeśli możemy twierdzić, że wahania częstości zdarzenia A wraz ze wzrostem liczby doświadczeń maleją do zera, to oznacza to, iż częstości te, gdy n rośnie, skupią się coraz bardziej wokół pewnej stałej liczby. Ta stała liczba jest właśnie prawdopodobieństwem zdarzenia A .

PRZYKŁAD 2. Wykonano 20 rzutów dwoma monetami. Oto liczby wyrzuconych orłów w poszczególnych rzutach:

$$1, 0, 0, 2, 1, 1, 2, 1, 1, 2, 2, 1, 1, 1, 0, 1, 0, 1, 1, 0.$$

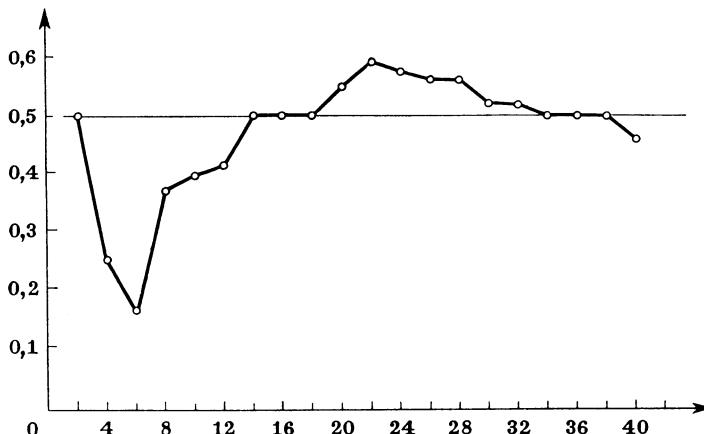
Dodając do siebie kolejne liczby wyrzuconych orłów i dzieląc przez łączną liczbę rzutów monet zbadajmy zachowanie się częstości wyrzucenia orła, gdy liczba rzutów monet rośnie. Częstości te przedstawiają się następująco:

$$\begin{aligned} &0,500, 0,250, 0,167, 0,375, 0,400, 0,417, 0,500, 0,500, 0,500, 0,550, \\ &0,591, 0,583, 0,577, 0,571, 0,533, 0,531, 0,500, 0,500, 0,500, 0,475. \end{aligned}$$

Na rysunku 1 (str. 168) widzimy, że łamana linia częstości w miarę wzrostu n wykazuje coraz mniejsze odchylenia od linii poziomej równoległej do osi odciętych. Punkt przecięcia tej linii z osią rzędnych wyznacza uzyskane w drodze doświadczeń prawdopodobieństwo p wyrzucenia orła monetą, przy użyciu której przeprowadzono opisane doświadczenie.

Wprowadzimy obecnie nowe, ważne pojęcie probabilistyczne. Pojęciem tym jest zbieżność stochastyczna.

⁽¹⁾ Patrz 2.1.



Rys. 1

Niech będzie dany ciąg zmiennych losowych

$$(1) \quad X_1, X_2, \dots, X_n, \dots$$

OKREŚLENIE 1. Jeżeli istnieje taka zmienna losowa X , że dla każdej dodatniej liczby ε spełniona jest relacja

$$(2) \quad \lim_{n \rightarrow \infty} P\{|X_n - X| < \varepsilon\} = 1,$$

to powiadamy, że ciąg zmiennych losowych $X_1, X_2, \dots, X_n, \dots$ jest *stochastycznie zbieżny* do zmiennej losowej X .

OKREŚLENIE 2. Jeśli istnieje taki ciąg stałych $C_1, C_2, \dots, C_n, \dots$, że dla każdej dodatniej liczby ε

$$(3) \quad \lim_{n \rightarrow \infty} P\{|X_n - C_n| < \varepsilon\} = 1,$$

to o ciągu zmiennych losowych $X_1, X_2, \dots, X_n, \dots$ mówimy, że czyni on zadość *prawu wielkich liczb*.

WNIOSEK. *Każdy ciąg zmiennych losowych zbieżny stochastycznie do stałej C czyni zadość prawu wielkich liczb.*

Twierdzenie odwrotne nie jest prawdziwe.

Przystąpimy obecnie do udowodnienia systemu twierdzeń noszących wspólną nazwę *prawa wielkich liczb*.

5.1.2. Nierówność Czebyszewa

Udowodnimy ważną nierówność zwaną nierównością Czebyszewa. Ma ona doniosłe znaczenie w rachunku prawdopodobieństwa, gdyż za jej pomocą dowodzi się wielu twierdzeń, do których przede wszystkim zaliczyć należy prawo wielkich liczb.

TWIERDZENIE. Jeżeli X jest dowolną zmienną losową o skończonej wariancji, to dla dowolnej liczby $\varepsilon > 0$ zachodzi tzw. nierówność Czebyszewa

$$(1) \quad P\{|X - E(X)| \geq \varepsilon\} \leq \frac{V(X)}{\varepsilon^2}.$$

Dowód. Oznaczmy dystrybuantę zmiennej losowej X symbolem $F(x)$. Wówczas

$$\begin{aligned} (2) \quad P\{|X - E(X)| \geq \varepsilon\} &= 1 - P\{|X - E(X)| < \varepsilon\} = \\ &= 1 - P\{-\varepsilon < X - E(X) < \varepsilon\} = \\ &= 1 - P\{E(X) - \varepsilon < X < E(X) + \varepsilon\} = \\ &= \int_{-\infty}^{E(X) - \varepsilon} dF(x) + \int_{E(X) + \varepsilon}^{\infty} dF(x). \end{aligned}$$

Jak widać ze wzoru (2), całkowanie rozciąga się na obszar, w którym

$$X \leq E(X) - \varepsilon, \quad X \geq E(X) + \varepsilon,$$

czyli

$$X - E(X) \leq -\varepsilon, \quad X - E(X) \geq \varepsilon.$$

Mnożąc obie strony pierwszej z tych nierówności przez -1 otrzymujemy

$$-[X - E(X)] \geq \varepsilon, \quad X - E(X) \geq \varepsilon,$$

co można zapisać krócej w postaci

$$|X - E(X)| \geq \varepsilon.$$

Stąd otrzymujemy natychmiast, że w przedziale całkowania

$$(3) \quad \frac{|X - E(X)|}{\varepsilon} \geq 1.$$

Wobec tego

$$\begin{aligned} P\{|X - E(X)| \geq \varepsilon\} &= \int_{-\infty}^{E(X) - \varepsilon} dF(x) + \int_{E(X) + \varepsilon}^{\infty} dF(x) \leq \\ &\leq \int_{-\infty}^{E(X) - \varepsilon} \frac{[x - E(X)]^2}{\varepsilon^2} dF(x) + \int_{E(X) + \varepsilon}^{\infty} \frac{[x - E(X)]^2}{\varepsilon^2} dF(x) \leq \\ &\leq \int_{-\infty}^{\infty} \frac{[x - E(X)]^2}{\varepsilon^2} dF(x) = \frac{1}{\varepsilon^2} \int_{-\infty}^{\infty} [x - E(X)]^2 dF(x) = \frac{V(X)}{\varepsilon^2}. \end{aligned}$$

5.1.3. Twierdzenie Czebyszewa

TWIERDZENIE. Jeżeli $X_1, X_2, \dots, X_n, \dots$ oznacza ciąg niezależnych zmiennych losowych, przy czym wariancje tych zmiennych są ograniczone wspólną stałą C , tzn.

$$V(X_1) < C, \quad V(X_2) < C, \quad \dots, \quad V(X_n) < C, \quad \dots,$$

to dla każdej dodatniej liczby ε

$$(1) \quad \lim_{n \rightarrow \infty} P \left\{ \left| \frac{X_1 + X_2 + \dots + X_n - [E(X_1) + E(X_2) + \dots + E(X_n)]}{n} \right| < \varepsilon \right\} = 1.$$

Dowód. Oznaczmy

$$\frac{X_1 + X_2 + \dots + X_n}{n} = Y_n.$$

W takim razie

$$E(Y_n) = \frac{E(X_1) + E(X_2) + \dots + E(X_n)}{n}.$$

Na mocy nierówności Czebyszewa mamy

$$(2) \quad P \{ |Y_n - E(Y_n)| \geq \varepsilon \} \leq \frac{V(Y_n)}{\varepsilon^2},$$

czyli

$$(3) \quad P \{ |Y_n - E(Y_n)| < \varepsilon \} > 1 - \frac{V(Y_n)}{\varepsilon^2}.$$

Ale

$$V(Y_n) = V \left(\frac{X_1 + X_2 + \dots + X_n}{n} \right) = \frac{V(X_1) + V(X_2) + \dots + V(X_n)}{n^2} \leq \frac{nC}{n^2} = \frac{C}{n},$$

stąd

$$P \{ |Y_n - E(Y_n)| < \varepsilon \} > 1 - \frac{V(Y_n)}{\varepsilon^2} \geq 1 - \frac{C}{n\varepsilon^2}.$$

Gdy $n \rightarrow \infty$, to

$$\lim_{n \rightarrow \infty} P \{ |Y_n - E(Y_n)| < \varepsilon \} \geq 1.$$

Widzimy, że ciąg zmiennych losowych Y_n jest stochastycznie zbieżny do swej wartości przeciętnej. Ponieważ prawdopodobieństwo nie może być większe od jedności, więc zastępując znak \geq znakiem $=$ otrzymamy

$$\lim_{n \rightarrow \infty} P \{ |Y_n - E(Y_n)| < \varepsilon \} =$$

$$= \lim_{n \rightarrow \infty} \left\{ \left| \frac{X_1 + X_2 + \dots + X_n - [E(X_1) + E(X_2) + \dots + E(X_n)]}{n} \right| < \varepsilon \right\} = 1.$$

5.1.4. Twierdzenie Bernoulliego

TWIERDZENIE. Niech p oznacza prawdopodobieństwo zajścia zdarzenia A , które nazywać będziemy sukcesem. W takim razie, jeżeli X oznacza liczbę sukcesów w n niezależnych doświadczeniach, to dla każdej dodatniej liczby ε

$$(1) \quad \lim_{n \rightarrow \infty} P \left\{ \left| \frac{X}{n} - p \right| < \varepsilon \right\} = 1 .$$

Dowód. Oznaczmy

$$\frac{X}{n} = Y_n .$$

W takim razie

$$E(Y_n) = \frac{1}{n} E(X) = \frac{1}{n} np = p .$$

Na mocy nierówności Czebyszewa mamy

$$P \left\{ |Y_n - E(Y_n)| < \varepsilon \right\} > 1 - \frac{V(Y_n)}{\varepsilon^2} .$$

Ale

$$V(Y_n) = V \left(\frac{X}{n} \right) = \frac{1}{n^2} V(X) = \frac{npq}{n^2} = \frac{pq}{n} ,$$

wobec tego

$$(2) \quad P \left\{ \left| \frac{X}{n} - p \right| < \varepsilon \right\} > 1 - \frac{pq}{n\varepsilon^2} .$$

Stąd

$$\lim_{n \rightarrow \infty} P \left\{ \left| \frac{X}{n} - p \right| < \varepsilon \right\} = 1 .$$

5.1.5. Twierdzenie Poissona

TWIERDZENIE. Jeżeli $X_1, X_2, \dots, X_n, \dots$ jest ciągiem niezależnych zmiennych losowych takich, że

$$E(X_1) = E(X_2) = \dots = E(X_n) = m$$

oraz

$$V(X_1) \leq C, \quad V(X_2) \leq C, \quad \dots, \quad V(X_n) \leq C, \quad \dots,$$

to dla każdej dodatniej liczby ε

$$(1) \quad \lim_{n \rightarrow \infty} P \left\{ \left| \frac{1}{n} \sum_{i=1}^n X_i - m \right| < \varepsilon \right\} = 1 .$$

Twierdzenia tego dowodzić nie będziemy, gdyż jest ono szczególnym przypadkiem udowodnionego twierdzenia Czebyszewa.

Z twierdzenia Poissona wynika, że przy spełnieniu założeń, przyjętych w twierdzeniu, zmienna losowa, będąc średnią arytmetyczną zmiennych losowych o wspólnej wartości przeciętnej, jest stochastycznie zbieżna do tej wartości przeciętnej.

5.1.6. Twierdzenie Chinczyna

TWIERDZENIE. Jeżeli $X_1, X_2, \dots, X_n, \dots$ jest ciągiem niezależnych zmiennych losowych o tym samym rozkładzie i o wspólnej wartości przeciętnej $E(X_i) = m$, to

$$(1) \quad \lim_{n \rightarrow \infty} P \left\{ \left| \frac{1}{n} \sum_{i=1}^n X_i - m \right| < \varepsilon \right\} = 1.$$

Dowód tego twierdzenia pomijamy.

Z twierdzenia Chinczyna wynika, że do tego, aby średnia arytmetyczna zmiennych losowych o tej samej wartości przeciętnej była stochastycznie zbieżna do tej wartości przeciętnej, nie jest konieczne, aby wariancje tych zmiennych były ograniczone, tak jak to postulowały twierdzenia Czebyszewa i Poissona. W twierdzeniu Chinczyna korzysta się z założenia, że zmienne losowe $X_1, X_2, \dots, X_n, \dots$ mają taki sam rozkład, natomiast na wariancje tych zmiennych nie nakłada się żadnych ograniczeń. Podamy obecnie parę przykładów na zastosowanie udowodnionych twierdzeń.

PRZYKŁAD 1. Przy założeniu, że prawdopodobieństwo urodzenia chłopca $p = 1/2$, znaleźć prawdopodobieństwo, iż częstość urodzeń chłopca będzie co do wartości bezwzględnej różnić się od $1/2$ mniej niż o 0,01, jeżeli liczba urodzeń wynosi 100 000.

Zapiszmy symbolicznie to, co jest w zadaniu dane, i to, czego szukamy. Zapis ten przybierze postać następującą:

$$p = \frac{1}{2}, \quad \varepsilon = 0,01, \quad n = 100\,000, \quad P = ?$$

Na podstawie wzoru (2) z 5.1.4 mamy

$$P \left\{ \left| \frac{X}{n} - \frac{1}{2} \right| < 0,01 \right\} > 1 - \frac{\frac{1}{2} \cdot \frac{1}{2}}{100\,000 \cdot (0,01)^2} = 1 - \frac{1}{4 \cdot 10} = 0,975,$$

czyli szukane prawdopodobieństwo P jest większe niż 0,975.

PRZYKŁAD 2. Jak liczną próbkę należy pobrać z partii zboża siewnego, aby z prawdopodobieństwem nie mniejszym niż 0,9 można było twierdzić, że udział ziaren zdolnych do kiełkowania w próbce będzie różnił się od udziału ziaren zdolnych do kiełkowania w całej populacji mniej niż o 0,01?

Zadanie, które mamy do rozwiązania, jest bardzo ciekawe, gdyż w zadaniu tym niczego nie zakłada się o populacji. Istotnie, w poprzednim przykładzie znaleźliśmy $p = 1/2$, natomiast w przykładzie niniejszym ten parametr populacji nie jest znany. Mamy bowiem

$$p = ?, \quad \varepsilon = 0,01, \quad n = ?, \quad P \geq 0,9.$$

Zauważmy, że zgodnie z warunkami zadania

$$0,9 \leq 1 - \frac{pq}{ne^2}.$$

Nierówności tej nie możemy rozwiązać względem n , gdyż nie znamy p . Ponieważ jednak $pq=p(1-p)$ osiąga maksimum, gdy $p=q=1/2$ ⁽¹⁾, przeto możemy napisać niewymagający objaśnienia łańcuch nierówności

$$1 - \frac{pq}{ne^2} \geq 1 - \frac{1}{4ne^2} \geq 0,9.$$

Podstawiając $\varepsilon=0,01$ rozwiązujemy nierówność

$$0,9 \leq 1 - \frac{10\,000}{4n}, \quad \text{czyli} \quad 3,6n - 4n \leq -10\,000,$$

skąd

$$n \geq 25\,000.$$

Widzimy więc, że próbka powinna zawierać co najmniej 25000 ziaren. Liczba ta obliczona została dla najbardziej niekorzystnego przypadku, gdy $p=q=1/2$. Gdyby p , a tym samym i q były znane, liczność próbki, być może, byłaby mniejsza.

PRZYKŁAD 3. Jak wielką próbkę należy pobrać z partii śrub, aby z prawdopodobieństwem nie mniejszym niż 0,9 średni rozmiar skoku śrub w próbce różnił się co do bezwzględnej wartości od średniego rozmiaru skoku śrub w populacji o mniej niż 0,01 mm, jeżeli odchylenie standardowe w populacji $\sigma=0,05$ mm?

Mamy

$$V(X)=0,0025 \text{ mm}^2, \quad \varepsilon=0,01, \quad n=? , \quad P \geq 0,9.$$

Na mocy wzoru (3) z 5.1.3

$$P \left\{ \left| \frac{1}{n} \sum X - \frac{1}{n} \sum E(X) \right| < 0,01 \text{ mm} \right\} \geq 1 - \frac{0,0025 \text{ mm}^2}{n \cdot (0,01 \text{ mm})^2} \geq 0,9.$$

Stąd

$$1 - \frac{25}{n} \geq 0,9, \quad \text{czyli} \quad n \geq 250.$$

Otrzymany wynik należy interpretować w sposób następujący: jeśli z populacji, o której mowa, pobierzemy próbkę liczącą 250 śrub i obliczymy średni skok śrub w próbce równy \bar{x} , to możemy twierdzić, że

$$P \{ \bar{x} - 0,01 \text{ mm} < \mu < \bar{x} + 0,01 \text{ mm} \} \geq 0,9,$$

to znaczy że prawdopodobieństwo tego, iż nieznana wartość średniego skoku śruby μ w całej populacji zostanie objęta przedziałem

$$\langle \bar{x} - 0,01, \bar{x} + 0,01 \rangle,$$

jest nie mniejsze niż 0,9.

(1) Czytelnik przekona się o tym z łatwością, obliczając i przyrównując do zera pochodną wyrażenia $p(1-p)$.

5.2. TWIERDZENIE MOIVRE'A-LAPLACE'A

Omawiając rozkład normalny wspomnieliśmy, że dla dużych n rozkład dwumianowy może być zastąpiony rozkładem normalnym. Udowodnimy twierdzenie, które uzasadnia takie postępowanie.

Niech p oznacza prawdopodobieństwo realizacji zdarzenia losowego A (zdarzenie to nazywać będziemy sukcesem), natomiast m niech oznacza liczbę sukcesów w n niezależnych doświadczeniach. W takim razie zmienna losowa

$$T_{m,n} = \frac{X_{m,n} - p}{\sqrt{\frac{pq}{n}}}$$

czyni zadość schematowi Bernoulliego (patrz 3.3.2, określenie 1). Symbolem $X_{m,n}$ oznaczono częstość występowania zdarzenia A .

TWIERDZENIE. Jeżeli $0 < p < 1$, to

$$\lim_{n \rightarrow \infty} P\{a < T_{m,n} < b\} = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-t^2/2} dt,$$

gdzie a i b są dowolnymi liczbami rzeczywistymi.

Twierdzenie to można by wysłowić w sposób następujący: prawdopodobieństwo tego, że zmienna losowa $T_{m,n}$ przyjmie wartość spełniającą nierówność $a < t < b$, dąży do

$$\frac{1}{\sqrt{2\pi}} \int_a^b e^{-t^2/2} dt,$$

gdy n nieograniczenie rośnie.

► **Dowód.** Ponieważ zmienna losowa $X_{m,n}$ ma rozkład dwumianowy, przeto

$$\begin{aligned} P\left\{X_{m,n} = \frac{m}{n}\right\} &= P\left\{T_{m,n} = \frac{\frac{m}{n} - p}{\sqrt{\frac{pq}{n}}}\right\} = C_n^m p^m q^{n-m} = \\ &= \frac{n!}{m!(n-m)!} p^m q^{n-m} = \frac{n!}{m!r!} p^m q^r, \end{aligned}$$

gdzie $r = n - m$.

Korzystając ze wzoru Stirlinga (patrz 1.1.6, wzór (2)) mamy

$$\begin{aligned} P\left\{X_{m,n} = \frac{m}{n}\right\} &= \frac{\sqrt{2\pi n} n^n e^{-n} p^m q^r}{\sqrt{2\pi m} \sqrt{2\pi r} m^m e^{-m} r^r e^{-r}} \cdot e^\theta = \\ &= \frac{n^{n+\frac{1}{2}} p^m q^r e^\theta}{m^{m+\frac{1}{2}} r^{r+\frac{1}{2}} \sqrt{2\pi}}, \end{aligned}$$

gdzie $\theta \rightarrow 0$, gdy $n \rightarrow \infty$. Logarytmując tę równość otrzymujemy

$$\ln P \left\{ X_{m,n} = \frac{m}{n} \right\} = (n + \frac{1}{2}) \ln n + m \ln p + r \ln q - (m + \frac{1}{2}) \ln m - (r + \frac{1}{2}) \ln r - \ln \sqrt{2\pi} + \theta.$$

Ponieważ

$$t = \frac{\frac{m}{n} - p}{\sqrt{\frac{pq}{n}}},$$

przeto

$$(1) \quad m = np + tn \sqrt{\frac{pq}{n}} = np \left(1 + t \sqrt{\frac{q}{np}} \right).$$

Korzystając z zależności $r = n - m$, otrzymujemy

$$(2) \quad \begin{aligned} r &= n - np - nt \sqrt{\frac{pq}{n}} = n(1-p) - nt \sqrt{\frac{pq}{n}} = \\ &= nq - nt \sqrt{\frac{pq}{n}} = nq \left(1 - t \sqrt{\frac{p}{nq}} \right). \end{aligned}$$

Po zlogarytmowaniu równań (1) i (2) mamy

$$(3) \quad \ln m = \ln n + \ln p + \ln \left(1 + t \sqrt{\frac{q}{np}} \right),$$

$$(4) \quad \ln r = \ln n + \ln q + \ln \left(1 - t \sqrt{\frac{p}{nq}} \right).$$

Z analizy wiadomo [19], że

$$\ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \dots + (-1)^{n+1} \frac{x^n}{n} + \dots;$$

stąd

$$\ln(1+x) = x - \frac{x^2}{2} [1 + O(x)],$$

przy czym $O(x) \rightarrow 0$, gdy $x \rightarrow 0$. Wobec tego

$$(5) \quad \ln \left(1 + t \sqrt{\frac{q}{np}} \right) = t \sqrt{\frac{q}{np}} - \frac{qt^2}{2np} - \frac{qt^2}{2np} O(n^{-\frac{1}{2}})$$

oraz

$$(6) \quad \ln \left(1 - t \sqrt{\frac{p}{nq}} \right) = -t \sqrt{\frac{p}{nq}} - \frac{pt^2}{2nq} - \frac{pt^2}{2nq} O(n^{-\frac{1}{2}}).$$

Podstawiając odpowiednio równości (5) i (6) do wzorów (3) i (4) otrzymujemy

$$(7) \quad \ln m = \ln n + \ln p + t \sqrt{\frac{q}{np}} - \frac{qt^2}{2np} - \frac{qt^2}{2np} O(n^{-\frac{1}{2}}) = \ln n + \ln p + O(n^{-\frac{1}{2}}),$$

$$(8) \quad \ln r = \ln n + \ln q - t \sqrt{\frac{p}{nq}} - \frac{pt^2}{2nq} - \frac{pt^2}{2nq} O(n^{-\frac{1}{2}}) = \ln n + \ln q - O(n^{-\frac{1}{2}}).$$

Ponieważ

$$(9) \quad \ln P \left\{ X_{m,n} = \frac{m}{n} \right\} = n \ln n + \frac{1}{2} \ln n + m \ln p + r \ln q - \\ - m \ln m - r \ln r - \frac{1}{2} (\ln m + \ln r) - \ln \sqrt{2\pi} + \theta,$$

więc w celu obliczenia sumy stojącej po prawej stronie wzoru (9) obliczymy najpierw poszczególne jej składniki, a następnie je dodamy. Na mocy wzorów (1) i (2)

$$(10) \quad m \ln p = (np + t \sqrt{npq}) \ln p,$$

$$(11) \quad r \ln q = (nq - t \sqrt{npq}) \ln q.$$

Natomiast na mocy wzorów (7) i (8)

$$(12) \quad \frac{\ln m + \ln r}{2} = \frac{\ln q + \ln p}{2} + \ln n.$$

W powyższych wzorach pominięto człony zdążające do zera. W drodze prostych przekształceń łatwo wykazać, że

$$(13) \quad m \ln m = (np + t \sqrt{npq}) \left[\ln n + \ln p + t \sqrt{\frac{q}{np}} - \frac{qt^2}{2np} - \frac{qt^2}{2np} O(n^{-\frac{1}{2}}) \right] = \\ = np (\ln n + \ln p) + \sqrt{npq} (\ln n + \ln p + 1) t + \\ + \frac{qt^2}{2} - \frac{q^2 t^3}{2\sqrt{npq}} - \frac{qt^2}{2np} O(n^{-\frac{1}{2}}) np - \frac{qt^2}{2np} O(n^{-\frac{1}{2}}) \sqrt{npq} t = \\ = np (\ln n + \ln p) + \sqrt{npq} (\ln n + \ln p + 1) t + \frac{qt^2}{2} + O(n^{-\frac{1}{2}}).$$

I analogicznie

$$(14) \quad r \ln r = nq (\ln n + \ln q) - \sqrt{npq} (\ln n + \ln q + 1) t + \frac{pt^2}{2} + O(n^{-\frac{1}{2}}).$$

Wstawiając wyrażenia (10), (11), (12), (13) i (14) do wzoru (9) i dodając składniki stojące po prawej stronie znaku równości, przy jednoczesnym pomijaniu wyrażeń zdążających do zera przy n rosnącym do nieskończoności, otrzymujemy

$$\begin{aligned}
\ln P \left\{ X_{m,n} = \frac{m}{n} \right\} &= n \ln n + \frac{1}{2} \ln n + np \ln p + t \sqrt{npq} \ln p + \\
&\quad + nq \ln q - t \sqrt{npq} \ln q - np \ln n - np \ln p - \\
&\quad - t \sqrt{npq} \ln n - t \sqrt{npq} \ln p - t \sqrt{npq} - \frac{qt^2}{2} - \\
&\quad - nq \ln n - nq \ln q + t \sqrt{npq} \ln n + t \sqrt{npq} \ln q + \\
&\quad + t \sqrt{npq} - \frac{pt^2}{2} - \frac{1}{2} (\ln q + \ln p) - \ln n - \ln \sqrt{2\pi} + O(n^{-\frac{1}{2}}) = \\
&= \frac{1}{2} \ln n - \frac{1}{2} (\ln p + \ln q) - \frac{t^2}{2} - \ln n - \ln \sqrt{2\pi} + O(n^{-\frac{1}{2}}) = \\
&= -\frac{1}{2} (\ln n + \ln p + \ln q + \ln 2\pi) - \frac{t^2}{2} + O(n^{-\frac{1}{2}}) = \\
&= \ln \frac{1}{\sqrt{2\pi npq}} - \frac{t^2}{2} + O(n^{-\frac{1}{2}}).
\end{aligned}$$

Stąd

$$P \left\{ X_{m,n} = \frac{m}{n} \right\} = \frac{1}{\sqrt{2\pi npq}} e^{-t^2/2} (1 + r_n),$$

przy czym $r_n \rightarrow 0$, gdy $n \rightarrow \infty$. Wobec tego

$$\lim_{n \rightarrow \infty} P \left\{ X_{m,n} = \frac{m}{n} \right\} = \frac{1}{\sqrt{2\pi npq}} e^{-t^2/2}.$$

Oznaczając $\sqrt{npq} = \sigma$ mamy

$$\lim_{n \rightarrow \infty} P \left\{ X_{m,n} = \frac{m}{n} \right\} = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-t^2/2} = \frac{1}{\sigma} f(t),$$

gdzie

$$f(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2},$$

stąd zaś

$$\lim_{n \rightarrow \infty} P \{ a < T_{m,n} < b \} = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-t^2/2} dt,$$

co należało okazać. 

5.3. TWIERDZENIE CENTRALNE

Twierdzenie, którym zajmowaliśmy się w poprzednim paragrafie, po raz pierwszy zostało udowodnione przez Moivre'a w roku 1773. Wtedy również został przez niego odkryty rozkład normalny. Te dwa wybitne osiągnięcia naukowe przeszły jednak w owych czasach bez echo. Rozkład normalny po raz wtóry odkryty został przez Gaussa w 1809 roku i niezależnie od niego przez Laplace'a w 1812 roku. Obaj ci uczeni pracowali nad teorią błędów obserwacji. Prowadzone przez nich badania doprowadziły do odkrycia rozkładu normalnego.

Jak wiadomo, dokonując jakiegokolwiek pomiaru cechy ciągłą popełniamy zawsze pewien błąd, niezależnie od tego, z jaką dokładnością pomiar był przeprowadzony. Takie błędy noszą nazwę *błędów obserwacji*. Od czasów Gaussa i Laplace'a przyzwyczajono się uważać, że każda zmienna losowa ma rozkład normalny lub przynajmniej rozkład zbieżny do rozkładu normalnego. Mniemanie takie mogło się rozprzestrzenić, gdyż doświadczenie na ogół temu nie przeczyło. Prowadzone w XIX wieku z dużym rozmachem badania statystyczne zdawały się raczej mniemanie to potwierdzać. Zarówno w demografii, jak w biologii, agrobiologii, zootechnice, astronomii, fizyce znajdowano liczne przykłady zmiennych losowych o rozkładzie normalnym. Rozkład ten zaczęto uważać za uniwersalny rozkład statystyczny i dlatego właśnie nazwano rozkładem *normalnym*, dając tym niejako wyraz, że jest to rozkład, który zazwyczaj występuje w praktyce.

Wyjaśnieniom i tłumaczeniom, dlaczego każda (jak mniemano) zmienna losowa ma rozkład zbliżony do normalnego, poświęcono wiele wysiłku i prac. I jakkolwiek dzisiaj wiemy, że rozkład normalny nie jest jedynym rozkładem teoretycznym, który daje dobre przybliżenie rozkładów empirycznych, to jednak rozkład ten można bez wątpienia zaliczyć do rozkładów, które najczęściej mogą być stosowane do aproksymacji rozkładów doświadczalnych.

Najbardziej rozpowszechnionym poglądem, tłumaczącym przyczyny częstego występowania rozkładu normalnego, był pogląd wysunięty przez F. Bessela (1784 - 1846), wybitnego astronoma i matematyka niemieckiego. Według niego wartości obserwacji statystycznych zależą od wielu drobnych przyczyn. Przyczyny te są to swego rodzaju impulsy, które powodują, że wartości obserwacji statystycznych odchylają się od średniej arytmetycznej. Gdy impulsy te są tego rodzaju, że odchylenia mają jednakowy kierunek, to zaobserwowane odchylenie wartości cechy od średniej arytmetycznej jest duże, jeżeli natomiast impulsy wywołują odchylenia różnorodne, to sumaryczne odchylenie wartości cechy od średniej arytmetycznej jest małe.

Oczywiście prawdopodobieństwo tego, że znaczna ilość odchyleń będzie miała ten sam kierunek, jest małe, jeżeli przyjmie się nie budzące intuicyjnych sprzeciwów założenie, że odchylenia jednakowe co do wielkości bezwzględnej, a różniące się jedynie kierunkiem (znakiem), mają jednakowe prawdopodobieństwa występowania. Zakłada się również, że odchylenia są od siebie niezależne. Przy takich założeniach, na mocy twierdzenia o prawdopodobieństwie iloczynu zdarzeń niezależnych możemy uważać, że wystąpienie znacznej ilości odchyleń jednokierunkowych jest zdarzeniem mało prawdopodobnym. Prawdopodobieństwo tego zdarzenia równa się iloczynowi prawdopodobieństwa zdarzeń elementarnych.

Wynika stąd, że odchylenia duże co do bezwzględnej wartości (czyli odchylenia będące sumą małych odchyleń jednakowo skierowanych) mają małe prawdopodobieństwo występowania, a odchylenia małe mają prawdopodobieństwo duże.

Bessel traktuje odchylenia od średniej arytmetycznej tak samo, jak błędy obserwacji. Ponieważ błędy obserwacji, jak to wykazali Gauss i Laplace, mają rozkład normalny, przeto z wywodów Bessela wypływa wniosek, że każda cecha statystyczna ma rozkład normalny.

Dalsze badania szeregu wybitnych matematyków i statystyków (Lindeberg, Lévy, Lapunow, Feller, Chinczyn) doprowadziły do pełnego rozwiązania tego zagadnienia. Sformułowano i udowodniono kilka ważnych twierdzeń, podających warunki, które muszą być spełnione, aby zmienna losowa miała rozkład normalny lub zbieżny do normalnego. Twierdzenia te, nazwane nazwiskami autorów, znane są w rachunku prawdopodobieństwa pod wspólną nazwą *twierdzenia centralnego*. Nazwa ta podkreśla doniosłą rolę, jaką twierdzenie to odgrywa w rachunku prawdopodobieństwa i statystyce matematycznej.

Przystąpimy obecnie do udowodnienia twierdzenia, znanego pod nazwą *twierdzenia Lindeberga-Lévy'ego*.

Dana jest zmienna losowa

$$X = X_1 + X_2 + \dots + X_n,$$

będąca sumą n niezależnych zmiennych losowych o jednakowym rozkładzie, przy czym wartością przeciętną zmiennej X , jest m_r , a odchyleniem standardowym σ_r ($r = 1, 2, \dots, n$).

W takim razie $m_1 = m_2 = \dots = m_n$ i $\sigma_1 = \sigma_2 = \dots = \sigma_n$. Stąd

$$m = E(X) = nm_1,$$

$$\sigma^2 = V(X) = n\sigma_1^2.$$

Zmienna standaryzowana I wyraża się wzorem

$$(1) \quad I = \frac{X - m}{\sigma} = \frac{X - nm_1}{\sigma_1 \sqrt{n}} = \sum_{r=1}^n \frac{X_r - m_1}{\sigma_1 \sqrt{n}}.$$

TWIERDZENIE 1. *Gdy n rośnie do nieskończoności, rozkład zmiennej losowej I dąży do rozkładu normalnego, tzn.*

$$\lim_{n \rightarrow \infty} P\{a < I < b\} = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-t^2/2} dt.$$

Dowód. Niech

$$\varphi_1(t) = E\left(\exp\left[it \frac{X_1 - m_1}{\sigma_1 \sqrt{n}}\right]\right).$$

Symbol $\varphi_1(t)$ oznacza więc funkcję charakterystyczną zmiennej losowej, stojącej pod znakiem sumy we wzorze (1). Na mocy twierdzenia głoszącego, że funkcja charakterystyczna sumy niezależnych zmiennych losowych równa się iloczynowi funkcji charakterystycz-

nych tych zmiennych (tw. 1, § 4.6), możemy napisać, że $\varphi(t) = [\varphi_1(t)]^n$, przy czym $\varphi(t)$ oznacza funkcję charakterystyczną zmiennej losowej X . Rozwijając funkcję $\varphi_1(t)$ w szereg MacLaurina otrzymujemy

$$\varphi_1(t) = \varphi_1(0) + \frac{t}{1} \varphi'_1(0) + \frac{t^2}{2} \varphi''_1(0) + o(t^2),$$

gdzie $o(t^2)$ jest wielkością rzędu mniejszego od t^2 , jeśli $t \rightarrow 0$. Na mocy wzoru (12) z § 4.6

$$\varphi'_1(0) = iE\left(\frac{X_1 - m_1}{\sigma_1 \sqrt{n}}\right) = 0$$

oraz

$$\varphi''_1(0) = i^2 E\left(\frac{X_1 - m_1}{\sigma_1 \sqrt{n}}\right)^2 = -\frac{1}{n}.$$

Stąd

$$\varphi_1(t) = 1 - \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right),$$

a wobec tego

$$\varphi(t) = \left[1 - \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right) \right]^n.$$

Naruszając formalną poprawność dowodu pominiemy wielkość $o(t^2/n)$ jako niższego rzędu względem t^2/n . Uchroni nas to od wykonywania skomplikowanych przekształceń, a idea dowodu zostanie zachowana.

Z analizy wiadomo, że

$$\lim_{n \rightarrow \infty} \left[1 + \frac{a}{n} \right]^n = e^a;$$

w takim razie przy $a = -t^2/2$

$$\lim_{n \rightarrow \infty} \varphi(t) = e^{-t^2/2}.$$

Otrzymaliśmy funkcję charakterystyczną rozkładu normalnego. Ze wzoru (25), § 4.6, wynika więc, że zmienna losowa X ma rozkład normalny.

Z twierdzenia 1 wynika następujący

WNIOSEK 1. Jeżeli zmienne X_1, X_2, \dots, X_n są od siebie niezależne i posiadają jednakowy rozkład, przy czym $E(X_1) = m_1$ i $V(X_1) = \sigma_1^2$, to rozkład zmiennej losowej

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n}$$

dąży do rozkładu normalnego

$$N\left(m_1, \frac{\sigma_1}{\sqrt{n}}\right).$$

Niech

$$X = X_1 + X_2 + \dots + X_n,$$

przy czym zmienne losowe X_r , ($r=1, 2, \dots, n$) są od siebie niezależne i mają dowolne rozkłady. Założmy, że dla każdego r

$$W_r^3 = E(|X_r - m_r|^3) < \infty$$

i oznaczmy

$$W^3 = W_1^3 + W_2^3 + \dots + W_n^3,$$

$$\sigma^2 = \sigma_1^2 + \sigma_2^2 + \dots + \sigma_n^2.$$

TWIERDZENIE 2 (LAPUNOWA). Jeżeli⁽¹⁾

$$(2) \quad \lim_{n \rightarrow \infty} \frac{W}{\sigma} = 0,$$

to

$$\lim_{n \rightarrow \infty} P\{a < I < b\} = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-t^2/2} dt,$$

gdzie

$$I = \frac{X_1 + X_2 + \dots + X_n - (m_1 + m_2 + \dots + m_n)}{\sqrt{\sigma_1^2 + \sigma_2^2 + \dots + \sigma_n^2}}.$$

Dowód tego twierdzenia znajdzie czytelnik w książce Fisza [8], str. 214 - 218, lub Craméra [1], str. 211 - 212.

Z twierdzenia Lapunowa wynika niezmiernie ważny wniosek praktyczny: jeżeli liczba zmiennych losowych nieograniczenie wzrasta, to przy założeniu, że jest spełniony warunek (2), rozkład średniej arytmetycznej tych zmiennych dąży do rozkładu normalnego.

Twierdzenie Lapunowa wyjaśnia ze stanowiska teoretycznego dobrze znany z doświadczenia fakt, że empiryczne zmienne losowe mają na ogół rozkład zbliżony do normalnego.

Spotykane w praktyce zmienne losowe mogą być przeważnie traktowane jako sumy znacznej liczby zmiennych losowych, z których żadna nie ma dominującego wpływu na wielkość tej sumy. Wyobraźmy sobie, na przykład, że badamy rozkład wydajności pracy robotników jakiegoś dużego zakładu produkcyjnego. Oczywiście wydajność pracy osiągnięta przez poszczególnego robotnika zależy od wielu czynników, jak wiek robotnika, jego umiejętności, doświadczenie, siła fizyczna, stan zdrowia, samopoczucie, pracowitość, jakość narzędzi i surowca. Każdy z tych czynników wywiera wpływ na osiągniętą przez robotnika wydajność pracy. Wobec tego wydajność pracy jest efektem łącznego działania wszystkich czynników. Ponieważ każdy z czynników może być uważany za zmieniącą losową, więc wydajność pracy jest sumą tych zmiennych losowych, a tym samym jest

(1) Warunek ten w praktyce jest przeważnie spełniony.

także zmienną losową, przy czym na mocy twierdzenia Lapunowa rozkład tej zmiennej będzie zbliżony do rozkładu normalnego. Centralne twierdzenie ma tak doniosłe znaczenie teoretyczne i praktyczne, że z powodzeniem może być uznane za najważniejsze twierdzenie rachunku prawdopodobieństwa. Twierdzeniem tym zamykamy nasze rozważania z zakresu probabilistyki i przechodzimy do omówienia niektórych zagadnień ze statystyki matematycznej.

Pytania kontrolne i zadania

1. Podać określenie zbieżności stochastycznej.
2. Kiedy mówimy, że ciąg zmiennych losowych $X_1, X_2, \dots, X_n, \dots$ czyni zadość prawu wielkich liczb?
3. Czy prawdziwe jest twierdzenie, że jeśli ciąg zmiennych losowych $X_1, X_2, \dots, X_n, \dots$ czyni zadość prawu wielkich liczb, to ciąg ten jest stochastycznie zbieżny do stałej C ? Odpowiedź uzasadnić.
4. Jak brzmi nierówność Czebyszewa?
5. Podać twierdzenie Czebyszewa.
6. Podać twierdzenie Bernoulliego.
7. Podać twierdzenie Poissona.
8. Podać twierdzenie Chinczyna.
9. Rzucamy monetą. Przyjmując, że wyrzucenie orła ma taką samą szansę jak wyrzucenie reszki, znaleźć prawdopodobieństwo tego, że rzucając monetą 10 000 razy otrzymamyczęstość pojawienia się orła różną od teoretycznego prawdopodobieństwa co do bezwzględnej wartości mniej niż o 0,01, było większe od 0,9?
10. Ile razy należy rzucić monetę, aby prawdopodobieństwo tego, że częstość wyrzucenia orła będzie się różniła od teoretycznego prawdopodobieństwa co do bezwzględnej wartości mniej niż o 0,01, było większe od 0,9?
11. Jak brzmi twierdzenie Moivre'a-Laplace'a?
12. Jak brzmi twierdzenie Lindeberga-Lévy'ego?
13. Jak brzmi twierdzenie Lapunowa?
14. Jakie warunki muszą być spełnione w twierdzeniu Lapunowa?
15. Podać przykłady ilustrujące działanie prawa wielkich liczb.
16. Podać uzasadnienie, dlaczego w praktyce mamy tak często do czynienia z rozkładem normalnym.
17. Podać przykłady zmiennych losowych, o których można sądzić, że mają rozkład normalny.

Część III

STATYSTYKA MATEMATYCZNA

6.1. DEFINICJE I POJĘCIA STATYSTYCZNE

W latach 1949 - 1954 zarówno w Związku Radzieckim, jak i u nas toczyła się ożywiona dyskusja na temat przedmiotu i metody nauki statystyki. Podsumowanie wyników odbyło się w marcu 1954 roku w Moskwie. Na naradzie tej uznano, że statystyka matematyczna jest gałęzią matematyki.

Zasadniczym aparatem naukowym, jakim posługuje się statystyka matematyczna, jest rachunek prawdopodobieństwa. Oto co pisze W. S. Niemczynow w tej sprawie: „... statystyka matematyczna jest, rzecz oczywista, gałęzią matematyki, lecz gałęzią specyficzną, taką mianowicie, której kategorie i metody są uwarunkowane obiektywną specyfiką przedmiotu jej zastosowania i zadaniem poznania właściwości tego przedmiotu. Zadaniem tym jest dostarczenie uogólniającej, abstrakcyjno-liczbowej charakterystyki masowej zbiorowości przy całkowitym pominięciu realnej treści zjawisk ...”⁽¹⁾

Statystyka matematyczna bada prawidłowości w masowych zjawiskach i opisuje te prawidłowości za pomocą liczb. Przez *zjawiska masowe* należy rozumieć takie zjawiska, które mogą występować nieograniczoną ilość razy.

Przedmiotem badań statystyki matematycznej (która dalej nazywać będziemy krótko statystyką) są zbiory, których elementami są wszelkiego rodzaju obiekty materialne i zjawiska. Zbiory te nazywają się *populacjami statystycznymi*, a ich elementy – *jednostkami statystycznymi*.

Jednostki statystyczne mogą mieć wiele różnych właściwości. Właściwości, które podlegają badaniu statystycznemu, nazywają się *cechami statystycznymi*.

Od czasu do czasu przeprowadzane są w Polsce zakrojone na dużą skalę badania antropometryczne. Ludzie objęci badaniem stanowią populację statystyczną. Jednostką statystyczną jest, w tym przykładzie, każda zbadana osoba. Badaniem objęte są, między innymi, takie właściwości osób, jak wiek, płeć, zawód, waga, wzrost, szerokość ramion, obwód klatki piersiowej, barwa oczu, barwa włosów. Są to wszystko przykłady cech statystycznych.

Cechy statystyczne dzielą się na cechy mierzalne i cechy niemierzalne. *Cechą mierzalną* nazywamy taką cechę, która może być wyrażona za pomocą liczby, pochodzącej z pomiaru lub policzenia, natomiast *cechą niemierzalną* nazywamy cechę, która może być wyrażona jedynie za pomocą określenia słownego, nie może być natomiast wyrażona za pomocą liczby. Wiek, wzrost, waga osób są przykładami cech mierzalnych, natomiast płeć, zawód, barwa włosów – to przykłady cech niemierzalnych.

⁽¹⁾ Ученые записки по статистике I, Москва 1955, str. 5.

Cecha statystyczna u poszczególnych jednostek statystycznych przybiera różne wartości. Wartości te, zarejestrowane w trakcie badania statystycznego, nazywają się *obserwacjami statystycznymi*. Zbiór uzyskanych w czasie badania obserwacji statystycznych nazywa się *materiałem statystycznym*.

Jeżeli zbiór obserwacji statystycznych dotyczy cechy mierzonej i jeżeli obserwacje statystyczne zostały uszeregowane według wartości rosnących, to zbiór ten nazywa się *statystycznym szeregiem uporządkowanym* lub krótko *szeregiem uporządkowanym*.

W szeregu uporządkowanym wymienione są wszystkie obserwacje statystyczne. Ponieważ na ogół populacje statystyczne są bardzo liczne, przeto w skład szeregu uporządkowanego wchodzi zazwyczaj znaczna liczba obserwacji. Jest to powodem, że szereg taki jest nieczytelny, a tym samym nie nadaje się do jakiegokolwiek analizy naukowej. Znacznie większe znaczenie poznawcze ma tzw. *szereg rozdzielczy*. Budowa szeregu rozdzielczego polega na zaliczaniu poszczególnych obserwacji do utworzonych zawczasu przedziałów liczbowych. Przedziały te nazywają się *przedziałami klasowymi* lub *klasami szeregu rozdzielczego*. Liczba obserwacji wchodzących w skład poszczególnych klas nazywa się *liczebnością danej klasy*. Liczby określające wielkość poszczególnych przedziałów klasowych noszą nazwę *granic przedziału klasowego*. Suma granic danej klasy podzielona przez 2 nazywa się *środkiem przedziału klasowego* lub *wariantem klasowym*.

Oto przykład szeregu rozdzielczego:

Nr klasy szeregu rozdzielczego	Wiek uczniów szkół podstawowych	Liczliwość (w tys.) <i>n</i>
1	6,5 i mniej	20
2	6,5 - 7,5	573
3	7,5 - 8,5	529
4	8,5 - 9,5	386
5	9,5 - 10,5	387
6	10,5 - 11,5	375
7	11,5 - 12,5	373
8	12,5 - 13,5	351
9	13,5 - 14,5	150
10	14,5 - 15,5	49
11	15,5 i więcej	10

Powiedzieliśmy już, że liczebnością klasy jest liczba obserwacji wchodzących w skład danej klasy. Dzieląc liczebność danej klasy przez sumę wszystkich liczebności szeregu rozdzielczego otrzymamy *częstość*⁽¹⁾ tej klasy. Częstość mówi o tym, jaki jest udział obserwacji, należących do danej klasy, w ogólnej masie wszystkich obserwacji. Dodając do częstości danej klasy sumę częstości klas poprzednich otrzymamy *częstość skumulowaną*.

Szereg rozdzielczy nazywa się inaczej *rozkładem empirycznym*.

⁽¹⁾ Zwaną również *częstością względną*.

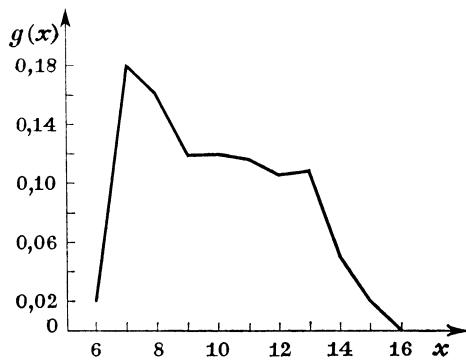
Oznaczając środek przedziału klasowego symbolem X , częstość – symbolem $g(x)$, zaś częstość skumulowaną – symbolem $G(x)$, szereg rozdzielczy można przedstawić w postaci tablicy 1.

Tablica 1

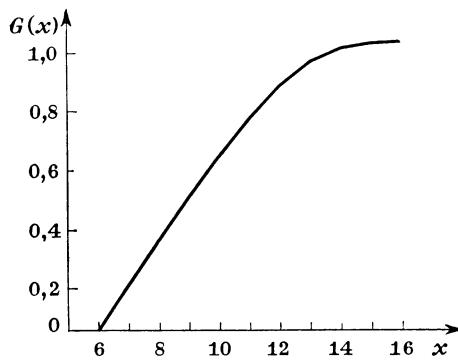
Nr	Przedziały klasowe	X	n	n skumulowane	Częstość $g(x)$	Częstość skumulowana $G(x)$
1	6,5 i mniej	6	20	20	0,006	0,006
2	6,5 - 7,5	7	573	593	0,179	0,184
3	7,5 - 8,5	8	529	1122	0,165	0,350
4	8,5 - 9,5	9	386	1508	0,121	0,471
5	9,5 - 10,5	10	387	1895	0,121	0,592
6	10,5 - 11,5	11	375	2270	0,117	0,709
7	11,5 - 12,5	12	373	2643	0,110	0,819
8	12,5 - 13,5	13	351	2994	0,109	0,928
9	13,5 - 14,5	14	150	3144	0,053	0,981
10	14,5 - 15,5	15	49	3193	0,016	0,997
11	15,5 i więcej	16	10	3203	0,003	1,000
			3203		1,000	

Szereg rozdzielczy może być też przedstawiony graficznie. Rysunek 1 wyobraża krzywą częstości, natomiast rysunek 2 przedstawia krzywą częstości skumulowanej.

Przyglądając się rozkładowi empirycznemu i obu wykresom dostrzegamy natychmiast zbieżność podstawowych pojęć statystycznych i probabilistycznych. Istotnie, odpowiednikiem pojęcia zmiennej losowej jest cecha statystyczna, odpowiednikiem rozkładu teore-



Rys. 1



Rys. 2

tycznego jest rozkład empiryczny, odpowiednikiem dystrybuanty jest częstość skumulowana. Wyjaśnia to związek, jaki zachodzi między rachunkiem prawdopodobieństwa i statystyką matematyczną.

