

**WYŻSZA SZKOŁA
INFORMATYKI STOSOWANEJ I ZARZĄDZANIA**

Seria: PODRĘCZNIKI WSISiZ

**Przemysław Grzegorzewski
Konstancja Bobecka
Anna Dembińska i Jerzy Pusz**

**RACHUNEK PRAWDOPODOBIEŃSTWA
I STATYSTYKA**

Wydanie 2 – zmienione i poprawione

**Podręcznik zgłoszony przez
Dziekana Wydziału Informatyki
Prof. dr hab. inż. Zbigniewa Nahorskiego**

Warszawa 2001

Pracę opiniował do druku:
Prof. dr hab. inż. Olgierd Hryniewicz

Wydanie 2, zmienione i poprawione

Spis treści

© Wyższa Szkoła Informatyki Stosowanej i Zarządzania
Warszawa 2001

ISBN 83-88311-30-1



Projekt graficzny okładki:
Jan Mlynarczyk

Druk:
Zakład Poligraficzny Jerzy Kosiński
Warszawa

S2 4708

Przedmowa	1
1 Podstawy rachunku prawdopodobieństwa	9
1.1 Wprowadzenie	9
1.2 Przestrzeń probabilistyczna	10
1.3 Własności prawdopodobieństwa	11
1.4 Metody obliczania prawdopodobieństwa	12
1.4.1 Schemat klasyczny	12
1.4.2 Uogólnienie schematu klasycznego	13
1.4.3 Prawdopodobieństwo geometryczne	13
1.5 Prawdopodobieństwo warunkowe	14
1.6 Prawdopodobieństwo całkowite i wzór Bayesa	15
1.7 Niezależność zdarzeń	16
1.8 Przykłady	18
1.9 Zadania	25
2 Zmienne losowe	31
2.1 Zmienne losowe jednowymiarowe	31
2.1.1 Określenie zmiennej losowej	31
2.1.2 Dystrybuanta zmiennej losowej	32
2.1.3 Zmienne losowe typu skokowego i ciągłego	33
2.1.4 Parametry zmiennych losowych jednowymiarowych	34
2.1.5 Własności wartości oczekiwanej i wariancji	39

6 Spis treści

5.11 Przykłady	165
5.12 Zadania	184
6 Tablice statystyczne	193
6.1 Dystrybuanta rozkładu normalnego	194
6.2 Kwantyle rozkładu normalnego	195
6.3 Kwantyle rozkładu chi-kwadrat	196
6.4 Kwantyle rozkładu t-Studenta	197
6.5 Kwantyle rozkładu F-Snedecora	198
6.6 Wartości krytyczne testu znaków	200
6.7 Wartości krytyczne testu rangowanych znaków	201
6.8 Wartości krytyczne testu Wilcoxona	202
6.9 Współczynniki dla testu Shapiro-Wilka	203
6.10 Wartości krytyczne testu Shapiro-Wilka	205
6.11 Wartości krytyczne testu Kolmogorowa	206
6.12 Wartości krytyczne testu Kolmogorowa-Smirnowa	207
Literatura	209

Przedmowa

Niniejszy podręcznik jest nieco rozszerzonym zapisem wykładów z rachunku prawdopodobieństwa i statystyki, wygłaszańskich od kilku lat dla studentów Wydziału Informatyki Wyższej Szkoły Informatyki Stosowanej i Zarządzania w Warszawie. Książka składa się z pięciu podstawowych rozdziałów, z których pierwsze dwa poświęcone są rachunkowi prawdopodobieństwa, natomiast dalsze trzy dotyczą podstawowych zagadnień statystyki, a więc statystyki opisowej, estymacji punktowej i przedziałowej oraz weryfikacji hipotez.

Dla przejrzystości wykładu, a zarazem dla wygody Czytelnika, przyjęto jednolity układ całej pracy: w każdym rozdziale przedstawiamy najpierw teorię (podstawowe definicje, własności omawianych pojęć, wybrane twierdzenia) oraz narzędzia służące do modelowania niepewności o charakterze losowym oraz podejmowania decyzji na podstawie danych obarczonych takim rodzajem niepewności. Następnie przytaczamy pewną liczbę w pełni rozwiązań przykładów, mających pomóc Czytelnikom w przyswojeniu sobie materiału teoretycznego. Na końcu każdego rozdziału zamieszczono także zestaw zadań do samodzielnego rozwiązania (z odpowiedziami). Tematykę przykładów i zadań dobrano w ten sposób, aby pokazać różnorodne zastosowania rachunku prawdopodobieństwa i statystyki w konkretnych problemach (i to nie tylko technicznych).

Ponieważ zasadniczym celem tej publikacji jest zwarta prezentacja podstawowych metod rachunku prawdopodobieństwa i statystyki, pominięto w niej wyprowadzenia wzorów matematycznych oraz dowody cytowanych twierdzeń. Czytelnicy zainteresowani pogłębieniem swej wiedzy w tym zakresie powinni sięgnąć do podręczników Bartoszewicza [1], Fella [3], Fisza

2.1.6 Podstawowe rozkłady prawdopodobieństwa zmiennych losowych	40	4.5.1 Pojęcie przedziału ufności	110
2.2 Wielowymiarowe zmienne losowe	47	4.5.2 Przedziały ufności dla wartości oczekiwanej	111
2.2.1 Dwuwymiarowe zmienne losowe	47	4.5.3 Przedziały ufności dla wariancji i odchylenia standardego	112
2.2.2 Niezależność zmiennych losowych	51	4.5.4 Przedział ufności dla wskaźnika struktury	113
2.2.3 Kowariancja i współczynnik korelacji	52	4.5.5 Przedziały ufności dla średniego czasu zdatności	113
2.2.4 Pewne rozkłady sum niezależnych zmiennych losowych	55	4.5.6 Estymacja przedziałowa o zadanej precyzyji	115
2.3 Prawa wielkich liczb i twierdzenia graniczne	56	4.6 Estymacja nieparametryczna	117
2.4 Przykłady	58	4.6.1 Uwagi wstępne	117
2.5 Zadania	72	4.6.2 Estymacja gęstości rozkładu	117
3 Statystyka opisowa	79	4.6.3 Estymacja dystrybuanty	118
3.1 Wprowadzenie	79	4.7 Przykłady	118
3.2 Podstawowe pojęcia	79	4.8 Zadania	130
3.3 Rozkład empiryczny cechy	81	5 Weryfikacja hipotez	133
3.3.1 Uwagi wstępne	81	5.1 Wprowadzenie	133
3.3.2 Metody opisu danych jakościowych	81	5.2 Pojęcia podstawowe	133
3.3.3 Metody opisu danych ilościowych	82	5.3 Algorytmy testowania hipotez	136
3.4 Syntetyczne charakterystyki próby	84	5.4 Testy dla wartości oczekiwanej	137
3.4.1 Uwagi wstępne	84	5.4.1 Testy dla pojedynczej próby	137
3.4.2 Miary położenia	85	5.4.2 Testy dla dwóch prób niezależnych	139
3.4.3 Miary rozproszenia	87	5.4.3 Test dla obserwacji parami zależnych	142
3.4.4 Charakterystyki kształtu	89	5.5 Testy dla mediany	142
3.4.5 Miary korelacji	90	5.5.1 Testy dla pojedynczej próby	142
3.5 Użyteczne wykresy statystyki opisowej	91	5.5.2 Test dla dwóch prób niezależnych	145
3.5.1 Uwagi wstępne	91	5.5.3 Testy dla obserwacji parami zależnych	146
3.5.2 Wykres skrzynkowy	91	5.6 Testy dla wariancji	146
3.5.3 Wykres łodygowo-liściowy	92	5.6.1 Testy dla pojedynczej próby	146
3.6 Przykłady	93	5.6.2 Testy dla dwóch prób niezależnych	147
3.7 Zadania	98	5.7 Testy dla wskaźnika struktury	149
4 Estymacja	101	5.7.1 Testy dla pojedynczej próby	149
4.1 Wprowadzenie	101	5.7.2 Testy dla dwóch niezależnych prób	150
4.2 Podstawowe własności estymatorów	102	5.8 Testy zgodności	151
4.3 Metody wyznaczania estymatorów	105	5.8.1 Uwagi wstępne	151
4.3.1 Wstęp	105	5.8.2 Test zgodności chi-kwadrat	152
4.3.2 Metoda momentów	105	5.8.3 Test Kolmogorowa	153
4.3.3 Metoda największej wiarogodności	106	5.8.4 Testy normalności	155
4.4 Przegląd estymatorów	107	5.8.5 Test Kolmogorowa-Smirnowa	158
4.4.1 Uwagi wstępne	107	5.8.6 Test Kruskala-Wallisa	158
4.4.2 Estymatory wartości oczekiwanej	108	5.9 Testowanie niezależności	159
4.4.3 Estymatory wariancji	109	5.9.1 Test niezależności chi-kwadrat	159
4.4.4 Estymatory odchylenia standardowego	109	5.9.2 Test dla współczynnika korelacji rangowej	161
4.4.5 Estymator wskaźnika struktury	110	5.10 Testy dla współczynnika korelacji	162
4.5 Przedziały ufności	110	5.10.1 Test hipotezy o braku korelacji liniowej	162
		5.10.2 Test dla współczynnika korelacji liniowej	164

[4], Jakubowskiego i Sztencla [8], Lehmanna [13], czy Zielińskiego [18]. Tych zaś, którym bliskie są rozważania z pogranicza matematyki i filozofii odsyłamy do książki Rao [15].

Ufamy, że niniejsza praca spotka się z życzliwym przyjęciem Czytelników i okaze się przydatna nie tylko dla słuchaczy wykładów z rachunku prawdopodobieństwa i statystyki na Wydziale Informatyki WSISiZ, ale także słuchaczy innych kierunków oraz wszystkich tych, którzy zainteresowani są praktycznymi zastosowaniami metod stochastycznych.

Przemysław Grzegorzewski
Warszawa, 8 grudnia 2000

1

Podstawy rachunku prawdopodobieństwa

1.1 Wprowadzenie

Rachunek prawdopodobieństwa to dział matematyki zajmujący się badaniem zjawisk losowych. Ze zjawiskami takimi mamy do czynienia, gdy wykonujemy doświadczenie, którego wyniku z góry nie znamy, ale możemy je wielokrotnie powtarzać w tych samych warunkach. Dzięki temu jesteśmy w stanie przewidzieć różne zakończenia tego doświadczenia i określić ich szanse. Np. jeśli rzucamy monetą, to z góry nie wiemy, czy wypadnie orzeł, czy reszka. Jednak po wielu powtórzeniach możemy nabrać podejrzeń, że szanse wypadnięcia orła są takie same, jak wypadnięcia reszki, czyli każde z tych zjawisk zachodzi z prawdopodobieństwem $\frac{1}{2}$.

Rachunek prawdopodobieństwa zajmuje się też zjawiskami masowymi. Np. interesuje nas, kto wygra wybory prezydenckie. Oczywiście wyniku z góry nie znamy, ale na podstawie sondaży przedwyborczych możemy przewidywać jakie są szanse każdego z kandydatów na objęcie urzędu.

Widzimy więc, że z określeniem prawdopodobieństw różnych zdarzeń mamy do czynienia na co dzień. W tym miejscu pragniemy zaznaczyć, że przedstawione powyżej rozumienie "losowości" nie jest jedynym możliwym. Mimo posługiwania się tym samym słowem "prawdopodobieństwo", czym innym jest ono w pytaniu "jakie jest prawdopodobieństwo wyrzucenia orła?", oraz w pytaniu "jakie jest prawdopodobieństwo, że jutro będzie ładna pogoda?". Udzielenie odpowiedzi na drugie z tych pytań nie należy do obszaru zainteresowania klasycznej teorii prawdopodobieństwa. Co więcej, warto pamiętać, że losowość nie jest również jedynym rodzajem spotykanej

niepewności – innym jest, przykładowo, brak precyzji. Raz więc jeszcze podkreślamy, że poniżej przedstawione rozważania dotyczą modelowania i badania zjawisk losowych rozumianych tylko i wyłącznie jako zjawiska masewne oraz doświadczenia, które mogą być wielokrotnie powtarzane w takich samych warunkach.

Aby uściślić to intuicyjne rozumienie prawdopodobieństwa wprowadza się pojęcie przestrzeni probabilistycznej.

1.2 Przestrzeń probabilistyczna

Rozważmy doświadczenie losowe, którego najprostsze możliwe wyniki ω nazywać będziemy **zdarzeniami elementarnymi**. Zbiór wszystkich zdarzeń elementarnych Ω nazywamy **przestrzenią zdarzeń elementarnych**. Interpretujemy ją jako kompletną i rozłączną listę możliwych wyników rozważanego doświadczenia losowego. Przestrzeń zdarzeń elementarnych może być zbiorem skończonym albo nieskończonym, przeliczalnym albo nieprzeliczalnym.