W przenośni można powiedzieć, iż rachunek prawdopodobieństwa to statystyka matematyczna, odziana w szatę abstrakcji matematycznej. Korzystając ze związku, jaki

łączy obie nauki – będziemy często kategorie statystyczne zastępować kategoriami probabilistycznymi, nie czyniąc przy tym żadnych omówień, jeśli nie będzie to grozić nieporozumieniem.

6.2. UWAGI O BADANIU CZEŚCIOWYM I METODZIE REPREZENTACYJNEJ⁽¹⁾

Badania statystyczne można podzielić na dwa rodzaje:

1. badania wyczerpujące⁽²⁾, czyli całkowite,
2. badania niewyczerpujące, czyli częściowe.

Przez *badanie wyczerpujące* rozumie się tego rodzaju badanie statystyczne, w którym badaniu podlega cała populacja. Jakkolwiek jest to może nieoczekiwane, jednak tego rodzaju badanie ma miejsce w praktyce stosunkowo rzadko. Znacznie częściej natomiast stosuje się drugi rodzaj badania, to znaczy badanie niewyczerpujące, czyli częściowe. Przez *badanie częściowe* rozumieć należy takie badanie statystyczne, w którym badaniu podlega jedynie część populacji.

Cała interesująca nas populacja statystyczna nazywa się *populacją generalną*, natomiast jej część nosi nazwę *populacji próbnej* lub krótko *próbki*. Wobec tego możemy powiedzieć, że z badaniem całkowitym mamy do czynienia wtedy, gdy badaniem objęta jest populacja generalna, natomiast badanie częściowe ma miejsce wtedy, gdy badaniem objęta jest populacja próbna.

Zajmiemy się obecnie omówieniem ważniejszych przyczyn częstszego korzystania w praktyce statystycznej z badania częściowego niż z badania całkowitego.

1. **Populacje nieograniczone.** Statystyka matematyczna jest nauką, która zajmuje się wykrywaniem, badaniem i opisem prawidłowości w zjawiskach masowych. Jak wiadomo, przez zjawiska masowe rozumiemy zasadniczo takie zjawiska, które mogą występować nieograniczoną ilość razy. Już z definicji przedmiotu statystyki matematycznej wynika, że ma ona do czynienia z populacjami nieograniczonymi, gdyż tylko takie populacje sensu stricto spełniają postulat masowości.

Większość podstawowych twierdzeń rachunku prawdopodobieństwa i statystyki matematycznej zakłada nieograniczoność populacji (prawo wielkich liczb, twierdzenia graniczne). Statystyka nie ma niczego do powiedzenia, gdy chodzi o zjawiska indywidualne. Wszelkie rozumowanie statystyczne ma sens jedynie wówczas, gdy obserwacje statystyczne czynią zadość warunkowi masowości.

Jest rzeczą zrozumiałą, że populacja nieograniczona (tzn. zawierająca nieskończoną ilość jednostek statystycznych) nie może być zbadana całkowicie. Jeśli populacja jest nieograniczona, to każde badanie, bez względu na to, jak wielka ilość jednostek była objęta badaniem, jest zawsze badaniem częściowym.

Rozpatrzmy dla ilustracji parę przykładów. Statystyka demograficzna zajmuje się, między innymi, badaniem struktury ludności ze względu na płeć, wiek i stan rodzinny, inte-

⁽¹⁾ Patrz [35].

⁽²⁾ Inaczej stuprocentowe.

resuje się ruchem naturalnym ludności, tzn. liczbą urodzeń i zgonów oraz liczbą małżeństw. Populacja, z którą ma do czynienia statystyka demograficzna, jest nieograniczona czasowo; populację tę stanowią wszyscy ludzie, którzy żyją lub będą żyli w przyszłości na danym terytorium. Podobnie rzeczą się przedstawia, jeśli idzie o badania statystyki rolnej, gdy trudni się ona np. ustaleniem wpływu opadów, temperatury, głębokości orki czy nawożenia na wysokość plonów. Tu również mamy do czynienia z populacją nieskończoną. Nawet bardzo duża ilość zbadanych roślin nie upoważnia nas do wydawania sądów kategorycznych, gdyż sądy te są zawsze sądami wydawanymi na podstawie próbki. Z analogiczną sytuacją mamy również do czynienia w fizyce statystycznej, gdy przedmiotem badania są np. takie zajwiska, jak rozpad atomów ciała promieniotwórczego.

2. Aktualność wyników badania statystycznego. W punkcie pierwszym była mowa o tym, że badaniem częściowym posługujemy się wtedy, gdy przedmiotem badania statystycznego jest populacja nieograniczona, która, rzeczą oczywistą, w całości zbadana być nie może.

W praktyce jednak zdarza się często, że nawet ograniczone populacje, jeśli są bardzo liczne, nie mogą być również zbadane w całości. Praktyka domaga się od statystyki rozwiązania szeregu zagadnień, przy czym zagadnienia te mają być rozwiązyane w stosunkowo krótkim czasie, gdyż w przeciwnym razie, ze względu na szybkie zmiany przedmiotu badania, wyniki badań mogłyby ulec dezaktualizacji.

Podamy przykłady z zakresu statystyki ekonomicznej. Dla właściwego kierowania całokształtem zjawisk społeczno-gospodarczych, dla spręzystego zarządzania aparatem gospodarczym władze państwowe muszą dysponować prawdziwym, obfitym i naukowo opracowanym materiałem sprawozdawczym. Materiału tego dostarcza statystyka, która gromadzi go w trakcie swych badań, prowadzonych bieżąco lub okresowo. Jedną z form tych badań są spisy, np. powszechny spis ludności czy też spis środków trwałych. Ze względu na rozległe rozmiary badanej populacji – opracowanie materiałów spisowych jest rzeczą żmudną, kosztowną i długotrwałą. Uzyskanie wszystkich danych informacyjnych z powszechnego spisu ludności trwa około 10 lat. Okres ten dzieli jeden spis od drugiego. Jest rzeczą zrozumiałą, że gdyby wszystkie informacje, które zawierają materiał spisowy, były uzyskane w drodze badania wyczerpującego, większość z nich stałaby się nieaktualna. Państwowe organa kierownicze, żądając od instytucji statystycznych szeregu informacji, wymagają nie tylko prawdziwości tych informacji, lecz domagają się również, aby informacje te były dostarczone w jak najkrótszym czasie. Chcąc tym trudnym wymogom uczynić zadość, statystyka ucieka się do badania częściowego.

3. Względy ekonomiczne. W wielu przypadkach na przeszkodzie w przeprowadzeniu badania całkowitego stoją względy natury ekonomicznej. Proces obserwacji statystycznej jest często bardzo kosztowny. Ma to miejsce w szczególności przy dokonywaniu precyzyjnych pomiarów. Wielu przykładów dostarcza tutaj praktyka przemysłowa. Przy produkcji masowej, w szczególności gdy obejmuje ona drobne, tanie przedmioty lub części składowe aparatów, maszyn i urządzeń montowanych z tych części (np. śruby, nity, nakrętki, kółka zębata, bolce, walce, kulki do łożysk kulkowych itp.), koszt pomiarów, których celem jest sprawdzenie, czy produkcja odpowiada wymogom technicznym, jest w wielu przypadkach wyższy od kosztu wyrobu mierzonego obiektu. Gdyby badanie nie było

przeprowadzone w ogóle – istniałaby poważna groźba, że niekontrolowana pod względem jakości produkcja zawierałaby znaczny procent braków, co rzecz prosta oznaczałoby stratę nie tylko dla przedsiębiorstwa, lecz także stratę społeczną. W ten sposób powstaje dylemat: z jednej strony, w razie badania wyczerpującego – ponoszenie kosztów badania wyższych od kosztów produkcji, z drugiej strony, w razie rezygnacji z jakichkolwiek badań statystycznych – nieuchronność poważnych strat wskutek znacznej ilości braków. Istnieje jedyne rozwiązanie tego dylematu: zrezygnować z badania wyczerpującego nie rezygnując jednak z wszelkiego badania statystycznego, tzn. zadowolić się badaniem częściowym. Spotykamy również inne sytuacje, gdzie względy ekonomiczne zmuszają nas do korzystania z badania częściowego. Do sytuacji takich zaliczyć należy te przypadki, gdy w trakcie kontroli jakości wyrobu sprawdzany przedmiot ulega zniszczeniu. Oczywiście w takich przypadkach o badaniu wyczerpującym nie może być mowy, gdyż doprowadziłoby to do zniszczenia całej populacji. Oto parę przykładów. Spółdzielnia produkcyjna otrzymuje zboże na siew. Ma być ono poddane próbie kiełkowania, którego celem jest ustalenie procentu nasion zdolnych do kiełkowania. Jeżeli bowiem siła kiełkowania nadesłanego zboża byłaby mała, wyrażająca się np. liczbą 50%, to po wysianiu takiego ziarna połowa jego ilości uległaby zniszczeniu. Aby uzmysłowić sobie wielkość strat, jakie groziłyby, gdyby nie kontrolowano siły kiełkowania, warto wyobrazić sobie 10 wagonów pszenicy, z których 5 zostaje zniszczonych w czasie wysiewu. Znaczne koszty, związane ze stratą tych 5 wagonów zboża, zostają wybitnie powiększone kosztami samego wysiewu.

Badanie siły kiełkowania odbywa się w ten sposób, że z partii zboża siewnego pobiera się próbę i ziarnom, które do niej trafiły, stwarza się warunki sprzyjające procesowi wegetacji. Po upływie czasu wystarczającego do wykiełkowania ziaren, liczy się ziarna, które nie wykiełkowały, i ustala się w ten sposób siłę kiełkowania ziaren w próbce. Otrzymane wyniki stanowią podstawę do oceny partii zboża. Badanie takie oczywiście niszczą ziarna, wobec tego badaniem tym może być objęta jedynie część (i to niewielka) partii zboża, przeznaczonego na siew.

Przykład następny. Mamy zbadać jakość 1000 żarówek. Badanie polega na poddaniu żarówek wstrząsom mechanicznym i wahaniom napięcia prądu aż do przepalenia żarówki. Ponieważ żarówki w trakcie badania ulegają zniszczeniu, przeto badanie wyczerpujące jest niemożliwe i należy zadowolić się badaniem częściowym.

Z analogiczną sytuacją mamy do czynienia przy sprawdzaniu jakości konserw. Aby zbadać jakość konserwy, należy otworzyć puszkę. Ponieważ wszystkich puszek otwierać nie sposób, więc kontrola jakości produkcji konserw musi korzystać z badania częściowego.

Podobnych przykładów można przytoczyć wiele. Wszystkie one uzasadniają konieczność stosowania badania częściowego.

Przy korzystaniu z badania częściowego wyniki otrzymane na podstawie populacji próbnej uogólnia się na populację generalną. Oczywiście, sądy wypowiadane o populacji na podstawie znajomości stosunków, panujących w próbce, mogą być prawdziwe lub fałszywe. Wynika stąd, że gdy korzystamy z badania częściowego, wydanie prawdziwego sądu jest zdarzeniem losowym, któremu odpowiada jakieś prawdopodobieństwo. To pra-

wdopodobieństwo jest tym większe, im stosunki panujące w populacji próbnej są bardziej zbliżone do stosunków panujących w populacji generalnej, czyli krótko mówiąc, im lepiej próbka reprezentuje populację. Próbka dobrze reprezentująca populację nazywa się *próbką reprezentatywną*. Aby próbka była reprezentatywna, muszą być spełnione dwa warunki:

1º Każdy element populacji powinien mieć jednakową szansę trafienia do próbki. Oznacza to, że elementy muszą być pobierane do próbki w sposób losowy.

2º Próbka powinna być dostatecznie liczna.

Przyjrzyjmy się temu na przykładzie. Przypuśćmy, że mamy wydać sąd o dostarczonej partii cegieł. Partia ta zawiera pewien procent cegieł złych; procent ten nazywać będziemy *frakcją braków*. Oczywiście dla dokładnego ustalenia frakcji braków w całej populacji musielibyśmy poddać badaniu wszystkie cegły, wchodzące w skład partii. Ze względów natury ekonomicznej jest to niemożliwe. Należy więc uciec się do badania częściowego. Wobec tego pobieramy próbkę i badamy frakcję braków w próbce. Jeżeli cegły pobierano do próbki w sposób losowy, to mamy prawo sądzić, że do próbki trafiły zarówno cegły dobre, jak i złe, przy czym stosunek cegieł dobrych do cegieł złych w próbce powinien być zbliżony odpowiednio do stosunku w populacji. Jeżeli bowiem wszystkie cegły mają jednakową szansę trafienia do próbki, natomiast w populacji więcej jest cegieł dobrych niż złych, to w próbce także powinno się znaleźć odpowiednio więcej cegieł dobrych niż złych. Po ustaleniu liczby złych sztuk w próbce i po obliczeniu frakcji braków możemy wydać sąd o frakcji braków w całej populacji. Sąd ten wypowiadam na ogół w ten sposób, że twierdzimy, iż frakcja braków w populacji zawiera się w pewnym przedziale, np. w przedziale od 3 do 5 %. Prawdziwość tego sądu jest zdarzeniem losowym, któremu odpowiada określone prawdopodobieństwo. Wielkość tego prawdopodobieństwa zależy od

1º liczności próbki, wzrastając wraz ze wzrostem próbki; równa się ona jedności, gdy próbka obejmuje całą populację;

2º dokładności wypowiadanego sądu, która w naszym przykładzie jest funkcją wielkości przedziału, w którym zawiera się frakcja braków w populacji; im ten przedział jest węższy, tym prawdopodobieństwo wydania trafnego sądu jest mniejsze.

Całokształt zagadnień, związanych z wypowiadaniem sądów o populacji generalnej na podstawie próbki, wchodzi w zakres działu statystyki matematycznej, który nazywa się *metodą reprezentacyjną*.

6.3. ZWIĄZEK MIEDZY POPULACJĄ GENERALNA I POPULACJĄ PRÓBNĄ

W paragrafie poprzednim była mowa o tym, że gdy nie jesteśmy w stanie zbadać populacji generalnej, możemy ograniczyć się do zbadania próbki i otrzymane wyniki badań uogólnić na całą populację. Oczywiście postępowanie takie jest możliwe tylko dlatego, że między populacją generalną i próbką zachodzi pewien związek. Związek ten wynika stąd, że próbka jest częścią populacji generalnej, przy czym część ta posiada ważną własność polegającą na tym, że rozkład wartości cechy w próbce reprezentacyjnej jest zbliżony do rozkładu wartości cechy w populacji generalnej.

Jak wiadomo, rozkład zmiennej losowej jest najpełniejszą charakterystyką tej zmiennej. Umówmy się, że rozkład cechy w populacji generalnej będziemy nazywać *rozkładem teoretycznym*, natomiast rozkład cechy w próbce nazywać będziemy *rozkładem empirycznym*. Dla oznaczenia dystrybuanty rozkładu teoretycznego używać będziemy symbolu $F(x)$, natomiast na oznaczenie dystrybuanty rozkładu empirycznego (tzn. częstości skumulowanej) będziemy posługiwać się symbolem $G(x)$.

Wysławienie istoty związku, łączącego populację generalną i populację próbную, wprowadza się do podania matematycznej interpretacji zależności, jaka zachodzi między rozkładem cechy w populacji generalnej a rozkładem cechy w próbce, lub inaczej – między rozkładem teoretycznym a rozkładem empirycznym. Zagadnieniem tym zajmowało się wielu matematyków i statystyków, na czele których wymienić należy Kołmogorowa, Smirnowa i Gliwenkę.

Oznaczmy symbolem $\sup_{-\infty < x < \infty} |G(x) - F(x)|$ kres górnny⁽¹⁾ bezwzględnej różnicy między dystrybuantą empiryczną a dystrybuantą teoretyczną. Gliwenko udowodnił ważne twierdzenie, które poniżej cytujemy.

TWIERDZENIE 1. *Niech $F(x)$ oznacza dystrybuantę teoretyczną, natomiast $G_n(x)$ niech oznacza dystrybuantę wartości cechy w próbce liczącej n elementów.*

Jeżeli wyniki losowania elementów populacji pobieranych do próbki są zdarzeniami niezależnymi, to

$$(1) \quad P\left\{\sup_{-\infty < x < \infty} |G_n(x) - F(x)| \rightarrow 0\right\} = 1.$$

Dowód tego twierdzenia znajdzie czytelnik w pracy [11].

Sens twierdzenia Gliwenki można potocznie wypowiedzieć w sposób następujący: gdy próbka jest dostatecznie liczna, to możemy uważać, że z prawdopodobieństwem bliskim jedności rozkład empiryczny mało różni się od rozkładu teoretycznego. Z twierdzenia Gliwenki wynika następująca ważna wskazówka praktyczna: próbka tym lepiej reprezentuje populację, im jest bardziej liczna. Fakt ten od dawna znany był z doświadczenia. Twierdzenie Gliwenki podaje wyjaśnienie tego faktu ze stanowiska teoretycznego.

Warto podkreślić, że obok klasycznego już dziś twierdzenia Gliwenki istnieją także inne twierdzenia, nie tylko wyjaśniające związek między dystrybuantą teoretyczną i empiryczną, ale pozwalające zmierzyć stopień podobieństwa między tymi dwoma charakterystykami rozkładu zmiennej losowej. Przytoczymy tu niektóre z tych twierdzeń.

Niech

$$D_n = \sup_{-\infty < x < \infty} |G_n(x) - F(x)|, \quad Q_n(\lambda) = \begin{cases} P(D_n \sqrt{n} < \lambda) & \text{dla } \lambda > 0, \\ 0 & \text{dla } \lambda \leq 0. \end{cases}$$

Zachodzi następujące

TWIERDZENIE 2. *Jeżeli $G_n(x)$ jest dystrybuantą empiryczną w n -elementowej próbce pobranej za pomocą losowania zwrotnego z populacji, w której zmienna losowa X ma ciągłą*

(1) Patrz 1.5.3.

dystrybuantę $F(x)$, to

$$Q(\lambda) = \lim_{n \rightarrow \infty} Q_n(\lambda) = \begin{cases} \sum_{k=-\infty}^{\infty} (-1)^k \exp(-2k^2 \lambda^2) & \text{dla } \lambda > 0, \\ 0 & \text{dla } \lambda \leq 0. \end{cases}$$

Twierdzenie to nosi nazwę *twierdzenia Kolmogorowa*. Trudny dowód tego twierdzenia pomijamy. Ideę dowodu znajdzie czytelnik w [8], a tablice granicznych prawdopodobieństw na końcu książki (tabl. VI).

Oznaczmy symbolami $G_{n_1}(x)$ i $G_{n_2}(x)$ dystrybuanty empiryczne dwóch próbek pobranych niezależnie jedna od drugiej w losowaniu ze zwracaniem z populacji generalnej o ciągłej dystrybuancie $F(x)$, przy czym liczebność pierwszej próbki niech wynosi n_1 , drugiej zaś n_2 .

Niech

$$D_{n_1 n_2} = \sup_{-\infty < x < \infty} |G_{n_1}(x) - G_{n_2}(x)|$$

oraz

$$Q_{n_1 n_2}(\lambda) = \begin{cases} P(D_{n_1 n_2} \sqrt{n} < \lambda) & \text{dla } \lambda > 0, \\ 0 & \text{dla } \lambda \leq 0. \end{cases}$$

Prawdziwe jest następujące

TWIERDZENIE 3.

$$\lim_{\substack{n_1 \rightarrow \infty \\ n_2 \rightarrow \infty}} Q_{n_1 n_2}(\lambda) = \begin{cases} \sum_{k=-\infty}^{\infty} (-1)^k \exp(-2k^2 \lambda^2) & \text{dla } \lambda > 0, \\ 0 & \text{dla } \lambda \leq 0. \end{cases}$$

Twierdzenie to jest znane pod nazwą *twierdzenia Smirnowa*. Zarys jego dowodu można znaleźć w [8], natomiast tablice wartości granicznych prawdopodobieństw w [28].

Zwracamy uwagę czytelnika, że twierdzenia Gliwenki, Kołmogorowa i Smirnowa są twierdzeniami granicznymi, co oznacza, że w praktyce możemy z nich korzystać tylko w przypadku próbek o dużych liczebnościach. Gdy próbki są małe, a to jest sytuacja, z której mamy przeważnie do czynienia w praktyce, konieczna jest znajomość nie prawdopodobieństw granicznych, lecz prawdopodobieństw dokładnych. Tablice takich prawdopodobieństw opracował Massey.

Najpełniejszą informację o populacji generalnej daje rozkład wartości cechy w populacji. W praktyce zdarza się jednak często, że nie jest nam potrzebna tak dokładna znajomość populacji. Przypuśćmy na przykład, że interesuje nas wartość przeciętna i odchylenie standardowe cechy X w populacji generalnej. Powstaje pytanie, czy dla określenia wartości tych parametrów niezbędna jest znajomość rozkładu cechy X w populacji? Twierdzenie Gliwenki daje odpowiedź na to pytanie: ponieważ dla dostatecznie dużych wartości n rozkład empiryczny różni się mało od rozkładu teoretycznego, przeto i interesujące nas parametry obu rozkładów muszą również mało się różnić między sobą⁽¹⁾. Wobec tego,

⁽¹⁾ Jest to słuszne przy założeniu, że populacja ma taki rozkład, w którym interesujące nas parametry istnieją. Statystyka zna bowiem rozkłady, w których nie istnieje żaden moment skończony. Takim rozkładem jest np. rozkład Cauchy'ego (patrz [8], str. 168 - 170).

aby określić wartość jakiegoś parametru w populacji generalnej (np. średniej arytmetycznej), należy ustalić, czemu równa się wartość tego parametru w próbce i uznać tę wartość jako przybliżenie nieznanej wartości parametru w populacji generalnej. Chcąc np. oszacować nieznaną wartość parametru $E(X)$ należy obliczyć średnią arytmetyczną \bar{x} . Z twierdzenia Gliwenki wynika, że jeśli próbka będzie dostatecznie duża, to z zadany z góry prawdopodobieństwem mamy prawo sądzić, że średnia z próbki będzie mało różnić się od średniej z populacji. Wynika stąd, że parametry obliczone na podstawie próbki mogą być używane do oceny parametrów populacji generalnej.

Doszliśmy do nowego, ważnego pojęcia statystycznego. Pojęciem tym są tzw. *estymatory* (od słowa angielskiego *to estimate* – oceniać, szacować). Estymatory są to wielkości wyznaczone na podstawie próbki, za pomocą których ocenia się wartości nieznanych parametrów populacji generalnej. Dział statystyki matematycznej, zajmujący się badaniem estymatorów, nazywa się *teorią estymacji*.

6.4. WYBRANE ZAGADNIENIA Z TEORII ESTYMACJI

6.4.1. Estymatory i ich klasyfikacja

Przypuśćmy, że stoi przed nami zadanie oceny nieznanego parametru Q w populacji generalnej. Aby zadanie to w jakiś sposób rozwiązać, pobieramy próbki i znajdujemy odpowiednią wartość tego parametru w próbce. Oczywiście wartość ta zależy od wartości elementów próbki, jest ich funkcją. Założymy, że liczebność próbki została ustalona i wynosi n . Ponumerujmy elementy próbki liczbami naturalnymi od 1 do n . Gdybyśmy losowanie powtarzali wielokrotnie, element próbki oznaczony liczbą 1 przybierałby różne wartości z różnym prawdopodobieństwem, byłby więc zmienną losową. Podobnie przedstawia się sprawa ze zbiorem wartości, jakie może przybierać drugi, trzeci i w końcu n -ty element próbki. Jeżeli próbka przybrała wartości x_1, x_2, \dots, x_n , to wartości te można traktować jako realizacje n -wymiarowej zmiennej losowej, która jest uogólnieniem poznanej przez nas w § 3.6 zmiennej dwuwymiarowej. Widzimy więc, że parametr z próbki, który nazywać będziemy *estymatorem* parametru Q z populacji, jest funkcją n -wymiarowej zmiennej losowej (patrz 4.7.6), a tym samym jest również zmienną losową, mającą własny rozkład. Na oznaczenie estymatora stosować będziemy symbol \hat{Q} , który czytamy „estymator parametru Q ”.

OKREŚLENIE 1. *Estymatorem parametru Q* nazywamy funkcję $\hat{Q}_n = U(X_1, X_2, \dots, X_n)$, która ma tę własność, że prawdopodobieństwo zdarzenia $\hat{Q}_n \approx Q$ jest tym bliższe jedności, im większa jest liczebność próbki.

Określenie to nasuwa przypuszczenie, że istnieje wiele sposobów konstruowania różnych estymatorów tego samego parametru Q . Łatwo się o tym przekonać. Chcemy np. oszacować parametr $m = E(X)$ w populacji generalnej. Możemy to uczynić obliczając średnią arytmetyczną w próbce. Nie jest to jednak jedyna możliwość. Zamiast \bar{x} możemy bowiem jako estymatora wartości przeciętnej populacji użyć obliczonych na podstawie

próbki średniej harmonicznej, średniej geometrycznej, średniej potęgowej dowolnego stopnia, mediany lub dominanty. Jedne z tych estymatorów spełnią lepiej swą rolę, a inne gorzej.

Z określenia 1 wynika, że estymatorem parametru Q w populacji może być każda taka funkcja \hat{Q}_n wartości wylosowanych do próbki, że dla arbitralnie obranej, lecz nie koniecznie dowolnie małej dodatniej liczby c zachodzi relacja

$$(1) \quad \lim_{n \rightarrow \infty} P\{|\hat{Q}_n - Q| < c\} = 1.$$

Jest rzeczą oczywistą, że estymator tym lepiej spełnia swą rolę, im dla mniejszych wartości c może on czynić zadość równości (1). Jak z tego widać, można mówić o estymatorach lepszych i gorszych. Zajmiemy się obecnie sformułowaniem kryteriów, pozwalających na dokonanie klasyfikacji estymatorów. Celem tej klasyfikacji jest wyodrębnienie z nieskończonego zbioru rozmaitych estymatorów małego podzbioru estymatorów najbardziej przydatnych do zastosowań.

W teorii estymacji wyodrębnia się następujące grupy estymatorów:

1. estymatory zgodne,
2. estymatory nieobciążone,
3. estymatory najefektywniejsze,
4. estymatory asymptotycznie najefektywniejsze.

OKREŚLENIE 2. *Estymatorem zgodnym* nazywa się taki estymator $\hat{Q}_n = U(X_1, X_2, \dots, X_n)$, dla którego

$$(2) \quad \lim_{n \rightarrow \infty} P\{|\hat{Q}_n - Q| < \varepsilon\} = 1,$$

gdzie ε jest dowolnie małą liczbą dodatnią.

Widzimy więc, że estymatorami zgodnymi nazywają się estymatory, które wraz ze wzrostem liczebności próbki są stochastycznie zbieżne (patrz 5.1.1, określenie 1) do wartości parametru estymowanego.

OKREŚLENIE 3. *Estymatorem nieobciążonym* nazywa się taki estymator, który czyni zadość następującej równości:

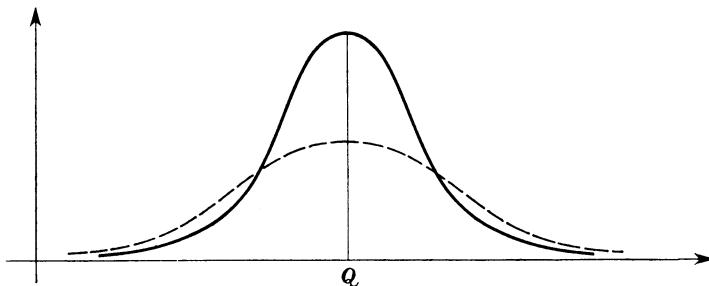
$$(3) \quad E(\hat{Q}_n) = Q.$$

Z powyższego określenia wynika, że estymatorem nieobciążonym nazywa się taki estymator, którego wartość przeciętna równa się parametrowi estymowanemu.

OKREŚLENIE 4. *Estymatorem najefektywniejszym* nazywa się taki estymator nieobciążony, który ma najmniejszą wariancję.

Jak wiemy, estymatory są zmiennymi losowymi, a więc mają swą wariancję. Okazuje się, że dokonując wyboru estymatora nie wystarcza zażądać, aby estymator był zgodny i nieobciążony. Można łatwo dać przykłady różnych estymatorów wartości przeciętnej w populacji generalnej, przy czym estymatory te będą zarówno zgodne, jak i nieobciążone. Ważnym kryterium, które umożliwia dokonanie wyboru najlepszego spośród tych estymatorów, jest wielkość wariancji. Jest rzeczą zrozumiałą, że estymator jest tym lepszy,

im ma mniejszą wariancję. Najłatwiej wyjaśnić to na rysunku. Wyobraźmy sobie, że mamy dwa estymatory \hat{Q}_1 i \hat{Q}_2 . Oba te estymatory są zgodne i nieobciążone. Założymy, że estymator \hat{Q}_1 ma mniejszą wariancję niż estymator \hat{Q}_2 . Na rysunku 1 przedstawione są rozkłady obu estymatorów. Linia ciągła wyobraża krzywą gęstości estymatora \hat{Q}_1 , natomiast linia przerywana przedstawia krzywą gęstości estymatora \hat{Q}_2 .



Rys. 1

Widzimy, że zmienne losowe \hat{Q}_1 i \hat{Q}_2 mają tę samą wartość przeciętną Q . Należy odpowiedzieć na pytanie, który z tych estymatorów jest lepszy. Spoglądając na rysunek 1 widzimy, że wartości zmiennej losowej \hat{Q}_1 są silniej skoncentrowane wokół Q niż wartości zmiennej losowej \hat{Q}_2 . Oznacza to, że średni błąd, jaki popełnimy przy wielokrotnym szacowaniu wartości parametru Q za pomocą estymatora \hat{Q}_1 , okaże się mniejszy od średniego błędu, jaki otrzymamy posługując się estymatorem \hat{Q}_2 . Możemy więc powiedzieć, że im estymator jest bardziej efektywny, tym dokładniejsza jest ocena nieznanego parametru populacji generalnej. Wybór najefektywniejszego estymatora umożliwia następującą nierówność:

$$(4) \quad V(\hat{Q}) \geq \frac{1}{nE\left\{\left[\frac{\partial \ln f(x, Q)}{\partial Q}\right]^2\right\}},$$

zwana *nierównością Rao-Craméra*⁽¹⁾. W powyższym wzorze symbolem $f(x, Q)$ oznaczono funkcję gęstości w populacji generalnej, natomiast literą n oznaczono liczebność próbki.

Dowód nierówności Rao-Craméra podany jest w pracy [8], str. 486 - 489.

Oznaczmy estymator najefektywniejszy parametru Q symbolem \hat{Q}_0 . Jeśli estymator ten istnieje oraz jeśli znamy jego wariancję $V(\hat{Q}_0)$, to wariancję tę można użyć jako wielkość porównawczą przy badaniu efektywności innych estymatorów parametru Q . Przypuszcmy, że \hat{Q} jest również estymatorem parametru Q . Chcemy zmierzyć efektywność tego parametru. Możemy to uczynić posługując się miarą *efektywności estymatora*:

$$(5) \quad e(\hat{Q}) = \frac{V(\hat{Q}_0)}{V(\hat{Q})}.$$

⁽¹⁾ Dla istnienia estymatora najefektywniejszego muszą być spełnione pewne warunki (patrz [8], str. 486). W praktyce rzadko są one spełnione.

Z określenia estymatora najefektywniejszego wynika, że $e(\hat{Q}) \leq 1$. Jest rzeczą zrozumiałą, że najbardziej przydatnymi dla celów praktycznych są estymatory najefektywniejsze. Niestety, estymatory te nie zawsze istnieją. Gdy nie ma estymatorów najefektywniejszych, należy zbadać, czy istnieją estymatory asymptotycznie najefektywniejsze.

OKREŚLENIE 5. *Estymatorem asymptotycznie najefektywniejszym* nazywamy taki estymator \hat{Q}_n , który spełnia relację

$$(6) \quad \lim_{n \rightarrow \infty} e(\hat{Q}_n) = 1.$$

Granica, do której dąży $e(\hat{Q}_n)$, gdy n rośnie do nieskończoności, nazywa się *efektywnością asymptotyczną* estymatora. Wobec tego estymatorem asymptotycznie najefektywniejszym nazywa się taki estymator, którego asymptotyczna efektywność równa się jedności.

Nowe pojęcia, wprowadzone w związku z klasyfikacją estymatorów, zostaną bliżej wyjaśnione przy omawianiu szacowania wartości przeciętnej i wariancji w populacji generalnej.

6.4.2. Estymacja wartości przeciętnej

Mamy oszacować nieznaną wartość m w populacji generalnej. Losując ze zwracaniem pobieramy próbki liczące n elementów i obliczamy \bar{x} . Za pomocą obliczonej wartości \bar{x} szacujemy m , tzn. przyjmujemy, że zachodzi przybliżona równość

$$\bar{x} \approx m.$$

Wartość przeciętna w próbce jest estymatorem zgodnym wartości przeciętnej w populacji generalnej. Wynika to bezpośrednio z twierdzenia Chinczyna (patrz 5.1.6). Wykażemy, że wartość przeciętna w próbce jest również estymatorem nieobciążonym.

Istotnie,

$$E(\bar{X}) = E\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n} \sum_{i=1}^n E(X_i) = E(X) = m,$$

czyli

$$(1) \quad E(\bar{X}) = m.$$

Aby zbadać, czy wartość przeciętna w próbce jest estymatorem najefektywniejszym wartości przeciętnej w populacji generalnej, musimy uczynić założenie o rozkładzie cechy w populacji, gdyż we wzorze (4), 6.4.1 występuje symbol gęstości rozkładu. Założmy więc, że cecha w populacji generalnej ma rozkład normalny. W takim razie

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

gdzie $\sigma^2 = V(X)$. Stąd

$$\ln f(x) = -\ln \sigma \sqrt{2\pi} - \frac{(x-m)^2}{2\sigma^2}, \quad \frac{\partial \ln f(x)}{\partial m} = \frac{x-m}{\sigma^2}.$$

Na mocy wzoru (4) powinno być

$$V(\bar{X}) \geq \frac{1}{nE\left(\frac{(X-m)^2}{\sigma^4}\right)} = \frac{1}{n} \frac{\sigma^2}{\sigma^4} = \frac{\sigma^2}{n},$$

ale

$$(2) \quad V(\bar{X}) = V\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n^2} \sum_{i=1}^n V(X_i) = \frac{\sigma^2}{n}.$$

Widzimy więc, że \bar{X} jest estymatorem najefektywniejszym parametru m . Gdy populacja generalna ma rozkład normalny, można dla oszacowania m użyć zamiast \bar{X} mediany Me . W rozkładzie normalnym (patrz np. [8]) mediana równa się wartości przeciętnej (oś symetrii w rozkładzie normalnym przechodzi bowiem przez m). Mediana z próbki jest estymatorem zgodnym i nieobciążonym wartości przeciętnej w populacji. Efektywność tego estymatora jest jednak gorsza od efektywności \bar{X} . Można wykazać, że wariancja mediany w próbce pobranej z populacji normalnej wynosi:

$$V(Me) = \frac{\pi\sigma^2}{2n},$$

wiemy zaś, że wariancja wartości przeciętnej w próbce wylosowanej z populacji normalnej

$$V(\bar{X}) = \frac{\sigma^2}{n};$$

stąd efektywność mediany

$$e(Me) = \frac{V(\bar{X})}{V(Me)} = \frac{\sigma^2}{n} : \frac{\pi\sigma^2}{2n} = \frac{2}{\pi} \approx 0,64.$$

Można łatwo udowodnić, że jeśli populacja ma rozkład dwumianowy lub rozkład Poissona, to \bar{X} jest zgodnym, nieobciążonym i najefektywniejszym estymatorem parametru m .

6.4.3. Estymacja wariancji

Dana jest populacja generalna, w której należy oszacować wariancję $V(X) = \sigma^2$. W tym celu losując ze zwracaniem pobieramy próbki, liczącą n elementów i obliczamy s^2 , gdzie

$$(1) \quad s^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2.$$

Obliczona na podstawie próbki wartość s^2 jest estymatorem nieznanej wartości σ^2 w populacji generalnej. Możemy napisać przybliżoną równość

$$s^2 \approx \sigma^2.$$

Oczywiście s^2 jest realizacją zmiennej losowej S^2 .

Łatwo wykazać, że S^2 jest estymatorem zgodnym σ^2 . Wprowadźmy w tym celu zmienną losową $Z_i = (X_i - \bar{X})^2$. Wtedy

$$(2) \quad S^2 = \frac{1}{n} \sum_{i=1}^n Z_i.$$

Widzimy więc, że S^2 jest wartością przeciętną zmiennych Z_i , mających ten sam rozkład i tę samą wartość przeciętną. Warunki wymagane w twierdzeniu Chinczyna są więc spełnione. Stosując to twierdzenie otrzymujemy, że S^2 jest estymatorem zgodnym σ^2 .

Przystąpimy obecnie do zbadania, czy S^2 jest estymatorem nieobciążonym, tzn. sprawdzimy, czy $E(S^2) = \sigma^2$.

Mamy

$$\begin{aligned} E(S^2) &= E\left[\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2\right] = E\left[\frac{1}{n} \sum_{i=1}^n (X_i - m + m - \bar{X})^2\right] = \\ &= E\left[\frac{1}{n} \sum_{i=1}^n [(X_i - m) - (\bar{X} - m)]^2\right] = \\ &= E\left[\frac{1}{n} \sum_{i=1}^n [(X_i - m)^2 - 2(X_i - m)(\bar{X} - m) + (\bar{X} - m)^2]\right] = \\ &= E\left[\frac{1}{n} \sum_{i=1}^n (X_i - m)^2 - 2(\bar{X} - m) \frac{1}{n} \sum_{i=1}^n (X_i - m) + \frac{1}{n} n(\bar{X} - m)^2\right] = \\ &= E\left[\frac{1}{n} \sum_{i=1}^n (X_i - m)^2 - 2(\bar{X} - m)^2 + (\bar{X} - m)^2\right] = \\ &= E\left[\frac{1}{n} \sum_{i=1}^n (X_i - m)^2 - (\bar{X} - m)^2\right] = \\ &= \frac{1}{n} \sum_{i=1}^n E(X_i - m)^2 - E(\bar{X} - m)^2 = \\ &= E(X - m)^2 - E(\bar{X} - m)^2 = V(X) - V(\bar{X}). \end{aligned}$$

Upřednio wykazaliśmy, że

$$V(\bar{X}) = \frac{\sigma^2}{n};$$

stąd

$$(3) \quad E(S^2) = \sigma^2 - \frac{1}{n} \sigma^2 = \sigma^2 \left(1 - \frac{1}{n}\right) = \frac{n-1}{n} \sigma^2.$$

Widzimy więc, że $E(S^2) \neq \sigma^2$. Oznacza to, że S^2 jest estymatorem obciążonym parametru σ^2 . Dzieląc obie strony wzoru (3) przez $\frac{n-1}{n}$ otrzymamy

$$\frac{n}{n-1} E(S^2) = \sigma^2.$$

Nieobciążonym estymatorem parametru σ^2 jest więc estymator dany wzorem:

$$(4) \quad S_1^2 = \frac{n}{n-1} S^2 = \frac{n}{n-1} \cdot \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2.$$

Zauważmy, że jeśli wartość przeciętna populacji generalnej $E(X)=m$ jest znana, to nieobciążony estymator parametru σ^2 wyraża się wzorem

$$(5) \quad S_2^2 = \frac{1}{n} \sum_{i=1}^n (X_i - m)^2.$$

Można to łatwo sprawdzić:

$$\begin{aligned} E(S_2^2) &= E\left[\frac{1}{n} \sum_{i=1}^n (X_i - m)^2\right] = E\left(\frac{1}{n} \sum_{i=1}^n X_i^2 - 2 \frac{m}{n} \sum_{i=1}^n X_i + m^2\right) = \\ &= \frac{1}{n} \sum_{i=1}^n E(X_i^2) - 2m \frac{1}{n} \sum_{i=1}^n E(X_i) + m^2 = \\ &= E(X^2) - 2mm + m^2 = E(X^2) - m^2 = \\ &= E(X^2) - E^2(X) = V(X) = \sigma^2 \end{aligned}$$

(patrz twierdzenie 1, 4.3.1).

Zbadajmy obecnie, czemu równa się najefektywniejszy estymator parametru σ^2 przy założeniu, że populacja ma rozkład normalny.

Przypominamy wzór (4), 6.4.1,

$$V(\hat{Q}) \geq \frac{1}{nE\left\{\left[\frac{\partial \ln f(x, Q)}{\partial Q}\right]^2\right\}}.$$

W naszym wypadku

$$f(x, Q) = \frac{1}{\sqrt{2\pi Q}} \exp\left[-\frac{(x-m)^2}{2Q}\right],$$

przy czym $Q=\sigma^2$. Mamy dalej

$$\ln f(x, Q) = -\ln \sqrt{2\pi} - \ln \sqrt{Q} - \frac{(x-m)^2}{2Q} Q^{-1}$$

oraz

$$\frac{\partial \ln f(x, Q)}{\partial Q} = -\frac{1}{2Q} + \frac{(x-m)^2}{2Q^2}.$$

W takim razie

$$\begin{aligned} E\left[\frac{(X-m)^2}{2Q^2} - \frac{1}{2Q}\right]^2 &= \frac{1}{4Q^4} E[(X-m)^2 - Q]^2 = \\ &= \frac{1}{4Q^4} [E(X-m)^4 - 2QE(X-m)^2 + Q^2]. \end{aligned}$$

W § 4.5 obliczyliśmy, że moment centralny czwartego rzędu w rozkładzie normalnym równa się

$$E[X - E(X)]^4 = 3\sigma^4;$$

stąd

$$\begin{aligned} E\left[\frac{(X-m)^2}{2Q^2} - \frac{1}{2Q}\right]^2 &= \frac{1}{4Q^4} [E(X-m)^4 - Q^2] = \\ &= \frac{1}{4Q^4} [3Q^2 - Q^2] = \frac{1}{2Q^2}. \end{aligned}$$

Ostatecznie więc

$$(6) \quad V(S^2) \geq \frac{1}{nE\left\{\left[\frac{\partial \ln f(x, Q)}{\partial Q}\right]^2\right\}} = \frac{2Q^2}{n} = \frac{2\sigma^4}{n}.$$

Otrzymaliśmy godny zapamiętania wynik: minimalna wariancja parametru S^2 , obliczonego na podstawie próbki wylosowanej z populacji generalnej o rozkładzie normalnym, wyraża się prostym wzorem

$$\min V(S^2) = \frac{2\sigma^4}{n}.$$

6.4.4. Estymacja wariancji wartości przeciętnej w próbce

Obliczona na podstawie próbki wartość przeciętna \bar{X} jest zmienną losową. Ma ona pewną wariancję. Zajmiemy się zbadaniem, w jaki sposób można dokonać estymacji tej wariancji.

Mamy

$$V(\bar{X}) = V\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n^2} V\left(\sum_{i=1}^n X_i\right).$$

Dalsze przekształcenia musimy prowadzić osobno dla przypadku, gdy zmienne losowe stojące pod znakiem sumy są niezależne, i osobno dla przypadku, gdy zmienne te są zależne. Dla większej przejrzystości oznaczmy te przypadki literami A i B.

Przypadek A – losowanie ze zwracaniem. Z przypadkiem tym mamy do czynienia wtedy, gdy pobieramy próbki z populacji nieograniczonej lub gdy pobieramy próbki z populacji ograniczonej i korzystamy z losowania ze zwracaniem. Jak wiadomo (patrz 4.3.2, twierdzenie 4), wariancja sumy niezależnych zmiennych losowych równa się sumie wariancji tych zmiennych. Wobec tego

$$V(\bar{X}) = \frac{1}{n^2} V\left(\sum_{i=1}^n X_i\right) = \frac{1}{n^2} \sum_{i=1}^n V(X_i).$$

Ponieważ wszystkie zmienne X_i mają ten sam rozkład, więc

$$\frac{1}{n^2} \sum_{i=1}^n V(X_i) = \frac{1}{n^2} nV(X_i) = \frac{1}{n} V(X_i).$$

Stąd ostatecznie

$$(1) \quad V(\bar{X}) = \frac{1}{n} V(X_i) = \frac{1}{n} \sigma^2.$$

Estymacja występującej w tym wzorze nieznanej wartości $V(X)$ została już omówiona uprzednio (patrz 6.4.3).

Przypadek B – losowanie bez zwracania. Z przypadkiem tym mamy do czynienia wówczas, gdy próbki pobieramy z populacji ograniczonej i posługujemy się losowaniem bez zwracania. Przy takiej metodzie losowania skład populacji zmienia się w trakcie losowania. Wyniki poszczególnych losowań są zdarzeniami zależnymi.

Zgodnie z definicją wariancji mamy

$$(2) \quad V(\bar{X}) = E[\bar{X} - E(\bar{X})]^2 = E(\bar{X} - m)^2.$$

► Przypuśćmy, że populacja generalna zawiera N elementów, u których cecha przybiera wartości

$$x_1, x_2, \dots, x_r, \dots, x_N$$

z odpowiednimi prawdopodobieństwami.

Niech Z oznacza zmienną losową o rozkładzie zero-jedynkowym (patrz 3.3.1). Umówmy się, że zmienna Z_r przybiera wartość 1, gdy element populacji o wartości cechy x_r trafił do próbki, oraz że zmienna Z_r przybiera wartość 0, gdy element ten do próbki nie trafił. Zgodnie z tą umową

$$x_r \cdot Z_r = x_r,$$

gdy element populacji o wartości cechy x_r trafił do próbki, oraz

$$x_r \cdot Z_r = 0,$$

gdy element populacji o wartości cechy x_r nie trafił do próbki. W takim razie

$$(3) \quad \bar{X} - m = \frac{1}{n} \sum_{i=1}^n X_i - m = \frac{1}{n} \sum_{r=1}^N x_r Z_r - m = \frac{1}{n} \sum_{r=1}^N x_r Z_r - m = \frac{1}{n} \sum_{r=1}^N Z_r,$$

gdziż

$$(4) \quad \sum_{r=1}^N Z_r = n.$$

Wobec tego

$$(5) \quad \bar{X} - m = \frac{1}{n} \sum_{r=1}^N (x_r - m) Z_r.$$

Wstawiając prawą stronę powyższej równości do wzoru (2) otrzymamy

$$\begin{aligned} E(\bar{X} - m)^2 &= E \left[\frac{1}{n} \sum_{r=1}^N (x_r - m) Z_r \right]^2 = \frac{1}{n^2} E \left[\sum_{r=1}^N (x_r - m) Z_r \right]^2 = \\ &= \frac{1}{n^2} E \left[\sum_{r=1}^N (x_r - m)^2 Z_r^2 + \sum_{r \neq s} (x_r - m)(x_s - m) Z_r Z_s \right], \end{aligned}$$

gdzie $r, s = 1, 2, \dots, N$.

Ponieważ x_r oraz x_s dla każdego r i s są wielkościami stałymi, przeto mamy dalej

$$(6) \quad E(\bar{X} - m)^2 = \frac{1}{n^2} \left[\sum_{r=1}^N (x_r - m)^2 E(Z_r)^2 + \sum_{r \neq s} (x_r - m)(x_s - m) E(Z_r \cdot Z_s) \right].$$

Ale

$$(7) \quad E(Z_r)^2 = 1^2 \cdot p + 0^2 \cdot q = p = \frac{n}{N},$$

gdzię p jest prawdopodobieństwem trafienia elementu populacji generalnej do próbki.

Dla obliczenia $E(Z_s \cdot Z_r)$ można posłużyć się następującą tabelką:

$Z_s \backslash Z_r$	Z_r	1	0
1	1 · 1	1 · 0	
0	0 · 1	0 · 0	

Z tabelki tej widzimy, że

$$(8) \quad E(Z_s \cdot Z_r) = 1 \cdot 1 \cdot P(Z_s = 1, Z_r = 1) =$$

$$= P(Z_s = 1) \cdot P[(Z_r = 1) | (Z_s = 1)] = \frac{n}{N} \cdot \frac{n-1}{N-1}.$$

Wstawiając (7) i (8) do wzoru (6) otrzymamy

$$\begin{aligned} E(\bar{X} - m)^2 &= \frac{1}{n^2} \left[\sum_{r=1}^N (x_r - m)^2 \frac{n}{N} + \sum_{r \neq s} (x_r - m)(x_s - m) \cdot \frac{n}{N} \cdot \frac{n-1}{N-1} \right] = \\ &= \frac{1}{n} \sum_{r=1}^N \frac{(x_r - m)^2}{N} + \frac{n-1}{N(N-1)} n \sum_{r \neq s} (x_r - m)(x_s - m). \end{aligned}$$

Ale

$$\sum_{r=1}^N \frac{(x_r - m)^2}{N} = \sigma^2,$$

natomiast

$$\begin{aligned} \sum_{r \neq s} (x_r - m)(x_s - m) &= \sum_{r=1}^N \sum_{s=1}^N (x_r - m)(x_s - m) - \sum_{r=1}^N (x_r - m)^2 = \\ &= \sum_{r=1}^N (x_r - m) \sum_{s=1}^N (x_s - m) - \sum_{r=1}^N (x_r - m)^2 = -N\sigma^2. \end{aligned}$$

Stąd

$$\begin{aligned} E(\bar{X} - m)^2 &= \frac{1}{n} \sigma^2 - \frac{1}{n} \cdot \frac{n-1}{N-1} \sigma^2 = \\ &= \frac{1}{n} \sigma^2 \left(1 - \frac{n-1}{N-1}\right) = \frac{1}{n} \sigma^2 \frac{N-n}{N-1} = \frac{1}{n} \sigma^2 \frac{1 - \frac{n}{N}}{1 - \frac{1}{N}}. \end{aligned}$$

Ostatecznie więc

$$(9) \quad V(\bar{X}) = \frac{\sigma^2}{n} \cdot \frac{1 - \frac{n}{N}}{1 - \frac{1}{N}}.$$

Gdy ułamek n/N jest liczbą małą (tzn. gdy próbka jest mała w stosunku do populacji), to zachodzi równość przybliżona

$$V(\bar{X}) \approx \frac{\sigma^2}{n}.$$

Otrzymaliśmy wzór analogiczny do wzoru na wariancję wartości przeciętnej w próbce, wyprowadzonego przy omawianiu przypadku A. Jeśli przy ustalonej liczbeności próbki n liczbeność populacji generalnej N nieograniczenie wzrasta, to

$$\lim_{N \rightarrow \infty} V(\bar{X}) = \lim_{N \rightarrow \infty} \frac{\sigma^2}{n} \cdot \frac{1 - \frac{n}{N}}{1 - \frac{1}{N}} = \frac{\sigma^2}{n}.$$

Widzimy więc, że jeżeli populacja generalna jest populacją nieograniczoną, to bez względu na to, czy korzystamy z losowania ze zwracaniem, czy też bez zwracania, zawsze możemy przyjąć, że zachodzi wzór

$$V(\bar{X}) = \frac{1}{n} \sigma^2.$$

Ze wzoru (9) wynika, że przy szacowaniu wartości przeciętnej w populacji ograniczonej należy stosować schemat losowania bez zwracania, wtedy bowiem nie tylko proces losowania jest łatwiejszy, ale i efektywność szacunku jest lepsza. Przekonuje nas o tym nierówność

$$(10) \quad \frac{\sigma^2}{n} \cdot \frac{1 - \frac{n}{N}}{1 - \frac{1}{N}} \leq \frac{\sigma^2}{n},$$

wynikająca z tego, że dla $n > 0$

$$\frac{1 - \frac{n}{N}}{1 - \frac{1}{N}} \leq 1.$$

Przy omawianiu klasyfikacji estymatorów była mowa o tym, że estymator jest tym lepszy, im ma mniejszą wariancję. Jak wynika z nierówności (10), wariancja wartości przeciętnej \bar{X} , obliczonej na podstawie próbki pobranej w losowaniu bezzwotnym, jest mniejsza od wariancji przeciętnej \bar{X} , obliczonej na podstawie próbki pobranej w losowaniu ze zwracaniem. Płynie stąd ważna wskazówka praktyczna: przy szacowaniu wartości przeciętnej w populacji generalnej za pomocą wartości przeciętnej w próbce należy stosować schemat losowania bez zwracania.

Wzór (9) podający wariancję wartości przeciętnej w próbce, pobranej z populacji ograniczonej metodą losowania bez zwracania, ulega pewnej modyfikacji, jeżeli cecha elementów populacji generalnej może przybierać jedynie dwie wartości $X = x_1$ oraz $X = x_2$, odpowiednio z prawdopodobieństwem p i $1 - p$. W tym przypadku spełnione są bowiem warunki rozkładu hipergeometrycznego (patrz 3.3.3).

Przypuśćmy, że populacja generalna liczy N elementów. Z populacji tej w drodze losowania bez zwracania pobieramy n elementów. W takim razie prawdopodobieństwo tego, że w próbce znajdzie się k elementów, których wartość cechy równa się x_1 oraz $n - k$ elementów, których wartość cechy równa się x_2 , wyraża się wzorem (patrz 3.3.3, wzór (1)):

$$P(X = k) = \frac{C_R^k C_{N-R}^{n-k}}{C_N^n},$$

gdzie R oznacza liczbę elementów populacji generalnej, mających wartość cechy $X = x_1$.

W punkcie 4.3.1 (wzór (11)) wykazaliśmy, że wariancja zmiennej losowej o rozkładzie hipergeometrycznym ma postać następującą:

$$(11) \quad V(X) = npq \frac{1 - \frac{n}{N}}{1 - \frac{1}{N}}.$$

Nietrudno dostrzec analogię między wzorem (9) i (11). We wzorze (11) zmienna losowa X jest sumą n zmiennych losowych Z o rozkładzie zero-jedynkowym. Położmy więc w tym wzorze $X = \sum Z$. Otrzymujemy

$$V(X) = V(\sum Z),$$

czyli

$$V(\sum Z) = npq \frac{1 - \frac{n}{N}}{1 - \frac{1}{N}}.$$

Dzieląc obie strony przez n^2 mamy

$$(12) \quad \frac{1}{n^2} V(\sum Z) = V\left(\frac{\sum Z}{n}\right) = V(\bar{Z}) = \frac{pq}{n} \cdot \frac{1 - \frac{n}{N}}{1 - \frac{1}{N}}.$$

Z przekształcenia tego widzimy, że wzór (11) jest szczególnym przypadkiem wzoru (9). Istotnie, podstawiając we wzorze (9) $V(X)=pq$ mamy

$$V(\bar{X}) = \frac{pq}{n} \cdot \frac{1 - \frac{n}{N}}{1 - \frac{1}{N}},$$

czyli to samo, co wyraża wzór (12).

6.4.5. Estymacja wariancji w populacji generalnej o znanym rozkładzie za pomocą odchylenia przeciętnego z próbki

Przypuśćmy, że mamy oszacować wariację w populacji generalnej, przy czym rozkład populacji jest znany. Jak wiemy z punktu 6.4.3, możemy w tym celu pobrać próbę i obliczyć wariację w próbce według wzoru (patrz 6.4.3, wzór (4)):

$$s_1^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2.$$

Obliczona w ten sposób wariacja w próbce jest estymatorem nieobciążonym wariacji w populacji generalnej. Jeżeli próbka jest duża, korzystanie z tego wzoru nastarcza jednak pewne trudności rachunkowe, polegające na konieczności obliczenia różnic typu $x_i - \bar{x}$, podnoszenia tych różnic do kwadratu i sumowania. Trudności te zmniejszają się znacznie, gdy wariację w populacji generalnej oszacujemy za pomocą odchylenia przeciętnego z próbki, które wyraża się wzorem

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

Wartością \bar{x} można podzielić zbiór wartości x_1, x_2, \dots, x_n w próbce na dwa podzbiory w ten sposób, że do pierwszego z tych podzbiorów zalicza się wartości mniejsze lub równe \bar{x} , do drugiego zaś pozostałe wartości. W takim razie oznaczając symbolem x_{i_l} ($l=1, 2, \dots, k$) wartości należące do pierwszego z tych podzbiorów możemy napisać nierówność

$$x_{i_l} \leq \bar{x}$$

i analogicznie, oznaczając symbolem x_{i_s} ($s=1, 2, \dots, r$) wartości należące do drugiego podzbioru możemy napisać nierówność

$$x_{i_s} > \bar{x}.$$

Oczywiście $k+r=n$.

Oznaczmy symbolem \bar{x}_I średnią arytmetyczną wartości pierwszego podzbioru, a symbolem \bar{x}_{II} średnią arytmetyczną wartości drugiego podzbioru, tzn.

$$\bar{x}_I = \frac{1}{k} \sum_{l=1}^k x_{i_l}, \quad \bar{x}_{II} = \frac{1}{r} \sum_{s=1}^r x_{i_s}.$$

TWIERDZENIE 1. *Mamy*

$$(1) \quad d = \frac{2k}{n} (\bar{x} - \bar{x}_I) = \frac{2r}{n} (\bar{x}_{II} - \bar{x}).$$

Dowód. Oznaczając $z_i = x_i - \bar{x}$ możemy napisać następującą równość:

$$(2) \quad \sum_{l=1}^k (-z_{i_l}) = \sum_{s=1}^r z_{i_s},$$

gdzie każde $z_{i_l} \leq 0$, zaś każde $z_{i_s} > 0$.

Odchylenie przeciętne wynosi

$$d = \frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}| = \frac{1}{n} \sum_{i=1}^n |z_i| = \frac{1}{n} \left(\sum_{l=1}^k |z_{i_l}| + \sum_{s=1}^r |z_{i_s}| \right),$$

a wobec równości (2):

$$(3) \quad d = \frac{2}{n} \sum_{l=1}^k (-z_{i_l}) = \frac{2}{n} \sum_{s=1}^r z_{i_s}.$$

Zbadajmy obecnie, czemu równa się $\bar{x}_{II} - \bar{x}$. Mamy

$$\begin{aligned} \bar{x}_{II} - \bar{x} &= \frac{1}{r} \sum_{s=1}^r (\bar{x} + z_{i_s}) - \bar{x} = \frac{r\bar{x} + \sum_{s=1}^r z_{i_s}}{r} - \bar{x} = \\ &= \frac{r\bar{x} + \sum_{s=1}^r z_{i_s} - r\bar{x}}{r} = \frac{1}{r} \sum_{s=1}^r z_{i_s}. \end{aligned}$$

Tezę twierdzenia otrzymamy, gdy podstawimy ten wynik do wzoru (3). Mianowicie powieźaż

$$\sum_{s=1}^r z_{i_s} = r(\bar{x}_{II} - \bar{x})$$

i podobnie

$$\sum_{l=1}^k (-z_{i_l}) = k(\bar{x} - \bar{x}_I),$$

zatem

$$(4) \quad d = \frac{2k}{n} (\bar{x} - \bar{x}_I) = \frac{2r}{n} (\bar{x}_{II} - \bar{x}).$$

Wyjaśnimy na przykładzie sposób obliczania odchylenia przeciętnego za pomocą wzoru (4).

PRZYKŁAD 1. W trakcie statystycznej kontroli jakości śrub, produkowanych na automacie RK-26, pobiera się co godzinę próbki śrub liczącą 10 sztuk i mierzy się średnicę każdej śruby w próbce. Poniżej podane są wyniki pomiarów w jednej z takich próbek. W oparciu o ten materiał liczbowy obliczono tradycyjnym sposobem odchylenie przeciętne.

Nr	x	$ x - \bar{x} $
1	1,15	0,143
2	1,16	0,133
3	1,20	0,093
4	1,22	0,073
5	1,27	0,023
6	1,27	0,023
7	1,36	0,067
8	1,41	0,117
9	1,44	0,147
10	1,45	0,157
$\bar{x} = 1,293; \quad d = 0,0976 \approx 0,098$		0,976

Obliczenie odchylenia przeciętnego, nawet sposobem tradycyjnym, tzn. za pomocą wzoru $\frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}|$, zabiera znacznie mniej czasu niż obliczenie odchylenia standaryzowanego. Uproszczenie rachunków staje się jednak naprawdę bardzo poważne, gdy przy obliczeniu odchylenia przeciętnego korzystamy ze wzoru (4). Ilustruje to następująca tabela:

Nr	x	$x > \bar{x}$
1	1,15	
2	1,16	
3	1,20	
4	1,22	
5	1,27	
6	1,27	
7	1,36	1,36
8	1,41	1,41
9	1,44	1,44
10	1,45	1,45
$\bar{x} = 1,293, \quad \bar{x}_{11} = 5,66 : 4 = 1,415$		5,66
$d = \frac{2 \cdot 4}{10} (1,415 - 1,293) = 0,8 \cdot 0,122 = 0,0976 \approx 0,098.$		

Ze względu na duże udoskonalenia rachunkowe, posługujemy się często odchyleniem przeciętnym z próbki dla oceny nieznanej wartości wariancji i odchylenia standaryzowanego w populacji generalnej. Przypuśćmy, że populacja generalna ma rozkład normalny. W celu

oszacowania wariancji pobieramy próbę liczącą n elementów i posługując się wzorem (4) obliczamy odchylenie przeciętne d . Oczywiście odchylenie przeciętne z próbki jest zmienną losową. Z twierdzenia Chinczyna (patrz 5.1.6) wynika, że zmienna ta jest stochastycznie zbieżna do D , gdzie D oznacza odchylenie przeciętne w populacji generalnej:

$$D = \frac{1}{\sigma \sqrt{2\pi}} \int_{-\infty}^{\infty} |x - m| \exp\left[-\frac{(x-m)^2}{2\sigma^2}\right] dx = \frac{2}{\sigma \sqrt{2\pi}} \int_m^{\infty} (x - m) \exp\left[-\frac{(x-m)^2}{2\sigma^2}\right] dx = \sigma \sqrt{\frac{2}{\pi}}.$$

Stąd

$$V(X) = \sigma^2 = \frac{\pi}{2} D^2.$$

Ogólnie możemy napisać, że

$$V(X) = CD^2,$$

gdzie C jest pewnym współczynnikiem zależnym od rozkładu populacji generalnej.

6.5. PRZEDZIAŁY UFNOŚCI

6.5.1. Sformułowanie problemu

Rozpatrzmy następujące zagadnienie: zmienna losowa X ma rozkład normalny

$$N\left(m, \frac{1}{\sqrt{n}} \sigma\right).$$

Należy znaleźć

$$P\{a_1 < \bar{X} < a_2\},$$

gdzie a_1 i a_2 są dowolnymi liczbami rzeczywistymi, przy czym $a_1 < a_2$.

Oczywiście

$$P\{a_1 < \bar{X} < a_2\} = F(a_2) - F(a_1) = \frac{\sqrt{n}}{\sigma \sqrt{2\pi}} \int_{a_1}^{a_2} \exp\left(-\frac{\sqrt{n}(\bar{x}-m)^2}{2\sigma^2}\right) d\bar{x}$$

(patrz § 3.4, wzór (4), oraz 3.5.3).

Aby ułatwić czytelnikowi opanowanie treści dalszych rozważań, obok symboli ogólnych wprowadzimy także liczby konkretne. Niech

$$m = 30, \quad \frac{\sigma}{\sqrt{n}} = 2, \quad a_1 = 28, \quad a_2 = 34.$$

W takim razie

$$P\{28 < \bar{X} < 34\} = \frac{1}{2\sqrt{2\pi}} \int_{28}^{34} \exp\left[-\frac{(\bar{x}-30)^2}{2 \cdot 4}\right] d\bar{x}.$$

Aby obliczyć całkę stojącą po prawej stronie znaku równości, wprowadzimy zmienną standaryzowaną $T = \frac{\bar{X} - 30}{2}$, która pozwoli nam posłużyć się tablicami rozkładu normalnego (tablica I), podanymi na końcu książki.

Wobec tego

$$P\{\bar{X} < 34\} = \frac{1}{\sqrt{2\pi}} \int_{t_1}^{t_2} e^{-v^2/2} dv,$$

gdzie

$$t_1 = \frac{a_1 - m}{\sigma/\sqrt{n}} = \frac{28 - 30}{2} = -1, \quad t_2 = \frac{a_2 - m}{\sigma/\sqrt{n}} = \frac{34 - 30}{2} = 2.$$

W takim razie

$$P\{\bar{X} < 34\} = \Phi(2) - \Phi(-1).$$

W tablicach znajdujemy, że $\Phi(2) = 0,4773$. W celu obliczenia $\Phi(-1)$ zauważmy, że na mocy symetrii funkcji gęstości rozkładu normalnego wokół osi $x=m$ zachodzi równość

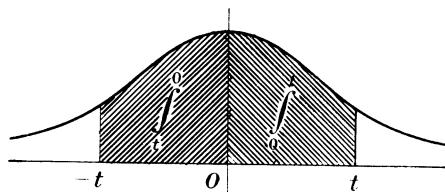
$$\Phi(t) = \frac{1}{\sqrt{2\pi}} \int_0^t e^{-v^2/2} dv = \frac{1}{\sqrt{2\pi}} \int_{-t}^0 e^{-v^2/2} dv$$

(patrz rys. 1). Ale

$$\frac{1}{\sqrt{2\pi}} \int_{-t}^0 e^{-v^2/2} dv = - \frac{1}{\sqrt{2\pi}} \int_0^{-t} e^{v^2/2} dv = -\Phi(-t),$$

skąd

$$\Phi(-t) = -\Phi(t).$$



Rys. 1

Ponieważ $\Phi(1) = 0,3413$, więc

$$P\{\bar{X} < 34\} = 0,4773 + 0,3413 = 0,8186$$

lub ogólnie

$$P\{a_1 < \bar{X} < a_2\} = \Phi(t_2) - \Phi(t_1) = \alpha.$$

Z tych rozważań wynika, że znając a_1 i a_2 potrafimy określić jednoznacznie liczbę α . Zbadajmy, czy dzieje się również odwrotnie, tzn. czy znając α potrafimy wyznaczyć liczby

a_1 i a_2 . Łatwo przekonać się, że tak nie jest. W równaniu

$$\Phi(t_2) - \Phi(t_1) = \alpha$$

znamy tylko α , nie znamy natomiast $\Phi(t_2)$ i $\Phi(t_1)$. Mamy więc jedno równanie o dwóch niewiadomych. Równanie takie ma nieskończoną liczbę rozwiązań. Aby otrzymać tylko jedno rozwiązanie, musimy uczynić jakieś dodatkowe założenie, np. że

$$(1) \quad t_2 = -t_1$$

lub, co na jedno wychodzi, że

$$(2) \quad \Phi(t_2) = \Phi(-t_1).$$

Wtedy otrzymamy

$$(3) \quad 2\Phi(t_2) = \alpha, \quad \text{czyli} \quad \Phi(t_2) = \frac{1}{2}\alpha.$$

Stąd już łatwo wyznaczymy t_2 posługując się tablicami rozkładu normalnego. W tablicach tych znajdujemy prawdopodobieństwo $\alpha/2$. W boku tablicy podane są różne wartości parametru t . Szukaną wartością t_2 jest wartość znajdująca się w wierszu odpowiadającym prawdopodobieństwu $\alpha/2$.

Ponieważ

$$(4) \quad \alpha = \Phi(t_2) - \Phi(t_1) = P \left\{ t_1 < \frac{\bar{X} - m}{\sigma_{\bar{x}}} < t_2 \right\},$$

przeto pomijając ze względu na założenie (1) indeksy stojące przy t , mamy

$$(5) \quad \alpha = 2\Phi(t) = P \left\{ -t < \frac{\bar{X} - m}{\sigma_{\bar{x}}} < t \right\}.$$

Równość ta ma pierwszorzędne znaczenie praktyczne, gdyż dzięki niej znając α i średnią \bar{X} w próbce, możemy napisać przedział, który z prawdopodobieństwem α pokryje nieznaną wartość parametru m w populacji generalnej. Istotnie, mnożąc przez -1 wszystkie wyrazy znajdujące się w nawiasach nierówności, otrzymamy

$$\alpha = P \left\{ t > \frac{m - \bar{X}}{\sigma_{\bar{x}}} > -t \right\}$$

lub inaczej

$$\alpha = P \left\{ -t < \frac{m - \bar{X}}{\sigma_{\bar{x}}} < t \right\},$$

to zaś można przekształcić dalej, tak że otrzymamy w końcu wzór

$$(6) \quad \alpha = P \{ \bar{X} - t\sigma_{\bar{x}} < m < \bar{X} + t\sigma_{\bar{x}} \}.$$

Treść zawartą w tym wzorze można wyrazić następująco: jeżeli w populacji rozkład

cechy jest znany⁽¹⁾, to możemy twierdzić, że z zadanym z góry prawdopodobieństwem α przedział liczbowy $(\bar{X} - t\sigma_{\bar{x}}, \bar{X} + t\sigma_{\bar{x}})$ pokryje nieznaną wartość parametru m .

Przedział $(\bar{X} - t\sigma_{\bar{x}}, \bar{X} + t\sigma_{\bar{x}})$ nazywa się *przedziałem ufności*, natomiast prawdopodobieństwo α nazywa się *poziomem ufności* (lub *współczynnikiem ufności*).

Zanim przejdziemy do rozpatrzenia przykładu liczbowego, dla ułatwienia czytelnikowi przyswojenia nowo wprowadzonych pojęć sięgniemy najpierw do przykładu poglądowego. Wyobraźmy sobie, że na sznurze do suszenia bielizny wieszamy uprana chustkę. Przy wieszaniu chustki dbamy o to, aby sznur przechodził przez środek chustki, równolegle do dwóch boków, których długość równa się 30 cm. Gdzieś na sznurze znajduje się węzeł. Jeżeli znamy długość sznura, która wynosi 300 cm, to z prawdopodobieństwem $\frac{30}{300} = 0,1$ możemy twierdzić, że wieszając chustkę w sposób losowy nakryjemy nią węzeł znajdujący się na sznurze (zakładamy, że wymiary węzła są tak małe, że można je pominać). Prawdopodobieństwo 0,1 jest w naszym przykładzie odpowiednikiem poziomu ufności, natomiast długość boku chustki, która wynosi 30 cm, wyobraża wielkość przedziału ufności. Nieznane położenie węzła na sznurze odpowiada nieznanej wartości parametru m w populacji generalnej. Przykład ten dobrze ilustruje związek, zachodzący między wielkością poziomu ufności i rozpiętością przedziału ufności. Jeżeli długość sznura jest stała, to żądając, aby prawdopodobieństwo pokrycia chustką węzła wynosiło nie 0,1, lecz 0,5, musimy użyć takiej chustki, która po zawieszeniu jej na sznurze w sposób opisany wyżej pokryje nie 30 cm, lecz 150 cm sznura. Widzimy więc, że *ze wzrostem poziomu ufności zwiększa się rozpiętość przedziału ufności*.

Przejdzmy teraz do przykładu liczbowego. W populacji generalnej o rozkładzie normalnym nie znamy wartości przeciętnej m , znamy natomiast odchylenie standardowe $\sigma=20$. Należy znaleźć przedział ufności, o którym moglibyśmy twierdzić, że pokryje on nieznaną wartość parametru m w populacji generalnej, przy poziomie ufności $\alpha=0,95$.

Czynności związane z konstruowaniem przedziału ufności można ująć w następujące punkty:

1° Pobieramy w sposób losowy próbki liczącą n elementów i obliczamy średnią arytmetyczną wartości cechy w próbce. Niech w naszym przykładzie średnia \bar{x} wynosi 120.

2° Obliczamy odchylenie standardowe średniej arytmetycznej z próbki. W punkcie 6.4.2, wzór (2), wykazaliśmy, że

$$V(\bar{X}) = \frac{\sigma^2}{n},$$

skąd

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}.$$

Załóżmy, że w naszym przykładzie liczebność próbki $n=100$. W takim razie

$$\sigma_{\bar{x}} = \frac{20}{\sqrt{100}} = 2.$$

⁽¹⁾ W rozważanym przez nas przypadku jest to rozkład normalny.

3º Odczytujemy w tablicach rozkładu normalnego wartość parametru t , odpowiadającą podanemu w warunkach zadania współczynnikowi ufności $\alpha=0,95$. Zauważmy, że zgodnie ze wzorem (5)

$$\alpha = 2\Phi(t),$$

skąd

$$\Phi(t) = \frac{\alpha}{2} = \frac{0,95}{2} = 0,4750.$$

W tablicach rozkładu normalnego odczytujemy, że $t = 1,96$.

4º Wyznaczamy przedział ufności. Ponieważ

$$\alpha = 0,95, \quad \bar{x} = 120, \quad \sigma_{\bar{x}} = 2,$$

zatem na mocy wzoru (6) mamy

$$P\{\bar{X} - t\sigma_{\bar{x}} < m < \bar{X} + t\sigma_{\bar{x}}\} = \alpha.$$

Podstawiając zamiast \bar{X} i $\sigma_{\bar{x}}$ odpowiednio liczby 120 i 2 otrzymujemy w naszym przykładzie następujący przedział ufności:

$$(120 - 1,96 \cdot 2, \quad 120 + 1,96 \cdot 2)$$

lub po wykonaniu rachunków

$$(116,08, \quad 123,92).$$

6.5.2. Wyznaczenie przedziału ufności dla oszacowania wartości przeciętnej w populacji generalnej o dowolnym rozkładzie za pomocą średniej arytmetycznej z dużej próbki

Aby wyznaczyć przedział ufności dla nieznanej wartości parametru m , trzeba w zasadzie znać rozkład wartości cechy w populacji generalnej. W praktyce rzadko się zdarza, że rozkład ten jest znany a priori, tzn. przed pobraniem i zbadaniem próbki. Na ogół wyjątkiem źródłem informacji o populacji generalnej jest próbka. Chcąc poznać rozkład cechy w populacji, należy zbadać rozkład cechy w próbce i na tej podstawie sformułować hipotezę statystyczną o rozkładzie cechy w populacji generalnej (omówimy to w punkcie 7.4.2).

Jak z tego widać, istnieją metody uzyskania dostatecznie wiarygodnych informacji o rozkładzie wartości cechy w populacji generalnej. Okazuje się jednak, że przy estymowaniu wartości przeciętnej nie potrzebujemy sięgać do tych metod, gdyż istnieje prostszy sposób postępowania. Sposób ten polega na wykorzystaniu wniosków, płynących z centralnego twierdzenia rachunku prawdopodobieństwa. Jak wiadomo (patrz twierdzenie Lapunowa, § 5.3), przy pewnych ogólnych założeniach, które są na ogół spełnione, rozkład średniej arytmetycznej zmiennych losowych o dowolnym rozkładzie zmierza do rozkładu normalnego, gdy liczba zmiennych losowych nieograniczenie wzrasta. Oznacza to, że rozkład średniej arytmetycznej z próbki, pobranej z populacji generalnej o dowolnym rozkładzie, różni się mało od rozkładu normalnego, jeśli tylko liczebność próbki jest

dostatecznie duża. W praktyce przyjmuje się, że rozkład średniej arytmetycznej z próbki może być zastąpiony rozkładem normalnym, gdy liczliwość próbki $n > 30$.

Próbki, których liczliwość przekracza 30 elementów, noszą nazwę *dużych próbek*.

Jak z tego wynika, jeśli tylko korzystamy z dużej próbki, możemy estymować wartość przeciętną w populacji generalnej nie znając nawet rozkładu cechy w populacji generalnej.

W punkcie 6.5.1 była mowa o tym, że przy szacowaniu parametru m wymagana jest znajomość rozkładu cechy w populacji generalnej oraz znajomość odchylenia standar-dowego σ . Postępujemy w sposób następujący:

1° Pobieramy próbke liczącą n elementów i obliczamy średnią arytmetyczną cechy w próbce. Niech ta średnia równa się 120.

2° Obliczamy odchylenie standardowe w próbce według wzoru

$$s = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}.$$

W punkcie 6.4.3, wzór (4), wykazaliśmy, że nieobciążony estymator wariancji w po-pulacji generalnej ma postać następującą:

$$\sigma^2 = \frac{n}{n-1} E(S^2).$$

Wobec tego możemy przyjąć, że zachodzi przybliżona równość

$$\sigma \approx \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}.$$

Gdy próbka jest dostatecznie duża, możemy również napisać, że

$$\sigma \approx \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}.$$

Przypuśćmy, że odchylenie standardowe s z próbki równa się 21. W takim razie przyj- mujemy, że zachodzi przybliżona równość $\sigma \approx 21$.

3° Po wyznaczeniu w ten sposób wartości odchylenia standar-dowego w populacji generalnej obliczamy

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{21}{\sqrt{100}} = 2,1.$$

4° Odczytujemy w tablicach rozkładu normalnego wartość t odpowiadającą obra-nemu poziomowi ufności α .

5° Wartości liczbowe \bar{x} , σ , n , t i α wstawiamy do wzoru

$$P \left[\bar{x} - t \frac{\sigma}{\sqrt{n}} < m < \bar{x} + t \frac{\sigma}{\sqrt{n}} \right] = \alpha.$$

6.5.3. Rozkład χ^2 . Przedział ufności dla oszacowania wariancji w populacji generalnej

Dana jest populacja generalna o rozkładzie normalnym $N(m, \sigma)$. Parametry populacji m i σ są nieznane. Należy oszacować wariancję populacji σ^2 . Z populacji generalnej pobieramy w sposób losowy próbki i obliczamy

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

Istnieje twierdzenie, z którego wynika, że jeśli X_1, X_2, \dots, X_k są to niezależne zmienne losowe o rozkładzie normalnym $N(0, 1)$, to zmienna losowa

$$(1) \quad \chi^2 = \sum_{i=1}^k X_i^2$$

ma rozkład, którego gęstość wyraża się wzorem

$$(2) \quad f(x) = \begin{cases} \frac{x^{k/2-1} e^{-x/2}}{\Gamma(k/2) \cdot 2^{k/2}} & \text{dla } x > 0, \\ 0 & \text{dla } x \leq 0. \end{cases}$$

Dowód tego twierdzenia znajdzie czytelnik w [5], str. 153 - 154.

W punkcie 3.5.4 podaliśmy definicję gęstości rozkładu gamma. Podstawiając do wzoru (1), 3.5.4, $a = \frac{1}{2}k - 1$ oraz $b = 2$, łatwo przekonać się, że rozkład χ^2 jest szczególnym przypadkiem rozkładu gamma.

Parametr k , występujący we wzorze (2), nosi nazwę *liczby stopni swobody*⁽¹⁾.

Tablice rozkładu χ^2 podane są na końcu niniejszej książki.

Z twierdzenia Lapunowa wiadomo, że wraz ze wzrostem liczby składników sumy niezależnych zmiennych losowych o pewnych własnościach rozkład tej sumy zmierza do rozkładu normalnego. Ponieważ stopnie swobody oznaczają właśnie liczbę składników sumy kwadratów zmiennych losowych o rozkładzie normalnym, przeto ze wzrostem liczby stopni swobody do nieskończoności rozkład χ^2 dąży do rozkładu normalnego. Gdy liczba stopni swobody jest większa od 30, rozkład normalny daje na ogół dostatecznie dobre przybliżenie rozkładu χ^2 . Dlatego właśnie największa liczba stopni swobody, występująca w tablicach, wynosi 30. Gdy liczba stopni swobody jest większa od 30, korzystamy z zależności

$$(3) \quad P\{\chi^2 > x\} \approx 0,5 - \Phi(t),$$

⁽¹⁾ Ten nieco dziwny termin wprowadzony został przez R. Fishera. Genezę tego terminu można wyjaśnić w ten sposób, że zgodnie z określeniem zmiennej χ^2 liczba k jest liczbą niezależnych zmiennych losowych X_i^2 . Jeśli zmienne losowe X_1, X_2, \dots, X_n są ze sobą powiązane r ($r \leq n$) związkami liniowymi typu

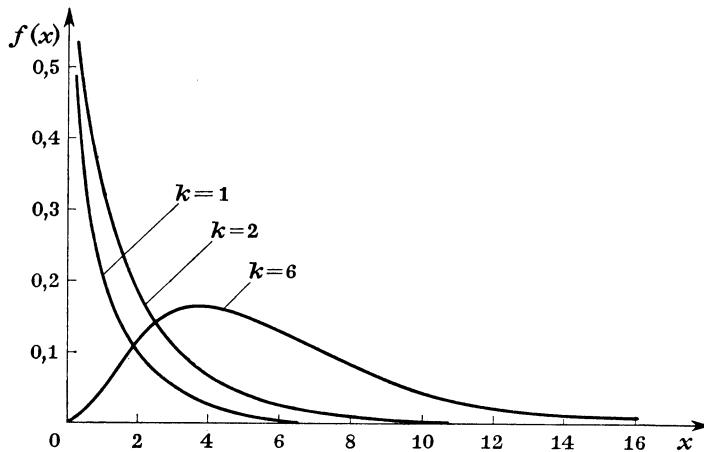
$$a_{j_1} X_1 + a_{j_2} X_2 + \dots + a_{j_r} X_n = a_j \quad (j = 1, 2, \dots, r),$$

to liczba stopni swobody wynosi $k = n - r$.

przy czym

$$(4) \quad x = \frac{(t + \sqrt{2k-1})^2}{2}.$$

Funkcja gęstości rozkładu χ^2 jest bardzo asymetryczna dla małych wartości k .



Rys. 1

Rysunek 1 przedstawia wykresy funkcji gęstości rozkładu χ^2 dla $k=1, 2$ i 6 . Widzimy, że w miarę wzrostu k asymetria rozkładu maleje i wykres rozkładu upodabnia się do rozkładu normalnego. Gdy $k \leq 2$, funkcja gęstości rozkładu χ^2 jest funkcją monotoniczną malejącą. Gdy $k > 2$, funkcja gęstości ma maksimum w punkcie $x=k-2$. Można udowodnić, że jeśli populacja generalna ma rozkład normalny $N(m, \sigma)$, to zmienna losowa $\frac{nS^2}{\sigma^2}$ ma rozkład χ^2 . Ponieważ zmienna losowa $S^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$, a $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$, przeto zmienne losowe $(X_i - \bar{X})^2$ łączy jeden związek liniowy, a mianowicie $\sum_{i=1}^n (X_i - \bar{X}) = 0$. Wobec tego liczba stopni swobody dla zmiennej losowej $\frac{nS^2}{\sigma^2}$ wynosi $k=n-1$.

Znajomość rozkładu zmiennej losowej χ^2 pozwoli znaleźć granice przedziału ufności, o którym możemy twierdzić, że z zadanym prawdopodobieństwem α pokryje on nieznaną wartość parametru σ^2 populacji generalnej. Sposób znajdowania granic przedziału ufności wyjaśnimy na przykładzie liczbowym.

Mamy wyznaczyć przedział ufności dla oszacowania nieznanej wartości σ^2 . Niech poziom ufności $\alpha=0,98$. Oznaczmy symbolem $P(x)$ prawdopodobieństwo tego, że χ^2 przyjmie wartość większą od x . W takim razie

$$(5) \quad P(x) = P\{\chi^2 > x\} = P\{\chi^2 \geq x\},$$

gdyż

$$P\{\chi^2 = x\} = 0.$$

Granice przedziału ufności oznaczmy symbolami x_1 oraz x_2 . Jak wiadomo, granice przedziału ufności określają jednoznacznie wartość poziomu ufności. Natomiast poziom ufności nie określa jednoznacznie granic przedziału ufności. Dla wyznaczenia tych granic potrzebne są nam dwa równania. Jednym z nich jest

$$(6) \quad P\{\chi^2 < x_1\} + P\{\chi^2 > x_2\} = 1 - \alpha$$

lub inaczej

$$(7) \quad P\{\chi^2 \geq x_1\} - P\{\chi^2 > x_2\} = \alpha.$$

Aby otrzymać drugie równanie, musimy przyjąć dodatkowe założenie⁽¹⁾, na przykład warunek

$$(8) \quad P\{\chi^2 < x_1\} = P\{\chi^2 > x_2\},$$

który można napisać inaczej, a mianowicie

$$(9) \quad 1 - P\{\chi^2 \geq x_1\} = P\{\chi^2 > x_2\}.$$

Wzory (7) i (9) pozwalają utworzyć następujący układ dwóch równań:

$$(10) \quad \begin{aligned} P(x_1) - P(x_2) &= \alpha, \\ P(x_1) + P(x_2) &= 1. \end{aligned}$$

Rozwiążując ten układ względem $P(x_1)$ i $P(x_2)$ otrzymujemy

$$(11) \quad P(x_2) = \frac{1 - \alpha}{2}, \quad P(x_1) = \frac{1 + \alpha}{2}.$$

Mówiliśmy poprzednio, że zmienna losowa $\frac{nS^2}{\sigma^2}$ ma rozkład χ^2 o $n-1$ stopniach swobody. W takim razie

$$(12) \quad P\left\{x_1 < \frac{nS^2}{\sigma^2} < x_2\right\} = 0,98$$

i dalej przekształcając

$$P\left\{\frac{x_1}{nS^2} < \frac{1}{\sigma^2} < \frac{x_2}{nS^2}\right\} = 0,98,$$

$$P\left\{\frac{nS^2}{x_1} > \sigma^2 > \frac{nS^2}{x_2}\right\} = 0,98,$$

$$(13) \quad P\left\{\frac{nS^2}{x_2} < \sigma^2 < \frac{nS^2}{x_1}\right\} = 0,98.$$

⁽¹⁾ Na ogół założenia te obiera się w ten sposób, aby otrzymać jak najwęższy przedział ufności, gdyż wtedy otrzymuje się największą dokładność oszacowania.

Przypuśćmy, że w celu oszacowania σ^2 pobraliśmy próbki liczącą 20 elementów i obliczyliśmy odchylenie standardowe s w próbce. Niech $s=2$. Liczba stopni swobody wynosi w naszym wypadku $k=20-1=19$. Podstawiając we wzorze (11) zamiast α liczbę 0,98 otrzymamy, że $P(x_2)=0,01$ oraz $P(x_1)=0,99$.

W tablicach rozkładu χ^2 na przecięciu wiersza, któremu odpowiada liczba stopni swobody $k=19$ i kolumny opatrzonej nagłówkiem 0,99 odczytujemy wartość $x_1=7,6$. W analogiczny sposób znajdujemy $x_2=36,2$. Wstawiając otrzymane wartości liczbowe do nierówności znajdującej się w nawiasach po lewej stronie znaku równości we wzorze (13) otrzymujemy

$$\frac{20 \cdot 4}{36,2} < \sigma^2 < \frac{20 \cdot 4}{7,6}$$

lub po wykonaniu działań

$$2,210 < \sigma^2 < 10,526.$$

Przedział liczbowy (2,210, 10,526) jest szukanym przedziałem ufności szacowanego parametru σ^2 w populacji generalnej.

6.5.4. Rozkład Studenta. Wyznaczanie przedziału ufności dla oszacowania wartości przeciętnej w populacji generalnej na podstawie małej próbki

Jak wiadomo, wyłącznym źródłem informacji, z którego korzystamy przy estymowaniu nieznanych wartości parametrów populacji generalnej, jest próbka. Im ta próbka jest bardziej liczna, tym większa jest dokładność estymacji. Pobieranie i badanie próbki jest w praktyce zawsze związane z pewnym kosztem. Koszt ten zależy od liczby elementów próbki. Zrozumiałe jest, że każdorazowo, gdy wymagana jest znajomość stosunków panujących w populacji generalnej, ścierają się ze sobą dwie przeciwnostawne tendencje. Jedną z tych tendencji jest dążność do otrzymania informacji jak najdokładniejszych, drugą natomiast jest pragnienie, aby koszt zdobycia informacji był najniższy. Nie można niestety uczynić zadość jednocześnie obu tendencjom. Dwie zmienne: dokładność badania i koszt badania są ze sobą powiązane w ten sposób, że gdy jedna rośnie, to i druga rośnie. Koszt badania jest jednoznacznie określony dokładnością badania – i na odwrót. W § 6.6 opisany zostanie sposób wyznaczania liczby elementów próbki, która pozwala wypowiadać sądy o populacji generalnej z określona z góry dokładnością i zadanym współczynnikiem ufności. Zdarza się często, że wystarczający dla potrzeb praktycznych zasób informacji o zbiorowości generalnej można otrzymać korzystając z próbki o małej liczbie elementów, nie przekraczającej 30 elementów. Wypowiadanie sądów o populacji na podstawie małej próbki pociąga za sobą poważne trudności teoretyczne, polegające na konieczności zrezygnowania ze wszystkich twierdzeń granicznych, przy których korzysta się z założenia, że liczba elementów próbki nieograniczenie wzrasta.

Wyznaczając przedział ufności dla oszacowania wartości przeciętnej w populacji generalnej korzystamy ze zmiennej

$$T = \frac{\bar{X} - m}{\sigma_{\bar{x}}} = \frac{\sqrt{n}}{\sigma} (\bar{X} - m),$$

o której wiemy, że jej rozkład zmierza wraz ze wzrostem próbki do rozkładu normalnego. Jeżeli jednak ze względu na koszty badania próbka jest mała i nie możemy jej powiększyć, to rozkład zmiennej T może znacznie różnić się od rozkładu normalnego. Gdy próbka jest mała, to nie możemy również zastąpić nieznanej wartości odchylenia standardowego σ w populacji generalnej odchyleniem standardowym s , obliczonym na podstawie próbki.

Powiedzieliśmy, że gdy liczebność próbki jest większa od 30, rozkład zmiennej T jest zbliżony do rozkładu normalnego. Im próbka jest mniejsza, tym rozbieżność między rozkładem T i rozkładem normalnym jest większa. Nie oznacza to jednak, że dysponując małą próbką nie potrafimy nigdy wyznaczyć przedziału ufności dla oszacowania parametru m . Statystyk angielski W. Gosset, występujący pod pseudonimem Student, udowodnił⁽¹⁾ następujące

TWIERDZENIE. *Jeśli niezależne zmienne losowe X_1, X_2, \dots, X_n mają rozkład normalny $N(m, \sigma)$, to zmienna losowa*

$$(1) \quad T = \frac{\sqrt{n-1}}{S} (\bar{X} - m) = \frac{\sqrt{k}}{S} (\bar{X} - m)$$

ma rozkład, którego funkcja gęstości wyraża się wzorem

$$(2) \quad f(t) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\Gamma\left(\frac{k}{2}\right)\sqrt{k\pi}} \left(1 + \frac{t^2}{k}\right)^{-(k+1)/2},$$

przy czym k oznacza liczbę stopni swobody.

Rozkład (2) znany jest w statystyce pod nazwą *rozkładu Studenta*. Rozkład ten jest symetryczny i wyglądem swym przypomina rozkład normalny. Gdy liczba stopni swobody k wzrasta do nieskończoności, rozkład Studenta zmierza do rozkładu normalnego. Tablice rozkładu Studenta podane są na końcu książki.

Jeśli cecha populacji generalnej ma rozkład normalny, to posługując się rozkładem Studenta potrafimy wyznaczyć granice przedziału ufności dla oszacowania wartości przeciętnej populacji generalnej nawet wtedy, gdy dysponujemy jedynie małą próbką. Wyjaśnimy to na przykładzie.

PRZYKŁAD 1. Cecha X populacji generalnej ma rozkład normalny. Z populacji tej pobrano próbkę, której liczebność $n=10$ elementów. Na podstawie danych z próbki należy oszacować wartość przeciętną m w populacji generalnej.

Obliczamy średnią arytmetyczną i odchylenie standardowe w próbce. Przypuśćmy, że $\bar{x}=24$, a $s=2,7$. Ustalamy poziom ufności α . Niech $\alpha=0,98$. W takim razie (patrz rys. 1, str. 220):

$$P\{|t_0\}| = P\{|T| > t_0\} = 1 - \alpha.$$

⁽¹⁾ Dowód znajdzie czytelnik u Fisza [8], str. 364 - 370.

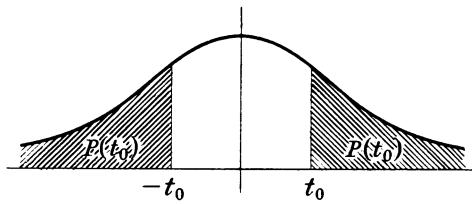
W tablicy III dla $\alpha=0,98$ i $k=10-1=9$ znajdujemy, że $t=2,821 \approx 2,8$. W związku z tym

$$P\{-t < T < t\} = P\left\{-t < \frac{\sqrt{n-1}}{S} (\bar{X} - m) < t\right\} = P\left\{\bar{X} - \frac{tS}{\sqrt{n-1}} < m < \bar{X} + \frac{tS}{\sqrt{n-1}}\right\}.$$

Ponieważ $\bar{x}=24$, $s=2,7$, $t=2,8$, więc szukany przedział ufności przybierze postać

$$\left(24 - \frac{2,8 \cdot 2,7}{3}, \quad 24 + \frac{2,8 \cdot 2,7}{3}\right),$$

a po wykonaniu działań arytmetycznych – postać ostateczną (21,48, 26,52).



Rys. 1

6.6. WYZNACZANIE WIELKOŚCI PRÓBKI⁽¹⁾

Estymując nieznaną wartość parametru populacji generalnej za pomocą parametru z próbki nie staramy się na ogół podać dokładnej wartości parametru estymowanego, lecz zadowalamy się wyznaczeniem przedziału, o którym z danym prawdopodobieństwem możemy twierdzić, że obejmie on nieznaną wartość parametru populacji. Jeśli cecha populacji jest cechą ciągłą – inaczej postąpić nie możemy. Wtedy bowiem prawdopodobieństwo tego, że wartość parametru z próbki będzie się dokładnie równała wartości parametru populacji generalnej, jest równa zeru.

Przypuśćmy, że chcemy oszacować wartość przeciętną m . Wiemy, że nie potrafimy wyznaczyć tej wartości dokładnie, lecz możemy jedynie dla określonego z góry prawdopodobieństwa podać granice przedziału ufności. O przedziale tym na poziomie ufności α możemy twierdzić, że pokryje on wartość przeciętną m .

Napiszemy stosowny wzór:

$$P\{\bar{X} - t\sigma_{\bar{x}} < m < \bar{X} + t\sigma_{\bar{x}}\} = \alpha.$$

Oznaczmy

$$(1) \quad \Delta_{\bar{x}} = t\sigma_{\bar{x}} = t \frac{\sigma}{\sqrt{n}}.$$

Wielkość $\Delta_{\bar{x}}$ nazywać będziemy *dopuszczalnym błędem oceny* lub inaczej *tolerancją*.

⁽¹⁾ Patrz [35], § 3.4.

Oczywiście przy ustalonej wartości poziomu ufności α estymacja parametru populacji będzie tym lepsza, im tolerancja będzie mniejsza. We wzorze (1) t i σ są to stałe. Wielkość t jest stała dlatego, że zależy ona od prawdopodobieństwa α , o którym powiedzieliśmy, że jest wielkością stałą. Natomiast σ jest to odchylenie standardowe populacji, a więc tym samym też jest wielkością stałą. We wzorze (1) wielkościami zmiennymi są $\Delta_{\bar{x}}$ i n .

Powiększając liczebność próbki n możemy błąd oceny $\Delta_{\bar{x}}$ dowolnie zmniejszyć. Otrzymaliśmy wynik zgodny z intuicją: *im większa próbka – tym większa dokładność oceny*.

W praktyce poziom ufności α oraz tolerancja $\Delta_{\bar{x}}$ są to liczby dane. Ustala się je tak, jak tego wymaga rozwiązywane zagadnienie. Znając α znamy również t (wartość t odczytujemy po prostu w tablicach rozkładu Studenta lub rozkładu normalnego). Mając te dane możemy wyznaczyć wielkość próbki. Ze wzoru (1) otrzymujemy

$$(2) \quad n = \frac{t^2 \sigma^2}{\Delta_{\bar{x}}^2}.$$

Jeśli próbkę pobieramy z populacji ograniczonej metodą losowania bez zwracania, to zgodnie ze wzorem (9) z 6.4.4

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \approx \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N}}.$$

Wstawiając prawą stronę tej równości do wzoru (1) otrzymujemy

$$(3) \quad \Delta_{\bar{x}} = t \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \approx \sigma t \sqrt{\frac{N-n}{Nn}}.$$

Stąd zaś po nietrudnych przekształceniach znajdujemy

$$(4) \quad n = \frac{N t^2 \sigma^2}{N \Delta_{\bar{x}}^2 + t^2 \sigma^2}.$$

Wyznaczanie rozmiarów próbki zilustrujemy na dwóch przykładach.

PRZYKŁAD 1. Wykonujemy pomiary grubości płytki platynowej. Mamy wyznaczyć ilość pomiarów, jeśli tolerancja $\Delta_{\bar{x}}=0,016$ mm, standardowy błąd pomiaru $\sigma=0,1$ mm, a poziom ufności $\alpha=0,95$.

Zgodnie ze wzorem (3) z 6.5.1, mamy $2\Phi(t)=\alpha=0,95$, $\Phi(t)=0,4750$. W tablicy rozkładu normalnego znajdujemy, że $t=1,96$. Wobec tego na mocy wzoru (2)

$$n = \frac{1,96^2 \cdot (0,1 \text{ mm})^2}{(0,016 \text{ mm})^2} = 150.$$

PRZYKŁAD 2. Jak wielką próbkę należy pobrać z partii zboża siewnego, aby z prawdopodobieństwem 0,9 można było twierdzić, że udział ziaren zdolnych do kiełkowania w próbie będzie się różnił od udziału ziaren zdolnych do kiełkowania w całej populacji o mniej niż 0,01?

Oznaczmy symbolem x/n częstość występowania ziaren zdolnych do kiełkowania w próbce. W zadaniu tym nieznanym parametrem populacji jest udział p ziaren zdolnych do kiełkowania. W zadaniu jest podane $\alpha=0,9$ oraz $A_{x/n}=0,01$. Nie znamy natomiast $\sigma_{x/n}$. Wiemy jednak, że w rozkładzie dwumianowym $V(X)=npq$ (patrz 4.3.1, wzór (9)). W naszym zadaniu zamiast zmiennej X występuje zmienna X/n . W punkcie 5.1.4 wykazaliśmy, że wariancja tej zmiennej równa się $p(1-p)/n$. Stąd $\sigma_{x/n}^2=p(1-p)/n$. Ponieważ iloczyn $p(1-p)$ osiąga maksimum, gdy $p=1-p=1/2$, przeto

$$\sigma_{x/n} = \sqrt{\frac{p(1-p)}{n}} \leq \frac{1}{2\sqrt{n}}.$$

W takim razie warunki zadania będą spełnione, jeśli przyjmiemy

$$A_{x/n} = \frac{t}{2\sqrt{n}}.$$

Stąd

$$(*) \quad n = \frac{t^2}{4A_{x/n}^2}.$$

W tablicach rozkładu normalnego znajdujemy, że prawdopodobieństwu $\alpha=0,4500$ odpowiada $t=1,64$. Podstawiając dane liczbowe zamiast t i $A_{x/n}$ do wzoru (*) mamy

$$n = \frac{1,64^2}{4 \cdot 0,01^2} = 6724.$$

Czytelnik zechce sprawdzić, że rozwiązyując to samo zadanie w punkcie 5.1.6 otrzymamy $n=25\,000$. Nietrudno wyjaśnić, skąd się wzięła ta rozbieżność wyników. W punkcie 5.1.6 rozwiązywaliśmy zadanie za pomocą nierówności Czebyszewa. Jak wiadomo, przy korzystaniu z tej nierówności nie wymaga się znajomości rozkładu populacji. Jest to ogromna zaleta nierówności Czebyszewa. Zaleta ta jednak ma znaczenie raczej teoretyczne. Do rozwiązywania zagadnień praktycznych nierówność Czebyszewa nie nadaje się, gdyż ocena prawdopodobieństwa uzyskana za pomocą tej nierówności jest mało dokładna. W praktyce dla wyznaczenia rozmiarów próbki wymagana jest znajomość rozkładu populacji. Korzyści, jakie płyną ze znajomości tego rozkładu, mają duże znaczenie. W naszym przykładzie sprowadzają się one do tego, że zamiast pobierać próbki liczącą 25 000 elementów, możemy poprzedzać na próbce liczącej 6724 elementy.

Gdy mowa o ziarnach pszenicy – daje to oszczędność jedynie na kosztach pobierania i badania próbki. Koszt ziaren pszenicy, ulegających zniszczeniu w trakcie procesu badania, jest tak znikomy, że może być pominięty. Zresztą koszty pobierania próbki i badania siły kiełkowania ziaren pszenicy są również niewysokie. Inaczej rzecz się przedstawia, gdy badanie próbki polega na precyzyjnych, a więc żmudnych i kosztownych pomiarach elementów próbki. Wtedy sprawa możliwie małej liczebności próbki nabiera pierwszorzędnego znaczenia gospodarczego we wszystkich zastosowaniach metody reprezentacyjnej.

6.7. ESTYMACJA PARAMETRÓW REGRESJI LINIOWEJ

6.7.1. Wprowadzenie

Nie ulega wątpliwości, że zagadnienia regresji są znacznie łatwiejsze w przypadku dwuwymiarowym niż w przypadku wielowymiarowym. Są one prostsze pojęciowo, bardziej intuicyjne, bo mogą być ilustrowane za pomocą wykresów, są również łatwiejsze pod względem numerycznym i wymagają prostszych metod i środków obliczeniowych. To wszystko sprawia, że analiza regresji w przypadku dwuwymiarowym ma znacznie wyższą rangę praktyczną niż analiza wielowymiarowa. Nadając się do rozpowszechniania i szerokiego spopularyzowania dwuwymiarowa analiza regresji może być zastosowana w skali masowej w naukach przyrodniczych, technicznych i społecznych (oraz, co natomiast szczególnie interesuje, może stać się potężnym narzędziem analizy gospodarczej i rachunku ekonomicznego), a przecież tylko masowe zastosowanie w praktyce określonej metody badawczej może przynieść korzyści zasługujące na uwagę i liczące się w skali społecznej.