W praktyce zainteresowani jesteśmy zazwyczaj nie tyle pojedynczymi zdarzeniami elementarnymi, ale ich zbiorami (a więc podzbiorami Ω). Taką rodzinę podzbiorów Ω , którą wyróżnia eksperymentator, nazywamy **przestrzenią zdarzeń losowych \mathcal{F}** . Jeżeli Ω jest zbiorem przeliczalnym, to \mathcal{F} składa się po prostu ze wszystkich podzbiorów Ω . Natomiast jeżeli Ω jest zbiorem nieprzeliczalnym, to \mathcal{F} jest pewną rodziną podzbiorów Ω , zwana σ -cięciem zdarzeń, spełniającą następujące warunki:

- $\Omega \in \mathcal{F}$;
- jeżeli $A \in \mathcal{F}$, to $A' \in \mathcal{F}$, gdzie A' oznacza zdarzenie przeciwnie do zdarzenia A ;
- jeżeli $A_1, A_2, A_3, \dots \in \mathcal{F}$, to $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.

Na zdarzeniach losowych, zwanych po prostu zdarzeniami, wykonuje się działania analogiczne do działań przeprowadzanych na zbiorach. Możemy więc mówić o sumie, przecięciu, różnicy, zawieraniu się zdarzeń itd. W przypadku gdy $\omega \in A$ mówimy, że zaszło zdarzenie A . Sytuację $\omega \in A \cup B$ interpretujemy jako zajście co najmniej jednego ze zdarzeń A i B . Gdy $\omega \in A \cap B$, to mówimy, że zaszło zdarzenie A i zdarzenie B . Jeśli $\omega \in A \setminus B$, to zaszło zdarzenie A i nie zaszło zdarzenie B . W sytuacji gdy $A \subset B$ mówimy, że zdarzenie A pociąga za sobą zdarzenie B . Przypadek $A \cap B = \emptyset$ oznacza, że zdarzenia A i B wykluczają się. Samo zdarzenie \emptyset oznacza tzw. zdarzenie niemożliwe.

Dochodzimy teraz do pytania, jak określić prawdopodobieństwo zajścia danego zdarzenia. W potocznym rozumieniu, prawdopodobieństwo jest

miarą szansy zajścia rozważanego zdarzenia. Taka intuicyjna definicja nie jest jednak wystarczająca i może prowadzić do różnych paradoksów (by wspomnieć choćby słynny paradoks Bertranda, por. np. [8]). Stąd też, aby móc rozwijać zwartą teorię opisującą prawa rządzące doświadczeniami losowymi, zwaną **teorią prawdopodobieństwa**, konieczne stało się sformułowanie definicji aksjomatycznej. Definicję taką podał w 1933 roku A. N. Kolmogorow.

Definicja 1 *Prawdopodobieństwem nazywamy funkcję $P : \mathcal{F} \rightarrow R$ przyporządkowującą każdemu zdarzeniu losowemu A liczbę $P(A)$, zwaną prawdopodobieństwem zajścia zdarzenia A , tak, że spełnione są następujące warunki:*

$$A1 \quad P(A) \geq 0, \text{ dla każdego zdarzenia } A \in \mathcal{F},$$

$$A2 \quad P(\Omega) = 1,$$

A3 jeżeli $A_1, A_2, \dots \in \mathcal{F}$ jest dowolnym ciągiem zdarzeń parami rozłącznych, tzn. $A_i \cap A_j = \emptyset$ dla $i \neq j$, to

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i). \quad (1.1)$$

Definicja 2 *Trójkę uporządkowaną (Ω, \mathcal{F}, P) , gdzie Ω jest przestrzenią zdarzeń elementarnych, \mathcal{F} - przestrzenią zdarzeń losowych, a P - prawdopodobieństwem, nazywamy przestrzenią probabilistyczną.*

1.3 Własności prawdopodobieństwa

Z warunków A1 – A3 z definicji Kolmogorowa, zwanych także aksjomatami rachunku prawdopodobieństwa, oraz z własnością działań na zdarzeniach, wynikają następujące własności prawdopodobieństwa:

$$W1 \quad P(\emptyset) = 0.$$

$$W2 \quad \forall A \in \mathcal{F} \quad P(A) \leq 1.$$

$$W3 \quad \text{Jeżeli } A \in \mathcal{F}, \text{ to } P(A') = 1 - P(A).$$

$$W4 \quad \text{Jeżeli } A, B \in \mathcal{F} \text{ i } A \subset B, \text{ to } P(A) \leq P(B).$$

$$W5 \quad \text{Jeżeli } A, B \in \mathcal{F} \text{ i } A \subset B, \text{ to } P(B \setminus A) = P(B) - P(A).$$

$$W6 \quad \text{Jeżeli } A \cap B = \emptyset, \text{ to } P(A \cup B) = P(A) + P(B).$$

$$W7 \quad \text{Jeżeli } A, B \in \mathcal{F}, \text{ to } P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

Własność W7 można uogólnić dla ciągu n zdarzeń $A_1, A_2, \dots, A_n \in \mathcal{F}$ i wówczas

$$\begin{aligned} P\left(\bigcup_{i=1}^n A_i\right) &= \sum_{1 \leq i \leq n} P(A_i) - \sum_{1 \leq i < j \leq n} P(A_i \cap A_j) \\ &\quad + \sum_{1 \leq i < j < k \leq n} P(A_i \cap A_j \cap A_k) \\ &\quad - \dots + (-1)^{n+1} P(A_1 \cap A_2 \cap \dots \cap A_n). \end{aligned} \quad (1.2)$$

Wzór (1.2) nazywamy **wzorem włączeń i wyłączeń**. Przykładowo, dla $n = 3$ przyjmuje on postać

$$\begin{aligned} P(A_1 \cup A_2 \cup A_3) &= P(A_1) + P(A_2) + P(A_3) \\ &\quad - P(A_1 \cap A_2) - P(A_1 \cap A_3) - P(A_2 \cap A_3) \\ &\quad + P(A_1 \cap A_2 \cap A_3). \end{aligned} \quad (1.3)$$

(patrz: **Przykład 1.7, 1.8, 1.9**)

1.4 Metody obliczania prawdopodobieństwa

1.4.1 Schemat klasyczny

Choć poznaliśmy już liczne własności prawdopodobieństwa, nie mówiliśmy – jak dotąd – w jaki sposób obliczać prawdopodobieństwa zajścia interesujących nas zdarzeń. Nie da się tu zresztą wskazać jednej, uniwersalnej metody, bowiem sposób wyznaczania prawdopodobieństwa powinien być uzależniony od przyjętego modelu matematycznego rozważanego doświadczenia losowego. W niniejszym podrozdziale przedstawimy trzy, w miarę ogólne, metody obliczania prawdopodobieństwa, które wszakże mogą być stosowane wyłącznie pod warunkiem spełnienia podanych założeń.

Najprostszą sytuację opisuje tzw. **schemat klasyczny**, zwany też „**klasyczną definicją**” prawdopodobieństwa. Mianowicie, jeżeli przestrzeń zdarzeń elementarnych Ω jest skończona, tzn. $\Omega = \{\omega_1, \dots, \omega_n\}$, a przy tym jeżeli wszystkie zdarzenia elementarne są jednakowo prawdopodobne, czyli

$$P(\{\omega_1\}) = \dots = P(\{\omega_n\}) = \frac{1}{n},$$

to prawdopodobieństwo zajścia dowolnego zdarzenia $A = \{\omega_{i_1}, \dots, \omega_{i_k}\}$, składającego się z k zdarzeń elementarnych, wyraża się wzorem

$$P(A) = \frac{\#A}{\#\Omega} = \frac{k}{n}, \quad (1.4)$$

gdzie $\#A$ oznacza liczbę zdarzeń elementarnych sprzyjających zdarzeniu A (w naszym przypadku k), zaś $\#\Omega$ – liczbę wszystkich zdarzeń elementarnych (w rozważanym przypadku n). Przestrzeń zdarzeń losowych \mathcal{F} jest w tym przypadku rodziną wszystkich podzbiorów przestrzeni Ω , tzn. $\mathcal{F} = 2^\Omega$. (patrz: **Przykład 1.1**)

Dawniej wzór (1.4) stanowił definicję prawdopodobieństwa. Nie jest to jednak właściwy sposób definiowania prawdopodobieństwa, bowiem do definiowania prawdopodobieństwa używa on samego pojęcia prawdopodobieństwa (a dokładniej, zdarzeń jednakowo prawdopodobnych). Przyjmując akcyjomatyczną definicję prawdopodobieństwa, wzór (1.4) otrzymujemy jako twierdzenie.

1.4.2 Uogólnienie schematu klasycznego

Schemat klasyczny jest stosunkowo restrykcyjny z uwagi na założenie skończonej liczby zdarzeń elementarnych, które na dodatek muszą być jednakowo prawdopodobne. Schemat ten można jednak w stosunkowo prosty sposób uogólnić.

Załóżmy, że przestrzeń zdarzeń elementarnych Ω składa się z przeliczalnej (a więc, być może, nieskończonej) liczby zdarzeń elementarnych, tzn. $\Omega = \{\omega_1, \omega_2, \dots\}$. Niech przestrzeń zdarzeń losowych \mathcal{F} będzie rodziną wszystkich podzbiorów przestrzeni Ω , tzn. $\mathcal{F} = 2^\Omega$. Przyjmijmy ponadto, że prawdopodobieństwo zajścia dowolnego zdarzenia elementarnego ω_i wynosi $p_i = P(\{\omega_i\})$, przy czym spełnione jest

$$p_i > 0 \quad \forall i, \quad (1.5)$$

$$\sum_i p_i = 1. \quad (1.6)$$

Wówczas prawdopodobieństwo zajścia dowolnego zdarzenia losowego $A \in \mathcal{F}$ dane jest wzorem

$$P(A) = \sum_{i: \omega_i \in A} p_i, \quad (1.7)$$

a więc jest sumą prawdopodobieństw zajścia wszystkich zdarzeń elementarnych sprzyjających A .

(patrz: **Przykład 1.2**)

1.4.3 Prawdopodobieństwo geometryczne

Czasem przestrzeń zdarzeń elementarnych Ω wygodnie jest modelować jako pewien podzbiór przestrzeni R^n (np. prostej, płaszczyzny). Wówczas prawdopodobieństwo zajścia dowolnego zdarzenia $A \subset \Omega$ obliczamy jako stosunek miar obszarów odpowiadających A i Ω , tzn. ze wzoru

$$P(A) = \frac{\text{miara } (A)}{\text{miara } (\Omega)}. \quad (1.8)$$

Przez miarę zbioru rozumiemy w tym miejscu jego długość, o ile A i Ω są podzbiorami prostej, pole obszaru – o ile A i Ω są podzbiorami płaszczyzny, itd. Z uwagi na geometryczne intuicje związane z tą metodą obliczania prawdopodobieństwa, została ona nazwana **prawdopodobieństwem geometrycznym**.

(patrz: Przykład 1.3)

1.5 Prawdopodobieństwo warunkowe

Zajście jednego zdarzenia może wpływać na prawdopodobieństwo zajścia innego zdarzenia. Przypuśćmy, że wiemy, iż zaszło zdarzenie B i interesuje nas, jakie jest prawdopodobieństwo zajścia zdarzenia A . Zauważmy, że w przypadku schematu klasycznego, licząc prawdopodobieństwo warunkowe $P(A | B)$ zajścia zdarzenia A , pod warunkiem zajścia zdarzenia B , nie rozważamy już przestrzeni wszystkich zdarzeń elementarnych lecz ograniczamy się jedynie do tych zdarzeń elementarnych, które sprzyjają zajściu zdarzenia B . Mamy więc:

$$P(A | B) = \frac{\#A \cap B}{\#\Omega \cap B} = \frac{\#A \cap B}{\#B} = \frac{\frac{\#A \cap B}{\#\Omega}}{\frac{\#B}{\#\Omega}} = \frac{P(A \cap B)}{P(B)}. \quad (1.9)$$

W przypadku ogólnym definicja prawdopodobieństwa warunkowego jest następująca:

Definicja 3 Niech (Ω, \mathcal{F}, P) będzie przestrzenią probabilistyczną, zaś $B \in \mathcal{F}$ dowolnym ustalonym zdarzeniem o dodatnim prawdopodobieństwie, tzn. $P(B) > 0$. **Prawdopodobieństwem warunkowym** zajścia zdarzenia $A \in \mathcal{F}$ pod warunkiem zajścia zdarzenia B nazywamy liczbę $P(A | B)$ określona wzorem

$$P(A | B) = \frac{P(A \cap B)}{P(B)}. \quad (1.10)$$

(patrz: Przykład 1.4)

Ze wzoru (1.10) wynika natychmiast użyteczny wzór na prawdopodobieństwo iloczynu dwóch zdarzeń $A, B \in \mathcal{F}$

$$P(A \cap B) = P(A | B) \cdot P(B). \quad (1.11)$$

Wzór (1.11) łatwo uogólnić dla iloczynu trzech zdarzeń $A, B, C \in \mathcal{F}$

$$P(A \cap B \cap C) = P(A | B \cap C) \cdot P(B | C) \cdot P(C), \quad (1.12)$$

czy też n zdarzeń losowych $A_1, \dots, A_n \in \mathcal{F}$

$$\begin{aligned} P(A_1 \cap \dots \cap A_n) &= P(A_1) \cdot P(A_2 | A_1) \cdot P(A_3 | A_1 \cap A_2) \dots \\ &\quad \dots P(A_n | A_1 \cap A_2 \cap \dots \cap A_{n-1}). \end{aligned} \quad (1.13)$$

1.6 Prawdopodobieństwo całkowite i wzór Bayesa

Często mamy do czynienia z doświadczeniami wieloetapowymi i interesuje nas, jak liczyć prawdopodobieństwa zdarzeń, których zajście zależy od innych zdarzeń, które zaszły we wcześniejszych etapach naszych doświadczeń. Posłuży nam do tego wzór na prawdopodobieństwo całkowite. Do jego sformułowania potrzebne nam będzie pojęcie układu zupełnego zdarzeń.