Regresja wielowymiarowa jest trudniejsza od dwuwymiarowej przede wszystkim dla tego, że nie podlega interpretacji za pomocą środków graficznych, a interpretacja geometryczna wymaga od czytelnika wyrobionej wyobraźni przestrzennej. Poza tym, że względu na zapis macierzowy, studiowanie zagadnień związanych z wielowymiarowymi zmiennymi losowymi wymaga dobrej znajomości podstaw algebry liniowej.

W niniejszym paragrafie omówione zostaną najpierw metody estymacji parametrów regresji w przypadku dwuwymiarowym, a dopiero potem zostanie przedstawione uogólnienie tych metod na wypadek k -wymiarowy. Ponieważ tematyka tego paragrafu wiąże się ściśle z problematyką paragrafu 4.7, przeto zaleca się, aby czytelnik studiował oba te paragrafy łącznie. Wiadomości zdobyte uprzednio przy przerabianiu paragrafu 4.7 okażą się obecnie bardzo przydatne.

Istnieje kilka sposobów estymacji parametrów dwuwymiarowej populacji generalnej w oparciu o dane liczbowe z próbki pobranej z populacji w sposób losowy. Do najważniejszych należy zaliczyć metodę maksimum wiarygodności, metodę minimalnej wariancji, metodę minimum χ^2 i metodę najmniejszych kwadratów. Jak dotąd, w teorii regresji stosuje się zwykle metodę najmniejszych kwadratów. Jest ona znana tak wśród astronomów jak i geodetów, zarówno wśród fizyków jak i biologów, techników i ekonomistów.

Metoda najmniejszych kwadratów ma bardzo cenną zaletę formalną, ważną w przypadku regresji liniowej. Istnieje twierdzenie znane pod nazwą twierdzenia Markowa, które głosi, że estymatory uzyskane tą metodą są zgodne, nieobciążone i najefektywniejsze. W twierdzeniu tym nie zakłada się normalności rozkładów zmiennych losowych, a nawet niezależności tych zmiennych.

Pomimo tych niewątpliwych zalet metody najmniejszych kwadratów – istnieją dwa wąględy, dla których obok tej metody warto poznać jeszcze jeden sposób estymacji parametrów regresji. Do czasu ostatecznego ustalenia nazwy nazywać go będziemy metodą punktową, zastrzegając się, że jest to termin tymczasowy. Otóż w przypadku regresji liniowej metoda punktowa daje również zgodne i nieobciążone estymatory parametrów re-

gresji, przy czym 1° rachunki związane z obliczeniem wartości liczbowych estymatorów są znacznie łatwiejsze od rachunków koniecznych w metodzie najmniejszych kwadratów oraz 2° do opanowania metody punktowej nie jest potrzebna znajomość rachunku ekstremum, wymagana przy metodzie najmniejszych kwadratów. Efektywność estymatorów uzyskanych za pomocą metody punktowej jest, co prawda, nieco gorsza od efektywności estymatorów otrzymanych metodą najmniejszych kwadratów, gdy jednak dysponujemy dużą próbką, względ ten nie ma zasadniczego znaczenia. Poważne korzyści, jakie daje wprowadzenie metody punktowej do teorii estymacji parametrów regresji liniowej, polegają przede wszystkim na tym, że metoda ta sprzyja w poważnym stopniu rozpowszechnieniu teorii regresji i korelacji wśród praktyków. Ma to szczególne znaczenie w odniesieniu do badań ekonomicznych. Główną przeszkodą w rozpowszechnieniu metod regresji i korelacji wśród ekonomistów są bez wątpienia trudności natury matematycznej, związane z wyznaczaniem parametrów regresji metodą klasyczną⁽¹⁾. Metoda punktowa w znacznym stopniu ułatwia pokonanie tej przeszkody.

6.7.2. Estymacja parametrów regresji liniowej za pomocą metody najmniejszych kwadratów w przypadku zmiennych losowych dwuwymiarowych

Rozważmy zmienną losową (X, Y) o rozkładzie normalnym. Jak wiadomo, regresja Y względem X oraz X względem Y jest w tym rozkładzie funkcją liniową, wobec tego równania regresji można zapisać następująco:

$$y = \alpha_{21} x + \beta_{20}, \quad x = \alpha_{12} y + \beta_{10}.$$

Parametry występujące w tych równaniach, jak również wszystkie parametry występujące w równaniu funkcji gęstości rozkładu normalnego są nieznane. Dla oszacowania tych parametrów realizuje się n -elementową serię eksperymentów, które dostarczają n par liczb

$$(1) \quad (x_1, y_1), \quad (x_2, y_2), \quad \dots, \quad (x_n, y_n)$$

stanowiących realizacje zmiennej losowej (X, Y) . Dla podkreślenia, że punkty eksperymentalne (1) otrzymane zostały w sposób doświadczalny, mówi się o nich, że tworzą *próbkę* lub (*populację próbną*) pobraną z *populacji generalnej*, którą stanowią wszystkie możliwe realizacje zmiennej losowej (X, Y) lub, mówiąc inaczej, wszystkie możliwe elementy przestrzeni punktów eksperymentalnych. Jeżeli pobieranie próbki przeprowadzone zostało w taki sposób, że

- 1° każdy element populacji miał zapewnione jednakowe prawdopodobieństwo trafienia do próbki;
- 2° doświadczenia generujące punkty eksperymentalne były niezależne, to mówi się, że próbka pobrana została z populacji w sposób *losowy*.

⁽¹⁾ Tzn. metodą najmniejszych kwadratów.

Informacje dostarczone przez dane eksperymentalne (1) służą do oszacowania parametrów α_{12} , α_{21} , β_{10} , β_{20} obu linii regresji. Zgodnie z przyjętą konwencją oznaczania estymatora danego parametru przez umieszczenie nad tym parametrem znaku „ $\hat{}$ ”, estymatory parametrów równań regresji można oznaczać symbolami $\hat{\alpha}_{12}$, $\hat{\alpha}_{21}$, $\hat{\beta}_{10}$, $\hat{\beta}_{20}$. Oprócz tej symboliki stosowane bywa także prostsze znakowanie, a mianowicie a_{12} , a_{21} , b_{10} , b_{20} . Te same zasady stosuje się przy oznaczaniu estymatorów takich parametrów, jak σ_{21} , σ_{12} , ρ , σ_1 , σ_2 . Estymatory te zwykle zapisywane są w postaci $\hat{\sigma}_{21}$, $\hat{\sigma}_{12}$, $\hat{\rho}$, $\hat{\sigma}_1$, $\hat{\sigma}_2$ lub w postaci s_{21} , s_{12} , r , s_1 , s_2 . Jedyne naruszenie tej konwencji występuje w przypadku estymatorów parametrów $m_1 = E(X)$ oraz $m_2 = E(Y)$, które zawsze oznacza się symbolami \bar{x} oraz \bar{y} .

Równania

$$\hat{y} = \hat{\alpha}_{21} x + \hat{\beta}_{20} \quad \text{i} \quad \hat{x} = \hat{\alpha}_{12} y + \hat{\beta}_{10}$$

nazywać będziemy *empirycznymi równaniami regresji Y względem X i X względem Y lub też równaniami regresji w próbce dla odróżnienia od równań hipotetycznych*

$$y = \alpha_{21} x + \beta_{20} \quad \text{i} \quad x = \alpha_{12} y + \beta_{10},$$

które określają również mianem *równań regresji w populacji*.

Przystępujemy obecnie do przedstawienia otrzymywania estymatorów a , b parametrów regresji α , β za pomocą metody najmniejszych kwadratów. Ponieważ zastosowane postępowanie będzie analogiczne dla przypadku regresji Y względem X jak i X względem Y , przeto nie zachodzi potrzeba rozróżniania tych przypadków za pomocą indeksów umieszczanych zwykle obok parametrów równań regresji. Tym razem indeksy te zostały pomijone.

Dla wyznaczenia wartości estymatorów a , b , za pomocą których można byłoby oszacować nieznane parametry α , β w równaniu regresji $y = \alpha x + \beta$, należy zminimizować wyrażenie

$$\sum_{i=1}^n (y_i - ax_i - b)^2.$$

W wyniku prostych przekształceń (patrz 4.7.1) otrzymuje się układ równań normalnych

$$\sum_{i=1}^n x_i y_i - a \sum_{i=1}^n x_i^2 - b \sum_{i=1}^n x_i = 0,$$

$$\sum_{i=1}^n y_i - a \sum_{i=1}^n x_i - nb = 0.$$

Rozwiązujeć ten układ równań względem a i b otrzymamy

(1)

$$b = \bar{y} - a\bar{x},$$

(2)

$$a = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

We wzorze (1) \bar{x} i \bar{y} są to średnie arytmetyczne w próbce, tzn.

$$(3) \quad \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i, \quad \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i.$$

Równanie regresji liniowej w próbce przybiera więc postać

$$(4) \quad \hat{y} = ax + b.$$

Wprowadźmy dalsze oznaczenia

$$(5) \quad s_1 = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}, \quad s_2 = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2}.$$

Są to odchylenia standardowe zmiennych X i Y w próbce. Kowariancja w próbce wyraża się wzorem

$$(6) \quad c(x, y) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}).$$

W takim razie parametry regresji Y względem X można zapisać w sposób następujący:

$$(7) \quad b_{20} = \bar{y} - a_{21} \bar{x},$$

$$(8) \quad a_{21} = \frac{c(x, y)}{s_1^2}.$$

Nietrudno dostrzec analogię wzorów (7) i (8) ze wzorami (27) i (29) z punktu 4.7.1.

Podobnie mamy dla regresji X względem Y :

$$(9) \quad b_{10} = \bar{x} - a_{12} \bar{y},$$

$$(10) \quad a_{12} = \frac{c(x, y)}{s_2^2}.$$

Parametry a_{21} i a_{12} nazywać będziemy współczynnikami regresji w próbce.

W przypadku regresji Y względem X standardowy błąd oceny w próbce, przez analogię do wzoru (35) w punkcie 4.7.1, definiuje się następująco:

$$(11) \quad s_{21} = \sqrt{s_2^2 - a_{21} c(x, y)}.$$

Podobnie dla regresji X względem Y

$$(12) \quad s_{12} = \sqrt{s_1^2 - a_{12} c(x, y)}.$$

Estymatorem współczynnika korelacji w populacji generalnej jest współczynnik korelacji r w próbce. Współczynnik r zdefiniujemy za pomocą wzoru

$$(13) \quad r^2 = a_{12} \cdot a_{21}$$

wykorzystując analogię do wzoru (6) z punktu 4.7.2.

6.7.3. Estymacja parametrów regresji liniowej za pomocą metody punktowej

Oznaczmy symbolem Ω dwuwymiarową populację generalną. Każdemu elementowi należącemu do tej populacji odpowiada para realizacji (x, y) zmiennej losowej (X, Y) . Zakłada się, że linie regresji pierwszego rodzaju w populacji generalnej są liniami prostymi, tzn.:

$$(1) \quad y = \alpha_{21} x + \beta_{20}$$

oraz

$$(2) \quad x = \alpha_{12} x + \beta_{10},$$

gdzie α_{21} , α_{12} , β_{20} i β_{10} są parametrami regresji w populacji generalnej. Z populacji pobiera się w sposób losowy próbki ω liczącą n elementów. Otrzymujemy n par liczb (x_i, y_i) ($i=1, 2, \dots, n$) odpowiadających wylosowanym elementom. Liczby te można interpretować jako współrzędne punktów eksperymentalnych na płaszczyźnie. Każdemu elementowi populacji Ω odpowiada taki punkt eksperymentalny.

Obliczamy

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i,$$

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i.$$

Dzielimy zbiór ω na dwa podzbiory w ten sposób, że do pierwszego podzbioru zaliczamy punkty o odciętej X nie większej od \bar{x} , natomiast do drugiego podzbioru – pozostałe punkty. Jeżeli do drugiego podzbioru trafi k punktów, to do pierwszego podzbioru trafi oczywiście $n - k$ punktów. Zwróćmy uwagę, że przy takim podziale zbioru ω na podzbiory wielkość k jest zmienną losową, która może przybierać wartości $1, 2, \dots, n-1$. Wprowadźmy następujące oznaczenia:

$$(3) \quad \begin{aligned} X_{(1)} &= X | X \leq \bar{x}, \\ Y_{\langle 1 \rangle} &= Y | X \leq \bar{x}, \\ X_{(2)} &= X | X > \bar{x}, \\ Y_{\langle 2 \rangle} &= Y | X > \bar{x}. \end{aligned}$$

Obliczamy

$$\bar{x}_{(1)} = \frac{1}{n-k} \sum x_{(1)}, \quad \bar{x}_{(2)} = \frac{1}{k} \sum x_{(2)},$$

$$\bar{y}_{\langle 1 \rangle} = \frac{1}{n-k} \sum y_{\langle 1 \rangle}, \quad \bar{y}_{\langle 2 \rangle} = \frac{1}{k} \sum y_{\langle 2 \rangle}.$$

Można udowodnić następujące

TWIERDZENIE. *Mamy*

$$(4) \quad \begin{vmatrix} \bar{x}_{(1)} & \bar{y}_{\langle 1 \rangle} & 1 \\ \bar{x}_{(2)} & \bar{y}_{\langle 2 \rangle} & 1 \\ \bar{x} & \bar{y} & 1 \end{vmatrix} = 0.$$

Z twierdzenia tego wynika, że punkty $(\bar{x}_{(1)}, \bar{y}_{\langle 1 \rangle})$, $(\bar{x}_{(2)}, \bar{y}_{\langle 2 \rangle})$, (\bar{x}, \bar{y}) leżą na jednej prostej. Za estymator parametru α_{21} można przyjąć współczynnik kątowy tej prostej, który da się wyrazić za pomocą jednego z następujących trzech wzorów:

$$(5) \quad a_{21} = \frac{\bar{y}_{\langle 2 \rangle} - \bar{y}}{\bar{x}_{(2)} - \bar{x}},$$

$$(6) \quad a_{21} = \frac{\bar{y} - \bar{y}_{\langle 1 \rangle}}{\bar{x} - \bar{x}_{(1)}},$$

$$(7) \quad a_{21} = \frac{\bar{y}_{\langle 2 \rangle} - \bar{y}_{\langle 1 \rangle}}{\bar{x}_{(2)} - \bar{x}_{(1)}}.$$

Estymator parametru β_{20} można również zapisać w jeden z następujących trzech sposobów:

$$(8) \quad b_{20} = \bar{y}_{\langle 2 \rangle} - a_{21} \bar{x}_{(2)},$$

$$(9) \quad b_{20} = \bar{y}_{\langle 1 \rangle} - a_{21} \bar{x}_{(1)},$$

$$(10) \quad b_{20} = \bar{y} - a_{21} \bar{x}.$$

Aby otrzymać analogiczne wzory estymatorów parametrów α_{12} i β_{10} , należy podzielić zbiór punktów ω na dwa podzbiory w ten sposób, że do pierwszego podzbioru zalicza się punkty, których rzędna Y jest nie większa od \bar{y} , a do drugiego podzbioru – pozostałe punkty.

Wprowadzimy dalsze oznaczenia

$$(11) \quad \begin{aligned} Y_{(1)} &= Y \mid Y \leq \bar{y}, \\ X_{\langle 1 \rangle} &= X \mid Y \leq \bar{y}, \\ Y_{(2)} &= Y \mid Y > \bar{y}, \\ X_{\langle 2 \rangle} &= X \mid Y > \bar{y}. \end{aligned}$$

Oto definicje następnych symboli:

$$\bar{y}_{(1)} = \frac{1}{n-m} \sum y_{(1)}, \quad \bar{y}_{(2)} = \frac{1}{m} \sum y_{(2)},$$

$$\bar{x}_{\langle 1 \rangle} = \frac{1}{n-m} \sum x_{\langle 1 \rangle}, \quad \bar{x}_{\langle 2 \rangle} = \frac{1}{m} \sum x_{\langle 2 \rangle}.$$

Litera m oznacza liczbę punktów, które w wyniku podziału zbioru ω na dwa podzbiory trafiły do drugiego podzbioru. Oczywiście m jest zmienną losową, która może przybierać

wartości $1, 2, \dots, n-1$. Przez zamianę liter we wzorach (5) - (10) otrzymuje się wzory na parametry regresji X względem Y :

$$(12) \quad a_{12} = \frac{\bar{x}_{(2)} - \bar{x}}{\bar{y}_{(2)} - \bar{y}},$$

$$(13) \quad a_{12} = \frac{\bar{x} - \bar{x}_{(1)}}{\bar{y} - \bar{y}_{(1)}},$$

$$(14) \quad a_{12} = \frac{\bar{x}_{(2)} - \bar{x}_{(1)}}{\bar{y}_{(2)} - \bar{y}_{(1)}},$$

$$(15) \quad b_{10} = \bar{x}_{(2)} - a_{12} \bar{y}_{(2)},$$

$$(16) \quad b_{10} = \bar{x}_{(1)} - a_{12} \bar{y}_{(1)},$$

$$(17) \quad b_{10} = \bar{x} - a_{12} \bar{y}.$$

Jak wynika z określenia tych parametrów regresji, cała umiejętność wyznaczania położenia linii regresji metodą punktową polega na przeprowadzeniu linii prostej przez dwa punkty. Gdy chcemy wyznaczyć położenie linii regresji Y względem X , prowadzimy linię przez dowolne dwa spośród trzech punktów $(\bar{x}_{(1)}, \bar{y}_{(1)})$, $(\bar{x}_{(2)}, \bar{y}_{(2)})$, (\bar{x}, \bar{y}) . Podobnie gdy chcemy określić położenie linii regresji X względem Y , prowadzimy linię przez dowolne dwa spośród trzech punktów $(\bar{y}_{(1)}, \bar{x}_{(1)})$, $(\bar{y}_{(2)}, \bar{x}_{(2)})$, (\bar{y}, \bar{x}) .

6.7.4. Technika rachunkowa związana z obliczaniem parametrów regresji

Korzystanie z wzorów podanych w punkcie 6.7.3 jest bardzo łatwe. Zilustrujemy to na przykładzie. W przykładzie tym, obejmującym materiał liczbowy pochodzący z okresu dwóch lat (24 pozycje – mała próbka), parametry regresji zostały obliczone dwiema metodami, tzn. metodą najmniejszych kwadratów i metodą punktową. Umożliwia to przedstawienie w sposób poglądowy korzyści, jakie daje metoda punktowa. Porównując tablice obliczeniowe obu metod można się przekonać, że metoda punktowa jest prostsza od metody klasycznej nie tylko pod względem pojęciowym, lecz również i pod względem techniki rachunkowej.

PRZYKŁAD. Tablica 1 na str. 230 zawiera liczby przejechanych kilometrów i zużytych kWh prądu w Miejskich Zakładach Komunikacyjnych we Wrocławiu (dane miesięczne). W oparciu o dane liczbowe tej tablicy chcemy obliczyć parametry regresji metodą najmniejszych kwadratów i metodą punktową. Wykonanie postawionego zadania rozpoznajemy od zaokrąglenia liczb do dziesiątek tysięcy km i dziesiątek tysięcy kWh.

Tablica 2 pokazuje przebieg obliczenia parametrów regresji za pomocą metody najmniejszych kwadratów. Mamy

$$\bar{x} = 131, \quad \bar{y} = 99,5, \quad u = 130, \quad w = 100,$$

$$\Delta_u = \bar{x} - u = 1, \quad \Delta_w = \bar{y} - w = -0,5.$$

Tablica 1

Nr	x km	y kWh	Nr	x km	y kWh
1	1 162 697	885 999	13	1 327 516	1 055 148
2	1 033 608	803 399	14	1 221 159	961 312
3	1 093 926	819 134	15	1 372 107	1 091 060
4	1 080 507	782 863	16	1 302 451	1 056 694
5	1 209 917	857 770	17	1 401 363	1 092 946
6	1 128 658	867 890	18	1 495 300	1 094 767
7	1 201 090	917 318	19	1 498 257	1 060 927
8	1 215 048	953 802	20	1 503 663	1 046 036
9	1 190 704	955 560	21	1 479 019	1 258 528
10	1 242 228	996 482	22	1 575 782	1 133 920
11	1 212 823	865 628	23	1 597 701	1 184 790
12	1 252 190	882 888	24	1 617 143	1 237 667

Tablica 2

Nr	x 10 tys. km	y 10 tys. kWh	x - u ⁽¹⁾		y - w		(x - u) ²	(y - w) ²	(x - u)(y - w)	
			+	-	+	-			+	-
1	116	89		14		11	196	121	154	
2	103	80		27		20	729	400	540	
3	109	82		21		18	441	324	378	
4	108	78		22		21	484	441	462	
5	121	86		9		14	81	196	126	
6	113	87		17		13	289	169	221	
7	120	92		10		8	100	64	80	
8	121	95		9		5	81	25	45	
9	119	96		11		4	121	16	44	
10	124	100		6	0	0	36	0	0	
11	121	87		9		13	81	160	117	
12	125	88		5		12	25	144	60	
13	133	106	3		6		9	36	18	
14	122	96		8		4	64	16	32	
15	137	109	7		9		49	81	63	
16	130	106	0	0	6		0	36	0	
17	140	109	10		9		100	81	90	
18	150	109	20		9		400	81	180	
19	150	106	20		6		400	36	120	
20	150	105	20		5		400	25	100	
21	148	126	18		26		324	676	468	
22	158	113	28		13		784	169	364	
23	160	118	30		18		900	324	540	
24	162	124	32		24		1024	576	768	
	3140.	2387	188	168	131	143	7118	4206	4970	

⁽¹⁾ Liczby u i w zostały obrane arbitralnie dla ułatwienia rachunków.

$$a_{21} = \frac{\sum_{i=1}^n (x_i - u)(y_i - w) - nA_u A_w}{\sum_{i=1}^n (x_i - u)^2 - nA_u^2} = 0,699,$$

$$a_{12} = \frac{\sum_{i=1}^n (x_i - u)(y_i - w) - nA_u A_w}{\sum_{i=1}^n (y_i - w)^2 - nA_w^2} = 1,180,$$

$$r^2 = 0,699 \cdot 1,180 = 0,83,$$

$$r = 0,91.$$

Przebieg obliczenia parametrów regresji metodą punktową przedstawia tablica 3.

Tablica 3

Nr	x	y	$x_{(2)}$	$y_{(2)}$	$y_{(2)}$	$x_{(2)}$
1	116	89				
2	103	80				
3	109	82				
4	108	78				
5	121	86				
6	113	87				
7	120	92				
8	121	95				
9	119	96				
10	124	100 ^u			100	124
11	121	87				
12	125	88				
13	133*	106 ^u	133	106	106	133
14	122	96				
15	137*	109 ^u	137	109	109	137
16	130	106 ^u			106	130
17	140*	109 ^u	140	109	109	140
18	150*	109 ^u	150	109	109	150
19	150*	106 ^u	150	106	106	150
20	150*	105 ^u	150	105	105	150
21	148*	126 ^u	148	126	126	148
22	158*	113 ^u	158	113	113	158
23	160*	118 ^u	160	118	118	160
24	162*	124 ^u	162	124	124	162
	3140	2387	1488	1125	1331	1742

$$\bar{x} = 130,8, \quad \bar{y} = 99,5, \quad \bar{x}_{(2)} = 148,8, \quad \bar{y}_{(2)} = 112,5, \quad \bar{y}_{(2)} = 110,9, \quad \bar{x}_{(2)} = 145,2,$$

$$a_{21} = 0,72, \quad a_{12} = 1,26,$$

$$r^2 = 0,72 \cdot 1,26 = 0,91,$$

$$r = 0,95.$$

Porównując tablicę 3 i tablicę 2 łatwo się przekonać, że rachunki związane z wyznaczeniem parametrów regresji metodą punktową są znacznie prostsze i mniej pracochłonne od rachunków, których wymaga metoda najmniejszych kwadratów.

Oto krótkie objaśnienie kolejnych operacji rachunkowych, które zostały wykonane dla wypełnienia tablicy 3 i znalezienia wartości parametrów regresji:

1. dodano liczby kolumn x i y i obliczono średnie arytmetyczne:

$$\bar{x} = \frac{3140}{24} = 130,8, \quad \bar{y} = \frac{2387}{24} = 99,5;$$

2. w kolumnie x wyszukano liczby większe od $\bar{x} = 131$ i opatrzone je znaczkiem *; wartości tych jest 10;

3. wyróżnione wartości x wpisano w kolumnie $x_{(2)}$, a odpowiadające im wartości y wpisano w kolumnie $y_{(2)}$;

4. w kolumnie y wyszukano liczby większe od $\bar{y} = 99,5$ i opatrzone je znaczkiem °; wartości tych jest 12;

5. wyróżnione wartości y wpisano w kolumnie $y_{(2)}$, a odpowiadające im wartości x – w kolumnie $x_{(2)}$;

6. obliczono średnie:

$$\bar{x}_{(2)} = \frac{1488}{10} = 148,8, \quad \bar{y}_{(2)} = \frac{1125}{10} = 112,5,$$

$$\bar{y}_{(2)} = \frac{1331}{12} = 110,9, \quad \bar{x}_{(2)} = \frac{1742}{12} = 145,2;$$

7. za pomocą wzorów (5) i (12), 6.7.3, obliczono a_{21} i a_{12} :

$$a_{21} = \frac{112,5 - 99,5}{148,8 - 130,8} = 0,722,$$

$$a_{12} = \frac{145,2 - 130,8}{110,9 - 99,5} = 1,263.$$

Znając estymatory parametrów regresji można oszacować wartość współczynnika korelacji. Mamy:

$$r^2 = a_{21} \cdot a_{12} = 0,722 \cdot 1,263 = 0,9119.$$

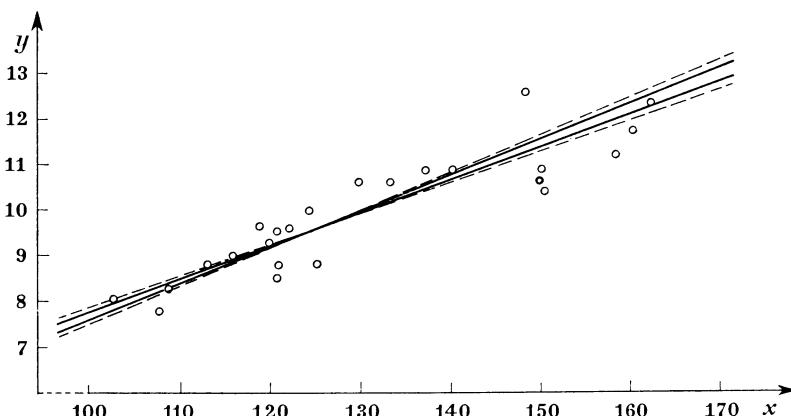
Stąd zaś

$$r = 0,95.$$

Wypełnienie tablicy obliczeniowej w metodzie punktowej nie wymaga wykonywania

żadnych operacji rachunkowych. Po prostu w odpowiednich kolumnach wpisujemy liczby wyróżnione znakami * i ^ i liczby skojarzone⁽¹⁾ z liczbami wyróżnionymi.

Odpada konieczność wykonywania odejmowania, podnoszenia do kwadratu i mnożenia liczb o różnych znakach, jak to miało miejsce w metodzie najmniejszych kwadratów. Jeżeli spełnione jest założenie o liniowym charakterze związku koreacyjnego między zmiennymi X i Y , to obie metody dają wyniki zbliżone, tak jak to widać w niniejszym przykładzie.



Rys. 1

Na rysunku 1 przedstawione są dwie linie regresji wyznaczone metodą najmniejszych kwadratów (linie przerywane) i dwie linie regresji otrzymane metodą punktową (linie ciągłe). Przebieg linii regresji otrzymanych obiema metodami różni się bardzo niewiele, pomimo że rozrzuć punktów jest dosyć znaczny.

6.7.5. Estymacja współczynnika zależności

W ustępie 4.7.3 został zdefiniowany współczynnik zależności (patrz 4.7.3, wzór (1) oraz 4.7.3, wzór (7)). Obecnie zilustrujemy na trzech przykładach liczbowych technikę obliczania tego współczynnika w oparciu o dane liczbowe z próbki.

PRZYKŁAD 1. Tablica 1 na str. 234 przedstawia dane ilustrujące rozkład urodzeń żywych w Polsce w roku 1964, przy czym zmienną losową X jest kolejne urodzenie dziecka u danej matki (tak więc X jest mierząłą, dyskretną zmienną losową). Zmienna Y jest wielkością niemierzalną i dotyczy podziału badanej populacji żywych urodzeń (w tysiącach) na miasto i wieś (patrz [38] str. 42).

(1) Słowo „skojarzone” jest tu użyte w sensie następującym: ponieważ mamy do czynienia ze zmieniącą się dwuwymiarową, przeto z każdą odciętą x_t jest skojarzona rzędna y_t — i na odwrót — z każdą rzędną y_t jest skojarzona odcięta x_t .

Tabela 1

	Kolejność urodzenia u matki								Razem
	1	2	3	4	5	6	7	8 i więcej	
Miasto	99,9	75,3	33,2	14,9	7,1	3,7	1,9	2,1	238,1
Wieś	91,5	83,3	61,3	38,4	22,1	12,4	7,2	8,5	324,7
Razem	191,4	158,6	94,5	53,3	29,2	16,1	9,1	10,6	562,8

W oparciu o dane w tablicy 1 znajdujemy macierz $\hat{\mathbf{P}}_1$ o elementach \hat{p}_{ij} oraz macierz $\hat{\mathbf{P}}_2$ o elementach $\hat{p}_i \hat{q}_j$ (porównaj (2), (3), z 4.7.3). Wyjaśniamy, że \hat{p}_{ij} jest estymatorem prawdopodobieństwa p_{ij} , natomiast $\hat{p}_i \hat{q}_j$ jest estymatorem $p_i q_j$. Estymatory te są to częstotliwości występowania zdarzeń opisanych liczebnosciami n_{ij} występującymi w odpowiednich klatkach tablicy 1. Mamy więc

$$(1) \quad \hat{p}_{ij} = \frac{n_{ij}}{\sum_{i=1}^r \sum_{j=1}^s n_{ij}},$$

$$(2) \quad \hat{p}_i = \frac{\sum_{j=1}^s n_{ij}}{\sum_{i=1}^r \sum_{j=1}^s n_{ij}},$$

$$(3) \quad \hat{q}_j = \frac{\sum_{i=1}^r n_{ij}}{\sum_{i=1}^r \sum_{j=1}^s n_{ij}}.$$

Elementy obu macierzy $\hat{\mathbf{P}}_1$ i $\hat{\mathbf{P}}_2$ zostały ujęte w jedną tablicę (tablica 2) przy czym liczba występująca w lewym górnym rogu klatki znajdującej się na przecięciu i -tego wiersza oraz j -tej kolumny — jest to częstotliwość \hat{p}_{ij} , natomiast liczba znajdująca się w prawym dolnym rogu tej klatki — jest to iloczyn częstotliwości $\hat{p}_i \hat{q}_j$.

Porównujemy obie liczby w każdej klatce i wyszukujemy klatki, dla których jest spełniona nierówność

$$(4) \quad \hat{p}_{ij} > \hat{p}_i \hat{q}_j.$$

Liczby w tych klatkach zostały w tekście wyróżnione kursywą.

Tabela 2

0,1775 0,1439	0,1338 0,1192	0,0590 0,0710	0,0265 0,0401	0,0126 0,0219	0,0066 0,0121	0,0034 0,0068	0,0037 0,0080	0,4231 0,4230
0,1626 0,1962	0,1480 0,1626	0,1089 0,0969	0,0682 0,0546	0,0393 0,0299	0,0220 0,0165	0,0128 0,0093	0,0151 0,0109	0,5769 0,5769
0,3401 0,3401	0,2818 0,2818	0,1679 0,1679	0,0947 0,0947	0,0519 0,0518	0,0286 0,0286	0,0162 0,0161	0,0188 0,0189	1,0000 0,9999

Obliczamy $\sum_{(i,j) \in M} \hat{p}_{ij}$, to znaczy sumę

0,1775
 0,1338
 0,1089
 0,0682
 0,0393
 0,0220
 0,0128
 0,0151
 $\overline{0,5776}$

oraz $\sum_{(i,j) \in M} \hat{p}_i \hat{q}_j$, czyli sumę

0,1439
 0,1192
 0,0969
 0,0546
 0,0299
 0,0165
 0,0093
 0,0109
 $\overline{0,4812}$

Wyjaśniamy, że literą M oznaczono zbiór klatek spełniających nierówność (4) (to znaczy klatek, w których liczby zostały wyróżnione kursywą).

Współczynnik zależności estymujemy za pomocą wzoru

$$(5) \quad \hat{d} = \sqrt{\frac{\sum_{(i,j) \in M} \hat{p}_{ij} - \sum_{(i,j) \in M} \hat{p}_i \hat{q}_j}{1 - \frac{1}{\min(r, s)}}}.$$

Ponieważ w naszym przykładzie $r=2$, a $s=8$, więc

$$(6) \quad \hat{d} = \sqrt{\frac{0,5776 - 0,4812}{1 - \frac{1}{2}}} = 0,44.$$

PRZYKŁAD 2. Zakłady produkujące odbiorniki radiowe i telewizyjne wyprodukowały próbna partię telewizorów w celu zbadania, czy najnowszy projekt rozwiązania plastycznego skrzynki zdobędzie uznanie w oczach potencjalnych nabywców i spowoduje wzrost popytu na odbiorniki telewizyjne pomimo wyższej ceny nowego modelu.

Tablica 3 przedstawia wyniki przeprowadzonego przez zakłady badania. Liczność badanej partii wynosiła 1000 telewizorów, z których 400 stanowiły odbiorniki starego typu, a 600 typu nowego.

Tablica 3

	B (nowe)	\bar{B} (stare)	Σ
A sprzedane	600	—	600
\bar{A} nie sprzedane	—	400	400
Σ	600	400	1000

Można przypuszczać, że nowoczesny wygląd telewizora wywarze wystarczająco silny wpływ na nabywców i skłoni ich w trakcie aktu zakupu odbiornika do przedkładania modelu nowego nad stary. Tablica 3 przypuszczenie to całkowicie potwierdza.

Można oczekwać, że miara zależności stochastycznej, jeżeli ma się zawierać w przedziale $\langle 0, 1 \rangle$, powinna w sytuacji takiej jak przedstawiona w niniejszym przykładzie przybrać wartość bliską 1. Zbadajmy, czemu się równa współczynnik zależności w rozważanym przypadku.

Częstości potrzebne do obliczenia współczynnika zależności przedstawia tablica 4.

Tablica 4

	<i>B</i>	\bar{B}	\sum
<i>A</i>	0,60 0,36	0,00 0,24	0,60 0,60
\bar{A}	0,00 0,24	0,40 0,16	0,40 0,40
\sum	0,60 0,60	0,40 0,40	1 1

Stąd

$$\hat{d} = \sqrt{\frac{1-0,52}{1-\frac{1}{2}}} = 0,98 .$$

Uwaga. Okrągle i wygodne dla celów obliczeniowych liczby, jakie występują w tablicy 3, zostały dobrane celowo, aby ułatwić proces obliczeniowy; sam przykład odpowiada jednak w całości pełni spotykanym często w praktyce zagadnieniom związanym z techniką przeprowadzania badania popytu za pomocą tzw. sondaży rynkowych.

PRZYKŁAD 3. Tablica 5 przedstawia dane statystyczne (w tys.) dotyczące wieku par małżeńskich w Polsce w 1964 r. (patrz [39]).

Tablica 5

Wiek mężczyzn	Wiek kobiet							Ogółem
	19 lat i mniej	20 - 24	25 - 29	30 - 34	35 - 39	40 - 49	50 lat i więcej	
19 lat i mniej	7 242	3 030	300	54	19	3	1	10 649
20 - 24	37 724	47 969	6 843	1 140	236	59	5	93 976
25 - 29	17 824	37 629	14 140	3 501	971	218	13	74 296
30 - 34	2 351	7 793	6 823	4 017	1 627	522	34	23 167
35 - 39	324	1 607	2 247	2 472	1 883	888	61	9 482
40 - 49	82	410	861	1 430	2 114	2 274	405	7 576
50 lat i więcej	18	85	181	465	1 055	3 948	5 777	11 529
Razem	65 565	98 523	31 395	13 079	7 905	7 912	6 296	230 675

Częstości potrzebne do obliczenia współczynnika zależności przedstawia tablica 6.

Tablica 6

0,0314 0,0131	0,0131 0,0197	0,0013 0,0063	0,0002 0,0026	0,0001 0,0016	0,0000 0,0016	0,0000 0,0013	0,0461 0,0462
0,1635 0,1158	0,2080 0,1740	0,0297 0,0554	0,0049 0,0231	0,0010 0,0140	0,0003 0,0140	0,0000 0,111	0,4074 0,4074
0,0773 0,0915	0,1631 0,1376	0,0613 0,0438	0,0152 0,0183	0,0042 0,0110	0,0009 0,0110	0,0001 0,0088	0,3221 0,3220
0,0102 0,0285	0,0338 0,0429	0,0296 0,0137	0,0174 0,0057	0,0071 0,0034	0,0023 0,0034	0,0001 0,0027	0,1005 0,1003
0,0014 0,0117	0,0070 0,0176	0,0097 0,0056	0,0107 0,0023	0,0082 0,0014	0,0038 0,0014	0,0003 0,0011	0,0411 0,0411
0,0004 0,0093	0,0018 0,0140	0,0037 0,0045	0,0062 0,0019	0,0092 0,0011	0,0099 0,0011	0,0018 0,0009	0,0330 0,0328
0,0001 0,0142	0,0004 0,0214	0,0008 0,0068	0,0020 0,0028	0,0046 0,0017	0,0171 0,0017	0,0250 0,0014	0,0500 0,0500
0,2843 0,2841	0,4272 0,4272	0,1361 0,1361	0,0566 0,0567	0,0344 0,0342	0,0343 0,0342	0,0273 0,0273	1,0002 0,9998

Współczynnik zależności wynosi

$$\hat{d} = \sqrt{\frac{0,7876 - 0,5277}{1 - \frac{7}{49}}} = 0,55.$$

Jak widać z przytoczonych trzech przykładów, współczynnik zależności może być stosowany we wszystkich możliwych sytuacjach: gdy obie zmienne są niemierzalne, gdy jedna jest mierzalna, a druga niemierzalna, gdy obie są mierzalne i dyskretne, gdy obie są mierzalne i ciągłe i, w końcu, gdy jedna jest dyskretna, a druga ciągła.

6.7.6. Wybrane wiadomości o estymacji parametrów regresji w przypadku zmiennych losowych wielowymiarowych

Rozważmy wektor losowy

$$\mathbf{X} = (X_1, X_2, \dots, X_k, Y),$$

o którym założymy, że ma rozkład normalny $N(\mathbf{m}, \boldsymbol{\Gamma})$. Parametry tego rozkładu są nieznane, a więc nieznane są również parametry $\alpha_0, \alpha_1, \dots, \alpha_k$ równania regresji

$$y = \alpha_1 x_1 + \alpha_2 x_2 + \dots + \alpha_k x_k + \alpha_0.$$

W celu oszacowania tych parametrów realizuje się serię doświadczeń; niech ich liczba równa się n . Doświadczenia te dostarczają $(k+1)$ -elementowych ciągów liczb x_1, x_2, \dots, x_k, y , będących współrzędnymi punktów eksperymentalnych $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n, \mathbf{y}$. Z tych współ-

rzędnych można utworzyć macierz $\hat{\mathbf{X}}$ i wektor $\hat{\mathbf{Y}}$:

$$(1) \quad \hat{\mathbf{X}} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1k} \\ x_{21} & x_{22} & \dots & x_{2k} \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix}, \quad \hat{\mathbf{Y}} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}.$$

O macierzy tej założymy, że jej rząd $R(\hat{\mathbf{X}})=k$. Założenie to należy rozumieć w ten sposób, że w każdym konkretnym badaniu po otrzymaniu za pośrednictwem eksperymentu danych liczbowych (1) sprawdza się, czy rząd $R(\hat{\mathbf{X}})=k$. Wszystkie dalsze rozważania dotyczą tylko tych przypadków, gdy sprawdzenie to daje wynik pozytywny, co zresztą w praktyce przeważnie ma miejsce.

W celu oszacowania parametrów $\alpha_0, \alpha_1, \dots, \alpha_k$ za pomocą metody najmniejszych kwadratów, minimizuje się wyrażenie

$$\sum_{i=1}^n (y_i - a_1 x_{i1} - a_2 x_{i2} - \dots - a_k x_{ik} - a_0)^2,$$

przy czym $a_j = \hat{\alpha}_j$ dla $j=0, 1, \dots, k$.

Znajdując pochodne cząstkowe i przyrównując je do zera otrzymujemy układ równań normalnych. Pierwsze z tych równań ma postać:

$$\sum_{i=1}^n y_i - a_1 \sum_{i=1}^n x_{i1} - a_2 \sum_{i=1}^n x_{i2} - \dots - a_k \sum_{i=1}^n x_{ik} - n a_0 = 0.$$

Wobec tego

$$a_0 = \bar{y} - \sum_{j=1}^k a_j \bar{x}_j.$$

Wszystkie dalsze działania związane z wyznaczaniem pozostałych niewiadomych a_1, a_2, \dots, a_k uproszczają się znacznie, jeżeli liczby $x_{i1}, x_{i2}, \dots, x_{ik}, y_i$ zastąpi się liczbami $x_{i1} - \bar{x}_1, x_{i2} - \bar{x}_2, \dots, x_{ik} - \bar{x}_k, y_i - \bar{y}$. Wtedy oczywiście $a_0 = 0$, a zredukowany układ równań normalnych, zawierający k równań o k niewiadomych, może być zapisany w postaci równania macierzowego

$$(2) \quad \hat{\Gamma}_k \hat{\mathbf{A}} = \hat{\Gamma}_0,$$

gdzie

$$\hat{\Gamma}_k = \begin{bmatrix} c_{11} & c_{12} & \dots & c_{1k} \\ c_{21} & c_{22} & \dots & c_{2k} \\ \dots & \dots & \dots & \dots \\ c_{k1} & c_{k2} & \dots & c_{kk} \end{bmatrix}, \quad c_{ij} = \hat{\gamma}_{ij} = \hat{C}(X_i, X_j),$$

$$\hat{\Gamma}_0 = \begin{bmatrix} c_{10} \\ c_{20} \\ \vdots \\ c_{k0} \end{bmatrix}, \quad c_{io} = \hat{y}_{io} = \hat{C}(X_i, Y),$$

$$\hat{\mathbf{A}} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_k \end{bmatrix}, \quad a_j = \hat{\alpha}_j.$$

Stąd rozwiązywanie równania (2) przybiera postać

$$(3) \quad \hat{\mathbf{A}} = \hat{\Gamma}_k^{-1} \hat{\Gamma}_0.$$

Porównując równanie (3) z równaniem (7) z 4.7.6 łatwo dostrzec, że (3) można otrzymać z (7) po prostu za pomocą formalnego zastosowania względem obu stron równania (7) operatora estymacji „ $\hat{\cdot}$ ”. Oznacza to, że $\hat{\mathbf{A}}$ można uważać za *wektorowy estymator A* i podobnie $\hat{\Gamma}_k$ jest *macierzowym estymatorem Γ_k* , a $\hat{\Gamma}_0$ jest *wektorowym estymatorem Γ_0* .

Widać stąd, że macierzowe estymatory są naturalnym uogólnieniem zwykłych estymatorów liczbowych.

O estymatorze macierzowym $\hat{\mathbf{A}}$ macierzy \mathbf{A} mówi się, że jest *zgodny, nieobciążony, najefektywniejszy (asymptotycznie najefektywniejszy)*, jeżeli elementy macierzy $\hat{\mathbf{A}}$ są zgodnymi, nieobciążonymi, najefektywniejszymi (asymptotycznie najefektywniejszymi) estymatorami odpowiednich elementów macierzy \mathbf{A} .

Zachodzi pytanie, co można powiedzieć o własnościach estymatora $\hat{\mathbf{A}}$ wektora \mathbf{A} parametrów regresji, jeżeli wiadomo, że $\hat{\mathbf{A}}$ otrzymano za pomocą metody najmniejszych kwadratów. Rzecz oczywista, własności te zależą od przyjętych założeń. Jak dotąd, korzystaliśmy z założenia, że wektor losowy \mathbf{X} ma rozkład normalny $N(\mathbf{m}, \mathbf{\Gamma})$. Przy tym założeniu można wykazać, że estymator $\hat{\mathbf{A}}$ jest zgodny, nieobciążony i najefektywniejszy. Niestety założenie to jest bardzo silne i poważnie ogranicza zakres stosowalności wielowymiarowej analizy regresji. Poza tym sprawdzenie, czy to założenie jest spełnione w każdym konkretnym przypadku, w praktyce jest bardzo kłopotliwe rachunkowo, a niekiedy w ogóle niemożliwe. Z tych właśnie względów czynione były liczne próby osłabienia założeń o rozkładzie wektora losowego \mathbf{X} . Próby te okazały się owocone i wykazały, że możliwe są w zasadzie dwa podejścia do problemu estymacji parametrów regresji liniowej metodą najmniejszych kwadratów.

W podejściu pierwszym, które nazywać będziemy *podejściem klasycznym* przyjmuje się następujące założenia:

1º Dana jest macierz $\hat{\mathbf{X}}$. Kolumny tej macierzy nie są zmiennymi losowymi, lecz zmiennymi w zwykłym sensie, przybierającymi raz na zawsze ustalony, skończony zbiór wartości, liczący n elementów.

2º Rząd $R(\hat{\mathbf{X}})$ macierzy $\hat{\mathbf{X}}$ jest równy $k < n$.

3º $\mathbf{Y} = \hat{\mathbf{X}}\mathbf{A} + \mathbf{Z}$, przy czym \mathbf{Y} i \mathbf{Z} są wektorami losowymi o n składowych.

4° $E(\mathbf{Z}) = \mathbf{O}$.

5° $E(\mathbf{Z}'\mathbf{Z}) = \sigma \mathbf{I}$, gdzie σ jest pewną stałą.

Analizując te założenia zauważymy łatwo, że zamiast silnego warunku, że zmienna losowa \mathbf{X} ma mieć rozkład normalny, przyjmuje się, że $\mathbf{X} = \hat{\mathbf{X}}$, gdzie $\hat{\mathbf{X}}$ jest macierzą o ustalonych elementach i o rzędzie równym k , oraz że wektor wartości oczekiwanych składowych wektora błędu losowego \mathbf{Z} jest równy zeru, a macierz kowariancji tych składowych jest macierzą diagonalną o stałych elementach. Ponieważ wektor losowy \mathbf{X} o rozkładzie normalnym spełnia również założenia 3°, 4°, 5°, więc istota osłabienia założeń w podejściu klasycznym polega na tym, że zamiast warunku, że wektor losowy \mathbf{X} ma mieć rozkład normalny, wprowadza się znacznie łatwiejsze do spełnienia w praktyce warunki 1° i 2°. Warto tu podkreślić, że przyjęcie tych założeń nie tylko ułatwia kontrolę ich zgodności z realiami praktyki, ale, co jest szczególnie ważne, bardzo często czyni tę kontrolę w ogóle niepotrzebną, jeżeli tylko doświadczenia dostarczające danych eksperymentalnych (l) organizuje się celowo w taki sposób, aby założenia te były spełnione. Zilustrujemy to na przykładzie.

Bada się wyrażony w sekundach czas Y reakcji kierowcy na bodziec zewnętrzny w zależności od wyrażonego w godzinach czasu trwania pracy kierowcy (zmienna x_1) i wyrażonej w stopniach temperatury otoczenia (zmienna x_2). Wielka litera Y wskazuje, że czas reakcji kierowcy na bodziec zewnętrzny jest zmienną losową, natomiast małe litery x_1 , x_2 informują, że czas pracy kierowcy i temperatura otoczenia nie są zmiennymi losowymi, lecz zmiennymi w zwykłym sensie. Rzeczywiście, jeżeli eksperiment badawczy przeprowadza się w warunkach laboratoryjnych, to dla całej n -elementowej serii doświadczeń można ustalić arbitralnie macierz $\hat{\mathbf{X}}$ o wymiarach $(n \times 2)$, utworzoną z wektorów kolumnowych \mathbf{x}_1 , \mathbf{x}_2 , których składowe są wartościami nielosowych zmiennych x_1 , x_2 . I tak, jeżeli pierwsza zmienna przybiera wartości

$$x_{1,1} = 1, \quad x_{2,1} = 2, \quad \dots, \quad x_{17,1} = 17,$$

to oznacza to, że obserwuje się szybkość reakcji kierowcy po jednej godzinie pracy, po dwóch godzinach itd., wreszcie – po 17 godzinach pracy. Podobnie, jeżeli druga zmienna przybiera wartości

$$x_{1,2} = -40, \quad x_{2,2} = -35, \quad \dots, \quad x_{17,2} = 40,$$

to wskazuje to, że obserwacje czasu Y szybkości reakcji kierowcy obserwuje się w temperaturach -40° , -35° , ..., -5° , 0° , $+5^\circ$, ..., $+35^\circ$, $+40^\circ$. Widać stąd, że zmienne x_1 , x_2 są nielosowe, gdyż wartości, jakie te zmienne przybierają, są z góry ustalone i znane. Zmiennymi losowymi są natomiast zmienne Y i Z , przy czym

$$Y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + Z.$$

Nie ma zatem potrzeby badania zgodności założeń 1° i 2° podejścia klasycznego z realiami praktyki, gdyż doświadczenia przeprowadza się celowo w taki sposób, aby założenia te były spełnione. Weryfikacji podlegają tylko założenia 3°, 4° i 5°. Sprawdzanie założeń 4° i 5° można wykonać za pomocą metod opisanych w § 6.4, natomiast weryfikację założenia 3° można przeprowadzić przy użyciu metod przedstawionych w § 7.4.