Definicja 4 Powiemy, że zdarzenia $H_1, \dots, H_n \in \mathcal{F}$ tworzą **układ zupełny zdarzeń** w przestrzeni probabilistycznej (Ω, \mathcal{F}, P) , jeśli spełniają następujące warunki:

- $H_1 \cup \dots \cup H_n = \Omega$,
- $H_i \cap H_j = \emptyset$ dla $i \neq j$ (czyli zdarzenia te są parami rozłączne),
- $P(H_i) > 0$, $i = 1, \dots, n$.

Jeśli zdarzenia H_1, \dots, H_n to możliwe wyniki pierwszego etapu doświadczenia, to zajście dowolnego zdarzenia w drugim etapie możemy wyliczyć z następującego wzoru:

Twierdzenie 5 (o prawdopodobieństwie całkowitym) Jeśli zdarzenia $H_1, \dots, H_n \in \mathcal{F}$ tworzą układ zupełny zdarzeń w przestrzeni probabilistycznej (Ω, \mathcal{F}, P) , to dla dowolnego zdarzenia $A \in \mathcal{F}$

$$P(A) = \sum_{i=1}^n P(A | H_i) \cdot P(H_i). \quad (1.14)$$

Powyzsze twierdzenie często interpretuje się w kategoriach "przyczyna – skutek". Otóż, jeżeli skutek A może zajść w wyniku jednej z n przyczyn H_1, \dots, H_n (jedynie możliwych i wzajemnie się wykluczających), to prawdopodobieństwo wystąpienia skutku A wyraża się wzorem (1.14).

Wyobraźmy sobie teraz, że znamy wynik drugiego etapu doświadczenia i pytamy o przebieg doświadczenia, czyli o to, co stało się w jego pierwszym etapie. W takich sytuacjach stosujemy wzór Bayesa.

Twierdzenie 6 (Bayesa) Niech zdarzenia $H_1, \dots, H_n \in \mathcal{F}$ tworzą układ zupełny zdarzeń w przestrzeni probabilistycznej (Ω, \mathcal{F}, P) i niech $A \in \mathcal{F}$ będzie dowolnym ustalonym zdarzeniem o dodatnim prawdopodobieństwie, tzn. $P(A) > 0$. Wówczas prawdziwy jest wzór

$$P(H_k | A) = \frac{P(H_k \cap A)}{P(A)} = \frac{P(A | H_k) \cdot P(H_k)}{\sum_{i=1}^n P(A | H_i) \cdot P(H_i)}, \quad (1.15)$$

gdzie $k = 1, \dots, n$.

W kategoriach "przyczyna – skutek" twierdzenie to można sformułować w następujący sposób: jeżeli skutek A nastąpił w wyniku zajścia jednej z n przyczyn H_1, \dots, H_n (jedynie możliwych i wzajemnie się wykluczających), to prawdopodobieństwo tego, że H_k była przyczyną zajścia A wyraża się wzorem (1.15), zwanym także **wzorem Bayesa**.

Prawdopodobieństwo $P(H_k | A)$ nazywamy czasem prawdopodobieństwem *a posteriori*, gdyż podaje ono szanse zrealizowania H_k dopiero po zajściu zdarzenia A , natomiast $P(H_k)$ nazywamy prawdopodobieństwem *a priori*.

(patrz: Przykład 1.5, 1.6)

1.7 Niezależność zdarzeń

Ważnym pojęciem teorii prawdopodobieństwa jest niezależność zdarzeń losowych. Odwołując się do intuicji powiemy, że zdarzenie A nie zależy od zdarzenia B jeśli informacja, że zaszło zdarzenie B nie wpływa na prawdopodobieństwo zajścia zdarzenia A . Korzystając ze wzoru na prawdopodobieństwo warunkowe można by to zapisać następująco

$$P(A | B) = P(A). \quad (1.16)$$

Przyjęcie wzoru (1.16) za definicję niezależności zdarzeń nie byłoby jednak najszczególniejszym rozwiązaniem z uwagi na brak symetrii, podczas gdy samo pojęcie niezależności zdarzeń tego wymaga (wyszak jeśli A i B są niezależne, to zarówno A nie zależy od B , jak i B nie zależy od A). Jednakże łatwe przekształcenie wzoru na prawdopodobieństwo warunkowe prowadzi już do satysfakcjonującej definicji niezależności. Mamy bowiem $P(A | B) = \frac{P(A \cap B)}{P(B)}$, a zatem na mocy (1.16) $P(A) = \frac{P(A \cap B)}{P(B)}$ skąd otrzymujemy następującą definicję zdarzeń niezależnych:

Definicja 7 Niech (Ω, \mathcal{F}, P) będzie przestrzenią probabilistyczną oraz niech $A, B \in \mathcal{F}$. Mówimy, że zdarzenia A i B są niezależne jeśli

$$P(A \cap B) = P(A) \cdot P(B). \quad (1.17)$$

Z definicji (1.17) oraz z własnością działań na zdarzeniach, wynikają następujące fakty:

- Jeżeli $P(A) > 0$, to $P(B | A) = P(B)$.
- Jeżeli $P(B) > 0$, to $P(A | B) = P(A)$.
- Jeżeli zdarzenia A i B są niezależne, to zdarzenia A' i B są również niezależne.
- Jeżeli zdarzenia A i B są niezależne, to zdarzenia A i B' są również niezależne.
- Jeżeli zdarzenia A i B są niezależne, to zdarzenia A' i B' są również niezależne.

Niezależności zdarzeń nie należy mylić z rozłącznością zdarzeń. Prawdziwe jest bowiem następujące twierdzenie

Twierdzenie 8 Niech (Ω, \mathcal{F}, P) będzie przestrzenią probabilistyczną oraz niech $A, B \in \mathcal{F}$. Jeżeli $0 < P(A) < 1$, $0 < P(B) < 1$ oraz jeśli zdarzenia A i B są niezależne, to

- A i B nie mogą być rozłączne,
- między A i B nie może zachodzić relacja inkluzji (czyli A nie może być podzbiorem B , ani B nie może być podzbiorem A).

W przypadku większej liczby zdarzeń niezależność definiujemy następująco:

Definicja 9 Niech (Ω, \mathcal{F}, P) będzie przestrzenią probabilistyczną oraz niech $A_1, \dots, A_n \in \mathcal{F}$. Mówimy, że zdarzenia A_1, \dots, A_n są niezależne (wzajemnie niezależne, zespołowo niezależne) jeśli dla każdego $k \leq n$ oraz dla każdego ciągu indeksów $1 \leq i_1 < i_2 < \dots < i_k \leq n$ zachodzi

$$P(A_{i_1} \cap A_{i_2} \cap \dots \cap A_{i_k}) = P(A_{i_1}) \cdot P(A_{i_2}) \cdot \dots \cdot P(A_{i_k}). \quad (1.18)$$

Przykładowo, w przypadku trzech zdarzeń A_1, A_2, A_3 powiemy, że są one niezależne jeśli spełnione są wszystkie poniższe warunki

$$\begin{aligned} P(A_1 \cap A_2) &= P(A_1)P(A_2), \\ P(A_1 \cap A_3) &= P(A_1)P(A_3), \\ P(A_2 \cap A_3) &= P(A_2)P(A_3), \\ P(A_1 \cap A_2 \cap A_3) &= P(A_1)P(A_2)P(A_3). \end{aligned}$$

(patrz: Przykład 1.7, 1.8, 1.9)

1.8 Przykłady

Przykład 1.1

Chcemy otworzyć sejf z zamkiem szyfrowym. Zamek składa się z trzech współśrodkowych tarcz, a każda podzielona jest na cztery sektory, z numerami od 1 do 4. Jakie jest prawdopodobieństwo, że uda nam się otworzyć sejf przy pierwszej próbie, jeśli nie znamy szyfru?

Rozwiązanie

Przestrzeń zdarzeń elementarnych składa się z wszystkich trójwyrazowych ciągów złożonych z cyfr od 1 do 4, tzn. $\Omega = \{111, 112, 134, 423, \dots\} = \{abc : \text{gdzie } a \text{ oznacza cyfrę ustawioną na pierwszej tarczy, } b - \text{na drugiej, } c - \text{na trzeciej, } a, b, c \in \{1, 2, 3, 4\}\}$. Ponieważ na każdej tarczy mamy cztery cyfry do wyboru, więc

$$\#\Omega = 4 \cdot 4 \cdot 4 = 64$$

Oznaczmy przez A zdarzenie odpowiadające takiemu ustawieniu cyfr na trzech tarczach, które otwiera zamek szyfrowy. Jest tylko jedno takie ustawienie, a więc $\#A = 1$. Ponieważ przestrzeń zdarzeń elementarnych jest skończona, a wszystkie ustawienia jednakowo prawdopodobne, zatem korzystając ze schematu klasycznego otrzymujemy

$$P(A) = \frac{\#A}{\#\Omega} = \frac{1}{64}.$$

Przykład 1.2

Rzucamy symetryczną monetą do momentu otrzymania pierwszego orła. Obliczyć prawdopodobieństwa następujących zdarzeń:

- a) A - wykonano mniej niż 4 rzuty,
- b) B - wykonano parzystą liczbę rzutów.

Rozwiązanie

Nasza przestrzeń zdarzeń elementarnych Ω jest tym razem zbiorem nieskończonym: $\Omega = \{\omega_1, \omega_2, \dots\}$, gdzie ω_1 oznacza zdarzenie elementarne – "orzeł w pierwszym rzucie", ω_2 oznacza zdarzenie elementarne – "reszka w pierwszym rzucie i orzeł w drugim rzucie" itd. Zakładając, że rzucamy monetą symetryczną, tzn. $P(\text{orzeł}) = P(\text{reszka}) = \frac{1}{2}$, obliczmy prawdopodobieństwa kolejnych zdarzeń elementarnych:

- $\omega_1 = \text{wyrzucono orła w pierwszym rzucie}$

$$p_1 = P(\{\omega_1\}) = \frac{1}{2},$$

- $\omega_2 = \text{wyrzucono reszkę w pierwszym rzucie i orła w drugim rzucie}$

$$p_2 = P(\{\omega_2\}) = \frac{1}{2} \cdot \frac{1}{2} = \left(\frac{1}{2}\right)^2,$$

- $\omega_3 = \text{wyrzucono reszki w pierwszym i drugim rzucie oraz orła w trzecim rzucie}$

$$p_3 = P(\{\omega_3\}) = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \left(\frac{1}{2}\right)^3,$$

-

- $\omega_k = \text{wyrzucono reszki w pierwszych } k-1 \text{ rzutach i orła w } k\text{-tym rzucie}$

$$p_k = P(\{\omega_k\}) = \left(\frac{1}{2}\right)^{k-1} \cdot \frac{1}{2} = \left(\frac{1}{2}\right)^k,$$

-

Zdarzenie A - wykonano mniej niż 4 rzuty, oznacza, że $A = \{\omega_1, \omega_2, \omega_3\}$, a zatem

$$P(A) = p_1 + p_2 + p_3 = \frac{1}{2} + \frac{1}{4} + \frac{1}{8} = \frac{7}{8}.$$

Z kolei zdarzenie B - wykonano parzystą liczbę rzutów, można opisać następująco: $B = \{\omega_2, \omega_4, \omega_6, \dots\}$. Korzystając ze wzoru na sumę nieskończonego ciągu geometrycznego otrzymamy

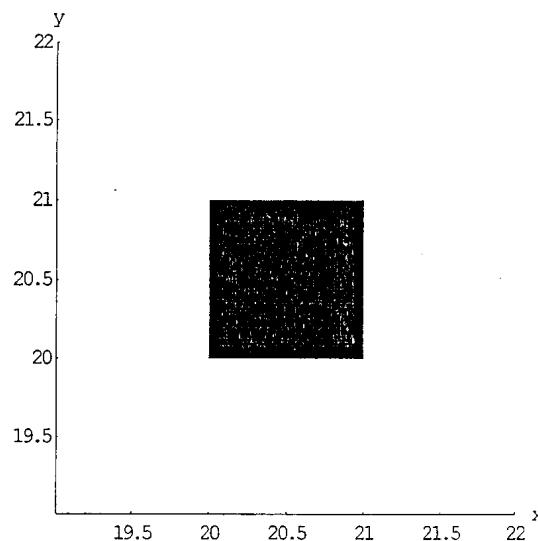
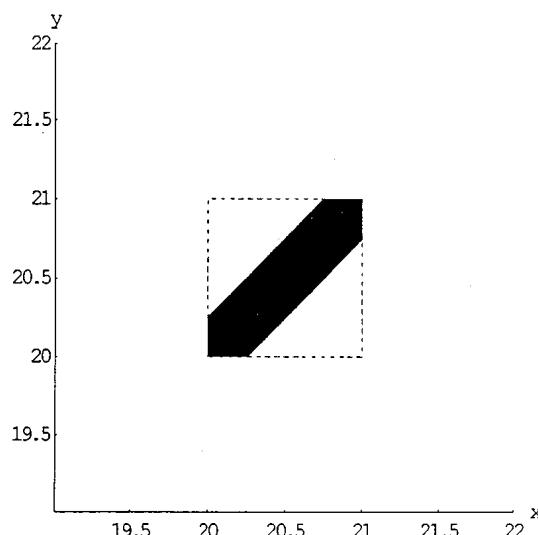
$$\begin{aligned} P(B) &= p_2 + p_4 + p_6 + \dots \\ &= \sum_{k=1}^{\infty} p_{2k} = \sum_{k=1}^{\infty} \left(\frac{1}{2}\right)^{2k} = \frac{\frac{1}{4}}{1 - \frac{1}{4}} = \frac{1}{3}. \end{aligned}$$

Przykład 1.3

Student i studentka umawiają się na spotkanie w "Harendzie" między 20:00 a 21:00. Momenty przybycia obu osób na spotkanie są losowe i niezależne od siebie. Każda z nich czeka 15 minut a potem odchodzi. Jakie jest prawdopodobieństwo, że się spotkają?

Rozwiązanie

Na osi X zaznaczmy moment przyjścia studentki, a na osi Y moment przyjścia studenta (por. rys. 1.1).

Rys. 1.1 Zbiór zdarzeń elementarnych Ω Rys. 1.2 Zdarzenie A

Widzimy, że Ω to zbiór wszystkich punktów o współrzędnych $(x, y) \in R^2$ takich, że $x \in [20, 21]$ i $y \in [20, 21]$, czyli jest to kwadrat o boku długości 1 (godzina). Aby doszło do spotkania, różnica czasów przybycia studentki i studenta nie może przekroczyć 15 minut tzn. $|x - y| \leq \frac{1}{4}$ (godziny).

Niech A oznacza zdarzenie polegające na tym, że student i studentka spotkają się (por. rys. 1.2). Zatem

$$\begin{aligned} A &= \left\{ (x, y) \in \Omega : |x - y| \leq \frac{1}{4} \right\} \\ &= \left\{ (x, y) \in \Omega : -\frac{1}{4} \leq x - y \leq \frac{1}{4} \right\}. \end{aligned}$$

Korzystając ze wzoru na tzw. prawdopodobieństwo geometryczne otrzymamy

$$P(A) = \frac{\text{pole } A}{\text{pole } \Omega} = \frac{1 - \left(\frac{3}{4}\right)^2}{1} = \frac{\frac{7}{16}}{1} = \frac{7}{16}.$$

Przykład 1.4

Po zakończeniu sesji stwierdzono, że 70% studentów II roku zdało egzamin z języka C++, natomiast 20% zdało egzamin z C++ i RPiS. Jakie jest prawdopodobieństwo, że student, który zdał egzamin z C++ zdał również egzamin z RPiS?

Rozwiązanie

Przez A, B oznaczmy następujące zdarzenia:
 A - wybrany losowo student zdał egzamin z C++,
 B - wybrany losowo student zdał egzamin z RPiS.
Z treści zadania wynika, że

$$\begin{aligned} P(A) &= 0.7, \\ P(A \cap B) &= 0.2. \end{aligned}$$

Interesuje nas prawdopodobieństwo zajścia zdarzenia B pod warunkiem, że zaszło zdarzenie A

$$\begin{aligned} P(B | A) &= \frac{P(B \cap A)}{P(A)} \\ &= \frac{P(A \cap B)}{P(A)} = \frac{0.2}{0.7} = \frac{2}{7}. \end{aligned}$$

Przykład 1.5

Pewien student codziennie je obiad w jednym z trzech barów mlecznych: w barze BAR, w barze KOLEJOWYM lub w barze RUSALKA. W barze BAR bywa 16 razy w miesiącu, w KOLEJOWYM dwa razy w miesiącu, zaś w RUSAŁCE 12 razy w miesiącu (przyjmujemy, że miesiąc ma 30 dni). Na każde 20 wizyt w barze BAR trzy kończą się zatruciem, na każde 15 wizyt w barze KOLEJOWYM jedna kończy się zatrutkiem i na każde 30 wizyt w RUSAŁCE cztery kończą się zatrutkiem.

- a) Jakie jest prawdopodobieństwo tego, że w losowo wybranym dniu student zatrudni się obiadem?
- b) Wiemy, że student się zatrudni. W którym z barów najprawdopodobniej jadł obiad?

Rozwiązańe

Jest to przykład doświadczenia dwuetapowego: najpierw student wybiera jeden z trzech barów, a potem je w nim obiad, którym może się zatrudnić lub nie. Zatem, przez H_1 oznaczmy zdarzenie "student wybrał bar BAR", przez H_2 zdarzenie "student wybrał bar KOLEJOWY", przez H_3 zdarzenie "student wybrał bar RUSAŁKA". Niech A oznacza zdarzenie "student zatrudni się obiadem". Zauważmy, że spełnione są założenia twierdzenia o prawdopodobieństwie całkowitym, tzn.

$$H_1 \cup H_2 \cup H_3 = \Omega,$$

$$H_1 \cap H_2 = \emptyset, H_1 \cap H_3 = \emptyset, H_2 \cap H_3 = \emptyset,$$

$$P(H_1) = \frac{16}{30} > 0, P(H_2) = \frac{2}{30} > 0, P(H_3) = \frac{12}{30} > 0.$$

a) Stosując wzór (1.14) na prawdopodobieństwo całkowite mamy:

$$P(A) = P(A | H_1)P(H_1) + P(A | H_2)P(H_2) + P(A | H_3)P(H_3),$$

gdzie $P(A | H_1)$ oznacza prawdopodobieństwo zdarzenia "student zatrudni się, pod warunkiem, że jadł obiad w barze BAR", a zatem

$$P(A | H_1) = \frac{3}{20}$$

i analogicznie:

$$P(A | H_2) = \frac{1}{15},$$

$$P(A | H_3) = \frac{4}{30}.$$

Stąd

$$P(A) = \frac{3}{20} \cdot \frac{16}{30} + \frac{1}{15} \cdot \frac{2}{30} + \frac{4}{30} \cdot \frac{12}{30} = \frac{31}{225} \approx 0.1377.$$

b) Wiemy, że student się zatrudni. Interesuje nas prawdopodobieństwo tego, że stało się to w barze BAR, KOLEJOWYM lub w RUSAŁCE. Szukamy więc, odpowiednio, następujących prawdopodobieństw: $P(H_1 | A)$, $P(H_2 | A)$, $P(H_3 | A)$. Stosując wzór Bayesa (wzór (1.15)) mamy:

$$P(H_1 | A) = \frac{P(A | H_1)P(H_1)}{P(A)} = \frac{\frac{3}{20} \cdot \frac{16}{30}}{\frac{31}{225}} = \frac{18}{31} \approx 0.5806,$$

$$P(H_2 | A) = \frac{P(A | H_2)P(H_2)}{P(A)} = \frac{\frac{1}{15} \cdot \frac{2}{30}}{\frac{31}{225}} = \frac{1}{31} \approx 0.03226,$$

$$P(H_3 | A) = \frac{P(A | H_3)P(H_3)}{P(A)} = \frac{\frac{4}{30} \cdot \frac{12}{30}}{\frac{31}{225}} = \frac{12}{31} \approx 0.3871.$$

Zatem najprawdopodobniej student zjadł obiad w barze BAR.

Przykład 1.6

70% kobiet i 15% mężczyzn oglądają brazylijskie seriale w telewizji. Z grupy złożonej z 1000 kobiet i 1500 mężczyzn wybrano losowo jedną osobę.

- a) Jakie jest prawdopodobieństwo, że wybrana osoba ogląda brazylijskie seriale?
- b) Okazało się, że wylosowana osoba ogląda brazylijskie seriale. Jakie jest prawdopodobieństwo, że jest to kobieta?

Rozwiązańe

W zadaniu tym mamy podział ze względu na dwie cechy: płeć i oglądanie lub nie oglądanie brazylijskich seriali. Wprowadźmy następujące oznaczenia:

H_1 - zdarzenie "wybrano kobietę",

H_2 - zdarzenie "wybrano mężczyznę",

A - zdarzenie "wybrana osoba ogląda brazylijskie seriale".

Spełnione są założenia twierdzenia o prawdopodobieństwie całkowitym:

$$H_1 \cup H_2 = \Omega,$$

$$H_1 \cap H_2 = \emptyset,$$

$$P(H_1) = \frac{1000}{2500} = 0.4 > 0, P(H_2) = \frac{1500}{2500} = 0.6 > 0.$$

a) Stosujemy wzór (1.14) na prawdopodobieństwo całkowite

$$P(A) = P(A | H_1)P(H_1) + P(A | H_2)P(H_2),$$

gdzie $P(A | H_1)$ oznacza prawdopodobieństwo zdarzenia "wybrana losowo kobieta ogląda seriale brazylijskie", a zatem $P(A | H_1) = 0.7$ i analogicznie: $P(A | H_2) = 0.15$. Stąd

$$P(A) = 0.7 \cdot 0.4 + 0.15 \cdot 0.6 = 0.37.$$

b) Szukamy prawdopodobieństwa $P(H_2 | A)$. Ze wzoru (1.15) mamy

$$P(H_2 | A) = \frac{P(A | H_2)P(H_2)}{P(A)} = \frac{0.15 \cdot 0.6}{0.37} = \frac{9}{37} \approx 0.2432.$$

Przykład 1.7

Pewien student mieszka pod Warszawą i na zajęcia dojeżdża najpierw kolejką WKD, a potem tramwajem. Jeśli pociąg lub tramwaj spóźni się, student nie zdąży na zajęcia. Oszacowano, że prawdopodobieństwa opóźnienia się pociągu i tramwaju wynoszą odpowiednio: 0.3 i 0.2. Jakie jest prawdopodobieństwo tego, że student przyjedzie o czasie na zajęcia?

Rozwiązańe

Niech A i B oznaczają następujące zdarzenia:

A - pociąg się nie spóźnił,
 B - tramwaj się nie spóźnił.

Z treści zadania oraz z własności prawdopodobieństwa mamy:

$$P(A) = 1 - P(A') = 1 - 0.3 = 0.7$$

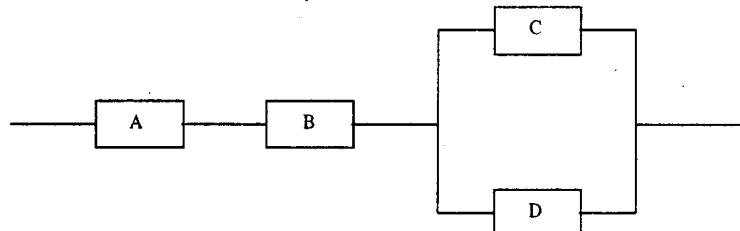
$$P(B) = 1 - P(B') = 1 - 0.2 = 0.8.$$

Można przyjąć, że spóźnienia kolejki i tramwaju są niezależne. Ponadto z treści zadania wynika, że aby student zdążył na zajęcia, kolejka i tramwaj muszą przyjechać punktualnie. Zatem szukane prawdopodobieństwo wynosi

$$P(A \cap B) = P(A)P(B) = 0.7 \cdot 0.8 = 0.56.$$

Przykład 1.8

Oblicz prawdopodobieństwo przekazania sygnału przez układ pokazany na rys. 1.3, składający się z czterech przekaźników A, B, C i D, działających niezależnie od siebie, jeśli prawdopodobieństwa działania każdego z przekaźników wynoszą odpowiednio: 0.7, 0.8, 0.9 i 0.6.