Dowodzi się (patrz [17]), że przy podejściu klasycznym estymatory parametrów regresji otrzymane metodą najmniejszych kwadratów są zgodne, nieobciążone i najefektywniejsze, tzn. mają wszystkie pożądane własności. Niestety istnieją niekiedy sytuacje, gdy nie można zastosować klasycznego podejścia do problemu estymacji parametrów regresji. Wracając do naszego przykładu, tak właśnie przedstawia się sprawa, gdy zależność szybkości reakcji kierowcy od długości czasu pracy i temperatury otoczenia bada się nie w warunkach laboratoryjnych, tzn. sztucznych, lecz w warunkach naturalnych, przeprowadzając kontrolę szybkości reakcji u wybieranych losowo kierowców, spotykanych na drogach przez ekipy kontrolne złożone z pracowników służby zdrowia i organów policji. Teraz już nie można ustalić arbitralnie ani długości czasu pracy kierowców, ani temperatury otoczenia, gdyż są to zmienne losowe. Dla podkreślenia tego faktu oznaczmy je nie jak poprzednio symbolami x_1, x_2 , lecz symbolami X_1, X_2 . Jeżeli nie chcemy lub nie możemy przyjąć założenia, że wektor losowy $\mathbf{X} = (X_1, X_2, Y)$ ma rozkład normalny, to dając do zapewnienia jak najlepszych własności estymatorom parametrów regresji, musimy uciec się do tzw. *podejścia stochastycznego*, przy którym wszystkie składowe wektora losowego $\mathbf{X} = (X_1, X_2, \dots, X_k)$ są zmiennymi losowymi (przypominamy, że w ujęciu klasycznym zmienne x_1, x_2, \dots, x_k były nielosowe; przyjmowały one ustalone z góry n -elementowy zbiór wartości i dlatego mogły być interpretowane jako dane wektory $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k$ o n składowych; tylko zmienna Y była zmienną losową).

Stochastyczne podejście do zagadnienia estymacji parametrów regresji liniowej oparte jest na następujących założeniach:

1º Dana jest macierz

$$\mathbf{X} = \begin{bmatrix} X_{11} & X_{12} & \dots & X_{1k} \\ X_{21} & X_{22} & \dots & X_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \dots & X_{nk} \end{bmatrix} = (X_1, X_2, \dots, X_k).$$

Wszystkie elementy tej macierzy są zmiennymi losowymi, przy czym zmienne należące do pierwszej kolumny mają taki sam rozkład jak zmienna X_1 , zmienne należące do drugiej kolumny mają taki sam rozkład jak zmienna X_2 i – wreszcie – zmienne należące do k -tej kolumny mają taki sam rozkład jak zmienna X_k . Oznacza to, że przedstawiona wyżej macierz i zapisany obok niej wektor losowy wyrażają tę samą treść, mogą być więc oznaczone tym samym symbolem \mathbf{X} .

Dane są również wektory

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = Y \quad \text{oraz} \quad \mathbf{Z} = \begin{bmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_n \end{bmatrix} = Z.$$

Wszystkie składowe wektora \mathbf{Y} są zmiennymi losowymi o rozkładzie takim jak rozkład zmiennej Y , natomiast wszystkie składowe wektora \mathbf{Z} są zmiennymi losowymi o rozkład-

dzię takim jak rozkład zmiennej Z . Tak więc wektory \mathbf{Y} i \mathbf{Z} są takim zapisem zmiennych losowych Y i Z , w którym widoczny jest fakt realizacji ciągu n niezależnych doświadczeń nad zmiennymi losowymi Y i Z . Losowe wyniki tych doświadczeń mogą być interpretowane jako zmienne losowe o rozkłach takich samych jak rozkłady zmiennych losowych Y i Z .

Zakłada się, że macierz \mathbf{X} oraz wektor \mathbf{Z} są stochastycznie niezależne, co należy rozumieć w ten sposób, że niezależne są pary zmiennych X_j, Z dla $j=1, 2, \dots, k$.

$$2^o E(\mathbf{Y}|\mathbf{X}) = \mathbf{XA} + E(\mathbf{Z}|\mathbf{X}).$$

$$3^o E(\mathbf{Z}|\mathbf{X}) = \mathbf{O}.$$

$$4^o E(\mathbf{Z}'\mathbf{Z}) = \sigma^2 \mathbf{I}.$$

5^o Macierz Γ_k jest nieosobliwa.

Przy tych założeniach można dowieść, że estymatory a_1, a_2, \dots, a_k parametrów regresji liniowej $\alpha_1, \alpha_2, \dots, \alpha_k$ są zgodne i nieobciążone. Te założenia nie wystarczają jednak do zapewnienia estymatorom maksymalnej efektywności. Dowodzi się, że estymatory są asymptotycznie najefektywniejsze, jeżeli wektor losowy \mathbf{Z} ma rozkład normalny. Założenie to jest już jednak bardzo silne i trudne do weryfikacji statystycznej i dlatego wynik ten ma ograniczoną przydatność praktyczną. Jak z tego wynika, najcenniejsze własności zapewnia estymatorom regresji podejście klasyczne. Podejście to, przy odpowiedniej organizacji prowadzenia doświadczeń i zbierania danych statystycznych niejako automatycznie zapewnia zgodność przyjętych założeń z realiami praktyki. Wszelkie inne podejścia opierają się już na znacznie silniejszych i trudniejszych do weryfikacji założeniach. Powoduje to, że ostatnio notuje się coraz liczniejsze próby badania i opisu zależności między zmiennymi losowymi nowymi sposobami, których wspólną charakterystyczną cechą jest to, że nie korzystają one z pojęcia regresji w rozumieniu funkcji analitycznej wyrażającej za pomocą równania zależność warunkowej nadziei matematycznej jednej zmiennej od zmiennych pozostałych. W próbach tych wykorzystuje się tzw. *zmienną losową dystansową* oraz *funkcje segmentowe*.

Zauważmy, że ponieważ estymatory parametrów regresji zależą od eksperymentalnych danych liczbowych z próbki (1), tzn. od wielkości losowych, więc same są zmiennymi losowymi. Aby dać temu wyraz, zapiszemy je przy użyciu wielkich liter: A_1, A_2, \dots, A_k . Wyjaśniamy, że liczby a_1, a_2, \dots, a_k są realizacjami tych zmiennych losowych. Na oznaczenie wektora losowego o składowych A_1, A_2, \dots, A_k użyjemy symbolu \mathbf{A}^* dla odróżnienia od wektora $\hat{\mathbf{A}}$ o składowych a_1, a_2, \dots, a_k .

Zajmiemy się obecnie wartością oczekiwana i wariancją wektora losowego \mathbf{A}^* . Jak wiemy, jeżeli przy estymacji \mathbf{A} stosuje się podejście klasyczne lub stochastyczne, to

$$E(\mathbf{A}^*) = \mathbf{A}.$$

W związku z tym wariancję $V(\mathbf{A}^*)$ można zapisać następująco:

$$V(\mathbf{A}^*) = E[(\mathbf{A}^* - \mathbf{A})(\mathbf{A}^* - \mathbf{A})^T].$$

Oznaczmy symbolem \mathbf{B} macierz $(\hat{\mathbf{X}}^T \hat{\mathbf{X}})^{-1} \hat{\mathbf{X}}^T$. Wobec tego

$$\mathbf{BY} = \mathbf{B}(\hat{\mathbf{X}}\mathbf{A} + \mathbf{Z}) = \mathbf{A} + \mathbf{BZ}.$$

Z drugiej strony jednak, jeżeli przy estymacji parametrów regresji stosuje się podejście klasyczne, to (por. (3)):

$$\mathbf{BY} = (\hat{\mathbf{X}}^T \hat{\mathbf{X}})^{-1} \hat{\mathbf{X}}^T \mathbf{Y} = \mathbf{A}^*.$$

Stąd

$$\mathbf{A}^* - \mathbf{A} = \mathbf{BZ}.$$

Tak więc

$$V(\mathbf{A}^*) = E[(\mathbf{A}^* - \mathbf{A})(\mathbf{A}^* - \mathbf{A})^T] = E(\mathbf{BZZ}^T \mathbf{B}^T) = \mathbf{BB}^T E(\mathbf{ZZ}^T).$$

Ale ponieważ

$$\mathbf{BB}' = (\hat{\mathbf{X}}^T \hat{\mathbf{X}})^{-1} \hat{\mathbf{X}}^T \hat{\mathbf{X}} (\hat{\mathbf{X}}^T \hat{\mathbf{X}})^{-1} = (\hat{\mathbf{X}}^T \hat{\mathbf{X}})^{-1},$$

więc

$$V(\mathbf{A}^*) = E(\mathbf{ZZ}^T) (\hat{\mathbf{X}}^T \hat{\mathbf{X}})^{-1} = V(Z) (\hat{\mathbf{X}}^T \hat{\mathbf{X}})^{-1} = \sigma^2 (\hat{\mathbf{X}}^T \hat{\mathbf{X}})^{-1},$$

gdzie zgodnie ze wzorem (14), 4.7.6,

$$V(Z) = V(Y) - \sum_{j=1}^k \alpha_j C(X_j, Y).$$

Można wykazać, że nieobciążonym estymatorem wariancji $V(Z)$ jest

$$(4) \quad \hat{V}(Z) = \frac{1}{n-k} \left(\sum_{i=1}^n y_i^2 - \sum_{j=1}^k \hat{\alpha}_j \sum_{i=1}^n x_{ij} y_i \right) = \hat{\sigma}^2 = s^2.$$

Nieobciążonym estymatorem $V(\mathbf{A}^*)$ jest więc

$$(5) \quad V(\mathbf{A}^*) = \sigma^2 (\hat{\mathbf{X}}^T \hat{\mathbf{X}})^{-1}.$$

Jeżeli oznaczymy symbolem g_i^2 element leżący na przecięciu i -tego wiersza oraz i -tej kolumny w macierzy $(\hat{\mathbf{X}}^T \hat{\mathbf{X}})^{-1}$, to

$$(6) \quad V(A_i) = \sigma^2 g_i^2.$$

Przypominamy, że zmienna losowa A_i jest to zależny od danych z próbki estymator parametru regresji α_i .

6.7.7. Technika obliczeniowa w przypadku regresji wielowymiarowej

Łatwo odgadnąć, że obliczenia związane z wyznaczeniem parametrów regresji liniowej stają się coraz bardziej kłopotliwe w miarę zwiększenia liczby n doświadczeń generujących punkty eksperymentalne oraz liczby k określającej ilość zmiennych występujących w roli argumentów w równaniu regresji. Zademonstrujemy na konkretnym przykładzie liczbowym przebieg tych obliczeń, informując jednocześnie czytelnika, że dziś, gdy mamy do dyspozycji maszyny elektroniczne, pracochłonność tych obliczeń nie jest trudną do pokonania przeszkodą na drodze do szerokiego rozpowszechnienia i wykorzystania niezwykle efektywnych metod analizy regresji.

Tablica 1 przedstawia wyrażone w odpowiednich jednostkach miary dane liczbowe dotyczące kwartalnych obrotów w sklepach odzieżowych i tekstylnych (zmienna Y) w zależności od udziału procentowego w ogólnej masie towarowej artykułów wykonanych z tkanin laminatowych i ortalionu (zmienna X_1), udziału procentowego artykułów wykonanych z elany i terylenu (zmienna X_2) oraz udziału procentowego artykułów wykonanych z modylu i tkanin podobnych (zmienna X_3).

Tablica 1

X_1	X_2	X_3	Y
10	9	4	2
11	7	4	3
10	10	4	4
11	8	5	6
15	7	5	10
12	15	4	12
16	20	8	20
16	15	9	25

W oparciu o dane tablicy 1 należy znaleźć parametry równania regresji

$$\hat{y} = a_1 x_1 + a_2 x_2 + a_3 x_3 + a_0.$$

Jeżeli równanie to przedstawimy w postaci

$$\hat{y} = a_1 x_1 + a_2 x_2 + a_3 x_3 + a_0 x_0,$$

gdzie x_0 jest zmienną, która może przybierać tylko wartości równe 1, to wektor $\hat{\mathbf{A}}$ o składowych a_1, a_2, a_3, a_0 może być wyznaczony z układu równań normalnych, który w postaci macierzowej ma postać następującą:

$$(1) \quad \hat{\mathbf{X}}^T \hat{\mathbf{X}} \hat{\mathbf{A}} = \hat{\mathbf{X}}^T \mathbf{Y},$$

gdzie $\hat{\mathbf{X}}$ w naszym przykładzie ma postać:

$$\begin{bmatrix} 10 & 9 & 4 & 1 \\ 11 & 7 & 4 & 1 \\ 10 & 10 & 4 & 1 \\ 11 & 8 & 5 & 1 \\ 15 & 7 & 5 & 1 \\ 12 & 15 & 4 & 1 \\ 16 & 20 & 8 & 1 \\ 16 & 15 & 9 & 1 \end{bmatrix}$$

Wykonując odpowiednie działania arytmetyczne otrzymujemy

$$\hat{\mathbf{X}}^T \hat{\mathbf{X}} = \begin{bmatrix} 1323 & 1200 & 574 & 101 \\ 1200 & 1193 & 534 & 91 \\ 574 & 534 & 259 & 43 \\ 101 & 91 & 43 & 8 \end{bmatrix}.$$

Ponieważ

$$(2) \quad \hat{\mathbf{A}} = (\hat{\mathbf{X}}^T \hat{\mathbf{X}})^{-1} \hat{\mathbf{X}}^T \hat{\mathbf{Y}},$$

przeto należy znaleźć macierz odwrotną $(\hat{\mathbf{X}}^T \hat{\mathbf{X}})^{-1}$ i wektor $\hat{\mathbf{X}}^T \hat{\mathbf{Y}}$. Wykonując odpowiednie obliczenia otrzymujemy

$$(\hat{\mathbf{X}}^T \hat{\mathbf{X}})^{-1} = \begin{bmatrix} 0,07628 & -0,00091 & -0,08372 & -0,50278 \\ -0,00091 & 0,01169 & -0,01780 & -0,02581 \\ -0,08372 & 0,01780 & 0,15802 & 0,41013 \\ -0,50278 & -0,02581 & 0,41013 & 4,56178 \end{bmatrix}$$

oraz

$$\hat{\mathbf{X}}^T \hat{\mathbf{Y}} = \begin{bmatrix} 1173 \\ 1152 \\ 549 \\ 82 \end{bmatrix}.$$

Stąd

$$\hat{\mathbf{A}} = \begin{bmatrix} 1,23788 \\ 0,51083 \\ 1,67448 \\ -20,26678 \end{bmatrix}.$$

Równanie regresji po zaokrągleniu wartości parametrów do dwóch miejsc po przecinku przybierze więc postać

$$\hat{y} = 1,24x_1 + 0,51x_2 + 1,67x_3 - 20,27.$$

Miarą dopasowania płaszczyzny regresji do punktów eksperymentalnych jest *średni błąd resztkowy*

$$(3) \quad s = \left[\frac{1}{n-k-1} \sum_{i=1}^n (y_i - \sum_{j=0}^k a_j x_{ij})^2 \right]^{\frac{1}{2}},$$

przy czym $a_j = \hat{\alpha}_j$ oraz $x_{i0} = 1$ dla $i = 1, 2, \dots, n$.

Zgodnie ze wzorem (3) mamy również:

$$(4) \quad s = \left[\frac{1}{n-k-1} \left(\sum_{i=1}^n y_i^2 - \sum_{j=0}^k a_j \sum_{i=1}^n x_{ij} y_i \right) \right]^{\frac{1}{2}}.$$

Zwracamy uwagę czytelnika, że we wzorze (4), 6.7.6, liczba stopni swobody wynosi $n-k$, gdyż liczba zmiennych wynosiła k , natomiast we wzorach (3) i (4) liczba zmiennych wynosi $k+1$ (doszła bowiem „zmienna” X_0), a stąd liczba stopni swobody w tych wzorach jest równa $n-k-1$.

Wykonując odpowiednie obliczenia znajdujemy, że w naszym przykładzie

$$s = 2,7002 .$$

Ze wzoru (4) wynika, że błąd resztkowy gwałtownie wzrasta, gdy liczba zmiennych k staje się bliska liczbie n punktów eksperymentalnych. Można z tego wyprowadzić praktyczną wskazówkę, że liczba n danych eksperymentalnych, za pomocą których estymuje się parametry funkcji regresji, musi być tak duża, aby wyrażenie

$$\frac{n-k}{n} = 1 - \frac{k}{n}$$

było bliskie jedności.

Zwróćmy uwagę, że obliczone wartości parametrów regresji

$$a_0 = -20,26678 , \quad a_1 = 1,23788 , \quad a_2 = 0,51083 , \quad a_3 = 1,67448$$

nie są wartościami nieznanych parametrów $\alpha_0, \alpha_1, \alpha_2, \alpha_3$, lecz jedynie realizacjami zmiennych losowych A_0, A_1, A_2, A_3 . Znajdziemy estymatory błędów standardowych tych zmiennych. Zauważmy w tym celu, że zgodnie ze wzorem (6) z 6.7.6 potrzebne nam są wartości elementów leżących na głównej przekątnej macierzy $(\hat{\mathbf{X}}^T \hat{\mathbf{X}})^{-1}$:

$$g_1^2 = 0,07628 , \quad g_1 = 0,2764 ,$$

$$g_2^2 = 0,01169 , \quad g_2 = 0,1082 ,$$

$$g_3^2 = 0,15802 , \quad g_3 = 0,3948 ,$$

$$g_0^2 = 4,56178 , \quad g_0 = 2,1331 .$$

Zakładając, że elementy macierzy $\hat{\mathbf{X}}$ są raz na zawsze ustalone⁽¹⁾, estymatory błędów parametrów regresji otrzymamy ze wzoru

$$s(a_i) = sg_i ,$$

gdzie

$$s(a_i) = \sqrt{V(a_i)} .$$

Tak więc, wykonując mnożenie i zaokrąglając wyniki do dwóch miejsc po przecinku, otrzymamy

⁽¹⁾ Jest to bardzo mocne założenie. Jeżeli prawdziwość tego założenia nie jest każdorazowo zweryfikowana w praktyce, to stosowanie analizy regresji jest niedopuszczalne.

$$s(a_1) = 0,75 ,$$

$$s(a_2) = 0,29 ,$$

$$s(a_3) = 1,07 ,$$

$$s(a_0) = 5,76 .$$

Zgodnie z konwencją wprowadzoną w 6.7.2, oznaczmy zmienną zależną w empiryczny równaniu regresji symbolem \hat{y} . Tak więc

$$\hat{y} = a_1 x_1 + a_2 x_2 + a_3 x_3 + a_0 .$$

Podstawiając w miejsce zmiennych x_1, x_2, x_3 składowe wektora \mathbf{x}_i , tj. x_{i1}, x_{i2}, x_{i3} , otrzymujemy

$$\hat{y}_i = a_1 x_{i1} + a_2 x_{i2} + a_3 x_{i3} + a_0 .$$

Współczynnik korelacji między wartościami y_i oraz \hat{y}_i nazywać będziemy *współczynnikiem korelacji wielowymiarowej* (inaczej *wielokrotniej* lub *wielorakiej*) i oznaczać symbolem $r(y, \hat{y})$ lub prościej r_0 . Parametr ten jest estymatorem parametru $\rho(Y, U)$ (patrz 4.7.6, określenie 5). Wykorzystując dane liczbowe naszego przykładu i wykonując stosowne obliczenia⁽¹⁾ znajdujemy, że

$$r_0 = r(y, \hat{y}) = 0,97 .$$

Przypuśćmy, że zostały wyznaczone parametry dwóch następujących równań regresji:

$$\hat{y} = b_2 x_2 + b_3 x_3 + b_0$$

oraz

$$\hat{x}_1 = b'_2 x_2 + b'_3 x_3 + b'_0 .$$

Niech

$$w_{i1} = y_i - b_2 x_{i2} - b_3 x_{i3} - b_0$$

oraz

$$w'_{i1} = x_{i1} - b'_2 x_{i2} - b'_3 x_{i3} - b'_0 ,$$

gdzie $i = 1, 2, \dots, n$.

Współczynnik korelacji między liczbami w_{i1} oraz w'_{i1} nazywa się *współczynnikiem korelacji cząstkowej* (patrz 4.7.6, określenie 6). Współczynnik ten będziemy oznaczać symbolem $r(w_1, w'_1)$ lub krócej r_1 . Parametr ten jest estymatorem parametru $\rho(W_1, W'_1)$. W naszym przykładzie

$$r_1 = r(w_1, w'_1) = 0,64 .$$

Oczywiście można również obliczyć wartości parametrów $r_2 = r(w_2, w'_2)$ i $r_3 = r(w_3, w'_3)$. W naszym przykładzie wynoszą one

$$r_2 = r(w_2, w'_2) = 0,65$$

⁽¹⁾ Obliczenia te wykonuje się dziś przy użyciu elektronicznej maszyny cyfrowej, więc ich nie przytaczamy.

oraz

$$r_3 = r(w_3, w'_3) = 0,61 .$$

Dla porównania podamy wartości współczynników korelacji r_{10}, r_{20} i r_{30} między zmienną Y a zmiennymi X_1, X_2, X_3 . Mamy:

$$r_{10} = r(X_1, Y) = 0,89 ,$$

$$r_{20} = r(X_2, Y) = 0,79 ,$$

$$r_{30} = r(X_3, Y) = 0,92 .$$

Zauważmy, że suma współczynników korelacji cząstkowej, która w naszym przykładzie wynosi

$$0,64 + 0,65 + 0,61 = 1,90 ,$$

jest większa zarówno od współczynnika korelacji wielowymiarowej $r_0 = 0,97$, jak i od 1. Oznacza to, że współczynniki korelacji cząstkowej nie nadają się do mierzenia relatywnego wpływu, jaki na zmienną Y wywierają zmienne X_1, X_2, X_3 . Do spełnienia takiego zadania najlepiej nadają się indywidualne wskaźniki pojemności informacji, dostarczanej przez te zmienne. Wskaźniki te, które oznaczać będziemy symbolami $h_j, j = 1, 2, \dots, k$, obliczamy za pomocą wzoru

$$h_j = \frac{r_{j0}^2}{1 + \sum_{i \neq j} |r_{ij}|}$$

(por. 4.7.6, określenie 7). W naszym przykładzie wartości tych wskaźników wynoszą:

$$h_1 = 0,33 , \quad h_2 = 0,27 , \quad h_3 = 0,34 .$$

Suma indywidualnych wskaźników pojemności informacji dostarczonej przez zmienne X_1, X_2, X_3 o zmiennej Y jest równa integralnemu wskaźnikowi pojemności informacji dostarczonej przez wszystkie zmienne razem. Wskaźnik ten, który oznaczymy symbolem h_0 , wyniesie więc

$$h_0 = 0,94 .$$

Jak widzimy, wskaźnik ten ma wartość mniejszą od jedności. Różnica $1 - h_0$, wynosząca w naszym przykładzie 0,06, mierzy wpływ wywierany na zmienną Y przez nieobjęte badaniem (a więc nieznane) wszelkie zmienne inne niż X_1, X_2, X_3 , wśród których znajduje się także zmienna zwana przez nas umownie *błędem losowym*. Klasyczne metody regresji nie pozwalają określić liczbowo, jaki jest względny udział w różnicy $1 - h_0$ błędu losowego. W praktyce postępuje się w ten sposób, że jeżeli różnica ta jest, tak jak w naszym przykładzie, bliska零, to w całości przypisuje się ją działaniu okoliczności losowych. W przeciwnym wypadku podejrzewa się działanie określonych zmiennych (losowych lub nie) i prowadzi się ich poszukiwanie. Poszukiwanie te umożliwiają w rezultacie ustalenie pewnej listy zmiennych, co do których zachodzi podejrzenie, że mogą wywierać wpływ na zmienną Y . I tu właśnie pojawia się kapitalny problem analizy regresji: jak dokonać podziału zmiennych na takie, które na liście tej powinny być zachowane i takie,

które z listy tej powinny być wykreślone. W literaturze przedmiotu spotyka się różne sposoby i kryteria przeprowadzenia takiego podziału. Do niedawna powszechnie stosowanymi kryteriami był średni błąd resztkowy s , współczynnik korelacji wielowymiarowej r_0 lub współczynnik zbieżności φ . Wszystkie te kryteria są równoważne i dają identyczne wskazania. Ich wadą jest to, że nie dają możliwości dokonania selekcji zmiennych, które powinny być wprowadzone do równania regresji. Wszystkie te kryteria prowadzą zawsze do trywialnego wniosku, że postąpimy tym lepiej, im więcej zmiennych weźmiemy pod uwagę. Tak więc posługując się tymi kryteriami zawsze stajemy przed koniecznością zaakceptowania pełnej listy zmiennych bez przeprowadzenia na niej jakichkolwiek skreśleń. Oczywiście bezkrytyczna akceptacja zawsze pełnej listy zmiennych jest nie do przyjęcia, a to już chociażby z dwóch następujących względów:

1º po obliczeniu parametrów równania regresji może się przecież okazać, że parametry regresji stojące przy niektórych zmiennych są równe lub bardzo bliskie zeru; oznacza to, że zmienne takie nie wywierają na zmienną Y wystarczająco silnego wpływu i powinny być z listy usunięte;

2º jeżeli na liście znajdą się zmienne, wykazujące silne wzajemne skorelowanie, to wpływ to na silne powiększenie błędów $s(a_i)$ parametrów regresji, a tym samym zredukuje wartość poznawczą równania regresji.

Z tych i innych względów, wśród których wymienić jeszcze wypada znaczne koszty obliczeń związanych z wyznaczeniem parametrów regresji w równaniach o znacznej liczbie zmiennych, podjęte zostały ostatnio próby wykorzystania innych kryteriów przeprowadzenia selekcji zmiennych. Jednym z takich kryteriów może być integralny wskaźnik pojemności informacji h_0 . Pokażemy na naszym przykładzie, jak postępujemy stosując to kryterium. Teoretyczna idea jest bardzo prosta: interpretujemy parametr h_0 jako funkcję dyskretną zmiennej $t=1, 2, \dots, 2^k-1$ (patrz 4.7.6) pisząc $h_0^{(t)}$ i znajdujemy maksymalną wartość tej funkcji. Przypuśćmy, że to maksimum funkcja $h_0^{(t)}$ osiąga dla $t=t^*$. W takim razie optymalną kombinację zmiennych będzie ta, która w uporządkowanym leksykograficznie ciągu ma numer t^* . W praktyce sprawa komplikuje się, gdy liczba k zmiennych $X_j, j=1, 2, \dots, k$, jest większa od 10, gdyż wtedy liczba kombinacji jest tak znaczna, że ich zbadanie, nawet przy użyciu elektronicznej maszyny cyfrowej, jest sprawą kłopotliwą, jeżeli nie wręcz niemożliwą. Oczywiście sprawa wyglądałaby inaczej, gdyby znany był sposób wyznaczenia maksimum funkcji $h_0^{(t)}$ bez konieczności badania wszystkich kombinacji. Niestety, jak dotąd sposób taki nie jest znany.

W naszym przykładzie liczba k jest mała i wynosi 3. Mała jest dzięki temu również liczba wszystkich możliwych kombinacji, jakie można utworzyć z elementów zbioru I (definicję symbolu I podano w 4.7.6). Liczba ta wynosi 7. Podajemy poniżej wartości parametru $h_0^{(t)}$ odpowiadające wartościom zmiennej $t=1, 2, \dots, 7$:

$$h_0^{(1)} = 0,80 ,$$

$$h_0^{(2)} = 0,62 ,$$

$$h_0^{(3)} = 0,86 ,$$

$$h_0^{(4)} = 0,90 ,$$

$$h_0^{(5)} = 0,90 ,$$

$$h_0^{(6)} = 0,88 ,$$

$$h_0^{(7)} = 0,94 .$$

Jak widać, maksimum parametru $h_0^{(t)}$ wynosi 0,94. Odpowiada mu kombinacja, której numer w uporządkowaniu leksykograficznym wynosi 7. Kombinacja ta ma postać (X_1, X_2, X_3) . Jest to kombinacja optymalna w sensie wyżej zdefiniowanym. Warto tu podkreślić, że w praktyce stosowania parametru h_0 jako kryterium wyboru optymalnej kombinacji zmiennych, które powinny wejść do równania regresji w roli zmiennych niezależnych, często zdarza się, że kryterium to preferuje kombinacje o małej liczbie elementów, a niekiedy wręcz kombinacje jednoelementowe. Sytuacja taka, jak w naszym przykładzie, że kombinacja optymalna objęła wszystkie zmienne X_1, X_2, \dots, X_k , jest raczej wyjątkowa. Tym się właśnie różni h_0 od współczynnika korelacji wielowymiarowej r_0 . Ten ostatni parametr, stosowany jako kryterium wyboru optymalnej kombinacji, zawsze preferuje kombinację zawierającą wszystkie zmienne X_1, X_2, \dots, X_k . Oto wartości współczynnika korelacji wielowymiarowej odpowiadające poszczególnym kombinacjom w rozpatrywanym przykładzie:

$$r_0^{(1)} = 0,90 , \quad r_0^{(2)} = 0,79 , \quad r_0^{(3)} = 0,93 , \quad r_0^{(4)} = 0,95 ,$$

$$r_0^{(5)} = 0,95 , \quad r_0^{(6)} = 0,95 , \quad r_0^{(7)} = 0,97 .$$

Maksymalna wartość współczynnika korelacji wielowymiarowej odpowiada, jak widzimy, kombinacji o numerze 7. Zupełnie analogiczny wynik dałoby zastosowanie jako kryterium wyboru optymalnej kombinacji innych stosowanych zwykle kryteriów, takich jak np. średni błąd resztkowy.

Pytania kontrolne i zadania

1. Co jest przedmiotem badań statystyki matematycznej?
2. Co to są populacje statystyczne? Podać przykłady.
3. Podać określenie cechy statystycznej.
4. Co to są cechy mierzalne i niemierzalne? Podać przykłady.
5. Co to jest szereg uporządkowany?
6. W roczniku statystycznym wyszukać przykłady szeregów rozdzielczych o cesze mierzalnej i niemierzalnej. Na tych przykładach omówić, co to jest przedział klasowy, środek przedziału klasowego, liczebność. Opisać sposób obliczania częstości, liczebności skumulowanej i częstości skumulowanej. Wybrane szeregi rozdzielcze przedstawić na wykresie. Omówić korzyści, jakie daje graficzna prezentacja szeregu rozdzielczego.
 7. Jak dzielimy badania statystyczne?
 8. Podać określenie populacji generalnej i populacji próbnej (próbki).
 9. Wyjaśnić, w jakich przypadkach jesteśmy zmuszeni korzystać z usług badania częściowego. Podać liczne przykłady.
10. Jakie warunki muszą być spełnione, aby próbka pobrana z populacji generalnej była reprezentacyjna?
11. Co to jest metoda reprezentacyjna?

12. Jak brzmi twierdzenie Gliwenki? Jaki jest sens praktyczny tego twierdzenia?
13. Podać określenie estymatora.
14. Podać określenie estymatora zgodnego, estymatora nieobciążonego i estymatora najefektywniejszego.
15. Co to jest efektywność estymatora?
16. Czy znasz jakiś estymator, który byłby jednocześnie estymatorem zgodnym, nieobciążonym i najefektywniejszym?
17. Na dowolnym przykładzie liczbowym wyjaśnić, jak oblicza się odchylenie przeciętne sposobem uproszczonym.
18. Estymujemy wartość przeciętną w populacji ograniczonej. Jaki schemat pobierania elementów do próbki powinniśmy zastosować: losowanie ze zwracaniem czy też losowanie bez zwarcania? Odpowiedź uzasadnić.
19. Podać określenie przedziału ufności.
20. W populacji generalnej o rozkładzie normalnym nie znamy wartości przeciętnej m . Wyznaczyć przedział ufności dla tego parametru, jeśli odchylenie standardowe populacji generalnej $\sigma = 9$, poziom ufności $\alpha = 0,90$, natomiast próbka liczy: a) 250 elementów, b) 16 elementów. Średnia arytmetyczna w próbce o 250 elementach i 16 elementach jest taka sama i wynosi $\bar{x} = 135$.
21. W populacji generalnej o rozkładzie normalnym $N(m, \sigma)$ parametry m i σ są nieznane. Z populacji pobrano próbkę liczącą 25 elementów i obliczono $\bar{x} = 640$ oraz $s = 16$. Wyznaczyć przedział ufności dla oszacowania σ , jeśli poziom ufności $\alpha = 0,95$. Przyjmując, że σ równa się górnej granicy otrzymanego przedziału ufności, wyznaczyć przedział ufności dla oszacowania \bar{x} , jeśli poziom ufności $\alpha = 0,90$.
22. Dla potrzeb przemysłu konfekcyjnego przeprowadza się w pewnym mieście zdjęcie antropometryczne ludności. Ile osób należy zmierzyć, aby można było twierdzić, że średni wzrost z próbki nie będzie się różnił od średniego wzrostu populacji generalnej co do bezwzględnej wartości więcej niż o 1 cm, jeśli odchylenie standardowe w populacji $\sigma = 8$ cm, zaś poziom ufności $\alpha = 0,95$?
23. Ile razy należy rzucić monetą, aby prawdopodobieństwo tego, że częstość wyrzucenia orła będzie się różniła od teoretycznego prawdopodobieństwa 0,5 co do bezwzględnej wartości mniej niż o 0,01, wynosiło 0,9? Porównaj wynik z wynikiem zadania 10 rozdział 15.
24. Podać wzory estymatorów parametrów regresji i współczynnika korelacji uzyskanych za pomocą metody najmniejszych kwadratów.
25. Zilustrować na przykładzie liczbowym sposób wyznaczania parametrów regresji i współczynnika korelacji metodą najmniejszych kwadratów.
26. W oparciu o dane liczbowe, które służyły do wyznaczania parametrów regresji metodą najmniejszych kwadratów, znaleźć te parametry za pomocą metody dwóch punktów; przedstawić na wykresie punkty empiryczne z próbki i linie regresji uzyskane obiema metodami; porównać położenie linii i skontrolować, czy nadaje się ona do opisu rozkładu punktów na wykresie korelacyjnym.
27. Podać przykłady ekonomicznych zastosowań linii regresji i współczynnika korelacji.
28. Dyrekcja sklepów spożywczych zleciła przeprowadzenie statystycznego badania zależności między wielkością odchylenia wagi towaru odważonego przez sprzedawcę od wagi żądanej przez klienta i wagą tego towaru. Badania te miały być przeprowadzone dla określonych grup towarów, takich jak mięso, tłuszcz, mąka, cukier, owoce i warzywa, słodycze; w każdej grupie towarów podlegała obserwacji praca wybranych losowo sprzedawców, przy czym rejestracja statystyczna obejmowała n aktów odważania towaru przez każdego ze sprzedawców. Sporządzić program organizacji eksperymentu i wskazać, jakich informacji dostarczyć może w tym przykładzie równanie linii regresji i współczynnik korelacji.
29. Wybrać z rocznika statystycznego przykłady tablic, w których
- obie cechy byłyby niemierzalne i przybierały tylko dwie wartości,
 - jedna cecha byłaby niemierzalna, a druga mierzalna dyskretna,
 - obie cechy byłyby niemierzalne dyskretne,
 - obie cechy byłyby ciągłe
- i obliczyć wartość współczynnika zależności.

30. Na papierze milimetrowym wykreślić okrąg promieniem $r = 100$ mm. Na okręgu opisać kwadrat, na którym rozpięto siatkę kwadratową. Bok oczka siatki wynosi 10 mm. W każdym oczku siatki wpisać liczbę odpowiadającą ilości kratek milimetrowych leżących w danym oczku siatki i takich, przez które przechodzi łuk okręgu. W oparciu o tak otrzymaną tablicę obliczyć współczynnik zależności między zmiennymi X i Y . Wyjaśnić, dlaczego współczynnik ten przybrał wartość bliską jedności. Jak zachowywałaby się wartość tego współczynnika, gdybyśmy zamiast papieru milimetrowego stosowali papier o kratkach coraz mniejszych (malejących do zera)? Jaka wartość w rozważanym przykładzie przyjąłby współczynnik korelacji? Dlaczego w tym przykładzie wskazania współczynnika zależności i współczynnika korelacji są tak diametralnie różne? Które z tych dwóch wskazań jest poprawne? Która z tych dwóch miar jest bardziej uniwersalna?

7.1. OKREŚLENIE HIPOTEZY STATYSTYCZNEJ WERYFIKACJA HIPOTEZY

OKREŚLENIE 1. Każdy sąd o populacji generalnej wydany bez przeprowadzania badania wyczerpującego nazywa się *hipotezą statystyczną*. Sądy wydawane o populacji mogą dotyczyć rozkładu populacji lub tylko niektórych parametrów rozkładu.

Sądy dotyczące parametrów populacji generalnej nazywają się *hipotezami parametrycznymi*, natomiast sądy dotyczące rozkładu populacji noszą nazwę *hipotez nieparametrycznych*.

Hipoteza statystyczna może być prawdziwa lub fałszywa. Jedynym sposobem rozstrzygnięcia, czy hipoteza jest prawdziwa czy fałszywa, jest poddanie zbadaniu wszystkich jednostek populacji generalnej, tzn. przeprowadzenie badania wyczerpującego. Często jest to jednak niemożliwe (patrz uwagi o stosowaniu metody reprezentacyjnej). W przypadkach, gdy nie możemy zbadać całej populacji generalnej, nie ma sposobu uzyskania absolutnej pewności, czy hipoteza jest prawdziwa, czy fałszywa. Wydawać by się mogło, że wtedy wypowiadanie hipotezy nie ma sensu. Otóż tak nie jest. Istnieją metody sprawdzania hipotez. Metody te nie dają absolutnej pewności, czy hipoteza jest słuszna, lecz pozwalają sprawdzać prawdziwość hipotezy z prawdopodobieństwem dowolnie bliskim jedności. Sprawdzanie hipotez statystycznych nosi nazwę *weryfikacji hipotez*.

OKREŚLENIE 2. *Testami statystycznymi* nazywamy sposoby weryfikacji hipotez statystycznych.

Testy statystyczne, służące do weryfikacji hipotez parametrycznych, nazywają się *testami parametrycznymi*, natomiast testy, za pomocą których weryfikuje się hipotezy nieparametryczne, noszą nazwę *testów nieparametrycznych*, zwanych także *testami zgodności*.

Sprawdzanie hipotez statystycznych ma doniosłe znaczenie praktyczne. Rozpatrzmy kilka przykładów, potwierdzających tę tezę.

PRZYKŁAD 1. Interesuje nas zagadnienie, kto dłużej żyje — kobiety czy mężczyźni. Całą populacji ludzkiej zbadać nie możemy, gdyż jest ona, praktycznie biorąc, nieograniczona w czasie. Ale nawet ustalając pewien okres czasu, nie możemy również objąć obserwacją wszystkich ludzi zamieszkujących naszą planetę, gdyż i takie badanie byłoby zbyt trudne do przeprowadzenia. Przypuśćmy, że wypowiadana została hipoteza, że przeciętna długość życia kobiet w Polsce jest obecnie taka sama, jak długość życia mężczyzn. Jeżeli nie znamy całej populacji generalnej, to nie znamy również długości życia wszystkich kobiet i mężczyzn w Polsce. Wygłoszona hipoteza o przeciętnej długości życia mężczyzn i kobiet może być prawdziwa lub fałszywa. Aby zamiast tego truizmu otrzymać odpowiedź pozytyczną, należy hipotezę zweryfikować.

PRZYKŁAD 2. Wynaleziono nowy antybiotyk. Przypuśćmy, że nadano mu nazwę preparatu X . Wy nalazcy twierdzą, że ma być on szczególnie skuteczny przy zwalczaniu gruźlicy. Ich zdaniem stosowanie preparatu X znacznie zmniejsza śmiertelność wywołaną przez gruźlicę. Mamy tu do czynienia z hipotezą statystyczną. Porównuje się populację osób chorych na gruźlicę i nieleczonych preparatem X z populacją osób chorych na gruźlicę i leczonych preparatem X . Oczywiście populacje te nie są znane. Znane są jedynie populacje próbne, na które składają się chorzy na gruźlicę, objęci badaniem statystycznym. Hipoteza głosi, że procent osób zmarłych na gruźlicę w populacji pierwszej jest wyższy od procentu osób zmarłych na gruźlicę w populacji drugiej. Nie jesteśmy w stanie stwierdzić kategorycznie, czy hipoteza ta jest słusza, czy nie. Możemy ją najwyżej poddać weryfikacji statystycznej.

PRZYKŁAD 3. Przedsiębiorstwo budowlane otrzymało transport cegieł liczący 200 000 sztuk. Dostawca twierdzi, że cegły są dobre: udział braków w partii towaru nie przekracza 2%. Oczywiście to twierdzenie dostawcy jest hipotezą statystyczną. Aby zdobyć absolutną pewność, czy jest ona słusza, należy poddać badaniu 200 000 sztuk cegieł. Ponieważ badanie jakości cegły łączy się ze zniszczeniem sztuki badanej, oznaczałoby to zniszczenie całej partii. Przypuśćmy, że odbiorca zgadza się przyjąć cegły, jeśli udział braków w partii cegieł nie przekracza 5%. Odbiorca musi podjąć decyzję, czy towar przyjmie, czy postawi do dyspozycji dostawcy. Gdyby zapewnienie dostawcy było prawdziwe, to dostarczona partia cegieł zostałaby przez odbiorcę przyjęta bez żadnych zastrzeżeń. Zapewnienia te mogą być jednak fałszywe, nawet bez złej woli dostawcy. Aby podjąć decyzję, czy partię cegieł przyjąć, czy nie, odbiorca musi zweryfikować hipotezę, że udział braków w dostarczonej partii nie przekracza 5%.

Zajmiemy się obecnie wyjaśnieniem, jak przeprowadzić weryfikację hipotez statystycznych. Przypuśćmy, że mamy zweryfikować hipotezę parametryczną. Hipoteza ta głosi, że nieznany parametr populacji generalnej $Q = Q_0$. Dalsze rozważania staną się bardziej zrozumiałe, jeśli będą poparte przykładem. Trzymamy w ręku monetę jednozłotową. Interesuje nas prawdopodobieństwo wyrzucenia orła tą monetą. Po dokładnym obejrzeniu monety dochodzimy do wniosku, że jest ona wykonana starannie i przedstawia krążek metalowy o jednakowej grubości. Wobec tego formułujemy hipotezę, że prawdopodobieństwo wyrzucenia orła naszą monetą wynosi 0,5.

W tym przykładzie nieznanym parametrem populacji jest prawdopodobieństwo wyrzucenia orła. Hipotetyczną populację generalną stanowią wyniki nieskończonej liczby rzutów, które teoretycznie rzecz biorąc można by wykonać trzymaną przez nas monetą. Oznaczmy hipotezę zerową, że parametr populacji $Q = Q_0$, symbolem H_0 i wprowadźmy zapis skrócony:

$$(1) \quad H_0(Q = Q_0),$$

który czytamy „hipoteza H_0 , że parametr populacji $Q = Q_0$ ”.

Jeśli nieznane prawdopodobieństwo wyrzucenia orła oznaczmy symbolem Q , to w naszym przykładzie skrócony zapis ma postać następującą:

$$H_0(Q = 0,5).$$

Zapis ten stwierdza, iż została sformułowana hipoteza, że nieznany parametr Q populacji równa się 0,5.

W celu zweryfikowania hipotezy H_0 , pobieramy próbki z populacji generalnej i wyznaczamy wartość jakiegokolwiek estymatora parametru Q . Przypuśćmy, że tym estymatorem jest \hat{Q} . W takim razie obliczamy, czemu równa się wartość \hat{Q} w próbce. Oznaczmy tę wartość symbolem q .

W naszym przykładzie odpowiednie postępowanie ma przebieg następujący: wykonujemy n rzutów monetą i ustalamy, czemu równa się częstość \hat{Q} wyrzucenia orła w próbce. Częstość \hat{Q} jest estymatorem prawdopodobieństwa Q . Przypuśćmy, że liczba rzutów monetą $n=100$, zaś liczba wyrzuconych orłów wynosi 65. Wobec tego $q=0,65$.

Niech α będzie pewną dodatnią liczbą rzeczywistą, spełniającą nierówność $0 < \alpha < 1$. Obierzmy α w ten sposób, że zdarzenia, których prawdopodobieństwo wystąpienia jest nie większe od α , są uznane za zdarzenia mało prawdopodobne. Oznaczmy symbolem t taką liczbę zależną od α , która przy założeniu, że hipoteza H_0 jest prawdziwa, spełnia relację

$$(2) \quad P\left\{\left|\frac{\hat{Q}-Q_0}{\sigma_{\hat{Q}}}\right| > t\right\} = \alpha.$$

W takim razie jeśli zmienna losowa \hat{Q} przybrała wartość q spełniającą nierówność

$$(3) \quad |q - Q_0| > t\sigma_{\hat{Q}},$$

to mamy do wyboru dwie alternatywy:

1° uznać, że zaszło zdarzenie mało prawdopodobne, uczynione założenie jest słuszne, tzn. hipoteza $H_0(Q=0,5)$ jest prawdziwa;

2° uznać, że nie zaszło zdarzenie mało prawdopodobne, lecz założenie jest niesłuszne, tzn. hipoteza H_0 jest fałszywa.

Wybieramy alternatywę drugą i hipotezę odrzucamy w oparciu o zasadę praktycznej pewności, że zdarzenia mało prawdopodobne nie występują.

Liczba α nazywa się *poziomem istotności hipotezy H_0* (lub *współczynnikiem istotności*). Zbiór wartości zmiennej losowej \hat{Q} , powodujących odrzucenie hipotezy H_0 , nazywa się *zbiorem krytycznym hipotezy H_0* . Zbiór ten oznaczać będziemy literą Z .

Powróćmy do naszego przykładu. Niech $\alpha=0,01$. Zmienną losową \hat{Q} jest częstość wyrzucenia orła w próbce liczącej 100 rzutów monetą. Przyjmując, że hipoteza $H_0(Q=0,5)$ jest słuszna, możemy łatwo obliczyć odchylenie standardowe $\sigma_{\hat{Q}}$ zmiennej losowej \hat{Q} :

$$\sigma_{\hat{Q}} = \sqrt{\frac{0,5(1-0,5)}{100}} = 0,05.$$

Zmienna losowa \hat{Q} ma rozkład dwumianowy. Ponieważ próbka jest dostatecznie liczna, rozkład ten można zastąpić rozkładem normalnym. W takim razie

$$\begin{aligned} \alpha &= P\left\{\left|\frac{\hat{Q}-0,5}{0,05}\right| > t\right\} = P\left\{\frac{\hat{Q}-0,5}{0,05} < -t \text{ albo } \frac{\hat{Q}-0,5}{0,05} > t\right\} \approx \\ &\approx 1 - \frac{1}{\sqrt{2\pi}} \int_{-t}^{+t} e^{-v^2/2} dv = 1 - \frac{2}{\sqrt{2\pi}} \int_0^t e^{-v^2/2} dv = 1 - 2\Phi(t). \end{aligned}$$

Stąd

$$(4) \quad \Phi(t) = \frac{1-\alpha}{2}.$$

Ponieważ $\alpha = 0,01$, przeto

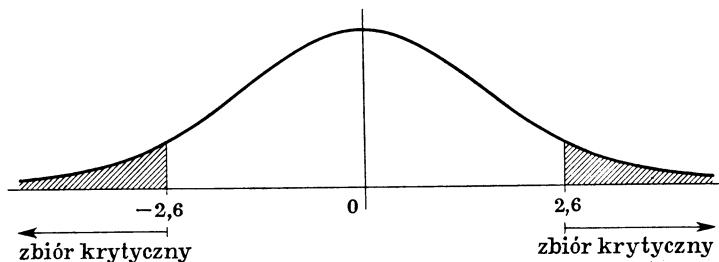
$$\Phi(t) = 0,4950.$$

W tablicach nie znajdujemy tej liczby. Bierzemy więc nieco większą 0,4953 i odczytujemy $t = 2,60$. Ponieważ zachodzi nierówność (patrz wzór (3)):

$$0,65 - 0,5 > 2,60 \cdot 0,05,$$

więc hipotezę H_0 odrzucamy⁽¹⁾.

Hipotezę H_0 odrzuciliśmy dlatego, że przyjmując ją musielibyśmy uznać, że zaszło zdarzenie mało prawdopodobne. Ponieważ zdarzenia takie występują bardzo rzadko, rzadziej niż przeciętnie raz na sto (założyliśmy bowiem, że $\alpha = 0,01$), przeto zamiast uznać, że właśnie w naszym wypadku natknęliśmy się na tak rzadkie zdarzenie, uczyniliśmy rozsądniej odrzucając hipotezę H_0 jako fałszywą. Oczywiście postępując w ten sposób możemy się mylić, lecz prawdopodobieństwo tej myłki jest nie większe niż $\alpha = 0,01$.



Rys. 1

Na rysunku 1 przedstawiona jest krzywa rozkładu normalnego. Początek układu współrzędnych umieszczono w punkcie $Q=0,5$. Jednostką skali na osi odciętych jest $\sigma\hat{Q}=0,05$.

Zakreskowana powierzchnia pod krzywą gęstości rozkładu normalnego przedstawia prawdopodobieństwo $\alpha=0,01$.

Hipotezę H_0 odrzucamy wtedy, gdy zmienna Q trafi do zbioru krytycznego, tzn. przybierze wartość większą od $0,5 + 2,6 \cdot 0,05 = 0,63$ lub mniejszą od $0,5 - 2,6 \cdot 0,05 = 0,37$. Liczby 0,63 i 0,37 oznaczać będziemy symbolami q_1^* i q_2^* i nazywać wartościąmi krytycznymi zmiennej \hat{Q} .

Zastanówmy się obecnie, jak należałoby postąpić, gdyby zmienna \hat{Q} nie przybrała wartości, upoważniającej do odrzucenia hipotezy H_0 . Należy podkreślić z naciskiem, że błędem byłoby przyjęcie hipotezy H_0 .

⁽¹⁾ W obliczeniach naszych wyznaczyliśmy $\sigma\hat{Q}$ zakładając, że $Q=0,5$. Jeśli $Q \neq 0,5$, to $\sigma\hat{Q} < 0,05$, gdyż wyrażenie

$$\sqrt{\frac{p(1-p)}{n}}$$

dla $0 < p < 1$ osiąga maksimum, gdy $p=0,5$. Wynika stąd, że nierówność $0,65 - 0,50 > 2,60 \sigma\hat{Q}$ jest słuszna dla każdej wartości $\sigma\hat{Q}$.

Gdy otrzymana z próbki wartość zmiennej losowej \hat{Q} nie trafi do zbioru krytycznego, wynajmniej nie oznacza to, że możemy przyjąć hipotezę H_0 . Oznacza to jedynie, że nie ma podstaw do odrzucenia tej hipotezy, co jednak nie jest wcale równoznaczne z uprawnieniem do jej przyjęcia.