Rys. 1.3 – Schemat układu

Rozwiązanie

Aby cały układ działał musi działać przekaźnik A i B i choć jeden z przekaźników C lub D. Zatem interesuje nas prawdopodobieństwo zajścia zdarzenia $S = A \cap B \cap (C \cup D)$, gdzie A, B, C, D oznaczają zdarzenia odpowiadające działaniu przekaźników A, B, C i D. Korzystając z własności rachunku zbiorów (rozłączność iloczynu względem sumy) możemy powyższe wyrażenie zapisać w postaci

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

Ze wzoru na prawdopodobieństwo sumy (własność W7) otrzymujemy dalej

$$P(S) = P(A \cap B \cap C) + P(A \cap B \cap D) - P(A \cap B \cap C \cap D).$$

Ponieważ przekaźniki działają niezależnie od siebie mamy:

$$\begin{aligned} P(S) &= P(A)P(B)P(C) + P(A)P(B)P(D) - P(A)P(B)P(C)P(D) \\ &= 0.7 \cdot 0.8 \cdot 0.9 + 0.7 \cdot 0.8 \cdot 0.6 - 0.7 \cdot 0.8 \cdot 0.9 \cdot 0.6 \\ &= 0.5376. \end{aligned}$$

Przykład 1.9

Aby zakwalifikować się do drugiego etapu teleturnieju trzeba odpowiedzieć poprawnie na przynajmniej jedno z trzech zadanych pytań (każde z pytań dotyczy innej dziedziny). Z dotychczasowych obserwacji wynika, że prawdopodobieństwo udzielenia poprawnej odpowiedzi na każde z pytań jest jednakowe i wynosi $\frac{1}{3}$. Jakie jest prawdopodobieństwo, że osoba, która zgłosiła się do udziału w teleturnieju zakwalifikuje się do drugiego etapu?

Rozwiązanie

Niech A_i oznacza zdarzenie: "osoba, która zgłosiła się do udziału w teleturnieju odpowiedziała poprawnie na i -te pytanie", $i = 1, 2, 3$. Z treści zadania wynika, że zdarzenia A_i są niezależne (pytania dotyczą różnych dziedzin) oraz $P(A_i) = \frac{1}{3}$, $i = 1, 2, 3$.

Prawdopodobieństwo tego, że osoba, która zgłosiła się do udziału w teleturnieju przejdzie do drugiego etapu, równe jest prawdopodobieństwu zajścia zdarzenia $A_1 \cup A_2 \cup A_3$ (czyli udzielenia poprawnej odpowiedzi na przynajmniej jedno z trzech zadanych pytań). Z własności prawdopodobieństwa (własność W7) oraz niezależności zdarzeń A_1, A_2, A_3 otrzymujemy:

$$\begin{aligned} P(A_1 \cup A_2 \cup A_3) &= P(A_1) + P(A_2) + P(A_3) \\ &\quad - P(A_1 \cap A_2) - P(A_1 \cap A_3) - P(A_2 \cap A_3) \\ &\quad + P(A_1 \cap A_2 \cap A_3) \\ &= \frac{1}{3} + \frac{1}{3} + \frac{1}{3} - \frac{1}{3} \cdot \frac{1}{3} - \frac{1}{3} \cdot \frac{1}{3} + \frac{1}{3} \cdot \frac{1}{3} \cdot \frac{1}{3} \\ &= \frac{19}{27} \simeq 0.7. \end{aligned}$$

1.9 Zadania

Zadanie 1.1

W torebce jest 10 cukierków, w tym 3 czekoladowe. Wyciągamy z tej torbki, w sposób losowy, 3 cukierki. Oblicz prawdopodobieństwo tego, że:

- a) wśród 3 wyjętych cukierków jest dokładnie jeden czekoladowy,
- b) wśród 3 wyjętych cukierków jest co najmniej jeden czekoladowy.

Zadanie 1.2

Mamy 4 swetry i w sposób losowy wkładamy je do szuflad w szafie. Szafa ma 7 szuflad.

- a) Jakie jest prawdopodobieństwo, że każdy sweter znajdzie się w innej szufladzie?
- b) Jakie jest prawdopodobieństwo, że wszystkie swetry włożyliśmy do tej samej szuflady?

Zadanie 1.3

Ania, Zosia, Kasia, Ela, Zenek, Mietek, Wacek i Stefan usiedli, w sposób losowy, przy okrągłym stole. Jakie jest prawdopodobieństwo, że dziewczęta nie siedzą obok siebie?

Zadanie 1.4

Miedzy A i B jest jeden tor i pociagi, przebywające tę trasę w 10 minut, kursują według schematu: pociąg z A do B i pociąg z B do A startują niezależnie od siebie w losowych momentach między 7⁰⁰ a 8⁰⁰. Jakie jest prawdopodobieństwo, że dotrą do celu?

Zadanie 1.5

Studenci stwierdzili, że prawdopodobieństwo zdania egzaminu z RPiS w pierwszym terminie wynosi 0.4, natomiast prawdopodobieństwo zdania egzaminu z C++ wynosi 0.7, przy czym zdarzenia te są niezależne. Jakie jest prawdopodobieństwo zdania przynajmniej jednego z tych dwóch egzaminów w pierwszym terminie?

Zadanie 1.6

Firma komputerowa przeprowadziła sondaż w urzędach gminnych, którego celem było zbadanie zainteresowania nową wersją programu biurowego. 80% respondentów wykazało chęć kupna nowej wersji programu. Spośród urzędów zainteresowanych kupnem nowej wersji programu, 40% wyraziło chęć modernizacji komputerów. Oblicz prawdopodobieństwo tego, że urząd gminny planuje modernizację komputerów i chce kupić nową wersję programu biurowego.

Zadanie 1.7

Zenek uwielbia konkursy organizowane przez stacje radiowe. Prawdopodobieństwo wygrania koszulki w konkusie radia RMF wynosi 0.1, natomiast prawdopodobieństwo wygrania koszulki w konkusie Radia Zet wynosi 0.2. Zakładając, że oba konkursy są niezależne obliczyć prawdopodobieństwo wygrania przez Zenka co najmniej jednej koszulki.

Zadanie 1.8

Pewna firma nabyła nowy serwer. Według zapewnienia sprzedawcy prawdopodobieństwo awarii tego serwera w pierwszym roku użytkowania wynosi 5%. Jeżeli w pierwszym roku nie nastąpi awaria, to z prawdopodobieństwem 10% nastąpi ona w drugim roku użytkowania. Natomiast jeżeli zarówno w pierwszym, jak i w drugim roku użytkowania nie będzie awarii serwera, to z prawdopodobieństwem 15% nastąpi ona w trzecim roku. Jakie jest prawdopodobieństwo, że serwer nie ulegnie awarii w ciągu pierwszych trzech lat użytkowania?

Zadanie 1.9

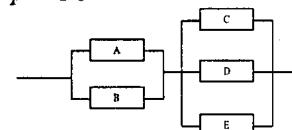
Udajemy się do Krakowa na Konferencję Statystyków. Podróż zaplanowaliśmy w następujący sposób: z domu na dworzec Warszawa Centralna jedziemy tramwajem, następnie wsiadziemy do pociągu do Krakowa,

a ostatni odcinek drogi - z Dworca Głównego w Krakowie na Uniwersytet Jagielloński - pokonamy autobusem. Jeżeli którykolwiek ze wspomnianych pojazdów spóźni się, nie zdążymy na tę interesującą i ważną konferencję. Oszacowaliśmy, że prawdopodobieństwa opóźnienia się tramwaju, pociągu i autobusu wynoszą, odpowiednio, 0.1, 0.4 i 0.2. Jakie jest prawdopodobieństwo, że zdążymy na rozpoczęcie konferencji?

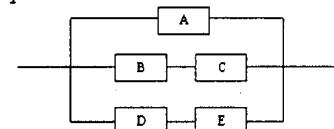
Zadanie 1.10

Oblicz prawdopodobieństwo przekazania sygnału przez układ, składający się z działających niezależnie przekaźników, jeśli prawdopodobieństwo działania każdego z nich wynosi p .

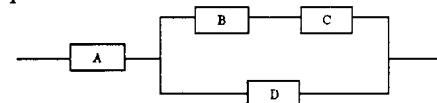
a) $p = 0.9$



b) $p = 0.7$



c) $p = 0.8$

**Zadanie 1.11**

Pewne zakłady metalowe współpracują z trzema odlewniami. Z poszczególnych odlewni pochodzi, odpowiednio, 10%, 30% i 60% odlewów. Z poczynionych obserwacji wynika, że 2% odlewów dostarczonych z pierwszej odlewni zawiera ukryte wady. Ukryte wady zawiera również 10% odlewów pochodzących z drugiej i 4% odlewów pochodzących z trzeciej odlewni. W trakcie obróbki pewnego odlewu stwierdzono, że jest on wadliwy. Z której odlewni najprawdopodobniej pochodzi ten odlew?

Zadanie 1.12

Stwierdzono, że awaryjność półprzewodnikowego układu scalonego zależy od stopnia domieszkowania w procesie produkcyjnym. I tak, prawdopodobieństwo awarii układu domieszkowanego w wysokim stopniu wynosi 0.1, układu średnio domieszkowanego 0.01, natomiast prawdopodobieństwo awarii układu scalonego domieszkowanego w małym stopniu wynosi 0.01. W pewnej partii półprzewodnikowych układów scalonych 20% stanowią układy domieszkowane w wysokim stopniu, 30% w średnim stopniu i 50% w małym

stopniu. Obliczyć prawdopodobieństwo awarii urządzenia zawierającego jeden układ scalony pochodzący z tej partii.

Zadanie 1.13

Około 70% kobiet i 90% mężczyzn posiada prawo jazdy. Z populacji liczącej 400 kobiet i 600 mężczyzn wybrano osobę posiadającą prawo jazdy. Oblicz prawdopodobieństwo, że był to mężczyzna.

Zadanie 1.14

W pewnym sklepie 45% sprzedawanych piw bezalkoholowych pochodzi z browaru w Warszawie, a 55% z browaru w Żywcu. Prawdopodobieństwo znalezienia muchy w butelce z piwem bezalkoholowym pochodzącym z browaru w Warszawie wynosi 0.01, zaś z browaru w Żywcu 0.02. Kupiono jedno piwo bezalkoholowe i okazało się, że pływa w nim mucha. Oblicz prawdopodobieństwo tego, że kupiona butelka z piwem pochodziła z browaru w Warszawie.

ODPOWIEDZI

Zadanie 1.1

- a) $\frac{21}{40} = 0.525$,
- b) $\frac{17}{24} \approx 0.71$.

Zadanie 1.2

- a) $\frac{120}{343} \approx 0.35$,
- b) $\frac{1}{343} \approx 0.003$.

Zadanie 1.3

$$\frac{3!4!}{7!} \approx 0.029.$$

Zadanie 1.4

$$\frac{25}{36} \approx 0.69.$$

Zadanie 1.5

0.82.

Zadanie 1.6

0.32

Zadanie 1.7

0.28.

Zadanie 1.8

0.72675.

Zadanie 1.9

0.432

Zadanie 1.10

- a) 0.98901,
- b) 0.92197,
- c) 0.7424.

Zadanie 1.11

Prawdopodobieństwa, że wadliwy odlew pochodzi z pierwszej, drugiej i trzeciej odlewni wynoszą, odpowiednio, $\frac{2}{56}$, $\frac{30}{56}$ i $\frac{24}{56}$. Zatem wadliwy odlew najprawdopodobniej pochodzi z drugiej odlewni.

Zadanie 1.12

0.0235.

Zadanie 1.13

$\frac{27}{41}$

Zadanie 1.14

$\frac{9}{31}$.

2

Zmienne losowe

2.1 Zmienne losowe jednowymiarowe

2.1.1 Określenie zmiennej losowej

W wielu zagadnieniach spotykamy się z wielkościami, których wartość liczbowa zależy od przypadku a więc od konkretnego zdarzenia elementarnego ω należącego do pewnej przestrzeni zdarzeń elementarnych Ω . W doświadczeniu z rzutem kostką sześcienną zdarzeniu elementarnemu ω_i reprezentującemu wypadnięcie i oczek można przyporządkować tę właśnie liczbę i . W doświadczeniu z rzutem parą kostek każdemu zdarzeniu elementarnemu - parze uporządkowanej (i, j) ($i, j = 1, 2, \dots, 6$) - można przyporządkować sumę oczek $i + j$. Przy n krotnym powtórzeniu doświadczenia z prawdopodobieństwem sukcesu p w każdym z doświadczeń, każdemu zdarzeniu elementarnemu można przyporządkować otrzymaną liczbę sukcesów. We wszystkich grach losowych (ruletka, Toto-Lotek, trzy karty, bingo...) każdemu zdarzeniu elementarnemu można przyporządkować liczbę będącą wypłatą. Powyższe przyporządkowania – inaczej mówiąc funkcje o wartościach rzeczywistych określone na przestrzeni zdarzeń elementarnych – są przykładami zmiennych losowych.

Jeżeli przestrzeń zdarzeń elementarnych Ω jest co najwyżej przeliczalna, to każdą funkcję $X(\omega)$ o wartościach rzeczywistych określoną na Ω nazywamy zmienną losową. W przypadku ogólnym, gdy przestrzeń zdarzeń elementarnych Ω jest nieprzeliczalna to na funkcję $X(\omega)$ musi być nałożony pewien dodatkowy warunek a mianowicie:

Definicja 10 Niech (Ω, \mathcal{F}, P) będzie dowolną przestrzenią probabilistyczną. Zmienną losową nazywamy dowolną funkcję rzeczywistą X określoną na przestrzeni zdarzeń elementarnych Ω taką, że dla każdej liczby rzeczywistej x :

$$\{\omega \in \Omega : X(\omega) \leq x\} \in \mathcal{F}. \quad (2.1)$$

Oznacza to, że przeciwbrazy przedziałów $(-\infty, x]$ są zdarzeniami losowymi.

2.1.2 Dystrybuanta zmiennej losowej

Z określenia zmiennej losowej X wynika, że podzbiory przestrzeni zdarzeń elementarnych określone wzorem (2.1) są zdarzeniami losowymi a więc można mówić o prawdopodobieństwach tych zdarzeń.

Definicja 11 Dystrybuantą zmiennej losowej X nazywamy funkcję rzeczywistą F , która jest określona dla wszystkich liczb rzeczywistych $x \in (-\infty, +\infty)$ wzorem:

$$F(x) = P(\omega \in \Omega : X(\omega) \leq x). \quad (2.2)$$

W dalszej części będziemy w skrócie pisali:

$$F(x) = P(X \leq x). \quad (2.3)$$

Dystrybuanta F posiada następujące własności:

W1 $\forall x \in R \quad 0 \leq F(x) \leq 1$,

W2 F jest funkcją niemalejącą,

W3 F jest funkcją co najmniej prawostronnie ciągłą,

W4 $\lim_{x \rightarrow -\infty} F(x) = 0$ oraz $\lim_{x \rightarrow +\infty} F(x) = 1$,

W5 $P(a < X \leq b) = F(b) - F(a)$,

W6 $P(X = x_0) = F(x_0) - F(x_0^-)$, gdzie $F(x_0^-)$ oznacza lewostronną granicę dystrybuanty w punkcie x_0 .

Twierdzenie 12 Jeżeli funkcja F ma własności W2, W3 i W4 to F jest dystrybuantą pewnej zmiennej losowej.

(patrz: Przykład 2.1, 2.5)

2.1.3 Zmienne losowe typu skokowego i ciągłego

Wyróżniamy dwa zasadnicze typy zmiennych losowych: zmienne losowe typu skokowego i zmienne losowe typu ciągłego.

Definicja 13 Zmienna losowa X jest typu skokowego (dyskretnego), jeżeli przyjmuje co najwyżej przeliczalną liczbę wartości $x_1, x_2, \dots, x_n, \dots$ oraz

$$P(X = x_i) = p_i > 0, \quad i = 1, 2, \dots, \quad (2.4)$$

przy czym

$$\sum_{i=1}^{\infty} p_i = 1, \quad (2.5)$$

gdzie górna granica sumowania wynosi n albo ∞ stosownie do tego czy zbiór wartości jest skończony czy też przeliczalny, ale nieskończony.

W takim przypadku rozkład prawdopodobieństwa wygodnie będzie zadać w postaci dwuwierszowej tablicy:

x_i	x_1	x_2	\dots	x_n	\dots
p_i	p_1	p_2	\dots	p_n	\dots

Dystrybuanta zmiennej losowej skokowej X wyraża się wzorem:

$$F(x) = \sum_{x_i \leq x} p_i. \quad (2.7)$$

(patrz: Przykład 2.2)

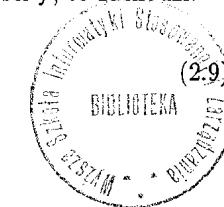
W przypadku zmiennej losowej X typu skokowego cała masa prawdopodobieństwa jest rozzielona pomiędzy punkty skokowe. W każdym punkcie skokowym x_i skupione jest niezerowe prawdopodobieństwo $p_i = P(X = x_i)$. Przedziemy teraz do przypadku, gdy w żadnym punkcie na prostej nie jest skupione dodatnie prawdopodobieństwo, a cała masa prawdopodobieństwa rozłożona jest w sposób ciągły.

Definicja 14 Zmienna losowa X jest typu ciągłego, jeżeli istnieje niejemna funkcja f - zwana gęstością - taka, że dystrybuantę tej zmiennej można przedstawić w postaci:

$$F(x) = \int_{-\infty}^x f(t)dt \quad \text{dla } x \in R. \quad (2.8)$$

Jeżeli zmieniona losowa X jest typu ciągłego o gęstości f , to zachodzi:

$$\int_{-\infty}^{+\infty} f(x)dx = F(+\infty) = 1 \quad (2.9)$$



oraz w punktach ciągłości gęstości f :

$$F'(x) = f(x). \quad (2.10)$$

(patrz: Przykład 2.5)

W wielu zagadnieniach praktycznych zachodzi konieczność wyznaczania rozkładów zmiennych losowych, które są funkcjami zadanych zmiennych losowych. Jeżeli X jest zmienną losową o rozkładzie skokowym $P(X = x_i) = p_i$ i jeżeli zmienna losowa $Y = g(X)$ to zmienna Y jest skokowa i jej rozkład zadany jest prawdopodobieństwami:

$$P(Y = y_k) = \sum_{i: g(x_i) = y_k} p_i, \quad (2.11)$$

gdzie sumowanie obejmuje te wskaźniki i , dla których $g(x_i) = y_k$. Jeżeli X jest zmienną losową ciągłą o gęstości f , to następujące twierdzenie pozwala na wyznaczenie gęstości zmiennej losowej $Y = g(X)$.

Twierdzenie 15 *Załóżmy, że X jest zmienną losową o gęstości f_X dodatniej w pewnym przedziale Δ i równej 0 poza tym przedziałem. Jeżeli funkcja rzeczywista g jest w przedziale Δ ścisłe monotoniczna, różniczkowalna, o pochodnej różnej od zera, to zmienna losowa $Y = g(X)$ ma gęstość:*

$$f_Y(y) = \begin{cases} f(h(y)) |h'(y)| & \text{dla } y \in \Delta_1 \\ 0 & \text{dla } y \notin \Delta_1 \end{cases} \quad (2.12)$$

gdzie Δ_1 jest przedziałem, na który funkcja g odwzorowuje przedział Δ , a funkcja h jest funkcją odwrotną do funkcji g .

2.1.4 Parametry zmiennych losowych jednowymiarowych

Każda zmienna losowa jest w pełni opisana przez jej rozkład prawdopodobieństwa (w ogólności przez jej dystrybuantę). Wzgłydy praktyczne dyktują potrzebę znalezienia pewnych charakterystyk liczbowych rozkładu, ponieważ są to krótkie opisy umożliwiające szybkie porównywanie rozkładów ze sobą. W tym punkcie zostaną zaprezentowane najważniejsze charakterystyki liczbowe dla zmiennych losowych typu skokowego i ciągłego. Będą to:

- **miary położenia** - wartość oczekiwana, kwantyl rzędu α (w szczególności mediana, kwartyl dolny i kwartyl górny), moda,
- **miary rozproszenia** - wariancja, odchylenie standardowe, odchylenie przeciętne, współczynnik zmienności,
- **charakterystyki kształtu** - współczynnik skośności i współczynnik skupienia.

Niech $g(X)$ będzie funkcją zmiennej losowej X .

Definicja 16 *Wartością oczekiwana zmiennej losowej $g(X)$, nazywamy liczbę*

$$Eg(X) = \sum_i g(x_i) p_i \quad (2.13)$$

w przypadku, gdy zmienna losowa X ma rozkład skokowy $P(X = x_i) = p_i$ i liczbe

$$Eg(X) = \int_{-\infty}^{+\infty} g(x) f(x) dx \quad (2.14)$$

jeżeli X ma rozkład ciągły o gęstości f przy założeniu, że szereg i całka są bezwzględnie zbieżne, tzn.,

$$\sum_{i=1}^{\infty} |g(x_i)| p_i < \infty, \quad \int_{-\infty}^{+\infty} |g(x)| f(x) dx < \infty. \quad (2.15)$$

Rozpatrzymy teraz szczególne przypadki funkcji g . Jeżeli $g(x) = x^k$ to:

Definicja 17 *Momentem zwykłym rzędu k zmiennej losowej X nazywamy liczbę*

$$EX^k = \sum_i x_i^k p_i, \quad (2.16)$$

w przypadku gdy zmienna losowa X ma rozkład skokowy i $P(X = x_i) = p_i$, bądź liczbe

$$EX^k = \int_{-\infty}^{+\infty} x^k f(x) dx, \quad (2.17)$$

w przypadku gdy zmienna losowa X ma rozkład ciągły o gęstości f .

Jeżeli $k = 1$ to wartością oczekiwana (wartością przeciętną) zmiennej losowej X nazywamy:

$$\mu = EX = \begin{cases} \sum_i x_i p_i & \text{gdy } X \text{ jest skokową zmienną losową} \\ \int_{-\infty}^{+\infty} x f(x) dx & \text{gdy } X \text{ jest ciągłą zmienną losową.} \end{cases} \quad (2.18)$$

(patrz: Przykład 2.1, 2.2, 2.5)

Należy zauważyć, że wartość oczekiwana zmiennej losowej X jest odpowiednikiem znanego z fizyki pojęcia *środkę ciężkości*, jeśli prawdopodobieństwa zinterpretujemy jako masy, a przyjęty układ jednostek jest taki, aby masa całkowita była równa 1. Innymi, ważnymi miarami położenia dla zmiennej losowej X są kwantyle.

Definicja 18 *Kwantylem rzędu α , ($0 < \alpha < 1$) zmiennej losowej X o dystrybuancie F nazywamy liczbę q_α spełniającą zależność:*

$$F(q_\alpha^-) \leq \alpha \leq F(q_\alpha). \quad (2.19)$$

W szczególności, kwantyl $q_{0.5}$ rzędu 0.5 nazywamy medianą zmiennej losowej X a liczby $q_{0.25}$ i $q_{0.75}$, odpowiednio, kwartylem dolnym i kwartylem górnym. Kwantyl q_α ma ważną interpretację, a mianowicie: $\alpha\%$ masy prawdopodobieństwa zmiennej losowej X nie przekracza liczby q_α .

(patrz: Przykład 2.7)

Jeżeli zmieniona losowa X jest typu ciągłego o gęstości f , to kwantyl rzędu α można wyznaczyć zgodnie ze wzorem:

$$\int_{-\infty}^{q_\alpha} f(x)dx = \alpha. \quad (2.20)$$

Z pozostałych parametrów położenia wymienimy **modę**.

Definicja 19 Jeżeli zmieniona losowa X ma rozkład skokowy to modą m_0 (wartością modalną, dominantą) nazywamy ten punkt skokowy x_i , różny od $\min(x_k)$ i $\max(x_k)$, dla którego prawdopodobieństwo p_i osiąga maksimum absolutne.

Jeżeli zmieniona losowa X ma rozkład ciągły o gęstości f to modą m_0 nazywamy odciętą maksimum absolutnego gęstości (przy założeniu, że punkt ten nie leży na brzegu nośnika gęstości).

Wymienimy teraz podstawowe miary rozproszenia dla zmiennych losowych. Rozważmy (2.13) i (2.14) z funkcją $g(x) = (x - \mu)^k$.

Definicja 20 *Momentem centralnym rzędu k zmiennej losowej X nazywamy liczbę*

$$E(X - \mu)^k = \sum_i (x_i - \mu)^k p_i, \quad (2.21)$$

w przypadku, gdy zmieniona losowa X ma rozkład skokowy i $P(X = x_i) = p_i$, bądź liczbę

$$E(X - \mu)^k = \int_{-\infty}^{+\infty} (x - \mu)^k f(x)dx, \quad (2.22)$$

w przypadku, gdy zmieniona losowa X ma rozkład ciągły o gęstości f . W obu przypadkach $\mu = E(X)$ jest wartością oczekiwana zmiennej losowej X .

Drugi moment centralny nazywamy **wariancją** zmiennej losowej X i obliczamy go zgodnie ze wzorem

$$Var X = \sigma^2 = E(X - \mu)^2 = \quad (2.23)$$

$$= \begin{cases} \sum_i (x_i - \mu)^2 p_i & \text{gdy } X \text{ jest skokową zmienią losową} \\ \int_{-\infty}^{+\infty} (x - \mu)^2 f(x)dx & \text{gdy } X \text{ jest ciągłą zmienią losową.} \end{cases}$$

Wariancja zmiennej losowej jest średniokwadratowym odchyleniem zmiennej losowej od jej wartości średniej i jest jednym z podstawowych parametrów, które mówią o rozproszeniu zmiennej losowej. Z wariancją zmiennej losowej związana jest inna miara rozproszenia, a mianowicie **odchylenie standardowe** zmiennej losowej

$$\sigma = \sqrt{Var X}. \quad (2.24)$$

(patrz: Przykład 2.1, 2.2, 2.5)

Za miarę rozproszenia może, obok odchylenia standardowego, służyć też tzw. odchylenie przeciętne.

Definicja 21 *Odchyleniem przeciętnym jest liczba*

$$d = \begin{cases} \sum_i |x_i - \mu| p_i & \text{gdy } X \text{ jest skokową zmienią losową} \\ \int_{-\infty}^{+\infty} |x - \mu| f(x)dx & \text{gdy } X \text{ jest ciągłą zmienią losową.} \end{cases} \quad (2.25)$$

Zarówno odchylenie standardowe jak i odchylenie przeciętne mówią o ile przeciętnie różnią się wartości zmiennej X od średniej μ . Dogodnie jest nieraz scharakteryzować rozproszenie nie za pomocą odchylenia standar-dowego, lecz za pomocą stosunku tego odchylenia do wartości oczekiwanej.