Aby uzyskać podstawę do przyjęcia hipotezy H_0 , sformułujemy *hipotezę alternatywną* $H_1(Q \neq Q_0)$ i zweryfikujemy ją na poziomie istotności β . Założymy, że parametr populacji Q może przybierać jedynie dwie wartości: Q_0 i Q_1 . W takim razie hipoteza alternatywna $H_0(Q = Q_0)$ jest równoważna hipotezie $H_1(Q = Q_1)$.

Ustalamy następującą regułę postępowania:

1. gdy zmienna losowa \hat{Q} przybierze wartość $q > q^*$, odrzucamy hipotezę $H_0(Q = Q_0)$ na poziomie istotności α i przyjmujemy hipotezę alternatywną $H_1(Q = Q_1)$;
2. gdy zmienna losowa \hat{Q} przybierze wartość $q \leq q^*$, odrzucamy hipotezę $H_1(Q = Q_1)$ na poziomie istotności β i przyjmujemy hipotezę $H_0(Q = Q_0)$.

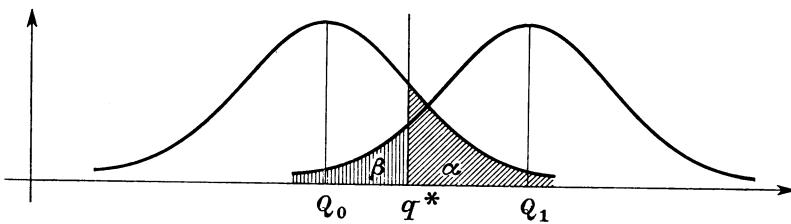
Na rysunku 2 przedstawione są dwa możliwe położenia parametru Q populacji generalnej. Zgodnie z przyjętym założeniem parametr Q może przybierać jedynie dwie wartości Q_0 i Q_1 . Gdy zmienna losowa \hat{Q} przybierze wartość $q > q^*$, to odrzucamy hipotezę H_0 . Mamy bowiem

$$P\{\hat{Q} > q^* | Q = Q_0\} = \alpha.$$

Prawdopodobieństwo to na rysunku 2 przedstawione jest za pomocą obszaru zakreślonego ukośnie. Gdy natomiast zmienna losowa przybierze wartość $q \leq q^*$, to odrzucamy hipotezę H_1 , gdyż

$$P\{\hat{Q} \leq q^* | Q = Q_1\} = \beta.$$

Prawdopodobieństwo β przedstawione jest na rysunku 2 za pomocą obszaru zakreślonego pionowo.



Rys. 2

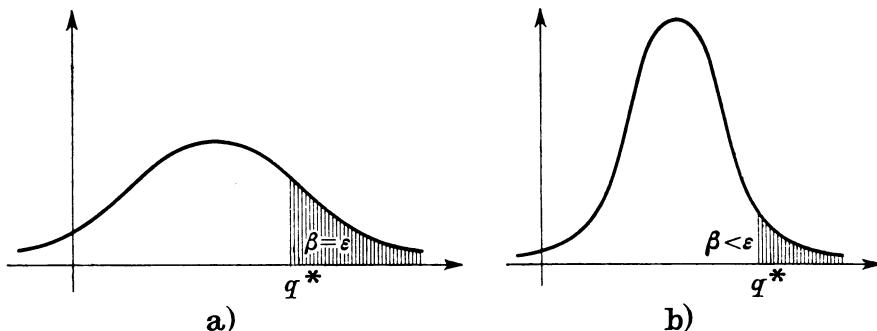
Prawdopodobieństwa α i β , na których poziomie sprawdza się hipotezę $H_0(Q = Q_0)$ i hipotezę alternatywną $H_1(Q = Q_1)$, mogą być obrane dowolnie, jeśli wielkość próbki n nie jest stała, lecz może się zmieniać. Łatwo to wyjaśnić. Przypuśćmy, że dla ustalonych wartości Q_0 , Q_1 i α chcemy sprawdzić, czy potrafimy dobrąć poziom istotności β w ten sposób, aby równał się on zadanej z góry liczbie ε , gdzie $0 < \varepsilon \leq 1$.

Przyjmijmy, że nieznanym parametrem populacji generalnej jest wartość przeciętna, tzn. $Q = m$. Parametr ten estymujemy za pomocą średniej arytmetycznej z próbki. Jak wiadomo,

odchylenie standardowe średniej arytmetycznej z próbki wyraża się wzorem

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}.$$

Widzimy więc, że im próbka jest większa, tym $\sigma_{\bar{x}}$ jest mniejsze.



Rys. 3

Na rysunku 3a przedstawiony jest rozkład średniej arytmetycznej z próbki liczącej n_1 elementów, a na rysunku 3b wyobrażony jest rozkład średniej arytmetycznej z próbki liczącej n_2 elementów. Ponieważ $n_2 > n_1$, przeto odchylenie standardowe w rozkładzie przedstawionym na rysunku 3b jest mniejsze niż odchylenie standardowe w rozkładzie przedstawionym na rysunku 3a i dlatego wykres drugiego rozkładu jest bardziej szpiczasty niż pierwszego. Na obu rysunkach, na osi odciętych, w tej samej odległości od początku układu umieszczony jest punkt krytyczny q^* . Punkt ten wyznacza pole pod krzywą gęstości rozkładu normalnego. Pole to na obu rysunkach jest zakreskowane i przedstawia poziom istotności β . Na rysunku 3b pole to jest mniejsze niż na rysunku 3a. Przyjmijmy, że pole na rysunku 3a wyobraża prawdopodobieństwo β równe ε . Czy można byłoby, postępując odpowiednio, tak zmienić rozkład na rysunku 3b, aby zakreskowane pole przedstawało również prawdopodobieństwo równe ε ? Oczywiście. Należały tylko zmniejszyć liczebność próbki n_2 . Jeśli bowiem n_2 będzie się równała n_1 , to pole pod krzywą gęstości na rysunku 3b będzie równe polu pod krzywą gęstości na rysunku 3a.

Widzimy więc, że manipulując odpowiednio wielkością próbki można zawsze doprowadzić do tego, że poziom istotności β będzie równy ε . Oznacza to, że jeśli liczebność próbki nie jest ustalona, to parametry Q_0, Q_1, α i β mogą być dobrane dowolnie.

Przy weryfikowaniu hipotezy, dotyczącej wartości przeciętnej m w populacji generalnej, liczebność próbki n , pozwalającą otrzymać zadane z góry wartości parametrów m_0, m_1, α, β , wyznaczamy z następującej relacji:

$$m_0 + t_0 \frac{\sigma}{\sqrt{n}} = m_1 - t_1 \frac{\sigma}{\sqrt{n}},$$

gdzie t_0 i t_1 są to liczby zależne od α i β . Stąd

$$n = \left[\frac{t_0 + t_1}{m_1 - m_0} \sigma \right]^2.$$

7.2. BŁĘDY PIERWSZEGO I DRUGIEGO RODZAJU

Sprawdzając hipotezę statystyczną możemy podjąć jedną z dwóch decyzji: hipotezę sprawdzaną uznać za prawdziwą i przyjąć ją lub hipotezę sprawdzaną uznać za fałszywą i odrzucić ją. Każda z tych decyzji może być słuszna lub niesłuszna. Wobec tego przy sprawdzaniu hipotez statystycznych mamy do czynienia z czterema sytuacjami, które przedstawimy za pomocą następującego schematu:

Sytuacja		H_0 prawdziwa	H_0 fałszywa
Decyzja			
H_0 przyjąć		decyzja słuszna	decyzja niesłuszna
H_0 odrzucić		decyzja niesłuszna	decyzja słuszna

Odrzucając hipotezę H_0 , każdorazowo narażeni jesteśmy na popełnienie błędu, polegającego na odrzuceniu hipotezy prawdziwej. Błąd ten nosi nazwę *błędu pierwszego rodzaju*. Przyjmując natomiast hipotezę H_0 każdorazowo narażeni jesteśmy na popełnienie błędu, polegającego na przyjęciu hipotezy fałszywej. Błąd ten nazywa się *błędem drugiego rodzaju*.

Hipotezę H_0 odrzucamy wtedy, gdy zmienna losowa \hat{Q} przybierze wartość większą od wartości krytycznej q^* . Prawdopodobieństwo tego zdarzenia przy założeniu, że $Q = Q_0$, wynosi α . Wynika stąd, że prawdopodobieństwo popełnienia błędu pierwszego rodzaju równa się α .

Hipotezę H_0 przyjmujemy wówczas, gdy odrzucamy hipotezę alternatywną H_1 . Prawdopodobieństwo odrzucenia hipotezy H_1 przy założeniu, że $Q = Q_1$, wynosi β . W takim razie prawdopodobieństwo przyjęcia hipotezy H_0 przy założeniu, że prawdziwa wartość parametru $Q = Q_1$, również wynosi β . Wobec tego prawdopodobieństwo popełnienia błędu drugiego rodzaju równa się β . Oczywiście powinniśmy dążyć do tego, aby prawdopodobieństwa popełnienia błędów pierwszego i drugiego rodzaju były jak najmniejsze. Im mniejsze jest bowiem α i β , tym większe jest prawdopodobieństwo, że nasza decyzja dotycząca przyjęcia lub odrzucenia hipotezy jest słuszna. Wyżej powiedzieliśmy, że α i β można obrać dowolnie, jeśli liczelność n próbki nie jest ustalona. Stąd nie wynika jednak, że α i β mogą się równać zeru, zgodnie bowiem z definicją poziomu istotności,

$$0 < \alpha < 1, \quad 0 < \beta < 1.$$

Im bliższe zero wartości chciemy nadać parametrom α i β , tym większa musi być liczelność próbki. Pobieranie próbki jest jednak związane z pewnymi trudnościami i kosztami. Koszty te wzrastają wraz ze zwiększeniem się liczelności próbki. W praktyce zatem zawsze dąży się do pobierania próbek o możliwie najmniejszej liczelnosci. Widzimy więc, że żądanie zmniejszenia poziomów istotności α i β jest sprzeczne z żądaniem zmniejszenia liczelnosci próbki. Sprzeczność ta wymaga rozstrzygnięcia kompromisowego. Sposób przeprowadzenia takiego rozstrzygnięcia, tzn. najważniejszego doboru liczb α , β i n , zależy od konkretnych warunków, z jakimi mamy do czynienia przy weryfikowaniu danej hipotezy statystycznej.

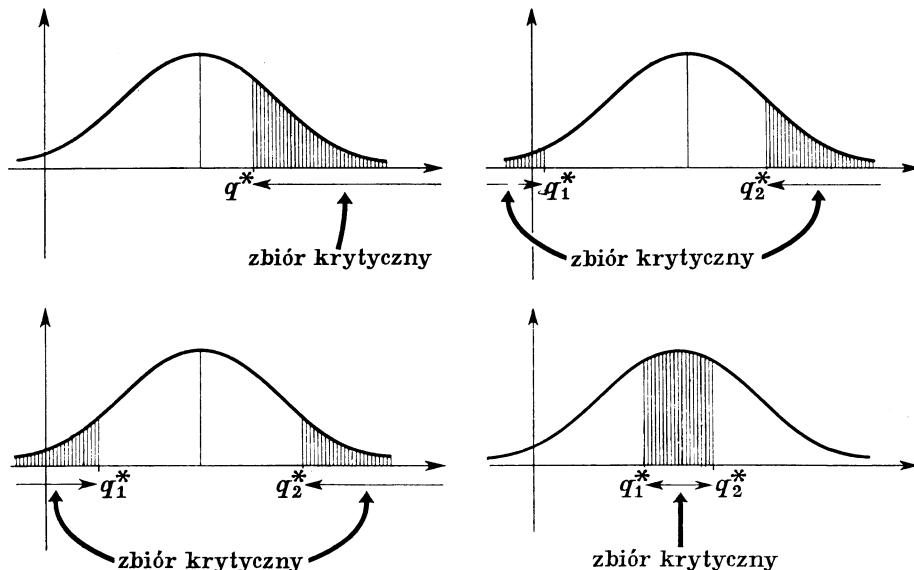
7.3. KRZYWA MOCY TESTU

Jak wiadomo, weryfikując hipotezę statystyczną wyznaczamy taki zbiór krytyczny Z , że jeśli zmienna losowa \hat{Q} przybierze wartość należącą do zbioru Z , to hipotezę H_0 odrzucamy, w przeciwnym przypadku hipotezę przyjmujemy.

Przy wyznaczaniu zbioru krytycznego Z , formalnie rzecz biorąc, jesteśmy związani tylko jednym warunkiem, który może być zapisany w postaci następującej:

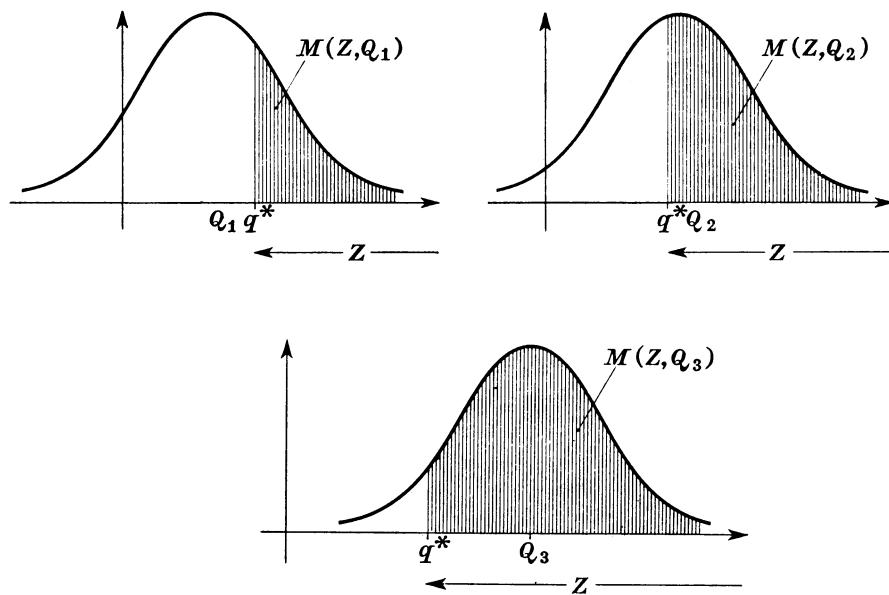
$$(1) \quad P\{\hat{Q} \in Z\} = \alpha.$$

Zapis ten czytamy tak: „prawdopodobieństwo tego, że zmienna losowa \hat{Q} przybierze wartość należącą do zbioru krytycznego Z , równa się α ”. Istnieje nieskończoność wiele sposobów takiego wyznaczania zbioru Z , aby była spełniona relacja (1). Na rysunku 1 przedstawione są cztery wykresy gęstości prawdopodobieństwa. Zakreskowana powierzchnia pod krzywą gęstości na każdym z tych wykresów przedstawia wartość poziomu istotności α . Pomimo tego, że we wszystkich czterech przypadkach poziom istotności jest taki sam (co wyraża się tym, że zakreskowane powierzchnie na wszystkich wykresach są jednakowe), zbiór krytyczny Z w każdym przypadku jest wyznaczony inaczej. Powstaje wobec tego pytanie, czym należy kierować się przy wyborze zbioru krytycznego. Odpowiedź jest następująca: zbiór krytyczny Z , jeśli to możliwe, należy wyznaczyć tak, aby przy ustalonym prawdopodobieństwie popełnienia błędu pierwszego rodzaju otrzymać najmniejszą wartość prawdopodobieństwa β popełnienia błędu drugiego rodzaju.

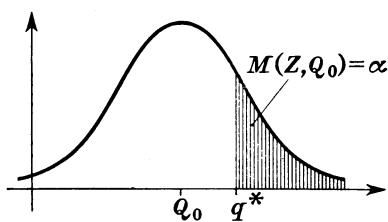


Rys. 1

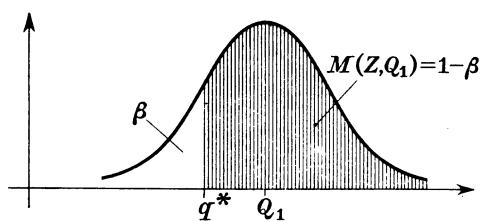
Oznaczmy symbolem $M(Z, Q)$ prawdopodobieństwo tego, że zmienna losowa \hat{Q} przybierze wartość należącą do zbioru Z przy założeniu, że nieznany parametr populacji generalnej równa się Q .



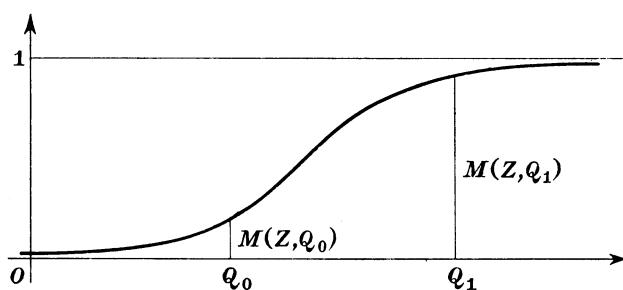
Rys. 2



Rys. 3



Rys. 4



Rys. 5

OKREŚLENIE 1. Prawdopodobieństwo $M(Z, Q)$ nazywa się *mocą testu*.

Oczywiście $M(Z, Q)$ przybiera różne wartości w zależności od Q , tzn. jest funkcją Q .

Rysunek 2 przedstawia 3 wykresy krzywej gęstości prawdopodobieństwa. Zakreskowana powierzchnia na tych wykresach jest geometrycznym obrazem mocy testu $M(Z, Q)$. Zbiór Z na każdym z tych wykresów jest taki sam.

Widzimy, że wielkość zakreskowanej powierzchni zależy od tego, jaką wartość ma parametr Q . Oczywiście (patrz rys. 3)

$$(1) \quad M(Z, Q_0) = \alpha$$

oraz (patrz rys. 4)

$$(2) \quad M(Z, Q_1) = 1 - \beta.$$

Przedstawiając na rysunku moc testu $M(Z, Q)$ nie za pomocą powierzchni, lecz za pomocą rzędnych, można sporządzić wykres *krzywej mocy testu* (rys. 5).

Test służący do weryfikowania hipotezy statystycznej powinien mieć tę własność, że jeśli $Q = Q_0$, to moc testu $M(Z, Q_0)$ jest mała, jeśli zaś $Q = Q_1$, to moc testu $M(Z, Q_1)$ jest duża.

Wyboru zbioru Z należy dokonać tak, aby ta własność testu była zapewniona.

7.4. WERYFIKACJA HIPOTEZ STATYSTYCZNYCH

7.4.1. Weryfikacja hipotezy parametrycznej o wartości przeciętnej w populacji generalnej

Przypuśćmy, że mamy dokonać oceny jakości partii zapałek, liczącej 100000 pudełek. Dostawca twierdzi, że w pudełku znajdują się średnio 54 zapałki.

Należy zweryfikować hipotezę $H_0(m=54)$. Jest to hipoteza parametryczna, gdyż nie dotyczy ona rozkładu populacji generalnej, lecz tylko jednego z parametrów tego rozkładu. Gdybyśmy wiedzieli, że rozkład cechy w populacji jest rozkładem normalnym – moglibyśmy przeprowadzić weryfikację hipotezy H_0 w oparciu o zbadanie małej próbki, gdyż jeśli populacja ma rozkład normalny (patrz 6.5.4), to rozkład średniej arytmetycznej z próbki o liczbie mniejszej niż 30 elementów jest rozkładem Studenta o $n-1$ stopniach swobody. Ponieważ jednak rozkład populacji nie jest znany, powinniśmy pobrać dużą próbkę. Za pobraniem dużej próbki przemawia również fakt, że nie znamy odchylenia standartowego σ w populacji generalnej. Proces badania próbki jest łatwy i mało kosztowny. Idzie przecież tylko o przeliczenie zapałek w każdym pudełku wylosowanym do próbki. Wzgłydy ekonomiczne nie stoją więc na przeszkodzie w zbadaniu dużej próbki.

Przypuśćmy, że liczbeność próbki ma stanowić 0,1% liczbeństwa populacji. W takim razie $n=100$. Średnia liczba zapałek w pudełku, obliczona na podstawie próbki, wynosi $\bar{x}=51,21$, a odchylenie standardowe w próbce równa się 2,45. Niech współczynnik istotności α , na poziomie którego weryfikujemy hipotezę H_0 , równa się 0,002. Wstawiając

$\alpha=0,002$ do wzoru (patrz § 7.1, wzór (4))

$$\Phi(t) = \frac{1-\alpha}{2}$$

mamy

$$\Phi(t) = 0,499.$$

W tablicach rozkładu normalnego odnajdujemy $t \approx 3$.

Ponieważ

$$(1) \quad P\{| \bar{X} - m | > t\sigma_{\bar{x}}\} = \alpha,$$

więc hipotezę $H_0(m=54)$ należy odrzucić wtedy, gdy zostanie spełniona nierówność

$$(2) \quad | \bar{X} - m | > t\sigma_{\bar{x}}.$$

W naszym zadaniu

$$\bar{x} = 51,21, \quad m = 54, \quad \sigma_{\bar{x}} \approx \frac{2,45}{\sqrt{100}} = 0,245, \quad t = 3.$$

Wstawiając te dane do wzoru (2) otrzymujemy

$$54 - 51,21 = 2,79 > 3 \cdot 0,245 = 0,735.$$

Wobec tego hipotezę $H_0(m=54)$ odrzucamy.

7.4.2. Weryfikacja hipotezy nieparametrycznej o postaci rozkładu cechy w populacji (test χ^2 , test λ Kołmogorowa-Smirnowa, test zgodności dla małej próbki, test serii)

W punkcie 7.4.1 omówiony został przykład weryfikowania hipotezy parametrycznej. Obecnie zajmiemy się sprawdzaniem hipotezy nieparametrycznej. Idea postępowania jest podobna: wypowiadamy sąd o rozkładzie cechy w populacji generalnej, a następnie pobieramy próbkę i w oparciu o rozkład cechy w próbce sąd ten poddajemy weryfikacji statystycznej. Weryfikacja ta polega na zbadaniu zgodności między hipotetycznym rozkładem w populacji generalnej a empirycznym rozkładem w próbce. Odpowiednio ustalona miara zgodności jest zmienną losową, która w zależności od wyników, otrzymanych w próbce, przybiera różne wartości z różnym prawdopodobieństwem. Rozkład tej zmiennej jest znany. Jeśli zgodność między rozkładem empirycznym z próbki a rozkładem hipotetycznym jest duża, nie ma podstaw do odrzucenia hipotezy. Jeśli natomiast jest mała, tak mała, że prawdopodobieństwo takiego zdarzenia równe jest poziomowi istotności α , to hipoteza zostaje odrzucona.

Przejdźmy do przykładów liczbowych.

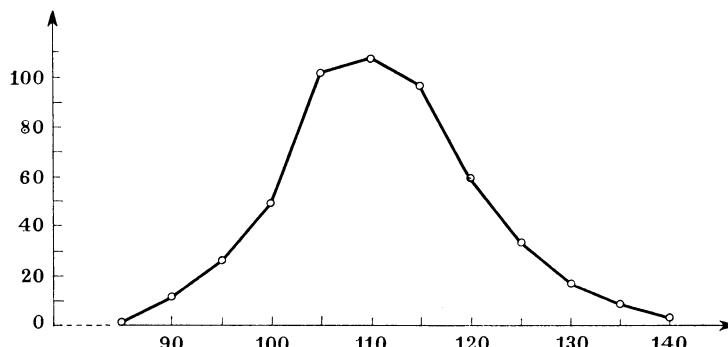
PRZYKŁAD 1. W pewnym zakładzie przemysłowym w trakcie analizy wykonania norm produkcyjnych otrzymano następujący szereg rozdzielczy:

Nr	Średni procent wykonania normy x	Liczba robotników n
1	85	2
2	90	8
3	95	22
4	100	47
5	105	101
6	110	110
7	115	96
8	120	55
9	125	29
10	130	15
11	135	10
12	140	5
		500

Wykres krzywej liczebności, sporządzony w oparciu o powyższy szereg rozdzielczy, podano na rysunku 1. Ocena wzrokowa wykresu nasuwa przypuszczenie, że rozkład empiryczny mało różni się od rozkładu normalnego. Wobec tego formułujemy hipotezę H_0 , że cecha X ma rozkład normalny. Aby zweryfikować tę hipotezę, musimy obliczyć, czemu równałyby się liczebności w naszym szeregu rozdzielczym, gdyby hipoteza H_0 była prawdziwa; liczebności te oznaczmy symbolem n' i nazywać będziemy *liczebnościami hipotetycznymi*. W tym celu obliczamy najpierw średnią artymetyczną i odchylenie standardowe w szeregu rozdzielczym. W naszym przykładzie otrzymujemy

$$\bar{x} = 111,3, \quad s = 9,54.$$

Ponieważ szereg zawiera 500 obserwacji, przeto mamy do czynienia z dużą próbką. Możemy więc przyjąć, że wartość przeciętna w populacji $m = 111,3$, a odchylenie standarde w populacji $\sigma = 9,54$. Wstawiając te wartości do wzoru funkcji gęstości rozkładu



Rys. 1

normalnego (patrz 3.5.3, wzór (1)) otrzymujemy

$$(1) \quad f(x) = \frac{1}{\sqrt{2\pi} \cdot 9,54} \exp \left[-\frac{(x-111,3)^2}{2(9,54)^2} \right].$$

Zgodnie z definicją gęstości (patrz § 3.4, wzór (5)), mamy

$$f(x) = \lim_{x_1 - x_2 \rightarrow 0^+} \frac{P\{x_1 \leq X < x_2\}}{x_2 - x_1},$$

w takim razie

$$(2) \quad P\{x_1 \leq X < x_2\} = f(x) \Delta x + \theta,$$

gdzie symbol $\Delta x = x_2 - x_1$ oznacza dowolny przedział liczbowy (który w szczególności może być przedziałem klasowym szeregu rozdzielczego), a θ oznacza ciąg, zdążający do zera wraz z Δx .

Liczewności hipotetyczne poszczególnych klas szeregu rozdzielczego otrzymamy mnożąc sumę liczebności $n = 500$ przez $P\{x_1 \leq X < x_2\}$. W takim razie

$$(3) \quad n' = n P\{x_1 \leq X < x_2\} = n f(x) \Delta x + n \theta \approx n f(x) \Delta x,$$

gdzie n oznacza sumę liczebności rozkładu empirycznego (w naszym przykładzie $n = 500$), $f(x)$ oznacza gęstość rozkładu normalnego o parametrach $m = 111,3$ i $\sigma = 9,54$ (patrz wzór (1)), Δx oznacza rozpiętość przedziału klasowego (w naszym przykładzie $\Delta x = 5$).

Aby uniknąć wykonywania żmudnych obliczeń, których wymaga wzór (1), wprowadzamy zmienną standaryzowaną.

Jak wiadomo (3.5.3, wzór (5)), zależność między gęstością zmiennej zwykłej i zmiennej standaryzowanej w rozkładzie normalnym wyraża się wzorem

$$f(t) = \frac{1}{\sigma} \varphi(t).$$

Wartości funkcji $\varphi(t)$ są stablicowane. Tablica funkcji gęstości zmiennej standaryzowanej o rozkładzie normalnym podana jest na końcu książki.

Ostateczna postać wzoru, za pomocą którego obliczamy liczebności hipotetyczne w szeregu rozdzielczym, przedstawia się następująco:

$$(4) \quad n' = \frac{n \cdot \Delta x \cdot \varphi(t)}{\sigma}.$$

Tablica 2 na str. 266 zawiera obliczone za pomocą tego wzoru liczebności hipotetyczne w naszym przykładzie.

Na rysunku 2 widzimy wykres krzywej liczebności hipotetycznej n' (linia ciągła). Na tym samym rysunku sporządzono również wykres krzywej liczebności n (linia przerywana). Umożliwia to dokonanie wzrokowej oceny stopnia zgodności rozkładu empirycznego z rozkładem normalnym. Obie krzywe przebiegają blisko siebie. Czyni to wrażenie, że zgodność między rozkładem empirycznym i rozkładem teoretycznym jest duża. Aby

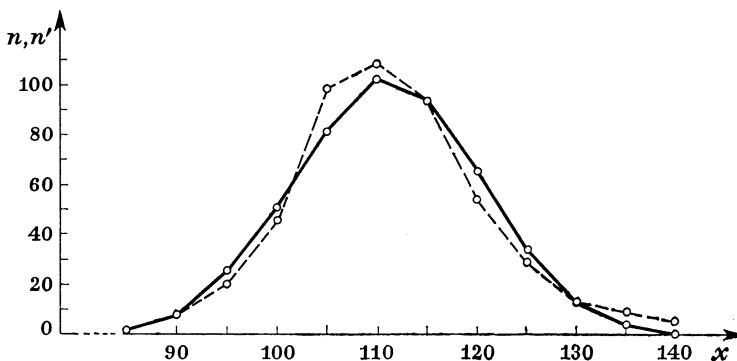
Tablica 1

x	$t = \frac{x-111,3}{9,54}$	$\varphi(t)$	n'
85	2,76	0,01	2,6
90	2,23	0,03	7,9
95	1,71	0,09	23,6
100	1,18	0,20	52,4
105	0,66	0,32	83,9
110	0,14	0,40	104,8
115	0,39	0,37	97,0
120	0,91	0,26	68,1
125	1,44	0,14	36,7
130	1,96	0,06	15,7
135	2,48	0,02	5,2
140	3,01	0,00	0,0
			497,9

przy ocenie zgodności nie opierać się jedynie na subiektywnych wrażeniach, wprowadzimy wielkość

$$(5) \quad \chi^2 = \sum_{i=1}^k \frac{(n_i - n'_i)^2}{n'_i},$$

gdzie k oznacza liczbę klas w szeregu rozdzielczym.



Rys. 2

Można wykazać, że zmienna losowa zdefiniowana wzorem (5) ma rozkład χ^2 o $k-3$ stopniach swobody. (Dowód znajdzie czytelnik u Fisza [8], str. 454 - 456. Bardzo przystępnie idea dowodu podana jest w książce [5], str. 316 - 319).

Zmienna χ^2 jest *miarą rozbieżności między rozkładami*. Zmienna ta przybiera tym większe wartości, im rozbieżność jest większa, czyli im zgodność mniejsza. Istotnie, w liczniku wyrażenia, stojącego pod znakiem sumy po prawej stronie znaku równości wzoru (5)

widzimy różnice między wartościami rozkładu empirycznego i rozkładu hipotetycznego. Różnice te mogą być dodatnie lub ujemne. Aby przy sumowaniu różnic o różnych znakach różnice te nie znosiły się wzajemnie, podnosi się je do kwadratu.

Aby sprawdzić hipotezę H_0 , że rozkład empiryczny nie różni się w sposób istotny od rozkładu teoretycznego, należy obrać dowolnie małą dodatnią liczbę α i odczytać z tablicy II wartość χ^2_0 , dla której

$$P\{\chi^2 > \chi^2_0\} = \alpha.$$

Następnie posługując się wzorem (5) należy wyznaczyć w oparciu o materiał liczbowy z próbki wartość χ^2 , stanowiącą miarę rozbieżności między rozkładem empirycznym a rozkładem teoretycznym. Hipotezę H_0 odrzucamy, jeśli zostanie spełniona nierówność

$$(6) \quad \chi^2 > \chi^2_0.$$

Wartość χ^2 obliczamy za pomocą tablicy 2.

Tablica 2

n	n'	$n-n'$	$(n-n')^2$	$\frac{(n-n')^2}{n'}$
$2 \} 8$	$2,6 \} 7,9$	10,5	-0,5	0,25
22	23,6	-1,6	2,56	0,11
47	52,4	-5,4	29,16	1,56
101	83,9	17,1	292,41	3,49
110	104,8	5,2	27,04	0,26
96	97,0	-1,0	1,00	0,01
55	68,1	-13,1	171,61	2,52
29	36,7	-7,7	59,29	1,62
15	15,7	-0,7	0,49	0,03
$10 \} 5$	$5,2 \} 0,0$	5,2	9,8	18,47
$\underline{500}$	$\underline{497,9}$			$\underline{\chi^2 = 27,09}$

Przyjmijmy poziom istotności $\alpha=0,05$. W tablicy II dla 7 stopni swobody (połączymy bowiem ze względu na małe liczbowości dwie pierwsze i dwie ostatnie klasy szeregu rozdzielczego, otrzymując w ten sposób 10 klas zamiast 12) odczytujemy $\chi^2_0=14,067$. Ponieważ

$$27,09 > 14,067,$$

przeto hipotezę H_0 odrzucamy. Zauważmy jednak, że na odrzucenie hipotezy decydujący wpływ miała ostatnia klasa, utworzona z połączenia dwóch mało licznych klas szeregu. Jest to właśnie słaba strona tego testu, że silnie reaguje na wpływ krańcowych, mało licznych klas.

Opisany sposób weryfikacji hipotezy nieparametrycznej o kształcie rozkładu cech w populacji nazywa się *testem χ^2* .

PRZYKŁAD 2. Prostym i wygodnym sposobem sprawdzania, że rozkład z próbki nie różni się istotnie od hipotetycznego rozkładu w populacji, jest sposób oparty na twierdzeniu Kołmogorowa (patrz paragraf 6.3). Z twierdzenia tego wynika, że jeżeli dystrybuanta $F(x)$ cechy X w populacji jest ciągła, a dystrybuantę empiryczną $G_n(x)$ otrzymano z próbki pobranej z populacji w drodze losowania ze zwracaniem, to znana jest relacja łącząca obie dystrybuanty, a mianowicie znane jest prawdopodobieństwo

$$P(\sqrt{n} D_n < \lambda) = P(\sqrt{n} \sup_{-\infty < x < \infty} |F(x) - G_n(x)| < \lambda).$$

Jeżeli λ_0 jest taką krytyczną wartością parametru λ , że

$$P(\sqrt{n} D_n \geq \lambda_0) = 1 - P(\sqrt{n} D_n < \lambda_0) = \alpha,$$

gdzie α jest tak dobranym współczynnikiem istotności, że zdarzenia, których prawdopodobieństwa realizacji nie przekraczają α , mogą być uznane za praktycznie niemożliwe, to zajście nierówności

$$\sqrt{n} D_n \geq \lambda_0$$

upoważnia do odrzucenia hipotezy H_0 , że dystrybuanta empiryczna $G_n(x)$ nie różni się istotnie od dystrybuanty teoretycznej $F(x)$.

Technikę rachunkową, związaną z tym sposobem weryfikacji hipotezy o kształcie rozkładu cechy w populacji, zilustrujemy na danych liczbowych poprzedniego przykładu. Warto zapamiętać, że sposób ten bywa w literaturze określany mianem *testu Kołmogorowa-Smirnowa*.

Obliczenia związane ze stosowaniem tego testu wygodnie jest ująć w następującą tablicę:

Tablica 3

Górna granica przedziału klasowego procentu wykonania normy	Górna granica przedziału klasowego (zestandardizowana)	$F(x)$	$G_n(x)$	$ F(x) - G_n(x) $
87,5	-2,50	0,006	0,004	0,002
92,5	-1,97	0,024	0,020	0,004
97,5	-1,45	0,074	0,064	0,010
102,5	-0,92	0,179	0,158	0,021
107,5	-0,40	0,345	0,360	0,015
112,5	0,12	0,548	0,580	0,032
117,5	0,65	0,742	0,772	0,030
122,5	1,17	0,879	0,882	0,003
127,5	1,70	0,955	0,940	0,015
132,5	2,22	0,987	0,070	0,017
137,5	2,74	0,997	0,990	0,007
142,5	3,27	0,999	1,000	0,001

Wyjaśniamy, że pierwsza kolumna tej tablicy, w odróżnieniu od pierwszej kolumny tablicy 2, zawiera nie środki przedziałów klasowych procentu wykonania normy, lecz

górnego granice przedziałów klasowych, które otrzymuje się przez dodanie do średników przedziałów połowy rozpiętości tych przedziałów, czyli przez dodanie $\Delta x/2$. Wprowadzenie górnych granic przedziałów klasowych stało się konieczne dla obliczenia wartości dystrybuanty teoretycznej, tzn. dystrybuanty rozkładu normalnego $F(x)$. Wartości te zawiera trzecia kolumna tablicy, podczas gdy w czwartej kolumnie podane są wartości dystrybuanty empirycznej $G_n(x)$. W ostatniej kolumnie znajdujemy, że

$$D_n = \max_x |F(x) - G_n(x)| = 0,032.$$

Stąd

$$\sqrt{n} D_n = \sqrt{500} \cdot 0,032 = 22,361 \cdot 0,032 = 0,716.$$

W tablicy VI na końcu niniejszej książki znajdujemy dla $\alpha=0,05$, czyli dla $Q(\lambda_0)=0,95$, że $\lambda_0=1,36$. Jak z tego wynika, iloczyn $\sqrt{n} D_n=0,716$ nie jest większy od $\lambda_0=1,36$, a więc nie ma podstaw do odrzucenia hipotezy H_0 . Jak widzimy, otrzymaliśmy wynik odmienny niż w przykładzie 1 i chyba trzeba przyjąć, że właśnie ten wynik jest poprawny. Przewaga testu Kołmogorowa-Smirnowa nad testem χ^2 polega właśnie na tym, że jest on niewrażliwy na działanie skrajnych, mało licznych klas szeregu rozdzielczego.

PRZYKŁAD 3. Rozważmy zmienną losową ciągłą X , której gęstość $f(x)$ jest znana. Przyjmujmy dla ustalenia uwagi, że X jest zmienną losową o rozkładzie jednostajnym, rozpiętym na odcinku $\langle 0, 1 \rangle$. Oczywiście zbiór wartości, jakie może przybierać zmienna X , jest nieskończony. Interpretując ten zbiór jako populację generalną rodzi się pytanie, jak z takiej populacji można byłoby wylosować n -elementową próbę?

Istnieją różne sposoby wylosowania takiej próbki. Najprostszy z nich polega na wylosowaniu próbki za pomocą tablic liczb losowych (patrz [28]). Gdy nie mamy tablic pod ręką, można postąpić w sposób następujący. Wykonujemy serię 10 rzutów monetą. Wyrzuceniu orła przypisujemy cyfrę 1, a wyrzuceniu reszki – cyfrę 0. Przypuśćmy, że zrealizowany został następujący ciąg: 0, 0, 1, 0, 1, 1, 1, 0, 1, 1. Ten dziesięcioelementowy ciąg utworzony z jedynek i zer interpretujemy jako liczbę zapisaną w systemie dwójkowym:

$$0010111011.$$

„Tłumaczymy” tę liczbę na liczbę w systemie dziesiętnym. Otrzymujemy

$$0 \cdot 2^9 + 0 \cdot 2^8 + 1 \cdot 2^7 + 0 \cdot 2^6 + 1 \cdot 2^5 + 1 \cdot 2^4 + 1 \cdot 2^3 + 0 \cdot 2^2 + 1 \cdot 2^1 + 1 \cdot 2^0 = \\ = 128 + 32 + 16 + 8 + 2 + 1 = 187.$$

Największa liczba, jaką można otrzymać za pomocą takiego postępowania, wynosi $2^{10}-1=1023$. Dzieląc zrealizowaną w doświadczeniu losowym liczbę 187 przez 1023 otrzymamy pierwszą realizację x_1 zmiennej losowej X o rozkładzie jednostajnym rozpiętym na przedziale $\langle 0, 1 \rangle$. Realizacja ta w naszym przykładzie wynosi $x_1=187/1023=0,1826$. Powtarzając takie postępowanie n razy otrzymamy ciąg liczb, który można interpretować jako ciąg realizacji zmiennej losowej X , tzn. jako n -elementową próbę wylosowaną z populacji generalnej o rozkładzie jednostajnym.

Wadą takiego sposobu losowania n -elementowej próbki z nieskończonej populacji generalnej jest to, że jest on dosyć uciążliwy pod względem rachunkowym oraz że do

wykonania obliczeń nie opłaca się angażować maszyny cyfrowej, gdyż szybkość generowania liczb losowych jest limitowana szybkością wykonywania rzutów monetą, odczytywania wyników tych rzutów i wprowadzania odpowiednich danych sukcesywnie do maszyny.

Zbadajmy następujący sposób produkowania liczb z przedziału $\langle 0, 1 \rangle$: Bierzemy dowolną parę liczb dziesięciocyfrowych z tego przedziału, np.

$$a_1 = 0,6500688123, \quad a_2 = 0,3232551075.$$

Liczby te dodajemy otrzymując liczbę

$$a_3 = 0,9733239198.$$

Gdyby liczba a_3 przybrała wartość równą lub większą od jedności – zatrzymalibyśmy tylko mantysę tej liczby.

Z kolei odrzucamy liczbę a_1 i dodajemy liczby a_2 i a_3 . Postępujemy w ten sposób tak długo, aż otrzymamy n -elementowy ciąg liczb a_3, a_4, \dots, a_{n+2} . Ciągi tego typu nazywają się *ciągami Fibonacciego*.

Z kolei w każdej liczbie Fibonacciego skreślamy pierwsze trzy cyfry po przecinku i trzy cyfry ostatnie. Cyfry, które pozostaną po takim zabiegu, zapisujemy jako cztery miejsca dziesiętne realizacji zmiennej losowej X o rozkładzie jednostajnym. Na przykład z liczby Fibonacciego

$$a_3 = 0,9733239198$$

otrzymujemy realizację

$$x_1 = 0,3239.$$

Otrzymywanie liczb za pomocą przedstawionego postępowania może być łatwo wykonane za pomocą maszyny cyfrowej, przy czym szybkość otrzymywania liczb jest bardzo duża. Nie trzeba chyba tłumaczyć, że liczby tak otrzymywane nie są jednak w rzeczywistości liczbami losowymi, lecz liczbami przypominającymi liczby losowe. Liczby tego typu noszą nazwę *liczb pseudolosowych*. Wyłania się tu ważny problem statystyczny badania dobrotą liczb pseudolosowych. Dobroć tę bada się za pomocą wielu różnych testów statystycznych, tak parametrycznych jak i nieparametrycznych. Przechodzimy do przedstawienia jednego z takich testów, który służy do orzekania, czy dany ciąg liczb można uznać za próbę wylosowaną z populacji o danym rozkładzie. Test ten odpowiada więc na to samo pytanie co test χ^2 czy też Kołmogorowa-Smirnowa, a cechą wyróżniającą ten test od dwóch poprzednich jest to, że może on być stosowany dla małych wartości n . Rozważmy zmienną losową X mającą tę własność, że dystrybuanta $F(x)$ tej zmiennej jest ciągła. Można wykazać (patrz [8]), że zmienna losowa

$$Y = F(X)$$

ma wtedy rozkład jednostajny określony na odcinku $\langle 0, 1 \rangle$.

Niech $F_1(y)$ będzie dystrybuantą zmiennej losowej Y . W takim razie

$$F_1(y) = P(Y < y) = P[F(X) < y] =$$

$$= \begin{cases} 0 & \text{dla } y \leq 0, \\ P[X < F^{-1}(y)] = F[F^{-1}(y)] = y & \text{dla } 0 < y < 1, \\ 1 & \text{dla } y \geq 1, \end{cases}$$

gdzie F^{-1} oznacza funkcję odwrotną względem F . Stąd

$$F'_1(y) = f_1(y) = \begin{cases} 1 & \text{dla } 0 < y < 1, \\ 0 & \text{dla } y \leq 0 \text{ lub } y \geq 1. \end{cases}$$

Jak widać, $f_1(y)$ jest gęstością rozkładu jednostajnego. Wykorzystamy to dla weryfikacji testu, który teraz przedstawimy.

Niech H_0 oznacza hipotezę, że nieznana dystrybuanta $F(x)$ zmiennej losowej X jest tożsamościowo równa pewnej funkcji $D(x)$. Oczywiście hipoteza ta jest równoważna hipotezie

$$H: F_1(y) = \int_0^y dy,$$

gdzie $F_1(y)$ jest symbolem dystrybuanty zmiennej losowej $Y = F(X)$. Hipoteza podlega sprawdzeniu w oparciu o ciąg realizacji zmiennej losowej x_1, x_2, \dots, x_n . Dla zweryfikowania hipotezy stosujemy następujące postępowanie:

Dzielimy odcinek jednostkowy $\langle 0, 1 \rangle$, który dalej oznaczać będziemy literą I , na m równych części I_1, I_2, \dots, I_m , które nazywać będziemy celami⁽¹⁾. Niech k oznacza liczbę cel pustych. Liczba ta jest oczywiście realizacją pewnej zmiennej losowej K , której dystrybuantę można łatwo wyznaczyć, jeżeli przyjmie się założenie, że hipoteza H_0 jest prawdziwa.

Jak wiadomo (patrz [6]), prawdopodobieństwo $p_k(n, m)$ zdarzenia polegającego na tym, że wśród m cel, w których losowo rozmiieszczono n kul, znajdzie się k cel pustych, wyraża się wzorem

$$p_k(n, m) = \binom{m}{k} \sum_{r=0}^{m-k} (-1)^r \binom{m-k}{r} \left(1 - \frac{k+r}{m}\right)^n.$$

Stąd dystrybuanta $P_k(n, m)$ zmiennej losowej K ma postać

$$P_k(n, m) = \sum_{s=0}^k \binom{m}{s} \sum_{r=0}^{m-s} (-1)^r \binom{m-s}{r} \left(1 - \frac{s+r}{m}\right)^n.$$

Niech teraz α oznacza małą liczbę dodatnią obraną tak, że zdarzenie losowe, którego prawdopodobieństwo realizacji jest nie większe od α , może być uważane za praktycznie niemożliwe. W takim razie hipotezę H_0 odrzucamy wtedy, gdy doświadczalnie zrealizowano taką wartość k_1 , że

$$P_{k_1}(n, m) < \alpha$$

⁽¹⁾ Od słowa: *cela*.

lub taką wartość k_2 , że

$$P_{k_2}(n, m) > 1 - \alpha.$$

Jeżeli zrealizowana w doświadczeniu wartość zmiennej losowej K spełnia pierwszą z tych nierówności, to hipoteza H_0 musi być odrzucona ze względu na zbyt wielką zgodność rozkładu empirycznego z hipotetycznym rozkładem jednostajnym. Warto w tym miejscu dać krótki komentarz. Zwrócić uwagę na to, że gdyby punkty A_1, A_2, \dots, A_n , których odciętymi są realizacje y_1, y_2, \dots, y_n zmiennej losowej Y , rozmieszczone zostały na odcinku I nie losowo, lecz równomiernie (a więc tendencyjnie), to liczba k cel pustych byłaby równa零. Jak z tego wynika, bardzo mała wartość zmiennej K i związana z tym bardzo bliska零 wartość prawdopodobieństwa $P_k(n, m)$ jest sygnałem, że rozkład empiryczny wykazuje zbyt wielką zgodność z rozkładem hipotetycznym, aby mogła się ona zdarzyć w praktyce. Hipoteza H_0 musi więc być odrzucona.

Jeżeli zrealizowana w doświadczeniu wartość zmiennej losowej K spełnia drugą nierówność, to hipoteza H_0 musi być również odrzucona, gdyż rozbieżność między rozkładem empirycznym a hipotetycznym rozkładem jednostajnym jest zbyt duża, aby mogła wystąpić w praktyce.

Odrzucenie hipotezy H_0 , że rozkład zmiennej losowej Y jest jednostajny, prowadzi do przyjęcia hipotezy alternatywnej H_1 , że rozkład zmiennej losowej Y nie jest jednostajny. To z kolei jest równoważne z przyjęciem hipotezy, że dystrybuanta $F(x)$ zmiennej losowej X różni się istotnie od funkcji $D(x)$.

Podany wyżej wzór służący do obliczania wartości dystrybuanty $P_k(n, m)$ jest uciążliwy do stosowania nawet przy małych wartościach parametrów m i n . Zachodzi więc konieczność opracowania odpowiedniej tablicy, której przykładem może być tablica VII zamieszczona na końcu książki. Tablica ta, opracowana dla czterech wartości parametru α , a mianowicie dla $\alpha_1 = 0,1$, $\alpha_2 = 0,05$, $\alpha_3 = 0,01$, $\alpha_4 = 0,005$, podaje wartości liczb k_1 i k_2 obliczone przy założeniu, że $m = n$ oraz $2 \leq n \leq 30$.

Dowodzi się (patrz [6]), że jeżeli m i n rosną nieograniczenie w taki sposób, że

$$\lambda = me^{k/m} < c,$$

gdzie c jest pewną stałą, to dla każdej ustalonej liczby k spełniona jest relacja

$$P(k, \lambda) = e^{-\lambda} \cdot \frac{\lambda^k}{k!}.$$

Jak z tego wynika, graniczne wartości prawdopodobieństwa $P_k(n, m)$ można otrzymać z rozkładu Poissona. To właśnie tłumaczy wybór liczby 30 jako górnej granicy przedziału zmienności parametru n w tablicy VII. Dla $n > 30$ można już uzyskać dobre przybliżenie prawdopodobieństw $P_k(n, m)$ za pomocą wzoru Poissona.

Zilustrujemy stosowanie opisanego testu, który nazywać będziemy *nieparametrycznym testem dla malej próbki*, za pomocą przykładu liczbowego dotyczącego badania dobroci ciągu liczb pseudolosowych. Liczby te podano w kolumnie drugiej tablicy 4. Należy zweryfikować hipotezę, że liczby te można uważać za realizacje zmiennej losowej o rozkładzie jednostajnym rozpiętym na odcinku $\langle 0, 1 \rangle$. W nagłówku kolumny drugiej umieszczono symbol y_i oznaczający realizację zmiennej losowej Y , której rozkład w myśl hipotezy

tezy ma być jednostajny. W kolumnie trzeciej, opatrzonej nagłówkiem I_j , podane są granice podprzedziałów, na jakie podzielony został przedział $\langle 0, 1 \rangle$. Podprzedziały te są właśnie celami, do których trafiać mają liczby y_i . Ponieważ liczb tych jest w naszym przykładzie 20, więc $i=1, 2, \dots, 20$. Cel jest także 20, więc $j=1, 2, \dots, 20$. Ostatnia, czwarta kolumna tablicy 4 zawiera znaki „+” i „-” umieszczone obok odpowiednich cel. Liczba znaków „+” odpowiada ilości przypadków trafienia realizacji y_i zmiennej losowej Y do odpowiedniej celi. Znak „-” oznacza, że do danej celi nie trafiła żadna realizacja.

Przyjmijmy, że $\alpha=0,005$. W takim razie, korzystając z tablicy VII zamieszczonej na końcu książki, w wierszu $n=20$ tej tablicy znajdujemy

$$k_1 = 3, \quad k_2 = 10.$$

Ponieważ w naszym przykładzie zrealizowana wartość zmiennej losowej K wynosi 2, gdyż w czwartej kolumnie tablicy 4 wystąpiły dwa znaki „-” wskazujące na to, że cele 0,80 - 0,85 i 0,95 - 1,00 są puste, przeto hipotezę H_0 należy odrzucić. Odrzucenie hipotezy jest spowodowane tym, że zrealizowana w doświadczeniu wartość zmiennej losowej K okazała się mniejsza od granicznej wartości $k_1 = 3$.