Definicja 22 *Stosunek odchylenia standardowego do wartości oczekiwanej nazywamy **współczynnikiem zmienności** i oznaczamy przez*

$$v = \frac{\sigma}{\mu}. \quad (2.26)$$

Współczynnik zmienności v jest odchyleniem standardowym, gdy wartość oczekiwana równa się jedności. Inaczej mówiąc, współczynnik zmienności jest miarą rozproszenia, gdy za jednostkę przyjmujemy wartość oczekiwana. Współczynnik ten mówi o zróżnicowaniu wartości zmiennej X względem wartości średniej μ . Często wyrażany jest w procentach.

Podamy teraz określenia dla parametrów kształtu rozkładu. Mówiąc o kształcie rozkładu, mamy tu na myśli kształt jego gęstości (w przypadku zmiennych losowych ciągłych), bądź kształt funkcji masy prawdopodobieństwa (w przypadku zmiennych losowych skokowych). Wiadomo, że jeżeli zmienność losowa ma rozkład symetryczny i skońzoną wartość oczekiwana, to wartość oczekiwana jest środkiem symetrii. Wynika stąd, że dla rozkładu symetrycznego momenty centralne rzędu nieparzystego są równe zeru. Niektóre potrzebne jest ustalenie stopnia asymetrii rozkładu.

Definicja 23 *Liczba*

$$\gamma = \frac{E(X - \mu)^3}{\sigma^3} \quad (2.27)$$

nazywa się **współczynnikiem asymetrii** (współczynnikiem skośności) zmiennej losowej X .

Dla rozkładu symetrycznego $\gamma = 0$. Jeżeli $\gamma > 0$, to rozkład jest prawoskoły (asymetria dodatnia), a jeżeli $\gamma < 0$, to rozkład jest lewoskoły (asymetria ujemna).

Drugim parametrem kształtu jest kurtoza, która mówi o koncentracji rozkładu wokół średniej.

Definicja 24 *Liczba*

$$\eta = \frac{E(X - \mu)^4}{\sigma^4} \quad (2.28)$$

nazywa się **kurtozą** (współczynnikiem spłaszczenia) zmiennej losowej X .

Im wyższa wartość η , tym większa wysmukłość rozkładu. Małe wartości tej miary oznaczają rozkład spłaszczony. Przyjmuje się, że dla rozkładu normalnego $\eta = 3$, dla spłaszczonego $\eta < 3$, a dla wysmukłego $\eta > 3$.

Przy porównywaniu dwóch rozkładów stosowana jest miara spłaszczenia nazywana **ekscesem**. Wartość ekscesu równa się wartości kurtozy pomniejszonej o 3. Jeżeli zatem $\eta - 3 = 0$, to rozkład ma kształt normalny (patrz: rozdz. 2.1.6), jeżeli $\eta - 3 > 0$, to rozkład jest spłaszczony w stosunku do rozkładu normalnego. Eksces informuje więc o tym, czy koncentracja wartości zmiennej wokół średniej jest mniejsza, czy też większa niż dla rozkładu normalnego.

2.1.5 Własności wartości oczekiwanej i wariancji

Wartość oczekiwana zmiennej losowej X ma następujące własności (przez a, b, c oznaczamy stałe):

$$E(c) = c \quad (2.29)$$

$$E(aX) = aEX \quad (2.30)$$

$$E(X + b) = EX + b \quad (2.31)$$

$$E(X + Y) = EX + EY \quad (2.32)$$

$$X \geq 0 \Rightarrow EX \geq 0 \quad (2.33)$$

$$X \geq Y \Rightarrow EX \geq EY \quad (2.34)$$

$$E|X| \geq EX \quad (2.35)$$

(patrz: Przykład 2.8, 2.9)

Warto też pamiętać, że jeżeli zmienność losowa X ma rozkład symetryczny względem punktu x_0 i jeżeli istnieje wartość oczekiwana EX , to wówczas $EX = x_0$.

Z kolei wariancja zmiennej losowej X ma następujące własności (jak poprzednio a, b, c są stałymi):

$$Var X \geq 0 \quad (2.36)$$

$$Var X = 0 \Leftrightarrow \text{istnieje } c, \text{ że } P(X = c) = 1 \quad (2.37)$$

$$Var(c) = 0 \quad (2.38)$$

$$Var(aX) = a^2 Var X \quad (2.39)$$

$$Var(X + b) = Var X \quad (2.40)$$

$$Var X = EX^2 - (EX)^2 \quad (2.41)$$

(patrz: Przykład 2.1, 2.2, 2.5, 2.8)

2.1.6 Podstawowe rozkłady prawdopodobieństwa zmiennych losowych

Przegląd podstawowych rozkładów zaczniemy od rozkładów skokowych.

- **Rozkład dwupunktowy – $Bern(p)$**

Zmienna losowa X ma rozkład dwupunktowy z parametrem p ($0 < p < 1$), jeżeli:

$$P(X = 1) = p, \quad P(X = 0) = q = 1 - p. \quad (2.42)$$

Wartość oczekiwana i wariancja dane są wzorami:

$$EX = p, \quad VarX = pq. \quad (2.43)$$

Zmienna losowa X związana jest z jednym doświadczeniem: przyjmuje wartość 1 (z prawdopodobieństwem p), jeżeli w danym doświadczeniu zaistniał "sukces" i wartość 0 (z prawdopodobieństwem q), jeżeli w danym doświadczeniu zaistniała "porażka".

Uwaga!

W literaturze doświadczenie opisywane rozkładem dwupunktowym nazywane jest zazwyczaj *doświadczeniem Bernoulliego*. Z tego też powodu w niniejszym podręczniku jako oznaczenie tego rozkładu przyjęto symbol $Bern$. Trzeba przy tym rozróżnić pojedyncze doświadczenie Bernoulliego od ciągu niezależnych powtórzeń takiego doświadczenia, opisywanego rozkładem dwumianowym (patrz poniżej), który w literaturze polskojęzycznej bywa często nazywany rozkładem Bernoulliego.

- **Rozkład dwumianowy – $Bin(n, p)$**

Zmienna losowa X ma rozkład dwumianowy z parametrami p ($0 < p < 1$) i n ($n \in N, n \geq 1$) jeżeli:

$$P(X = k) = \binom{n}{k} p^k q^{n-k}, \quad k = 0, 1, 2, \dots, n \quad (2.44)$$

gdzie $q = 1 - p$. Wartość oczekiwana i wariancja dane są wzorami:

$$EX = np, \quad VarX = npq. \quad (2.45)$$

Zmienna losowa X przyjmuje wartości równe liczbie sukcesów w n niezależnych doświadczeniach z prawdopodobieństwem sukcesu p w każdym z nich.

Czasem interesuje nas najbardziej prawdopodobna liczba sukcesów

w rozkładzie dwumianowym. W przypadku, gdy $(n+1)p$ nie jest liczbą całkowitą, wtedy najbardziej prawdopodobna liczba sukcesów wynosi $\lfloor (n+1)p \rfloor$ (zapis $[a]$ oznacza największą liczbę całkowitą nie przekraczającą wartości danej liczby a). Natomiast w przypadku gdy $(n+1)p$ jest liczbą całkowitą, wówczas istnieją dwie najbardziej prawdopodobne liczby sukcesów (przyjmowane z tym samym prawdopodobieństwem), a mianowicie liczby: $(n+1)p$ i $(n+1)p - 1$. (patrz: Przykład 2.3)

- **Rozkład ujemny dwumianowy – $NB(r, p)$**

Zmienna losowa X ma rozkład ujemny dwumianowy z parametrami p ($0 < p < 1$) i r ($r \in N, r \geq 1$) jeżeli:

$$P(X = k) = \binom{k-1}{r-1} p^r q^{k-r}, \quad k = r, r+1, \dots \quad (2.46)$$

Wartość oczekiwana i wariancja dane są wzorami:

$$EX = \frac{r}{p}, \quad VarX = \frac{rq}{p^2}. \quad (2.47)$$

Zmienna losowa X przyjmuje wartości równe liczbie niezależnych doświadczeń (z prawdopodobieństwem sukcesu p , w każdym z nich), potrzebnej do uzyskania r sukcesów.

- **Rozkład geometryczny – $G(p)$**

Zmienna losowa X ma rozkład geometryczny z parametrem p ($0 < p < 1$), jeżeli:

$$P(X = k) = pq^{k-1}, \quad k = 1, 2, \dots \quad (2.48)$$

Wartość oczekiwana i wariancja dane są wzorami:

$$EX = \frac{1}{p}, \quad VarX = \frac{q}{p^2}. \quad (2.49)$$

Rozkład geometryczny jest szczególnym przypadkiem ujemnego rozkładu dwumianowego, a mianowicie rozkładu $NB(1, p)$. Zmienna losowa X o rozkładzie geometrycznym przyjmuje wartości równe liczbie doświadczeń do uzyskania pierwszego sukcesu.

- **Rozkład hipergeometryczny – $HG(N, M, n)$**

Zmienna losowa X ma rozkład hipergeometryczny z parametrami N, M, n (gdzie N, M, n są liczbami naturalnymi, przy czym $M \leq N$ oraz $n \leq N$) jeżeli:

$$P(X = k) = \frac{\binom{M}{k} \binom{N-M}{n-k}}{\binom{N}{n}}, \quad k \leq \min(M, n), \quad n - k \leq N - M. \quad (2.50)$$

Zmienną losową o rozkładzie hipergeometrycznym można interpretować jako liczbę wyróżnionych elementów w n -elementowej próbce, jeżeli cała N -elementowa populacja zawiera M -elementów wyróżnionych. Wartość oczekiwana i wariancja dane są wzorami:

$$EX = \frac{Mn}{N}, \quad VarX = \frac{Mn(N-n)(N-M)}{N^2(N-1)} \quad (2.51)$$

• **Rozkład Poissona – $P(\lambda)$**

Zmienna losowa X ma rozkład Poissona z parametrem $\lambda > 0$, jeżeli:

$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}, \quad k = 0, 1, 2, \dots \quad (2.52)$$

Wartość oczekiwana i wariancja dane są wzorami:

$$EX = \lambda, \quad VarX = \lambda. \quad (2.53)$$

(patrz: Przykład 2.3, 2.6)

Do najczęściej spotykanych rozkładów ciągłych należą:

• **Rozkład jednostajny – $U(a, b)$**

Zmienna losowa X ma rozkład jednostajny na przedziale $[a, b]$, jeżeli jej gęstość f jest postaci:

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{dla } x \in [a, b] \\ 0 & \text{dla } x \notin [a, b]. \end{cases} \quad (2.54)$$

Wartość oczekiwana i wariancja dane są wzorami:

$$EX = \frac{a+b}{2}, \quad VarX = \frac{(b-a)^2}{12}. \quad (2.55)$$

• **Rozkład wykładniczy – $Exp(\lambda)$**

Zmienna losowa X ma rozkład wykładniczy z parametrem $\lambda > 0$, jeżeli jej gęstość f jest postaci:

$$f(x) = \begin{cases} \lambda \exp(-\lambda x) & \text{dla } x > 0 \\ 0 & \text{dla } x \leq 0. \end{cases} \quad (2.56)$$

Wartość oczekiwana i wariancja dane są wzorami:

$$EX = \frac{1}{\lambda}, \quad VarX = \frac{1}{\lambda^2}. \quad (2.57)$$

Rozkład wykładniczy odgrywa dużą rolę np. w teorii niezawodności związanej z czasem poprawnej pracy elementu, urządzenia itp. W wielu przypadkach zakłada się, że czas działania elementu ma rozkład wykładniczy.

W tym miejscu chcielibyśmy podać pewne praktyczne związki pomiędzy procesami zgłoszeń i rozkładami czasów między kolejnymi zgłoszeniami. Jeżeli zgłoszenia do systemu (np. zgłoszenia do centrali telefonicznej, zgłoszenia do procesora, awarie elementów w urządzeniach, zgłoszenia klientów do banku) nadchodzą niezależnie w przedziale czasu $[0, t]$ i liczba zgłoszeń w tym przedziale nie zależy od liczby zgłoszeń, które miały miejsce przed chwilą wystąpienia danego zgłoszenia, to można powiedzieć, że zmienna losowa X opisująca liczbę zgłoszeń do systemu w przedziale czasu $[0, t]$ ma rozkład Poissona:

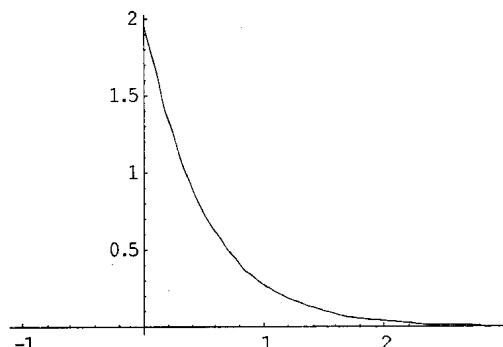
$$P(X = k) = \frac{(\lambda t)^k}{k!} e^{-\lambda t}, \quad k = 1, 2, \dots \quad (2.58)$$

gdzie λ jest intensywnością zgłoszeń. Dowodzi się wówczas, że czas T między kolejnymi zgłoszeniami ma rozkład wykładniczy z parametrem λ :

$$P(T \leq x) = 1 - e^{-\lambda x} \quad (2.59)$$

i na odwrót, jeżeli czas między kolejnymi zgłoszeniami ma rozkład wykładniczy, to liczba zgłoszeń w rozpatrywanym przedziale ma rozkład Poissona (patrz: Przykład 2.6).