Tablica 4

i	y_i	I_j	$y_i \in I_j$
1	0,0344	0,00	+
2	0,0902	0,05	+
3	0,1146	0,10	+
4	0,1803	0,15	+
5	0,2361	0,20	+
6	0,2705	0,25	+
7	0,3262	0,30	+
8	0,3607	0,35	+
9	0,4164	0,40	+
10	0,4721	0,45	+
11	0,5066	0,50	+
12	0,5623	0,55	+
13	0,6180	0,60	+
14	0,6525	0,65	+
15	0,7082	0,70	+
16	0,7426	0,75	+
17	0,7984	0,80	-
18	0,8541	0,85	+
19	0,8885	0,90	+
20	0,9442	0,95 1,00	-

Wyjaśniamy, że powodem odrzucenia hipotezy H_0 jest zbyt wielka zgodność rozkładu empirycznego z rozkładem teoretycznym, tj. zgodność tak duża, że przeczy to losowemu charakterowi danych liczbowych podlegających weryfikacji statystycznej. Warto podkreślić, że wskazania testu są prawidłowe, gdyż liczby zawarte w kolumnie 2 tablicy 4

nie są realizacjami zmiennej losowej o rozkładzie jednostajnym, lecz liczbami pseudolosowymi, tzw. *liczbami złotymi* Steinhauza odczytanymi z tablicy 30 zbioru tablic [28].

PRZYKŁAD 4. Sprawdzenie hipotezy, że liczby przytoczone w kolumnie drugiej tablicy 4 są realizacjami zmiennej losowej o rozkładzie jednostajnym, można przeprowadzić wieloma innymi sposobami, z których bardzo prostym i dlatego wygodnym do stosowania w praktyce jest tzw. *test serii*.

Oznaczmy literą A zdarzenie polegające na tym, że określona liczba w ciągu liczb podanych w kolumnie 2 tablicy 4 jest parzysta, i literą B , że jest nieparzysta. W takim razie ciągowi tych liczb można przyporządkować ciąg utworzony z następujących po sobie w pewnym porządku liter A i B . Korzystając z danych liczbowych tablicy 4 otrzymamy ciąg następującej postaci:

$$A \ A \ A \ B \ B \ B \ A \ B \ A \ B \ A \ B \ A \ B \ A \ A \ A \ B \ B \ A \ .$$

Każdy podciąg takiego ciągu, utworzony z możliwie maksymalnej ilości elementów tego samego rodzaju, nazywa się *serią*. Liczba elementów wchodzących w skład danej serii nazywa się *długością serii*. Oznaczmy liczbę serii literą U , natomiast długość serii – literą V . Oczywiście, jeżeli zdarzenia A i B są zdarzeniami losowymi, to U i V są zmiennymi losowymi.

Przy założeniu, że zdarzenia A i B są niezależne i że znane są prawdopodobieństwa $P(A)=p$ oraz $P(B)=1-p$, można stosunkowo łatwo znaleźć dystrybuanty zmiennych losowych U i V oraz wartości oczekiwane i wariancje tych zmiennych (patrz [8]). Na końcu książki zamieszczona została tablica opatrzona numerem VIII, która dla wybranych czterech poziomów istotności 0,025, 0,05, 0,95, 0,975 i dla ustalonej pary liczb n_1 oraz n_2 , gdzie n_1, n_2 są to liczby elementów jednego i drugiego rodzaju w badanym ciągu, podaje krytyczne wartości parametru u odpowiadające przyjętym wartościom poziomu istotności.

W naszym przykładzie liczba elementów A w ciągu wynosi $n_1=11$, natomiast liczba elementów B w tym ciągu wynosi $n_2=9$. Z tablicy VIII dla $\alpha=0,025$ znajdujemy, że liczba serii nie może być mniejsza od 6, a dla $\alpha=0,975$ nie może być większa od 15. W naszym przykładzie liczba serii wynosi 13 i mieści się w dopuszczalnym przedziale, co oznacza, że posługując się testem serii nie mamy podstaw do odrzucenia sprawdzanej hipotezy.

Zaletą testu serii jest jego prosta, natomiast jego wadą jest stosunkowo mała czułość, która, tak jak w niniejszym przykładzie, wyraża się tym, że test niekiedy nie daje podstaw do odrzucenia hipotez fałszywych. Zauważmy, że gdybyśmy przyjęli $\alpha=0,05$ oraz $\alpha=0,95$, to odpowiednie wartości krytyczne parametru u wyniosłyby odpowiednio 6 i 14, a więc i w tym przypadku nie byłoby podstaw do odrzucenia sprawdzanej hipotezy. Czytelnik spostrzegł niewątpliwie, że test zaprezentowany w poprzednim przykładzie był znacznie czulszy i pozwolił na odrzucenie hipotezy już przy wartości parametru $\alpha=0,005$.

7.4.3. Weryfikacja hipotezy nieparametrycznej o niezależności zmiennych losowych

PRZYKŁAD 1. Do fabryki przewodników elektrycznych zaczęły napływać reklamacje od użytkowników, użalających się na złą jakość izolacji przewodników. W reklamacjach wyrażano przypuszczenie, że łamliwość izolacji zwiększa się pod wpływem działania kwasów.

Izolacje przewodników produkowanych w fabryce sporządzane były z igelitu (produkt całkowitej polimeryzacji chlorku winylu). Chemicy zatrudnieni w fabryce oświadczyli kategorycznie, że łamliwość izolacji musiały wywołać inne przyczyny, gdyż igelit jest tworzywem kwasoodpornym, i że powtarzające się wielokrotnie w reklamacjach przypuszczenia w sprawie rzekomej przyczyny łamliwości izolacji jest fałszywe. Na potwierdzenie swego stanowiska powoływali się na gwarancję dostawcy folii igelitowej, który również zapewniał, że folia jest odporna na działanie kwasu.

Dla rozstrzygnięcia problemu zzewzano do pomocy statystyka. Pod jego kierunkiem wykonano następujący eksperiment. Pobrano losowo 1000 pasków folii. Każdy pasek miał jednakowe wymiary. Wylosowaną próbki podzielono na 10 części, po 100 pasków każda. Pierwszą część poddano działaniu kwasu w ciągu jednej godziny, drugą trzymano w kwasie przez 2 godziny, trzecią – 3 godziny itd. Ostatnią, dziesiątą część zanurzono w kwasie na okres 10 godzin. Po wyjęciu z kwasu każdy pasek został starannie opłukany wodą i wysuszony, a następnie poddany badaniu łamliwości, polegającemu na wielokrotnym zginaniu paska – aż do pęknięcia.

Wyniki eksperimentu przedstawia poniższa tablica, X oznacza liczbę godzin działania kwasem na folię, Y – liczbę zgięć w tysiącach:

Tablica 1

$X \backslash Y$	1	2	3	4	5	6	7	8	9	10	\sum
2,0 - 2,2	5	1	9	3	4	16	25	8	4	10	85
2,2 - 2,4	36	25	11	7	24	7	19	22	11	30	192
2,4 - 2,6	12	14	41	65	16	41	35	27	49	21	321
2,6 - 2,8	41	38	21	21	36	22	7	23	15	21	245
2,8 - 3,0	2	18	9	4	8	9	9	13	10	14	96
3,0 - 3,2	4	4	9	–	12	5	5	7	11	4	61
\sum	100	100	100	100	100	100	100	100	100	100	1000

Należy sprawdzić hipotezę nieparametryczną, że zmienne losowe X i Y są niezależne. Jak wiadomo (patrz 3.6.2 wzór (14)), warunek niezależności dwóch zmiennych losowych można przedstawić za pomocą następującej relacji:

$$(1) \quad P(x_i, y_j) = P(x_i) P(y_j)$$

lub krócej

$$p_{ij} = p_i q_j.$$

Materiał liczbowy potrzebny do weryfikacji tej hipotezy jest zawarty w przytoczonej wyżej tablicy 1. Tablica ta jest podobna do tablicy podanej na str. 93. Różnica między nimi polega na tym, że prawdopodobieństwa występujące w tablicy na str. 93 zostały teraz zastąpione liczebnosciami.

Oznaczmy symbolem n_{ij} ($i=1, 2, \dots, r$, $j=1, 2, \dots, s$) liczebność stojącą w tablicy na przecięciu i -tego wiersza oraz j -tej kolumny. Liczba r oznacza ilość wierszy w tablicy, natomiast s – ilość kolumn w tablicy.

Oznaczmy dalej

$$\sum_{i=1}^r \sum_{j=1}^s n_{ij} = n, \quad \sum_{i=1}^r n_{ij} = n_{\cdot j}, \quad \sum_{j=1}^s n_{ij} = n_{i \cdot}.$$

Dowodzi się (patrz [5]), że wielkość

$$(2) \quad \sum_{i=1}^r \sum_{j=1}^s \frac{(n_{ij} - np_{ij})^2}{np_{ij}}$$

ma w przybliżeniu rozkład χ^2 (patrz 6.5.3) o liczbie stopni swobody

$$k = (r-1)(s-1).$$

Występujący we wzorze (2) symbol p_{ij} wprowadzony został dla uproszczenia zapisu zamiaszt symbolu $P(x_i, y_j)$, którym posługiwano się w tablicy 1 w punkcie 3.6.2.

Przy założeniu, że hipoteza jest prawdziwa, spełnia się relacja (1), a więc możemy napisać

$$p_{ij} = p_i \cdot q_j.$$

W § 6.3 przytoczyliśmy twierdzenie Gliwenki. Twierdzenie to daje się uogólnić na zmienne wielowymiarowe. Z uogólnionego twierdzenia Gliwenki wynika, że prawdopodobieństwa p_i oraz q_j można estymować za pomocą częstości, przyjmując

$$p_i = \frac{n_{i \cdot}}{n}, \quad q_j = \frac{n_{\cdot j}}{n}.$$

W takim razie

$$(3) \quad p_{ij} = \frac{n_{i \cdot}}{n} \cdot \frac{n_{\cdot j}}{n}.$$

Podstawiając wyrażenie (3) do (2) otrzymujemy

$$(4) \quad \sum_{i=1}^r \sum_{j=1}^s \frac{\left(n_{ij} - \frac{n_{i \cdot} \cdot n_{\cdot j}}{n} \right)^2}{\frac{n_{i \cdot} \cdot n_{\cdot j}}{n}} = n \left(\sum_{i=1}^r \sum_{j=1}^s \frac{n_{ij}^2}{n_{i \cdot} \cdot n_{\cdot j}} - 1 \right).$$

Wzór (4) przedstawia zmienną losową, która ma rozkład zbliżony do rozkładu χ_0^2 o $(r-1)(s-1)$ stopniach swobody.

Niech α oznacza liczbę spełniającą nierówność $0 < \alpha < 1$. Obierzmy α w ten sposób, żeby zdarzenie, którego prawdopodobieństwo realizacji równa się α , można było uważać za zdarzenie praktycznie niemożliwe. W tablicy rozkładu χ^2 , w wierszu odpowiadającym liczbie stopni swobody $k = (r-1)(s-1)$ znajdujemy taką wartość χ_0^2 , że

$$P(\chi^2 > \chi_0^2) = \alpha.$$

Weryfikowaną hipotezę o niezależności zmiennych losowych X i Y odrzucamy, gdy spełniona zostanie nierówność

$$(5) \quad n \left(\sum_{i=1}^r \sum_{j=1}^s \frac{n_{ij}^2}{n_{i \cdot} \cdot n_{\cdot j}} - 1 \right) > \chi_0^2.$$

Do wyznaczenia podwójnej sumy, stojącej w nawiasach po lewej stronie powyższej nierówności, posłużymy się tablicą 2. W poszczególnych kratkach tej tablicy podane są wartości ułamka

$$\frac{n_{ij}^2}{n_i \cdot n_j}$$

stojącego pod znakiem podwójnej sumy we wzorze (5). I tak np. na przecięciu pierwszego wiersza i pierwszej kolumny znajduje się liczba 0,00294.

Tablica 2

	1	2	3	4	5	6	7	8	9	10	
1	0,00294	0,00012	0,00953	0,00106	0,00188	0,03012	0,07353	0,00753	0,00188	0,01176	0,14035
2	0,06750	0,03255	0,00630	0,00255	0,03000	0,00255	0,01881	0,02521	0,00630	0,04687	0,23863
3	0,00449	0,00611	0,05237	0,13162	0,00798	0,05237	0,03816	0,02271	0,07480	0,01374	0,40435
4	0,06861	0,05894	0,01800	0,01800	0,05290	0,01976	0,00200	0,02159	0,00918	0,01800	0,28698
5	0,00042	0,03375	0,00844	0,00167	0,00667	0,00844	0,00844	0,01760	0,01042	0,02042	0,11627
6	0,00262	0,00262	0,01328	—	0,02361	0,00410	0,00410	0,00803	0,01984	0,00262	0,08082
	0,14658	0,13409	0,10792	0,15490	0,12304	0,11734	0,14503	0,10267	0,12242	0,11341	1,26740

W tablicy 1 na stronie 275 odczytujemy np., że

$$n_{11}=5, \quad n_{1\cdot}=85, \quad n_{\cdot 1}=100.$$

Stąd po wykonaniu rachunków otrzymujemy liczbę 0,00294. Jak wiadomo, gdy liczba stopni swobody jest większa od 30, posługujemy się zamiast rozkładu χ^2 rozkładem normalnym; korzystamy wtedy z relacji

$$P\{\chi^2 > \chi_0^2\} \approx 0,5 - \Phi(t),$$

przy czym

$$\chi_0^2 = \frac{(t + \sqrt{2k-1})^2}{2}$$

(patrz 6.5.3, wzory (3) i (4)).

Przyjmijmy, że $\alpha=0,01$. W takim razie

$$P\{\chi^2 > \chi_0^2\} \approx 0,01,$$

czyli

$$0,5 - \Phi(t) = 0,01,$$

skąd

$$\Phi(t) = 0,49.$$

Z tablicy I wynika, że $t=2,325$, wobec tego

$$\chi_0^2 = \frac{(2,325 + \sqrt{89})^2}{2} \approx 69.$$

Ponieważ po wstawieniu wartości liczbowych do wzoru (5) otrzymujemy nierówność

$$267 > 67,$$

więc hipotezę o niezależności zmiennych losowych odrzucamy. Oznacza to, że nie można twierdzić, że działanie kwasu nie ma wpływu na trwałość (a dokładniej mówiąc – na kruchosć igelitowej folii).

Jak widać, badanie statystyczne potwierdziło przypuszczenie użytkowników.

Pytania kontrolne i zadania

1. Co to jest hipoteza statystyczna?
2. Wyjaśnić, na czym polega różnica między hipotezą parametryczną a hipotezą nieparametryczną?
3. Co to są testy statystyczne i jak się one dzielą?
4. Podać przykłady hipotez statystycznych.
5. Opisać tok postępowania przy weryfikacji hipotezy parametrycznej.
6. Co to jest poziom istotności?
7. Co to jest zbiór krytyczny?
8. Wyjaśnić, dlaczego przy weryfikacji hipotezy statystycznej nie możemy mieć nigdy absolutnej pewności, że nasza decyzja, dotycząca przyjęcia lub odrzucenia sprawdzanej hipotezy, okaże się słuszna?
9. Czy jeśli wyznaczona na podstawie próbki wartość zmiennej losowej nie trafi do zbioru krytycznego, to możemy uważać, że hipoteza H_0 jest prawdziwa?
10. Co to jest błąd pierwszego rodzaju i czemu równe się prawdopodobieństwo popełnienia tego błędu?
11. Co to jest błąd drugiego rodzaju i czemu równe się prawdopodobieństwo popełnienia tego błędu?
12. Co nazywamy mocą testu?
13. Co to jest krzywa mocy testu?
14. Wykonano 1536 rzutów siedmioma monetami i zapisano liczbę orłów, które pojawiły się w każdym rzucie. Otrzymano następujące wyniki:

X – liczba orłów	0	1	2	3	4	5	6	7
n – liczba rzutów	12	78	270	456	386	252	69	13

Zbadać, czy rozkład zmiennej X różni się istotnie od rozkładu dwumianowego, w którym $p=0,5$.

Wskazówka. Liczebności hipotetyczne n' oblicza się za pomocą wzoru

$$n' = n \cdot C_7^x \cdot \frac{1}{2^7}.$$

8.1. WYBRANE WIADOMOŚCI Z DZIEDZINY PROCESÓW STOCHASTYCZNYCH

Najnowszym działem rachunku prawdopodobieństwa i statystyki matematycznej, który w okresie ostatniego kwartału wykształcił się w samodzielną dyscyplinę naukową, jest tzw. *teoria procesów stochastycznych*. Nie sposób jednym zdaniem wyrazić, czym zajmuje się rozległa pod względem pojęciowym i trudna pod względem formalno-matematycznym dziedzina wiedzy, której poświęca się obszerne monografie naukowe (patrz np. [10], [26], [31], [33]). Jeżeli się jednak zrezygnuje z troski o daleko posuniętą poprawność i ścisłość naukową, to można byłoby powiedzieć, że teoria procesów stochastycznych zajmuje się uogólnianiem twierdzeń rachunku prawdopodobieństwa na przypadek zmiennych losowych, których rozkłady zależą od zmiennych parametrów, przy czym jeden z tych parametrów może być interpretowany jako czas. Oznacza to między innymi, że przedmiotem teorii procesów stochastycznych są zachodzące w czasie zdarzenia losowe, zwane dla podkreślenia tej ich wspólnej własności *procesami losowymi* lub *procesami stochastycznymi*.

Aby zaznajomić czytelnika w sposób jak najbardziej przystępny z podstawami teorii procesów stochastycznych, posłużymy się serią poglądowych przykładów.

PRZYKŁAD 1. Realizujemy eksperiment polegający na wykonaniu rzutu K monetami. Rzut ten prowadzi do wystąpienia K niezależnych zdarzeń losowych. W tym sensie eksperiment taki jest podobny do eksperimentu polegającego na wykonaniu serii rzutów jedną monetą. Różnica między tymi dwoma eksperimentami polega na tym, że w pierwszym przypadku otrzymujemy zbiór realizacji K zdarzeń losowych, w drugim ciąg realizacji tych zdarzeń. Oba eksperymenty byłyby równoważne, gdyby monety biorące udział w pierwszym eksperymencie były ponumerowane liczbami $k=1, 2, \dots, K$.

Wyobraźmy sobie teraz, że powtarzamy eksperiment polegający na rzucaniu K ponumerowanymi monetami $N \leq T$ razy, gdzie T jest wielkością skońzoną lub nie. Oznaczmy numer kolejnego eksperimentu literą t . Tak więc $t=1, 2, \dots, N$. Parametr t nazywać będziemy *czasem*. Czas jest z definicji zmienną nielosową, mogącą przybierać wartości liczb naturalnych lub całkowitych (mamy wtedy do czynienia z czasem *dyskretnym*), albo też wartości liczb rzeczywistych (co oznacza, że czas jest *ciągły*).

Każdej wartości parametru t odpowiada K liczb, przy czym K może również być wielkością skońzoną lub nie. Liczby te można interpretować jako realizacje pewnej zmiennej losowej X_t . Indeks t występujący przy zmiennej losowej wskazuje, że postać analityczna rozkładu tej zmiennej lub niektóre parametry tego rozkładu mogą zależeć od czasu t . W naszym pierwszym, a więc najprostszym przykładzie rozkład zmiennej X_t ,

nie zależy od czasu, co oznacza, że na dobrą sprawę indeks t mógłby być pominięty. Otrzymalibyśmy wtedy zmienną losową X rozumianą w dobrze nam znanym sensie. O zmiennych losowych, którymi zajmuje się teoria procesów stochastycznych, zakłada się jednak, że na ogół zależą one w jakiś sposób od t .

Zmienne losowe X_t mogą być dyskretne lub ciągłe. Prowadzi to w sposób naturalny do podziału procesów stochastycznych na następujące cztery klasy:

- | | |
|--|--|
| 1. procesy dyskretne o czasie dyskretnym,
2. procesy ciągłe o czasie dyskretnym
3. procesy dyskretne o czasie ciągłym,
4. procesy ciągłe o czasie ciągłym | } tzw. <i>ciągi losowe</i> ;
} tzw. <i>procesy stochastyczne właściwe</i> . |
|--|--|

Przykładem procesu dyskretnego o czasie dyskretnym może być liczba urodzeń chłopców w pewnej klinice położniczej w ciągu roku, obserwowana w okresie T lat. Ilustracją procesu dyskretnego o czasie ciągłym może być występująca w ewidencji tej kliniki liczba pacjentek znana dla każdego momentu t przez okres T lat. Zużycie prądu przez tę klinikę, odczytywane na liczniku co miesiąc przez T lat, jest przykładem procesu ciągłego o czasie dyskretnym, natomiast wahania napięcia w sieci, odczytywane na wykresie kreślonym przez okres T lat przez wskazówkę woltomierza współpracującego z urządzeniami anestezjologicznymi, jest ilustracją procesu ciągłego o czasie ciągłym.

PRZYKŁAD 2. Brygady spawaczy w stoczni otrzymały nowe typy aparatów spawalniczych i elektrody o zmienionym składzie. Początkowo spowodowało to spadek wydajności, która jednak, w miarę przyzwyczajania się spawaczy do posługiwania się nowym sprzętem, zaczęła szybko wzrastać. Stwierdzono mianowicie, że średnia dzienna wydajność, mierzona w metrach bieżących spawu, a obliczana co dekadę, wykazuje tendencję wzrostową, dającą się dobrze opisać za pomocą funkcji liniowej. Jeżeli znane jest równanie tej funkcji i średni błąd resztkowy (patrz 4.7.1), to interesujący nas proces wzrostu wydajności pracy może być łatwo symulowany za pomocą tzw. *metody Monte Carlo*, polegającej na wykorzystaniu generatorów liczb pseudolosowych do imitowania przebiegu w czasie danego zjawiska gospodarczego. Pokażemy, na czym polega idea takiej symulacji.

Przypuśćmy, że równanie funkcji liniowej opisującej wzrost wydajności pracy spawaczy ma postać

$$f(t) = 5t + 48 ,$$

natomiast niezależny od czasu średni błąd resztkowy $s=4$.

Założymy, że interesujący nas proces stochastyczny może być przedstawiony za pomocą następującej relacji:

$$(1) \quad X_t = f(t) + Z ,$$

gdzie Z jest zmienną losową o rozkładzie $N(0, s)$.

Przymijmy, że chcemy symulować kształtowanie się wydajności pracy czterech robotników: A, B, C, D w ciągu okresu obejmującego 9 dekad, czyli w okresie kwartału. W tym celu generujemy cztery ciągi liczb pseudolosowych o rozkładzie jednostajnym rozpiętym na odcinku $(0, 1)$, przy czym każdy z ciągów zawiera 9 elementów. Do otrzymywania liczb pseudolosowych albo wykorzystujemy odpowiednie tablice, albo odpowiednie generatory, np. generator Fibonacciego.

Oto przykłady takich ciągów:

0,2362	0,4001	0,2555	0,7871
0,1233	0,9200	0,3878	0,5552
0,3595	0,3201	0,6423	0,3423
0,4828	0,2401	0,0291	0,8975
0,8423	0,5602	0,6714	0,2398
0,3251	0,8003	0,7005	0,1373
0,1674	0,3605	0,3719	0,3771
0,4925	0,1608	0,0724	0,5144
0,6599	0,5213	0,3433	0,8915

Za pomocą przekształcenia przedstawionego w przykładzie 3, 7.4.2, znajdujemy wartości zmiennej losowej o rozkładzie normalnym $N(0, 1)$. W tym celu poszczególne liczby pseudolosowe interpretujemy jako wartości dystrybuanty $\Phi(z)$ rozkładu normalnego, a następnie korzystając z odpowiednio dokładnych tablic lub stosując interpolację, znajdujemy wartości zmiennej Z . Wartości te mnożymy przez liczbę $s=4$ otrzymując wartości zmiennej losowej Z o rozkładzie normalnym $N(0, 4)$. Oto te wartości:

-2,88	-1,00	-2,64	3,20
-7,52	5,64	-1,12	0,52
-1,44	-1,88	1,44	-1,64
-0,12	-2,74	-7,56	5,08
4,00	0,68	1,36	-2,84
-1,80	3,36	2,12	-4,36
-3,84	-1,44	-1,32	-1,24
-0,04	-3,96	-5,84	0,16
1,64	0,20	-1,12	4,92

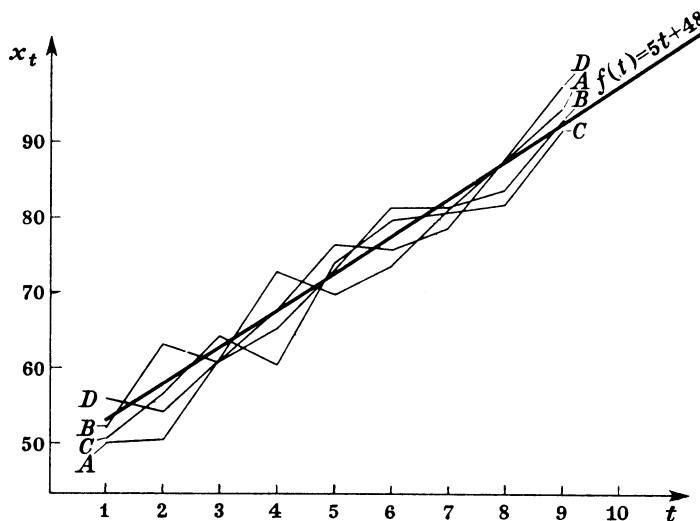
Z kolei znajdujemy wartości funkcji $f(t)$ w momentach czasu $t=1, 2, 3, 4, 5, 6, 7, 8, 9$. Otrzymujemy ciąg liczb:

$$53, 58, 63, 68, 73, 78, 83, 88, 93.$$

Dodając do poszczególnych elementów tego ciągu odpowiednie liczby z poszczególnych kolumn wartości zmiennej Z , otrzymujemy ostatecznie symulowane wartości wydajności pracy spawaczy A, B, C, D . Oto te wartości:

50,12	52,00	50,36	56,20
50,48	63,64	56,88	58,52
61,56	61,12	64,44	61,36
67,88	65,26	60,44	73,08
77,00	73,68	74,36	70,16
76,20	81,36	80,12	73,64
79,16	81,56	81,68	81,76
87,96	84,04	82,16	88,16
94,64	93,20	91,88	97,92

Te cztery ciągi liczb otrzymano za pomocą elektronicznych maszyn cyfrowych, przy czym szybkość generowania liczb jest bardzo duża. Sposób generowania liczb pseudolosowych bywa określany mianem *metody Monte Carlo*. Oto cztery krzywe łamane otrzymane przez połączenie punktów, których współrzędnymi są kolejne momenty czasu i wartości wydajności pracy kolejnych robotników A, B, C, D (patrz rys. 1). Przedstawione na rysunku cztery linie łamane są to tzw. *realizacje procesu stochastycznego*. Gdyby czas był zmienną ciągłą, a nie dyskretną, tak jak to jest w niniejszym przykładzie, realizacje procesu byłyby gładkimi krzywymi, bez załamań. W naszym przykładzie mamy do czynienia tylko z czterema realizacjami procesu, łatwo można sobie jednak wyobrazić sytuację, w których ilość realizacji procesu byłaby bardzo znaczna. Można nawet podać przykłady takich procesów o nieskończonym przeliczalnym lub nieprzeliczalnym zbiorze realizacji.



Rys. 1

Zauważmy, że zbiór wszelkich możliwych wartości realizacji procesu, odpowiadających dowolnemu, ustalonemu momentowi czasu t , może być interpretowany jako zbiór wartości zwykłej, tzn. niezależnej od czasu, zmiennej losowej X . Możemy więc wprowadzić następujące

OKREŚLENIE 1. *Procesem stochastycznym* $\{X_t\}$ nazywa się rodzinę zmiennych losowych X_t , których rozkłady zależą od parametru t , przy czym $t \in T$.

Obok nazwy proces stochastyczny używany bywa również termin *funkcja losowa*, przy czym niektórzy autorzy rezerwują nazwę proces stochastyczny wyłącznie na określenie takich przypadków, w których parametr t jest czasem. Tu nie będziemy stosowali takiego rozróżnienia.

Otrzymaną w rezultacie wykonania k -tego doświadczenia funkcję $x_t(k)$, której wartościami są zrealizowane w k -tym doświadczeniu wartości zmiennych losowych X_t , nazywać będziemy *realizacjami procesu stochastycznego* $\{X_t\}$.

W zasadzie można uważać proces X_t jako dobrze opisany, gdy dla dowolnego skończonego N znana jest N -wymiarowa dystrybuanta $P_{t_1, t_2, \dots, t_N}(x_1, x_2, \dots, x_N)$, przy czym $t_i \in T$, a x_i są to dane liczby rzeczywiste, $i = 1, 2, \dots, N$.

Tak pełna znajomość procesu jest rzadko możliwa. Dlatego też w praktyce stosujemy dwa wybiegi: albo rozpatrujemy pewne wyróżnione klasy procesów o prostych właściwościach, albo też opisując proces – opisujemy tylko niektóre jego własności (ważne, a przy tym dające się stosunkowo łatwo opisać). Często oba te sposoby postępowania stosujemy łącznie.

8.2. CHARAKTERYSTYKI PROCESU STOCHASTYCZNEGO

Do najważniejszych charakterystyk procesu stochastycznego należą: wartość średnia, wariancja i odchylenie standardowe, funkcja koreacyjna i współczynnik korelacji, trend i widmo procesu. Omówimy krótko te pojęcia.

OKREŚLENIE 1. *Wartością średnią procesu $\{X_t\}$* , zwaną także *wartością oczekiwana*, nazywa się wielkość

$$(1) \quad m(t) = E(X_t).$$

Czytelnik zwróci uwagę, że zgodnie z tym określeniem wartość oczekiwana procesu stochastycznego jest funkcją parametru t .

OKREŚLENIE 2. *Wariancją procesu $\{X_t\}$* nazywa się wielkość

$$(2) \quad V(t) = E[X_t - m(t)]^2.$$

Pierwiastek kwadratowy z wariancji procesu będziemy określali zamiennie jednym z dwóch terminów, a mianowicie *błędem standardowym* lub *średnim błędem resztowym procesu* i oznaczali symbolem $\sigma(t)$.

OKREŚLENIE 3. *Funkcją koreacyjną* procesu $\{X_t\}$ nazywa się wielkość

$$(3) \quad R(t_1, t_2) = E[X_{t_1} - m(t_1)][X_{t_2} - m(t_2)].$$

Łatwo dostrzec, że jeżeli $t_1 = t_2 = t$, to

$$(4) \quad R(t_1, t_2) = V(t).$$

OKREŚLENIE 4. *Współczynnikiem korelacji* procesu $\{X_t\}$ jest nazywana wielkość

$$(5) \quad \rho(t_1, t_2) = \frac{R(t_1, t_2)}{\sigma(t_1)\sigma(t_2)}.$$

Oczywiście jeżeli $t_1 = t_2 = t$, to

$$\rho(t, t) = 1.$$

Powróćmy do wzoru (1). Nielosowy składnik $f(t)$ występujący po prawej stronie znaku równości w wyrażeniu

$$X_t = f(t) + Z$$

jest na ogół nieznany. Funkcję tę nazywać będziemy *funkcją trendu hipotetycznego* lub krócej *trendem hipotetycznym* procesu $\{X_t\}$. Ponieważ trend hipotetyczny jest z reguły nieznany, przeto staramy się odgadnąć jego postać w oparciu o empiryczny materiał statystyczny, jakiego dostarczają realizacje procesu. Takie odgadywanie jest niełatwym zadaniem, gdyż dane statystyczne oprócz „czystej” informacji o wartościach trendu, jakie przybiera on w poszczególnych punktach osi czasowej, przekazują nam również „zanieczyszczenie” wywołane działaniem czynnika losowego Z . Niech H oznacza klasę funkcji, której elementy $h(t)$ mają tę własność, że hipoteza, iż zmienna $X_t - h(t)$, $t \in T$, jest zmienną losową o rozkładzie $N(0, s)$, gdzie s – pewna stała, nie może być odrzucona na poziomie α w oparciu o dane eksperymentalne dostarczone przez k realizacji procesu $\{X_t\}$, przy czym α jest bliską zeru liczbą dodatnią. Klasę H nazywać będziemy *klasą empirycznych funkcji trendu*, natomiast funkcje $h(t)$ nazywać będziemy *trendami empirycznymi*. Problem trątnego odgadnięcia, jak w przybliżeniu przebiega krzywa trendu hipotetycznego, jest zagadnieniem z dziedziny *aproksymacji stochastycznej*. Funkcję trendu hipotetycznego $f(t)$, jako funkcję przybliżaną, można określić mianem *funkcji aproksymowanej*, czyli *aproksymaty*, natomiast funkcję trendu empirycznego $h(t)$ przybliżającego funkcję $f(t)$ można nazwać *funkcją aproksymującą*, czyli *aproksymantą*. Istnieje wiele metod wyznaczania trendu empirycznego. Ponieważ do najczęściej stosowanych należą metody analogiczne do tych, jakie się stosuje do wyznaczania parametrów równań linii regresji, a więc np. metoda najmniejszych kwadratów, przeto zagadnienia tego, jako znanego czytelnikowi, nie będziemy omawiać. Umiejętność dobrego aproksymowania trendu $f(t)$ za pomocą funkcji $h(t)$ ma duże znaczenie praktyczne, gdyż umożliwia ona

1^o poznanie i zbadanie, w jaki sposób zmieniał się w przeszłości nielosowy składnik trendu $f(t)$ (jeżeli taki istnieje),

2^o przekształcenie procesu $\{X_t\}$ za pomocą prostej operacji odejmowania $X_t - f(t)$; przekształcenie to często sprowadza proces do postaci stacjonarnej (patrz określenie 1, 8.3),

3^o odgadnięcie przebiegu obserwowanego procesu stochastycznego w przyszłości.

Odgadnięcie to, zwane *prognozą* (patrz [16]), jest możliwe właśnie dlatego, że dzięki sprowadzeniu procesu do postaci stacjonarnej staje się on niezależny od czasu, a tym samym sondaż przyszłości może być dokonany za pomocą zwykłej ekstrapolacji trendu, opisującego, z dokładnością do losowego błędu, przebieg procesu $\{X_t\}$ w teraźniejszości.

8.3. STACJONARNOŚĆ PROCESU

Pojęcie stacjonarności procesu stochastycznego jest tak ważne, że poświęcimy mu nieco więcej uwagi. W rozumieniu potocznym termin „proces stacjonarny” oznacza, że ani kształt, ani parametry rozkładu N -wymiarowej zmiennej losowej

$$(X_{t_1}, X_{t_2}, \dots, X_{t_N}), \quad t_i \in T,$$

nie zależą od czasu. W teorii procesów stochastycznych rozróżnia się pojęcie procesu stacjonarnego w węższym i szerszym sensie. Przytoczymy formalne definicje tych pojęć.

OKREŚLENIE 1. Proces stochastyczny $\{X_t\}$ jest *stacjonarny w węższym sensie*, jeżeli dla każdego N oraz t_1, t_2, \dots, t_N, t takich, że $t_i + t \in T, i=1, 2, \dots, N$, dystrybuanta N -wymiarowej zmiennej losowej

$$(X_{t_1+t}, X_{t_2+t}, \dots, X_{t_N+t})$$

nie zależy od t .

Z określenia tego wynika, że badanie, czy jakiś konkretny, napotkany w praktyce proces stochastyczny jest stacjonarny, czy nie, byłoby na ogół zadaniem trudnym, jeżeli nie wręcz niemożliwym. Właśnie dlatego dla celów praktycznych stosowane jest na ogół

OKREŚLENIE 2. Proces stochastyczny $\{X_t\}$ jest *stacjonarny w szerszym sensie*, jeżeli

$$E(X_t^2) < \infty,$$

$$E(X_t) = m = \text{const.},$$

$$E[(X_{t_1} - m)(X_{t_2} - m)] = R(t_2 - t_1) = R(\tau), \quad \tau = t_2 - t_1.$$

8.4. ERGODYCZNOŚĆ PROCESU

W wielu naukach stosowanych, wśród nich szczególnie w naukach ekonomicznych, bardzo często spotykamy się z sytuacją, kiedy to obserwując jakiś proces stochastyczny nie możemy otrzymać więcej niż jedną realizację procesu. Taka pojedyncza realizacja procesu bywa nazywana *szeregiem statystycznym*, jeżeli czas przybiera wartości dyskretne. Wiele przykładów szeregów statystycznych znaleźć można w każdym roczniku statystycznym. Badanie własności procesu stochastycznego za pomocą jego jednej realizacji jest zadaniem, w ogólnym przypadku, niewykonalnym, gdyż dla każdej ustalonej wartości parametru t dysponujemy tylko jedną realizacją zmiennej losowej X_t , a więc tracą sens takie pojęcia, jak estymowana wartość oczekiwana procesu, estymowana wariancja procesu, czy też jego funkcja korelacyjna. Wyjście z tej niemilej sytuacji pojawia się wówczas, gdy proces ma tzw. własność ergodyczną. Przechodzimy do wprowadzenia formalnej definicji tego podstawowego pojęcia teorii procesów stochastycznych.

Niech $\{X_t\}$ będzie procesem stacjonarnym o czasie t dyskretnym przebiegającym wartości $1, 2, \dots$. Oznaczmy k -tą realizację procesu symbolem $x_i(k)$.

Przyjmijmy, że $T \geq N = nK$. Ponumerujemy kolejne momenty czasu $t=1, 2, \dots, nK$ wskaźnikiem podwójnym i, j , przy czym $i=1, 2, \dots, n$, natomiast $j=1, 2, \dots, K$.

Rozbijamy przedział o długości N na K podprzedziałów i dokonujemy przesunięcia podprzedziałów wzdłuż osi czasowej tak, aby punkty o jednakowej pierwszej składowej podwójnego wskaźnika (i, j) pokrywały się. W taki oto sztuczny sposób z jednej realizacji $x_i(k)$ procesu $\{X_t\}$ rozpiętej na przedziale $\langle 1, N \rangle$ otrzymujemy K realizacji rozpiętych na przedziale $\langle 1, n \rangle$, gdzie $n=N/K$. Średnią arytmetyczną

$$\bar{x}(k) = \frac{1}{n} \sum_{i=1}^n x_i(k)$$

nazwiemy *średnią k -tej realizacji procesu $\{X_t\}$ po czasie*.

Wyobraźmy sobie teraz sytuację, że mamy K realizacji

$$x_t(1), x_t(2), \dots, x_t(K)$$

procesu $\{X_t\}$, z których każda jest rozpięta na przedziale $\langle 1, n \rangle$. Średnią arytmetyczną

$$\bar{x} = \frac{1}{K} \cdot \frac{1}{n} \sum_{t=1}^n \sum_{k=1}^K x_t(k)$$

nazwiemy średnią procesu $\{X_t\}$ po przestrzeni.

OKREŚLENIE 1. Mówimy, że proces $\{X_t\}$ jest *ergodycznym*, jeżeli dla wszystkich $k = 1, 2, \dots, K$, gdy K jest skończone, i dla prawie wszystkich $k = 1, 2, \dots$, gdy K jest nieskończonym,

$$\lim_{n \rightarrow \infty} |\bar{x}(k) - \bar{x}| = 0.$$

Warto podkreślić, że zbadanie w praktyce, czy dany szereg czasowy może być uznany za realizację x_t procesu stochastycznego $\{X_t\}$, jest zadaniem bardzo trudnym, wymagającym wyznaczenia empirycznej funkcji korelacyjnej $R(\tau)$ i badania, czy w miarę jak τ wzrasta, $R(\tau)$ maleje do zera czy też zmierza do stałej różnej od zera. Ponieważ badanie tego typu implikuje dysponowanie szeregiem statystycznym, dotyczącym długiego okresu czasu, co w badaniach gospodarczych, w których dane statystyczne napływają raz w roku (a w najlepszym razie raz w miesiącu), zdarza się rzadko, przeto prowadząc studia nad szeregiem statystycznym nie sprawdza się, czy proces ma własność ergodyczności, lecz po prostu zakłada się, że tak jest. Redukuje to w poważnym stopniu zakres możliwości wykorzystania nowoczesnych metod analizy szeregów czasowych, które nadając się do opisu i studiowania przeszłości mają wielce problematyczną wartość w prognozowaniu przebiegu danego procesu w przyszłości. (O zagadnieniach konstruowania prognozy na podstawie szeregów czasowych patrz np. [16]).

Wiemy, że własność ergodyczności procesu jest dlatego tak pożądana, że umożliwia ona badanie procesu w oparciu o dane statystyczne szeregu czasowego, tzn. tylko jednej realizacji procesu.

I tak dla procesu ciągłego o czasie ciągłym estymatorem parametru $E(X_t) = m$ jest wielkość

$$(1) \quad \hat{m} = \bar{x} = \frac{1}{T} \int_0^T x_t dt,$$

gdzie x_t oznacza daną, pojedynczą realizację procesu $\{X_t\}$.

Oszacowaniem funkcji korelacyjnej $R(\tau)$ jest wielkość

$$(2) \quad R(\tau) = \frac{1}{T-\tau} \int_0^{T-\tau} (x_t - \bar{x})(x_{t+\tau} - \bar{x}) dt.$$

W przypadku procesów ciągłych lub dyskretnych o czasie dyskretnym znaki całek występujących w tych wzorach zamieniają się na znaki sum; otrzymujemy

$$(3) \quad \hat{m} = \bar{x} = \frac{1}{N} \sum_{t=1}^N x_t$$

oraz

$$(4) \quad R(\tau) = \frac{1}{N-\tau} \sum_{t=1}^{N-\tau} (x_t - \bar{x})(x_{t+\tau} - \bar{x}),$$

przy czym $\tau = 1, 2, \dots, N-1$.

Warto podkreślić, że obok dwóch omówionych już pożądanych własności procesów stochastycznych, a mianowicie stacjonarności i ergodyczności, istnieją i inne, równie pożądane. Wymienimy tu na pierwszym miejscu stochastyczną niezależność X_t i $X_{t+\tau}$ dla $t, t+\tau \in T$. W teorii procesów stochastycznych wykazuje się (patrz [10] str. 130) następujące ważne

TWIERDZENIE 1. *Ciąg niezależnych zmiennych losowych o jednakowym rozkładzie jest ergodyczny.*

Ponieważ niezależność stochastyczna jest własnością bardzo pożądaną, ale w praktyce rzadko spotykana, przeto zamiast niezależności zmiennych X_t i $X_{t+\tau}$ trzeba się kontentować własnością mniej cenną, a mianowicie brakiem korelacji między tymi zmiennymi. Gdy badany proces stochastyczny nie spełnia i tego warunku, to oczekujemy, że zachodzi przynajmniej relacja

$$\lim_{\tau \rightarrow \infty} R(\tau) = 0.$$

Doświadczenie uczy, że ta ostatnia własność występuje w praktyce dość często. Przejawia się ona w tym, że empiryczna funkcja korelacyjna obliczona według wzoru (2) lub (4) przyjmuje wartości bliskie zeru dla odpowiednio dużych wartości τ .

Jeżeli funkcja korelacyjna procesu $\{X_t\}$ jest różna od zera, to mówimy, że zachodzi zjawisko *autokorelacji*.

Jeżeli w szczególności proces stochastyczny spełnia relację

$$X_{t+1} = f(X_t, X_{t-1}, \dots, X_{t-m}) + Z_{t+1},$$

gdzie m – pewna dodatnia liczba naturalna, to mówimy, że w procesie stochastycznym występuje zjawisko *autoregresji*. Do najprostszych i najczęściej spotykanych przykładów autoregresji należą stochastyczne równania różnicowe liniowe. Oto takie równanie pierwszego rzędu:

$$X_{t+1} = aX_t + b + Z_{t+1},$$

gdzie a, b stałe parametry, natomiast Z_{t+1} oznacza zmienną losową o rozkładzie $N(0, s)$. Równanie takie generuje proces, jeżeli dana jest wartość początkowa x_0 .

A oto stochastyczne równanie różnicowe drugiego rzędu:

$$X_{t+2} = aX_{t+1} + bX_t + Z_{t+2},$$

przy czym Z_{t+2} jest zmienną losową o rozkładzie $N(0, s)$.

Równanie to generuje proces, jeżeli znane są wartości początkowe x_0 i x_1 . W praktyce wartości początkowe są na ogół równe zeru.

8.5. KANONICZNE ROZWINIĘCIE PROCESU

Rozważmy proces stochastyczny $\{X_t\}$, o którym założymy, że

$$(1) \quad X_t = \sum_{l=1}^L Z_l v_l(t),$$

gdzie Z_l są to niezależne zmienne losowe o wartościach oczekiwanych $E(Z_l)=0$ dla $l=1, 2, \dots, L$, a $v_l(t)$ są to dane nielosowe funkcje zmiennej t . O procesie stochastycznym, który został przedstawiony w postaci kombinacji liniowej (1) nielosowych funkcji $v_l(t)$ z losowymi współczynnikami, mówimy, że został sprowadzony do postaci *kanonicznej*. Funkcja korelacyjna procesu przedstawionego w postaci kanonicznej wyraża się wzorem

$$(2) \quad R(t_1, t_2) = \sum_{l=1}^L v_l(t_1) v_l(t_2) V(Z_l),$$

co wynika bezpośrednio z definicji funkcji korelacyjnej. Symbol $V(Z_l)$ oznacza wariancję zmiennej losowej Z_l . Wyrażenie (2) przedstawia tzw. *kanoniczną* postać funkcji korelacyjnej. Jak widać, znając postać kanoniczną procesu stochastycznego można wyznaczyć postać kanoniczną funkcji korelacyjnej – i na odwrót.

8.6. WIDMO (SPEKTRUM) PROCESU

Przypuśćmy, że dany jest proces stochastyczny $\{X_t\}$, o którym wiadomo, że jest stacjonarny w szerszym sensie, przy czym znana jest funkcja korelacyjna $R(\tau)$ tego procesu. Rozwijając tę funkcję w szereg Fouriera (patrz [19]) w przedziale $-\pi \leq \tau \leq \pi$ otrzymamy

$$(1) \quad R(\tau) = \sum_{n=0}^{\infty} (a_n \cos n\tau + b_n \sin n\tau),$$

gdzie

$$\begin{aligned} a_0 &= \frac{1}{2\pi} \int_{-\pi}^{\pi} R(\tau) d\tau, \\ a_n &= \frac{1}{\pi} \int_{-\pi}^{\pi} R(\tau) \cos n\tau d\tau, \quad n \neq 0, \\ b_n &= \frac{1}{\pi} \int_{-\pi}^{\pi} R(\tau) \sin n\tau d\tau. \end{aligned}$$

Podstawiając za τ we wzorze (1) $\frac{2\pi}{T}\tau$ otrzymamy

$$R(\tau) = 2 \sum_{n=0}^{\infty} (a_n \cos np\tau + b_n \sin np\tau),$$

przy czym $-\frac{1}{2}T \leq \tau \leq \frac{1}{2}T$, $p = \pi/T$ oraz

$$(2) \quad a_0 = \frac{1}{2T} \int_{-T/2}^{T/2} R(\tau) d\tau,$$

$$(3) \quad a_n = \frac{1}{T} \int_{-T/2}^{T/2} R(\tau) \cos np\tau d\tau,$$

$$b_n = \frac{1}{T} \int_{-T/2}^{T/2} R(\tau) \sin np\tau d\tau.$$

Ponieważ, jak to wynika z definicji, funkcja $R(\tau)$ jest parzysta, tzn.

$$R(\tau) = R(-\tau),$$

przeto

$$b_n = \frac{1}{T} \int_{-T/2}^{T/2} R(\tau) \sin np\tau d\tau = 0.$$

Stąd

$$(4) \quad R(\tau) = 2 \sum_{n=0}^{\infty} a_n \cos np\tau.$$

Przypomnijmy sobie, że $\tau = t_2 - t_1$, $t_1, t_2 \in T$.

Stąd

$$\cos np\tau = \cos np(t_2 - t_1) = \cos npt_2 \cos npt_1 + \sin npt_2 \sin npt_1.$$

Wobec tego

$$(5) \quad R(\tau) = \sum_{n=0}^{\infty} (a_n \cos npt_2 \cos npt_1 + a_n \sin npt_2 \sin npt_1).$$

Podstawiając

$$\cos npt_1 \cos npt_2 = v_l(t_1) v_l(t_2)$$

oraz

$$\sin npt_1 \sin npt_2 = v_{l+1}(t_1) v_{l+1}(t_2)$$

i przyjmując

$$(6) \quad a_n = V(Z_l) = V(Z_{l+1}), \quad \text{gdzie } l = 2n,$$

sprowadzamy (5) do (2) z § 8.5, otrzymując kanoniczną postać funkcji korelacyjnej. Stąd łatwo otrzymać postać kanoniczną procesu stochastycznego:

$$(7) \quad X_t - m = \sum_{n=0}^{\infty} (U_n \cos npt + W_n \sin npt),$$

przy czym U_n i W_n są to nieskorelowane zmienne losowe o parametrach

$$(8) \quad E(U_n) = E(W_n) = 0,$$

$$(9) \quad V(U_n) = V(W_n) = a_n.$$

Kanoniczna postać procesu stochastycznego, w której funkcje $v_i(t)$ są funkcjami trygonometrycznymi, nazywa się *spektralnym rozwinięciem procesu stochastycznego*. Zauważmy, że

$$V(X_t - m) = V \left[\sum_{n=0}^{\infty} (U_n \cos npt + W_n \sin npt) \right] = \sum_{n=0}^{\infty} a_n.$$

Jeżeli we wzorze (4) podstawimy

$$np = \omega_n \quad \text{oraz} \quad a_n = \varphi(\omega_n) \Delta \omega_n,$$

gdzie $\Delta \omega_n = p = 2\pi/T$, to przy $T \rightarrow \infty$ i $\Delta \omega_n \rightarrow 0$ otrzymamy

$$(10) \quad R(\tau) = 2 \int_0^{\infty} \varphi(\omega) \cos \omega \tau d\omega.$$

Podstawiając we wzorze (4)

$$\cos \omega_n \tau = \frac{1}{2} (e^{i\omega_n \tau} - e^{-i\omega_n \tau})$$

otrzymamy

$$R(\tau) = 2 \sum_{n=0}^{\infty} \frac{1}{2} a_n (e^{i\omega_n \tau} - e^{-i\omega_n \tau}).$$

Wprowadzając oznaczenie $-\omega_n = \omega_{-n}$ można to zapisać w postaci

$$\begin{aligned} (11) \quad R(\tau) &= \lim_{T \rightarrow \infty} \sum_{n=0}^T a_n e^{i\omega_n \tau} - \lim_{T \rightarrow \infty} \sum_{n=0}^{-T} a_n e^{i\omega_n \tau} = \\ &= \lim_{T \rightarrow \infty} \sum_{n=0}^T \varphi(\omega_n) e^{i\omega_n \tau} \Delta \omega_n - \lim_{T \rightarrow \infty} \sum_{n=0}^{-T} \varphi(\omega_n) e^{i\omega_n \tau} \Delta \omega_n = \\ &= \int_0^{\infty} \varphi(\omega) e^{i\omega \tau} d\omega - \int_0^{-\infty} \varphi(\omega) e^{i\omega \tau} d\omega = \int_{-\infty}^{\infty} \varphi(\omega) e^{i\omega \tau} d\omega = \\ &= \int_{-\infty}^{\infty} e^{i\omega \tau} d\Phi(\omega). \end{aligned}$$

Funkcja $\varphi(\omega)$ nazywa się *gęstością widmową (spektralną)*, natomiast funkcja $\Phi(\omega)$ nosi nazwę *funkcji widmowej (spektralnej)*.

Porównując wzór (11) ze wzorem (3), § 4.6, czytelnik dostrzeże z łatwością, że $R(\tau)$ można interpretować jako funkcję charakterystyczną. Na tym polega między innymi znaczenie analizy spektralnej, że umożliwia ona znajdowanie funkcji korelacyjnej, gdy znana jest funkcja widmowa – i na odwrót – znajdowanie funkcji widmowej, gdy znana

jest funkcja koreacyjna. Ta ostatnia transformacja odbywa się za pomocą wzoru

$$(12) \quad \Phi'(\omega) = \varphi(\omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} R(\tau) e^{-i\omega\tau} d\tau$$

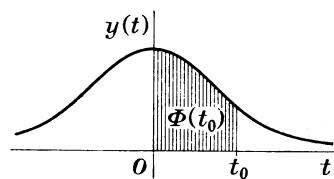
(porównaj wzór (27) z 4.6).

Zwracamy uwagę czytelnika, że analiza spektralna procesów stochastycznych ma znikome znaczenie w zastosowaniach ekonomicznych. Korzysta się z niej głównie w zastosowaniach fizycznych i technicznych, gdzie stanowi ona potężne narzędzie badania przebiegu zjawisk w czasie. W badaniach ekonomicznych do opisu i analizy procesów stochastycznych używa się metod korelacji i regresji oraz specjalnych metod, jak np. przedstawione w pracy [16].

TABLICE

Tablica I
 ROZKŁAD NORMALNY

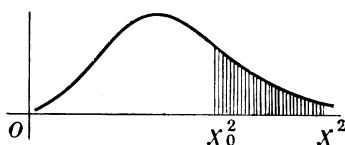
$$\Phi(t) = \frac{1}{\sqrt{2\pi}} \int_0^t e^{-v^2/2} dv$$



t	$\Phi(t)$	t	$\Phi(t)$	t	$\Phi(t)$
0,00	0,0000	1,00	0,3413	2,00	0,4773
0,05	0,0199	1,05	0,3531	2,05	0,4798
0,10	0,0398	1,10	0,3643	2,10	0,4821
0,15	0,0596	1,15	0,3749	2,15	0,4842
0,20	0,0793	1,20	0,3849	2,20	0,4861
0,25	0,0987	1,25	0,3944	2,25	0,4878
0,30	0,1179	1,30	0,4032	2,30	0,4893
0,35	0,1368	1,35	0,4115	2,35	0,4906
0,40	0,1554	1,40	0,4192	2,40	0,4918
0,45	0,1736	1,45	0,4265	2,45	0,4929
0,50	0,1915	1,50	0,4332	2,50	0,4938
0,55	0,2088	1,55	0,4394	2,55	0,4946
0,60	0,2257	1,60	0,4452	2,60	0,4953
0,65	0,2422	1,65	0,4505	2,65	0,4960
0,70	0,2580	1,70	0,4554	2,70	0,4965
0,75	0,2734	1,75	0,4599	2,75	0,4970
0,80	0,2881	1,80	0,4641	2,80	0,4974
0,85	0,3023	1,85	0,4678	2,85	0,4978
0,90	0,3159	1,90	0,4713	2,90	0,4981
0,95	0,3289	1,95	0,4744	2,95	0,4984
				3,00	0,4987

Tablica II
ROZKŁAD χ^2

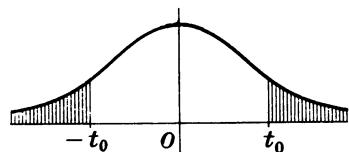
$$P\{\chi^2 > \chi_0^2\} = \int_{\chi_0^2}^{\infty} \frac{x^{k/2-1} e^{-x/2}}{\Gamma(\frac{1}{2}k) 2^{k/2}} dx$$



k	χ^2 jako funkcja k i P								
	P=0,99	0,95	0,90	0,80	0,50	0,20	0,10	0,05	0,01
1	0,000	0,004	0,016	0,064	0,455	1,642	2,706	3,841	6,635
2	0,020	0,103	0,211	0,446	1,386	3,219	4,605	5,991	9,210
3	0,115	0,352	0,584	1,005	2,366	4,642	6,251	7,815	11,341
4	0,297	0,711	1,064	1,649	3,357	5,989	7,779	9,488	13,277
5	0,554	1,145	1,610	2,343	4,351	7,289	9,236	11,070	15,086
6	0,872	1,635	2,204	3,070	5,348	8,558	10,645	12,592	16,812
7	1,239	2,167	2,833	3,822	6,346	9,803	12,017	14,067	18,475
8	1,646	2,733	3,490	4,594	7,344	11,030	13,362	15,507	20,090
9	2,088	3,325	4,168	5,380	8,343	12,242	14,684	16,919	21,666
10	2,558	3,940	4,865	6,179	9,342	13,442	15,987	18,307	23,209
11	3,053	4,575	5,578	6,989	10,341	14,631	17,275	19,675	24,725
12	3,571	5,226	6,304	7,807	11,340	15,812	18,549	21,026	26,217
13	4,107	5,892	7,042	8,634	12,340	16,985	19,812	22,362	27,688
14	4,660	6,571	7,790	9,467	13,339	18,151	21,064	23,685	29,141
15	5,229	7,261	8,547	10,307	14,339	19,311	22,307	24,996	30,578
16	5,812	7,962	9,312	11,152	15,338	20,465	23,542	26,296	32,000
17	6,408	8,672	10,085	12,002	16,338	21,615	24,769	27,587	33,409
18	7,015	9,390	10,865	12,857	17,338	22,760	25,989	28,869	34,805
19	7,633	10,117	11,651	13,716	18,338	23,900	27,204	30,144	36,191
20	8,260	10,851	12,443	14,578	19,337	25,038	28,412	31,410	37,566
21	8,897	11,591	13,240	15,445	20,337	26,171	29,615	32,671	38,932
22	9,542	12,338	14,041	16,314	21,337	27,301	30,813	33,924	40,289
23	10,196	13,091	14,848	17,187	22,337	28,429	32,007	35,172	41,638
24	10,856	13,848	15,659	18,062	23,337	29,553	33,196	36,415	42,980
25	11,524	14,611	16,473	18,940	24,337	30,657	34,382	37,652	44,314
26	12,198	15,379	17,292	19,820	25,336	31,795	35,563	38,885	45,642
27	12,879	16,151	18,114	20,703	26,336	32,912	36,741	40,113	46,963
28	13,565	16,928	18,939	21,588	27,336	34,027	37,916	41,337	48,278
29	14,256	17,708	19,768	22,475	28,336	35,139	39,087	42,557	49,588
30	14,953	18,493	20,599	23,364	29,336	36,250	40,256	43,773	50,892

ROZKŁAD STUDENTA

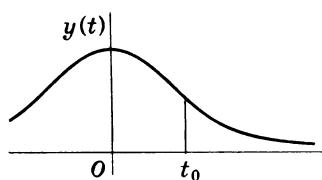
$$P\{|t| > t_0\} = 2 \int_{t_0}^{\infty} \frac{\Gamma(\frac{1}{2}(k+1))}{\Gamma(\frac{1}{2}k) \sqrt{k\pi}} \left(1 + \frac{v^2}{k}\right)^{-(k+1)/2} dv$$

**Tablica III**

k	t jako funkcja k i P								
	P=0,90	0,80	0,70	0,50	0,30	0,20	0,10	0,02	0,01
1	0,158	0,325	0,510	1,000	1,963	3,078	6,314	31,821	63,657
2	0,142	0,289	0,445	0,816	1,386	1,886	2,920	6,965	9,925
3	0,137	0,277	0,424	0,765	1,250	1,638	2,353	4,541	5,841
4	0,134	0,271	0,414	0,741	1,190	1,533	2,132	3,747	4,604
5	0,132	0,267	0,408	0,727	1,156	1,476	2,015	3,365	4,032
6	0,131	0,265	0,404	0,718	1,134	1,440	1,943	3,143	3,707
7	0,130	0,263	0,402	0,711	1,119	1,415	1,895	2,998	3,499
8	0,130	0,262	0,399	0,706	1,108	1,397	1,860	2,896	3,355
9	0,129	0,261	0,398	0,703	1,100	1,383	1,833	2,821	3,250
10	0,129	0,260	0,397	0,700	1,093	1,372	1,812	2,764	3,169
11	0,129	0,260	0,396	0,697	1,088	1,363	1,796	2,718	3,106
12	0,128	0,259	0,395	0,695	1,083	1,356	1,782	2,681	3,055
13	0,128	0,259	0,394	0,694	1,079	1,350	1,771	2,650	3,012
14	0,128	0,258	0,393	0,692	1,076	1,345	1,761	2,624	2,977
15	0,128	0,258	0,393	0,691	1,074	1,341	1,753	2,602	2,947
16	0,128	0,258	0,392	0,690	1,071	1,337	1,746	2,583	2,921
17	0,128	0,257	0,392	0,689	1,069	1,333	1,740	2,567	2,898
18	0,127	0,257	0,392	0,688	1,067	1,330	1,734	2,552	2,878
19	0,127	0,257	0,391	0,688	1,066	1,328	1,729	2,539	2,861
20	0,127	0,257	0,391	0,687	1,064	1,325	1,725	2,528	2,845
21	0,127	0,257	0,391	0,686	1,063	1,323	1,721	2,518	2,831
22	0,127	0,256	0,390	0,686	1,061	1,321	1,717	2,508	2,819
23	0,127	0,256	0,390	0,685	1,060	1,319	1,714	2,500	2,807
24	0,127	0,256	0,390	0,685	1,059	1,318	1,711	2,492	2,797
25	0,127	0,256	0,390	0,684	1,058	1,316	1,708	2,485	2,787
26	0,127	0,256	0,390	0,684	1,058	1,315	1,706	2,479	2,779
27	0,127	0,256	0,389	0,684	1,057	1,314	1,703	2,473	2,771
28	0,127	0,256	0,389	0,683	1,056	1,313	1,701	2,467	2,763
29	0,127	0,256	0,389	0,683	1,055	1,311	1,699	2,462	2,756
30	0,127	0,256	0,389	0,683	1,055	1,310	1,697	2,457	2,750

ROZKŁAD NORMALNY
(funkcja gęstości)

$$\varphi(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$$



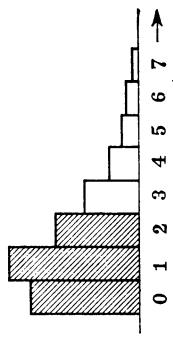
<i>t</i>	0	2	4	6	8
0,0	0,3989	0,3989	0,3986	0,3982	0,3977
0,1	0,3970	0,3961	0,3951	0,3939	0,3925
0,2	0,3910	0,3894	0,3876	0,3857	0,3836
0,3	0,3814	0,3790	0,3765	0,3739	0,3712
0,4	0,3683	0,3653	0,3621	0,3589	0,3555
0,5	0,3521	0,3485	0,3443	0,3410	0,3372
0,6	0,3332	0,3292	0,3251	0,3209	0,3166
0,7	0,3123	0,3079	0,3034	0,2989	0,2943
0,8	0,2897	0,2850	0,2803	0,2756	0,2709
0,9	0,2661	0,2613	0,2565	0,2516	0,2468
1,0	0,2420	0,2371	0,2323	0,2275	0,2227
1,1	0,2179	0,2131	0,2083	0,2036	0,1989
1,2	0,1942	0,1895	0,1849	0,1804	0,1758
1,3	0,1714	0,1669	0,1626	0,1582	0,1539
1,4	0,1497	0,1456	0,1415	0,1374	0,1334
1,5	0,1295	0,1257	0,1219	0,1182	0,1145
1,6	0,1109	0,1074	0,1040	0,1006	0,0973
1,7	0,0940	0,0909	0,0878	0,0848	0,0818
1,8	0,0790	0,0761	0,0734	0,0707	0,0681
1,9	0,0656	0,0632	0,0608	0,0584	0,0562
2,0	0,0540	0,0519	0,0498	0,0478	0,0459
2,1	0,0440	0,0422	0,0404	0,0387	0,0371
2,2	0,0355	0,0339	0,0325	0,0310	0,0297
2,3	0,0283	0,0270	0,0258	0,0246	0,0235
2,4	0,0224	0,0213	0,0203	0,0194	0,0184
2,5	0,0175	0,0167	0,0158	0,0151	0,0143
2,6	0,0136	0,0129	0,0122	0,0116	0,0110
2,7	0,0104	0,0099	0,0093	0,0089	0,0084
2,8	0,0079	0,0075	0,0071	0,0068	0,0063
2,9	0,0060	0,0056	0,0053	0,0050	0,0047
3,0	0,0044	0,0042	0,0039	0,0037	0,0035

Tablica IV

ROZKŁAD POISSONA

Tablica V

Tablice



$$P(x \leq k) = \sum_{r=0}^k \frac{(np)^r e^{-np}}{r!}$$

$np \setminus k$	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0,5	0,607	0,910	0,986	0,998												
0,6	0,549	0,878	0,977	0,997												
0,7	0,497	0,844	0,966	0,994	0,999											
0,8	0,449	0,809	0,953	0,991	0,999											
0,9	0,407	0,772	0,937	0,987	0,998											
1,0	0,368	0,736	0,920	0,981	0,996	0,999										
1,2	0,301	0,663	0,879	0,966	0,992	0,998										
1,4	0,247	0,592	0,833	0,946	0,986	0,997	0,999									
1,6	0,202	0,525	0,783	0,921	0,976	0,994	0,999									
1,8	0,165	0,463	0,731	0,891	0,964	0,990	0,997	0,999								
2,0	0,135	0,406	0,677	0,857	0,947	0,983	0,995	0,999								
2,2	0,111	0,355	0,623	0,819	0,928	0,975	0,993	0,998								
2,4	0,091	0,308	0,570	0,779	0,904	0,964	0,988	0,997	0,999							
2,6	0,074	0,267	0,518	0,736	0,877	0,951	0,983	0,995	0,999							
2,8	0,061	0,231	0,469	0,692	0,848	0,935	0,976	0,992	0,998	0,999						
3,0	0,050	0,199	0,423	0,647	0,815	0,916	0,966	0,988	0,996	0,999						
3,5	0,030	0,136	0,321	0,537	0,725	0,858	0,935	0,973	0,990	0,996	0,999					
4,0	0,018	0,092	0,238	0,433	0,629	0,785	0,889	0,949	0,979	0,992	0,997					
5,0	0,007	0,040	0,125	0,265	0,440	0,616	0,762	0,867	0,932	0,968	0,986	0,995	0,998			
6,0	0,002	0,017	0,062	0,151	0,285	0,446	0,606	0,744	0,847	0,916	0,957	0,980	0,991	0,999		

Tablica VI

ROZKŁAD GRANICZNY D_n KOŁMOGOROWA-SMIRNOWA

$$Q(\lambda) = \lim_{n \rightarrow \infty} P\left(D_n < \frac{\lambda}{\sqrt{n}}\right)$$

λ	Q	λ	Q	λ	Q	λ	Q
0,38	0,001	0,83	0,504	1,28	0,925	1,73	0,995
0,39	0,002	0,84	0,519	1,29	0,928	1,74	0,995
0,40	0,003	0,85	0,535	1,30	0,932	1,75	0,996
0,41	0,004	0,86	0,550	1,31	0,935	1,76	0,996
0,42	0,005	0,87	0,565	1,32	0,939	1,77	0,996
0,43	0,007	0,88	0,579	1,33	0,942	1,78	0,996
0,44	0,010	0,89	0,593	1,34	0,945	1,79	0,997
0,45	0,013	0,90	0,607	1,35	0,948	1,80	0,997
0,46	0,016	0,91	0,621	1,36	0,951	1,81	0,997
0,47	0,020	0,92	0,634	1,37	0,953	1,82	0,997
0,48	0,025	0,93	0,647	1,38	0,956	1,83	0,998
0,49	0,030	0,94	0,660	1,39	0,958	1,84	0,998
0,50	0,036	0,95	0,673	1,40	0,960	1,85	0,998
0,51	0,043	0,96	0,685	1,41	0,962	1,86	0,998
0,52	0,050	0,97	0,696	1,42	0,965	1,87	0,998
0,53	0,059	0,98	0,708	1,43	0,967	1,88	0,998
0,54	0,067	0,99	0,719	1,44	0,968	1,89	0,998
0,55	0,077	1,00	0,730	1,45	0,970	1,90	0,999
0,56	0,088	1,01	0,741	1,46	0,972	1,91	0,999
0,57	0,099	1,02	0,751	1,47	0,973	1,92	0,999
0,58	0,110	1,03	0,761	1,48	0,975	1,93	0,999
0,59	0,123	1,04	0,770	1,49	0,976	1,94	0,999
0,60	0,136	1,05	0,780	1,50	0,978	1,95	0,999
0,61	0,149	1,06	0,789	1,51	0,979	1,96	0,999
0,62	0,163	1,07	0,798	1,52	0,980	1,97	0,999
0,63	0,178	1,08	0,806	1,53	0,981	1,98	0,999
0,64	0,193	1,09	0,814	1,54	0,983	1,99	0,999
0,65	0,208	1,10	0,822	1,55	0,984	2,00	0,999
0,66	0,224	1,11	0,830	1,56	0,985	2,01	0,999
0,67	0,240	1,12	0,837	1,57	0,986	2,02	0,999
0,68	0,256	1,13	0,845	1,58	0,986	2,03	0,999
0,69	0,272	1,14	0,851	1,59	0,987	2,04	1,000
0,70	0,289	1,15	0,858	1,60	0,988	2,05	1,000
0,71	0,305	1,16	0,864	1,61	0,989	2,06	1,000
0,72	0,322	1,17	0,871	1,62	0,989	2,07	1,000
0,73	0,339	1,18	0,877	1,63	0,990	2,08	1,000
0,74	0,356	1,19	0,882	1,64	0,991	2,09	1,000
0,75	0,373	1,20	0,888	1,65	0,991	2,10	1,000
0,76	0,390	1,21	0,893	1,66	0,992	2,11	1,000
0,77	0,406	1,22	0,898	1,67	0,992	2,12	1,000
0,78	0,423	1,23	0,903	1,68	0,993	2,13	1,000
0,79	0,440	1,24	0,908	1,69	0,993	2,14	1,000
0,80	0,456	1,25	0,912	1,70	0,994	2,15	1,000
0,81	0,472	1,26	0,916	1,71	0,994	2,16	1,000
0,82	0,488	1,27	0,921	1,72	0,995	2,17	1,000

Tablica VII

ROZKŁAD K

$$P_k(n, n) = \sum_{s=0}^k \binom{n}{s} \sum_{r=0}^{n-s} (-1)^r \binom{n-s}{r} \left(1 - \frac{s+r}{n}\right)^n$$

n	$\alpha=0,1$		$\alpha=0,05$		$\alpha=0,01$		$\alpha=0,005$	
	k_1	k_2	k_1	k_2	k_1	k_2	k_1	k_2
2	—	—	—	—	—	—	—	—
3	—	—	—	—	—	—	—	—
4	0	2	—	2	—	—	—	—
5	0	2	0	3	—	3	—	3
6	0	3	0	3	—	4	—	4
7	0	3	0	4	0	4	—	4
8	1	4	0	4	0	5	0	5
9	1	4	1	5	0	5	0	5
10	1	5	1	5	0	6	0	6
11	2	5	1	6	1	6	0	6
12	2	6	2	6	1	7	1	7
13	2	6	2	6	1	7	1	7
14	2	6	2	7	1	8	1	8
15	3	7	2	7	2	8	1	8
16	3	7	3	8	2	9	2	9
17	3	8	3	8	2	9	2	9
18	4	8	3	9	2	9	2	10
19	4	9	4	9	3	10	2	10
20	4	9	4	9	3	10	3	10
21	5	9	4	10	3	11	3	11
22	5	10	5	10	4	11	3	12
23	5	10	5	11	4	12	4	12
24	6	11	5	11	4	12	4	13
25	6	11	5	12	4	13	4	13
26	6	11	6	12	5	13	4	13
27	7	12	6	12	5	13	5	14
28	7	12	6	13	5	14	5	14
29	7	13	6	13	6	14	5	15
30	8	13	7	14	6	15	6	15

Tablica VIII
ROZKŁAD SERII⁽¹⁾

Tablica podaje dla prawdopodobieństw $\alpha=0,025$, $\alpha=0,05$, $\alpha=0,095$, $\alpha=0,975$ takie największe liczby całkowite u_α , dla których $P(U \leq u_\alpha) \leq \alpha$, jeżeli $\alpha < 0,5$ oraz takie najmniejsze liczby całkowite u_α , dla których $P(U \geq u_\alpha) \geq \alpha$, jeżeli $\alpha > 0,5$, gdzie U jest liczbą serii w ciągu zdarzeń losowych o elementach A i B , przy czym liczba elementów A wynosi n_1 , a liczba elementów B wynosi n_2 .

$$P(U \leq u_{0,025}) = 0,025$$

$n_1 \backslash n_2$	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
4																			
5			2	2															
6			2	2	3	3													
7			2	2	3	3	3												
8			2	3	3	3	4	4											
9			2	3	3	4	4	5	5										
10			2	3	3	4	5	5	5	5	6								
11			2	3	4	4	5	5	5	6	6	6	7						
12			2	2	3	4	4	5	6	6	6	7	7	7	7				
13			2	2	3	4	5	5	6	6	6	7	7	8	8	8			
14			2	2	3	4	5	5	6	7	7	8	8	8	9	9	9		
15			2	3	3	4	5	6	6	7	7	8	8	8	9	9	9	10	
16			2	3	4	4	5	6	6	7	8	8	8	9	9	10	10	11	
17			2	3	4	4	5	6	7	7	8	9	9	9	10	10	11	11	11
18			2	3	4	5	5	6	7	8	8	8	9	9	10	10	11	11	12
19			2	3	4	5	6	6	7	8	8	9	10	10	11	11	12	12	13
20			2	3	4	5	6	6	7	8	9	9	10	10	11	12	12	13	13

$$P(U \leq u_{0,05}) = 0,05$$

$n_1 \backslash n_2$	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
3																			
4				2															
5			2	2	3														
6			2	3	3	3													
7			2	3	3	4	4												
8			2	2	3	3	4	4	5										
9			2	2	3	4	4	5	5	6									
10			2	3	3	4	5	5	6	6	6	6	6						
11			2	3	3	4	5	5	6	6	6	7	7	7					
12			2	3	4	4	5	6	6	7	7	7	8	8					
13			2	3	4	4	5	6	6	7	8	8	8	9	9				
14			2	3	4	5	5	6	7	7	8	8	8	9	9	9	10		
15			2	3	4	5	6	6	7	8	8	8	9	9	10	10	10	11	
16			2	3	4	5	6	6	7	8	8	8	9	10	10	11	11	11	
17			2	3	4	5	6	7	7	8	9	9	10	10	11	11	12	12	
18			2	3	4	5	6	7	8	8	8	9	10	10	11	11	12	12	
19			2	3	4	5	6	7	8	8	8	9	10	10	11	12	12	13	
20			2	3	4	5	6	7	8	9	9	10	11	11	12	12	13	13	14

(¹) Sadowski, W. i inni, *Tablice statystyczne*, Warszawa 1957.

Tablica VIII (cd.)

$$P(U < u_{0.95}) = 0.95$$

$n_2 \backslash n_1$	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
n_2	2	4																	
3	5	6																	
4	5	6	7																
5	5	7	8	8															
6	5	7	8	9	10														
7	5	7	8	9	10	11													
8	5	7	9	10	11	12	12												
9	5	7	9	10	11	12	13	13											
10	5	7	9	10	11	12	13	13	14										
11	5	7	9	11	12	13	14	14	15	15									
12	5	7	9	11	12	13	14	15	15	16	16								
13	5	7	9	11	12	13	14	15	15	16	17	17							
14	5	7	9	11	12	13	15	15	16	16	17	18	19						
15	5	7	9	11	13	14	15	15	16	17	18	18	19	20					
16	5	7	9	11	13	14	15	15	16	17	18	19	20	20	21				
17	5	7	9	11	13	14	15	17	17	18	19	20	20	21	21	22			
18	5	7	9	11	13	14	15	17	17	18	19	20	20	21	22	23	23		
19	5	7	9	11	13	14	15	17	17	18	19	20	21	22	23	24	24		
20	5	7	9	11	13	14	16	17	18	19	20	21	21	22	23	24	25		

$$P(U < u_{0.975}) = 0.975$$

$n_2 \backslash n_1$	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
n_2	2	4																	
3	5	6																	
4	5	7	8																
5	5	7	8	9															
6	5	7	8	9	10														
7	5	7	9	10	11	12													
8	5	7	9	10	11	12	13												
9	5	7	9	11	12	13	13	14											
10	5	7	9	11	12	13	14	15	15										
11	5	7	9	11	12	13	14	15	15	16									
12	5	7	9	11	12	13	15	15	15	16	17								
13	5	7	9	11	13	14	15	16	17	18	18	18							
14	5	7	9	11	13	14	15	16	17	17	18	19	19						
15	5	7	9	11	13	14	15	17	17	18	19	20	20	21					
16	5	7	9	11	13	15	16	17	18	19	20	20	21	22					
17	5	7	9	11	13	15	16	17	18	19	20	21	22	22	23	24			
18	5	7	9	11	13	15	16	17	18	19	20	21	22	23	24	24	25		
19	5	7	9	11	13	15	16	17	19	20	21	22	22	23	24	25	25		
20	5	7	9	11	13	15	16	17	19	20	21	22	23	24	24	25	26		

LITERATURA

- [1] Cramér, H., *Metody matematyczne w statystyce*, Warszawa 1958.
- [2] Czechowski, T., *Elementarny wykład rachunku prawdopodobieństwa*, Warszawa 1970.
- [3] Czechowski, T., *Rachunek różniczkowy i całkowy dla ekonomistów i statystyków*, Warszawa 1971.
- [4] Dlin, A. M. (Длин, А. М.), *Математическая статистика в технике*, Москва 1958.
- [5] Dunin-Barkowski, I. W., Smirnow, N. W. (Дунин-Барковский, И. В., Смирнов, Н. В.), *Теория вероятностей и математическая статистика в технике*, Москва 1955.
- [6] Feller, W., *Wstęp do rachunku prawdopodobieństwa i jego zastosowań* tom I, Warszawa 1966, tom II, Warszawa 1969.
- [7] Fichtenholz, G. M., *Rachunek różniczkowy i całkowy* tom I, Warszawa 1965, tom II, Warszawa 1969, tom III, Warszawa 1969.
- [8] Fisz, M., *Rachunek prawdopodobieństwa i statystyka matematyczna*, Warszawa 1969.
- [9] Gersternkorn, T., Śródka, T., *Ćwiczenia z kombinatoryki i rachunku prawdopodobieństwa*, Łódź 1968.
- [10] Gichman, I. I., Skorochod, A. W., *Wstęp do teorii procesów stochastycznych*, Warszawa 1968.
- [11] Gniedenko, B. W. (Гнеденко, Б. В.), *Курс теории вероятностей*, Москва 1954.
- [12] Greń, J., *Modele i zadania statystyki matematycznej*, Warszawa 1972.
- [13] Hald, A., *Statistical theory with engineering applications*, New York-London 1952.
- [14] Hannan, E. J., *Time series analysis*, London.
- [15] Hellwig, Z., *Regresja liniowa i jej zastosowanie w ekonomii*, Warszawa 1960.
- [16] Hellwig, Z., *Schemat budowy prognozy statystycznej metodą wag harmonicznych*, Przegląd statystyczny 2, XIV (1967).
- [17] Kendall, M. G., *The advanced theory of statistics*, London 1948.
- [18] Kuratowski, K., *Wstęp do teorii mnogości i topologii*, Warszawa 1972.
- [19] Leja, F., *Rachunek różniczkowy i całkowy ze wstępem do równań różniczkowych*, Warszawa 1971.
- [20] Mills, F. C., *Statistical methods*, New York.
- [21] Mostowski, A., Stark, M., *Elementy algebry wyższej*, Warszawa 1970.
- [22] Obalski, J., *Statystyczna kontrola jakości podczas produkcji*, Warszawa 1955.
- [23] Oderfeld, J., *Zarys statystycznej kontroli jakości*, Warszawa 1954.
- [24] Pawłowski, Z., *Wstęp do statystyki matematycznej*, Warszawa 1965.
- [25] Renyi, A., *Podstawowe problemy rachunku prawdopodobieństwa*, Myśl filozoficzna. Przekłady 34 (1955).
- [26] Rosenblatt, M., *Procesy stochastyczne*, Warszawa 1967.
- [27] Sadowski, W., *Statystyka matematyczna*, Warszawa 1965.
- [28] Sadowski, W., i inni, *Tablice statystyczne*, Warszawa 1957.
- [29] Sierpiński, W., *Wstęp do teorii mnogości i topologii*, Warszawa 1947.
- [30] Skrzywan, W., *Historia statystyki*, Warszawa 1951.
- [31] Swiesznikow, A. A., *Podstawowe metody funkcji losowych*, Warszawa 1965.
- [32] Szulc, S., *Metody statystyczne*, Warszawa 1963.
- [33] Wentzel, E. S. (Венцель, Е. С.), *Teoria вероятностей*, Москва 1962.

- [34] Wołodin, B. G., Ganin, M. P., Diner, I. J., Komarow, L. B., Swiesznikow, A. A., Starobin, K. B., *Problemy rachunku prawdopodobieństwa*, Warszawa 1966.
- [35] Zasępa, T., *Badania statystyczne metodą reprezentacyjną. Zarys teorii i praktyki*, Warszawa 1962.
- [36] Zubrzycki, S., *Wykłady z rachunku prawdopodobieństwa i statystyki matematycznej*, Warszawa 1970.
- [37] *Rocznik Statystyczny* 1955, Warszawa 1956.
- [38] *Rocznik Statystyczny* 1965, Warszawa 1966.

SKOROWIDZ

Aproksymacja stochastyczna 284
aproksymanta 284
aproksymata 284
autokorelacja 287
autoregresja 287

Badanie całkowite (wyczerpujące) 188
– częściowe (niewyczerpujące) 188
błąd losowy 133, 248
– obserwacji 178
– oceny dopuszczalny 220
– – standardowy 134, 226
– resztkowy 133
– – średni 134, 283
– rodzaju I 259
– – II 259
– standardowy 283
borełowskie ciało zbiorów 28
– – zdarzeń 39

Ciąka Eulera rodzaju I 24
– – – II 22
– Poissona 23

cecha mierzalna 185
– niemierzalna 185
– statystyczna 185

ciało borełowskie zdarzeń 39
– przeliczalnie addytywne 28
– zbiorów 28
– – borełowskie 28

ciąg Fibonacciego 270
– losowy 280

czas 279
– ciągły 279
– dyskretny 279

częstość klasy 186
– skumulowana 186
– względna 186
– zdarzenia 46

Diagram Eulera 38

długość przedziału 29
– serii 274

dopełnienie zbioru 27

dwumian Newtona 20
dystrybuanta 67, 95, 152
– brzegowa 95

Efektywność estymatora 196
– – asymptotyczna 197

elipsa jednakowej gęstości prawdopodobieństwa 145

ergodyczność procesu stochastycznego 286

estymator 194
– asymptotycznie najefektywniejszy 197, 239
– macierzowy 239
– najefektywniejszy 195, 239
– nieobciążony 195, 239
– parametru 194
– wektorowy 239
– zgodny 195, 239

Frakcja braków 191

funkcja beta 24
– charakterystyczna 118
– gamma 22
– koreacyjna 283, 288
– losowa 282
– segmentowa 242
– spektralna 290
– trendu hipotetycznego 284
– widmowa 290
– zbioru nieujemna 29

Gęstość prawdopodobieństwa 80, 95, 152
– spektralna 290
– widmowa 290
– w rozkładzie beta 89
– – – gamma 87
– – – normalnym 83
– – – trójkątnym 82

granica przedziału klasowego 186

Hipoteza alternatywna 257
– continuum 26
– nieparametryczna 253
– parametryczna 253
– statystyczna 253
– zerowa 254

- Iloczyn zbiorów 27
 - zdarzeń 36
- implikacja 37
- inkluzja 26
- Jednostka statystyczna 185
- Klasa empirycznych funkcji trendu 284
 - szeregu rozdzielczego 186
 - zbiorów 26
- kombinacja 14
 - z powtórzeniami 17
- korelacja dodatnia 136
 - ujemna 136
 - wielowymiarowa (wielokrotna, wieloraka) 158
- kowariancja 127
 - w próbce 226
- kres zbioru dolny 28
 - – górnny 28
- krzywa mocy testu 262
- Liczba kardynalna 26
 - pseudolosowa 270
 - stopni swobody 215
 - złota 274
- liczebność 186
 - hipotetyczna 264
- linia regresji rodzaju I 129
 - – – II 130
- losowanie bez zwracania 52
 - niezależne 153
 - ze zwracaniem 14, 52
- Metoda maksimum wiarygodności 223
 - minimalnej wariancji 223
 - Monte Carlo 280, 282
 - najmniejszych kwadratów 223
 - punktowa 223
 - reprezentacyjna 191
- miara 29
 - nieparametryczna 139
 - parametryczna 139
 - rozbieżności między rozkładami 266
 - skończona 29
 - unormowana 29
- mnogość 25
- moc testu 262
 - zbioru 26
- moment 116
 - absolutny 116, 126
 - centralny 117, 126
 - mieszany 126
 - warunkowy 128
 - względny 116, 125
- moment zwykły 116, 126
- Nadzieja matematyczna 99
- negacja 34, 37
- nierówność Czebyszewa 168
 - Rao-Craméra 196
- niezależność stochastyczna 153
 - zmiennych losowych 95
- Obserwacja statystyczna 186
- odchylenie przeciętne 99, 115
 - standardowe 99, 108
 - wartości zmiennej losowej 107
 - w próbce 217
- Parametr opisowy 99
 - regresji 131
- permutacja 11
 - z powtórzeniami 14, 16
- populacja dwuwymiarowa 92
 - generalna 14, 58, 188, 224
 - próbna 58, 188, 224
 - statystyczna 185
- poziom istotności 255
 - ufności 212
- prawdopodobieństwo 41, 46, 47, 49
 - całkowite 57
 - iloczynu zdarzeń 54
 - – – niezależnych 55
 - sumy zdarzeń 57
 - warunkowe 52
 - względne 52
 - zdarzenia 47
- prawo De Morgana 27
- małych liczb 78
- wielkich liczb 168
- predykta 155
- predyktanta 155
- problem komiwojażera 12
 - marszruty 12
- proces stochastyczny(losowy) 279, 280, 290
 - – ergodyczny 286
 - – stacjonarny w sensie szerszym 285
 - – – – – węższym 285
 - – właściwy 280
 - – –, postać kanoniczna 288
 - – –, rozwinięcie spektralne 290
- produkt kartezjański 28
- prognoza statystyczna 155, 284
- prosta regresji ortogonalnej 138
- próbka 14, 188
 - duża 214
 - reprezentacyjna 191

- przedział 27
 - domknięty 27
 - – z lewa 27
 - – – prawa 27
- klasowy 186
- nieskończony 27
- otwarty 27
- ufności 212
- przestrzeń 26
 - ośrodkowa 28
 - probabilistyczna 153
 - punktów empirycznych 153
 - zdarzeń elementarnych 35
- punkt eksperymentalny (empiryczny) 92, 153

Realizacja 65

- procesu stochastycznego 282
- zmiennej losowej 65
 - – – *k*-wymiarowej 153
- regresand** 155
- regresja** 129
 - liniowa 131
- regresor** 155
- reszta regresji 133
- rozkład beta 89
 - brzegowy 93
 - χ^2 215
 - dwumianowy 70
 - empiryczny 186, 192
 - gamma 87
 - hipergeometryczny 74
 - jednakowych prawdopodobieństw 82
 - jednostajny 81
 - łączny 94
 - normalny 83, 144, 178
 - Poissona 76
 - prawdopodobieństwa zmiennej losowej 66
 - prostokątny 81
 - rzadkich zdarzeń 78
 - Studenta 219
 - teoretyczny 192
 - warunkowy 94
 - zero-jedynkowy 69
 - zmiennej losowej ciągłe 81
 - – – skokowej 67, 69, 92
- równanie regresji empiryczne 225
 - – w populacji 225
 - – – próbce 225, 226
 - – I rodzaju 129
- różnica zbiorów 37
 - zdarzeń 37

Schemat Bernoulliego 73

- seria 274
- silnia 9
- stosunki korelacyjne 135
- suma zbiorów 27
 - zdarzeń 36
- szereg Maclaurina** 24
 - rozdzielczy 186
 - statystyczny 186, 285
 - uporządkowany 186

Średnia arytmetyczna w próbce 226

- procesu po czasie 285
- – – przestrzeni 285
- środek ciężkości populacji** 132
- przedziału klasowego 186

Tablica dwudzielna 93

- korelacyjna 93
- teoria estymacji** 194
 - mnogości 25
 - procesów stochastycznych 279
- test χ^2** 267
 - Kolmogorowa-Smirnowa 268
 - nieparametryczny 253
 - – małej próbki 272
 - parametryczny 253
 - serii 274
 - statystyczny 253
 - zgodności 253
- tolerancja** 220
- trend empiryczny** 284
 - hipotetyczny 284
- trójkąt Pascala** 18
- twierdzenie Bernoulliego** 171
 - centralne 179
 - Chinczyna 172
 - Czebyszewa 170
 - Kołmogorowa 193
 - Lapunowa 181
 - Lévy'ego 125
 - Lindberga-Lévy'ego 179
 - Markowa 223
 - Moivre'a-Laplace'a 174
 - o prawdopodobieństwie sumy zdarzeń 44
 - – wariancji 113
 - – wartości przeciętnej 104
 - Poissona 171
 - Smirnowa 193
 - złote 167

Układ równań normalnych 132

- Wariacja** 12
 – z powtórzeniami 14
wariancja procesu 283
 – resztkowa 133
 – warunkowa 128
 – zmiennej losowej 99, 107
wariant klasowy 186
wartość krytyczna 256
 – oczekiwana 99, 128, 283
 – przeciętna 99, 100, 103
 – średnia procesu 283
wektor losowy 92, 152, 153
weryfikacja hipotez 253
widmo rozkładu normalnego 145
wskaźnik pojemności indywidualnej 161
 – – integralnej 162
współczynnik istotności 212, 255
 – korelacji 136, 160, 283
 – – częstkowej 160, 247
 – – procesu 283
 – – w próbce 226
 – – wielokrotnej 158, 247
 – – wielorakiej 158, 247
 – – wielowymiarowej 158, 247
 – regresji 132
 – – w próbce 226
 – ufności 212
 – zależności 139, 141
 – zbieżności 160
wzór Bayesa 59
 – Eulera 25
 – na prawdopodobieństwo całkowite 57
 – – iloczynu zdarzeń 54, 55
- wzór na prawdopodobieństwo sumy zdarzeń** 57
 – Stirlinga 19
- Zależność absolutna** 140
zbieżność stochastyczna 168
zbiorowość 14
zbiory rozłączne 27
zbiór 25
 – borełowski 28
 – krytyczny hipotezy 255
 – mierzalny 29
 – nieprzeliczalny 26
 – pusty 26
 – uporządkowany 10
zdarzenia niezależne 51, 52
 – przeciwnie 37
 – równoważne 37
 – wyłączające się 38
zdarzenie losowe 35, 39
 – niemożliwe 37
 – pewne 37
zmienna losowa 65, 66, 152
 – – ciągła 66, 79, 95
 – – dwuwymiarowa 92
 – – dystansowa 242
 – – k -wymiarowa 152
 – – quasi-ciągła 143
 – – quasi-dyskretna 143
 – – skokowa (dyskretna) 65, 92
 – objaśniająca 155
 – objaśniana 155
 – standaryzowana 83
zmienne losowe nieskorelowane 135
 – – niezależne 98, 119, 153
 – – skorelowane 135

SPIS RZECZY

Od Autora	5
---------------------	---

Część I

WYBRANE ZAGADNIENIA Z ALGEBRY I ANALIZY MATEMATYCZNEJ

Rozdział 1

1.1. Kombinatoryka	9
1.1.1. Pojęcie silni	9
1.1.2. Permutacje	10
1.1.3. Wariacje	12
1.1.4. Kombinacje	14
1.1.5. Trójkąt Pascala	17
1.1.6. Uwagi o obliczaniu dużych wartości silni. Wzór Stirlinga	18
Pytania kontrolne i zadania	19
1.2. Dwumian Newtona	20
Pytania kontrolne i zadania	22
1.3. Całki Eulera	22
1.3.1. Funkcja gamma Eulera	22
1.3.2. Funkcja beta Eulera	24
1.4. Wzory Eulera	24
1.5. Zbiory	25
1.5.1. Ogólne wiadomości o zbiorach	25
1.5.2. Algebra zbiorów	27
1.5.3. Przedziały, produkty kartezjańskie	27
1.5.4. Ciało zbiorów	28
1.5.5. Miara	29

Część II

RACHUNEK PRAWDOPODOBIĘSTWA

Rozdział 2

2.1. Wiadomości z zakresu historii rachunku prawdopodobieństwa	33
2.2. O zdarzeniach	35
2.2.1. Klasifikacja zdarzeń	35
2.2.2. Algebra zdarzeń	36
Pytania kontrolne i zadania	40

2.3. Pojęcie prawdopodobieństwa	40
2.3.1. Klasyczna definicja prawdopodobieństwa	40
2.3.2. Wady klasycznej definicji prawdopodobieństwa	44
2.3.3. Geometryczna definicja prawdopodobieństwa	45
2.3.4. Statystyczna, czyli częstościowa definicja prawdopodobieństwa	46
2.3.5. Współczesna definicja prawdopodobieństwa. Aksjomatyka rachunku prawdopodobieństwa	48
Pytania kontrolne i zadania	50
2.4. Podstawowe twierdzenia rachunku prawdopodobieństwa	50
2.4.1. Wnioski z aksjomatyki prawdopodobieństwa	50
2.4.2. Zdarzenia niezależne	51
2.4.3. Prawdopodobieństwo iloczynu zdarzeń	53
2.4.4. Prawdopodobieństwo sumy zdarzeń	56
2.4.5. Wzór na prawdopodobieństwo całkowite i wzór Bayesa	57
2.4.6. O konieczności ścisłego formułowania zagadnień probabilistycznych	60
Pytania kontrolne i zadania	63

Rozdział 3

3.1. Zmienne losowe	65
3.1.1. Pojęcia ogólne	65
3.2. Rozkład i dystrybuanta zmiennej losowej skokowej	67
3.3. Niektóre rozkłady zmiennej losowej skokowej	69
3.3.1. Rozkład zero-jedynkowy	69
3.3.2. Rozkład dwumianowy	70
3.3.3. Rozkład hipergeometryczny	73
3.3.4. Rozkład Poissona	76
3.4. Dystrybuanta zmiennej losowej ciągłej. Gęstość prawdopodobieństwa	79
3.5. Niektóre rozkłady zmiennej losowej ciąglej	81
3.5.1. Rozkład prostokątny	81
3.5.2. Rozkład trójkątny	82
3.5.3. Rozkład normalny	83
3.5.4. Rozkład gamma	87
3.5.5. Rozkład beta	89
Pytania kontrolne i zadania	90
3.6. Zmienne losowe dwuwymiarowe	92
3.6.1. Sformułowanie zagadnienia	92
3.6.2. Dwuwymiarowa zmienna losowa skokowa	92
3.6.3. Dwuwymiarowa zmienna losowa ciągła	95

Rozdział 4

4.1. Parametry opisowe	99
4.2. Wartość przeciętna	100
4.2.1. Określenie wartości przeciętnej. Przykłady	100
4.2.2. Twierdzenia o wartości przeciętnej	104

4.3. Wariancja i odchylenie standardowe	107
4.3.1. Określenia i przykłady	107
4.3.2. Twierdzenia o wariancji	113
4.4. Odchylenie przeciętne	115
4.5. Momenty	116
4.6. Funkcje charakterystyczne	118
4.7. Momenty zmiennej losowej dwuwymiarowej	125
4.7.1. Regresja pierwszego i drugiego rodzaju	125
4.7.2. Korelacja. Stosunek korelacyjny i współczynnik korelacji	134
4.7.3. Współczynnik zależności	139
4.7.4. Dwuwymiarowy rozkład normalny	144
4.7.5. Związki między współczynnikiem korelacji i współczynnikiem zależności w rozkładzie normalnym	149
4.7.6. Uwagi o wielowymiarowych zmiennych losowych	151
Pytania kontrolne i zadania.	163

Rozdział 5

5.1. Prawo wielkich liczb	166
5.1.1. Wprowadzenie	166
5.1.2. Nierówność Czebyszewa	168
5.1.3. Twierdzenie Czebyszewa	170
5.1.4. Twierdzenie Bernoulliego	171
5.1.5. Twierdzenie Poissona	171
5.1.6. Twierdzenie Chinczyna	172
5.2. Twierdzenie Moivre'a-Laplace'a	174
5.3. Twierdzenie centralne	178
Pytania kontrolne i zadania	182

*Część III***STATYSTYKA MATEMATYCZNA***Rozdział 6*

6.1. Definicje i pojęcia statystyczne	185
6.2. Uwagi o badaniu częściowym i metodzie reprezentacyjnej	188
6.3. Związek między populacją generalną i populacją próbą	191
6.4. Wybrane zagadnienia z teorii estymacji	194
6.4.1. Estymatory i ich klasyfikacja	194
6.4.2. Estymacja wartości przeciętnej	197
6.4.3. Estymacja wariancji	198
6.4.4. Estymacja wariancji wartości przeciętnej w próbce	201
6.4.5. Estymacja wariancji w populacji generalnej o znanym rozkładzie za pomocą odchylenia przeciętnego z próbki	206
6.5. Przedziały ufności	209
6.5.1. Sformułowanie problemu	209

6.5.2. Wyznaczenie przedziału ufności dla oszacowania wartości przeciętnej w populacji generalnej o dowolnym rozkładzie za pomocą średniej arytmetycznej z dużej próbki	213
6.5.3. Rozkład χ^2. Przedział ufności dla oszacowania wariancji w populacji generalnej	215
6.5.4. Rozkład Studenta. Wyznaczanie przedziału ufności dla oszacowania wartości przeciętnej w populacji generalnej na podstawie małej próbki	218
6.6. Wyznaczanie wielkości próbki	220
6.7. Estymacja parametrów regresji liniowej	223
6.7.1. Wprowadzenie	223
6.7.2. Estymacja parametrów regresji liniowej za pomocą metody najmniejszych kwadratów w przypadku zmiennych losowych dwuwymiarowych	224
6.7.3. Estymacja parametrów regresji liniowej za pomocą metody punktowej	227
6.7.4. Technika rachunkowa związana z obliczaniem parametrów regresji	229
6.7.5. Estymacja współczynnika zależności	233
6.7.6. Wybrane wiadomości o estymacji parametrów regresji w przypadku zmiennych losowych wielowymiarowych	237
6.7.7. Technika obliczeniowa w przypadku regresji wielowymiarowej	243
Pytania kontrolne i zadania	250

Rozdział 7

7.1. Określenie hipotezy statystycznej. Weryfikacja hipotezy	253
7.2. Błędy pierwszego i drugiego rodzaju	259
7.3. Krzywa mocy testu	260
7.4. Weryfikacja hipotez statystycznych	262
7.4.1. Weryfikacja hipotezy parametrycznej o wartości przeciętnej w populacji generalnej	262
7.4.2. Weryfikacja hipotezy nieparametrycznej o postaci rozkładu cechy w populacji (test χ^2, test λ Kolmogorowa-Smirnowa, test zgodności dla małej próbki, test serii)	263
7.4.3. Weryfikacja hipotezy nieparametrycznej o niezależności zmiennych losowych	274
Pytania kontrolne i zadania	278

Rozdział 8

8.1. Wybrane wiadomości z dziedziny procesów stochastycznych	279
8.2. Charakterystyki procesu stochastycznego	283
8.3. Stacjonarność procesu	284
8.4. Ergodyczność procesu	285
8.5. Kanoniczne rozwinięcie procesu	288
8.6. Widmo (spektrum) procesu	288
Tablice	293
Literatura	303
Skorowidz	305

Do nabycia w księgarniach:

J. Bartoszewicz

Wykłady ze statystyki matematycznej

Cz. Bracha

Teoretyczne podstawy metody reprezentacyjnej

W. Krysicki, J. Bartos, W. Dyczka, K. Królikowska, M. Wasilewski

Rachunek prawdopodobieństwa i statystyka matematyczna w zadaniach, cz.I

W. Krysicki, J. Bartos, W. Dyczka, K. Królikowska, M. Wasilewski

Rachunek prawdopodobieństwa i statystyka matematyczna w zadaniach, cz.II

L.T. Kubik

Zastosowanie elementarnego rachunku prawdopodobieństwa do wnioskowania statystycznego

I. Stewart

***Czy Bóg gra w kości?
Nowa matematyka chaosu***

Książki PWN są do nabycia w księgarniach własnych PWN:

Warszawa, ul. Miodowa 10; **Gdańsk**, ul. Korzenna 33/35;

Katowice, ul. Dworcowa 9; **Kraków**, ul. Św. Tomasza 30;

Łódź, ul. Więckowskiego 13; **Poznań**, ul. Wodna 8/9;

Wrocław, ul. Kuźnicza 56.

Zamówienia telefoniczne i pisemne przyjmuje:

Dział Dystrybucji Wysyłkowej i Prenumerat

ul. Miodowa 10, 00-251 Warszawa,

fax 69 54 179, infolinia 0-800-20145 (połączenie bezpłatne).