Przykładowy wykres gęstości dla rozkładu wykładniczego $Exp(2)$ przedstawia rysunek rys. 2.1.



Rys. 2.1 – Gęstość rozkładu wykładniczego

• **Rozkład normalny – $N(\mu, \sigma)$**

Zmienna losowa X ma rozkład normalny z parametrami μ i $\sigma > 0$,

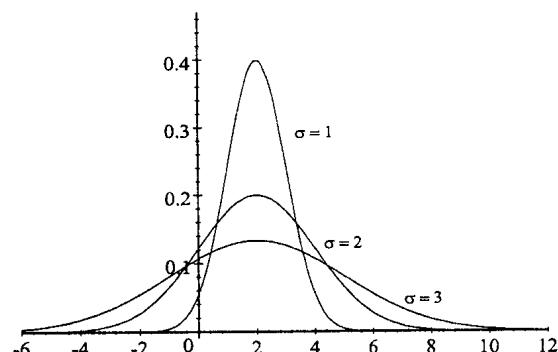
jeżeli jej gęstość f jest postaci:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right], \quad x \in R. \quad (2.60)$$

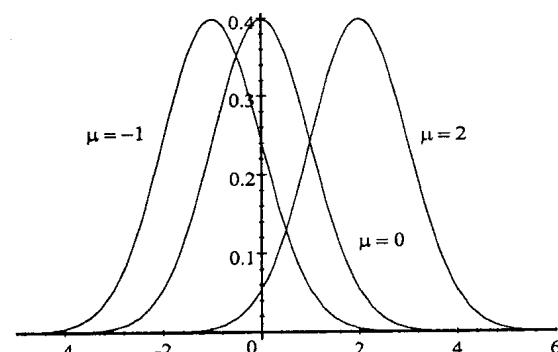
Wartość oczekiwana i wariancja dane są wzorami:

$$EX = \mu, \quad VarX = \sigma^2. \quad (2.61)$$

Na rysunku rys. 2.2 przedstawione są wykresy gęstości trzech rozkładów normalnych: $N(2, 1)$, $N(2, 2)$ i $N(2, 3)$. Ponieważ, dla tych rozkładów, wartości oczekiwane są takie same, więc ten rozkład, dla którego odchylenie standardowe jest mniejsze, jest mniej rozproszony (bardziej skupiony) względem wartości średniej. Jeżeli założymy, że odchylenia standardowe są takie same, to wykresy gęstości rozkładów normalnych różnią się tylko przesunięciem (rys. 2.3 przedstawia wykresy gęstości dla rozkładów normalnych $N(-1, 1)$, $N(0, 1)$ i $N(2, 1)$).



Rys. 2.2 – Rozkłady normalne o różnych odchyleniach standardowych



Rys. 2.3 – Rozkłady normalne o różnych wartościach oczekiwanych

Warto pamiętać, że za pomocą prostej operacji, zwanej *standardyzacją*, można dowolny rozkład normalny $N(\mu, \sigma)$ sprowadzić do rozkładu normalnego o zerowej wartości oczekiwanej i jednostkowym odchyleniu standardowym, tzn. $N(0, 1)$. Mianowicie, jeżeli zmienna losowa X ma rozkład $N(\mu, \sigma)$, wówczas zmienna losowa

$$Y = \frac{X - \mu}{\sigma} \quad (2.62)$$

ma rozkład $N(0, 1)$. Właśnie ta jest o tyle istotna, że rozkład $N(0, 1)$, zwany **rozkładem normalnym standardowym**, jest stablicowany, co bardzo ułatwia dokonywanie obliczeń (patrz: Przykład 2.7). Dodańnie, powszechnie stosowane są tablice wartości dystrybuanty tego rozkładu, oznaczanej zwyczajowo symbolem $\Phi(x)$ (patrz podrozdz. 6.1).

- **Rozkład logarytmiczno-normalny – $LN(\mu, \sigma)$**

Zmienna losowa X ma rozkład logarytmiczno-normalny z parametrami μ i $\sigma > 0$, jeżeli jej gęstość f jest postaci:

$$f(x) = \begin{cases} \frac{1}{x\sigma\sqrt{2\pi}} \exp\left[-\frac{(\ln x - \mu)^2}{2\sigma^2}\right] & \text{dla } x > 0 \\ 0 & \text{dla } x \leq 0. \end{cases} \quad (2.63)$$

Wartość oczekiwana i wariancja dane są wzorami:

$$EX = \exp(\mu + \frac{\sigma^2}{2}), \quad VarX = \exp(2\mu + 2\sigma^2) - \exp(2\mu + \sigma^2). \quad (2.64)$$

Jeżeli zmienna losowa X ma rozkład logarytmiczno-normalny $LN(\mu, \sigma)$, to zmienna losowa $Y = \ln(X)$ ma rozkład normalny $N(\mu, \sigma)$.

- **Rozkład gamma – $\Gamma(\alpha, \beta)$**

Zmienna losowa X ma rozkład gamma z parametrami $\alpha > 0$ i $\beta > 0$, jeżeli jej gęstość f jest postaci:

$$f(x) = \begin{cases} \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\beta x) & \text{dla } x > 0 \\ 0 & \text{dla } x \leq 0, \end{cases} \quad (2.65)$$

gdzie

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t} dt, \quad \alpha > 0. \quad (2.66)$$

Warto zauważać, że

$$\Gamma(1) = 1, \quad \Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1) \quad \text{dla } \alpha > 1.$$

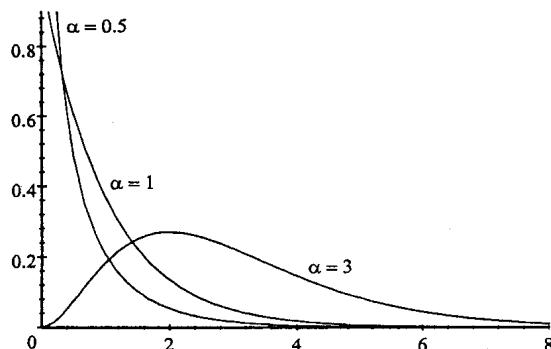
W szczególności

$$\Gamma(n) = (n - 1)! \quad \text{dla } n \in N.$$

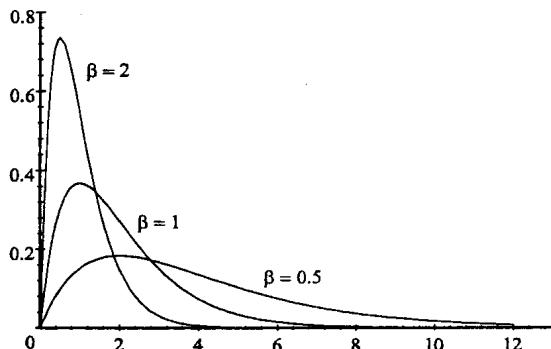
Wartość oczekiwana i wariancja rozkładu $\Gamma(\alpha, \beta)$ dane są wzorami:

$$EX = \frac{\alpha}{\beta}, \quad VarX = \frac{\alpha}{\beta^2}. \quad (2.67)$$

Parametr α nazywa się parametrem kształtu a parametr β jest parametrem skali. Rozkład wykładniczy jest szczególnym przypadkiem rozkładu gamma $\Gamma(1, \beta)$. Na rys. 2.4 przedstawione są wykresy gęstości dla trzech rozkładów gamma o tym samym parametrze skali $\beta = 1$ i różnych parametrach kształtu α , a mianowicie $\Gamma(0.5, 1)$, $\Gamma(1, 1)$, i $\Gamma(3, 1)$. Z kolei na rys. 2.5 pokazano, w jaki sposób zmiana parametru skali wpływa na kształt gęstości rozkładu gamma (wykonano wykresy gęstości dla rozkładów $\Gamma(2, 0.5)$, $\Gamma(2, 1)$ i $\Gamma(2, 2)$).



Rys. 2.4 – Rozkłady gamma o różnych parametrach kształtu



Rys. 2.5 – Rozkłady gamma o różnych parametrach skali

- **Rozkład chi-kwadrat – $\chi^2(n)$**

Zmienna losowa X ma rozkład chi-kwadrat o n ($n \in N$, $n \geq 1$) stopniach swobody, jeżeli jej gęstość f jest postaci:

$$f(x) = \begin{cases} \frac{1}{2^{n/2}\Gamma(n/2)} x^{(n-2)/2} \exp(-\frac{x}{2}) & \text{dla } x > 0, \\ 0 & \text{dla } x \leq 0. \end{cases} \quad (2.68)$$

Wartość oczekiwana i wariancja dane są wzorami:

$$EX = n, \quad VarX = 2n. \quad (2.69)$$

Rozkład chi-kwadrat jest szczególnym przypadkiem rozkładu gamma $\Gamma(\frac{n}{2}, \frac{1}{2})$.

- **Rozkład t-Studenta – $t(n)$**

Zmienna losowa X ma rozkład t-Studenta o n ($n \in N$, $n \geq 1$) stopniach swobody, jeżeli jej gęstość f jest postaci:

$$f(x) = \frac{\Gamma((n+1)/2)}{\sqrt{n\pi}\Gamma(n/2)} (1 + \frac{x^2}{n})^{-(n+1)/2}, \quad x \in R. \quad (2.70)$$

Wartość oczekiwana i wariancja, dla $n > 1$, dane są wzorami:

$$EX = 0, \quad VarX = \frac{n}{n-2}. \quad (2.71)$$

- **Rozkład F-Snedecora – $F(m, n)$**

Zmienna losowa X ma rozkład F-Snedecora z parametrami $m, n = 1, 2, \dots$, jeżeli jej gęstość f jest postaci:

$$f(x) = \begin{cases} \frac{\Gamma(\frac{1}{2}(n+m))}{\Gamma(\frac{1}{2}n)\Gamma(\frac{1}{2}m)} \left(\frac{m}{n}\right)^{m/2} x^{n/2-1} (x + \frac{m}{n})^{-(n+m)/2} & \text{dla } x \geq 0, \\ 0 & \text{dla } x < 0. \end{cases} \quad (2.72)$$

Wartość oczekiwana i wariancja dane są wzorami:

$$EX = \frac{m}{m-2}, \quad VarX = \frac{2m^2(n+m-2)}{n(m-2)(m-4)}. \quad (2.73)$$

2.2 Wielowymiarowe zmienne losowe

2.2.1 Dwuwymiarowe zmienne losowe

Zmienne losowe jednowymiarowe służą do modelowania takich doświadczeń losowych, których wyniki można przedstawić za pomocą liczb rzeczywistych. Istnieje jednak wiele doświadczeń losowych, których wyniki przedstawione są za pomocą par liczb rzeczywistych (np. gdy badamy prędkość

i drogę zatrzymania samochodu, gdy badamy wzrost i wagę człowieka, gdy badamy długość i wytrzymałość włókna bawełny), trójkę liczb rzeczywistych (np. gdy badamy ciśnienie, objętość i temperaturę gazu), lub ogólniej za pomocą ciągów n -wyrazowych liczb rzeczywistych. Do modelowania takich doświadczeń służą zmienne losowe dwuwymiarowe, trójwymiarowe lub ogólniej zmienne losowe n -wymiarowe.

W tej części podamy podstawowe określenia dotyczące zmiennych losowych dwuwymiarowych (uogólnienia określeń na przypadek zmiennych losowych wielowymiarowych o wymiarze większym niż dwa są analogiczne). Podobnie, jak w przypadku jednowymiarowym, podamy najpierw definicję dystrybuanty dla zmiennej losowej dwuwymiarowej. Niech X i Y będą zmiennymi losowymi określonymi niekoniecznie na tej samej przestrzeni probabilistycznej. Parę (X, Y) nazywamy **dwuwymiarową zmienną losową** lub **dwuwymiarowym wektorem losowym**, a X oraz Y jej współzależnymi.

Definicja 25 Dystrybuantą zmiennej losowej (X, Y) nazywamy funkcję rzeczywistą F , która jest określona dla wszystkich liczb rzeczywistych $x \in (-\infty, +\infty)$ i $y \in (-\infty, +\infty)$ wzorem:

$$F(x, y) = P(X \leq x, Y \leq y) \quad (2.74)$$

Dystrybuanta F posiada następujące własności:

W1 $\forall (x, y) \in R^2 \quad 0 \leq F(x, y) \leq 1,$

W2 F jest funkcją niemalejącą ze względu na każdy z argumentów,

W3 F jest funkcją co najmniej prawostronnie ciągłą ze względu na każdy z argumentów,

W4 dla każdego $x \in R$ zachodzi $\lim_{y \rightarrow -\infty} F(x, y) = 0$,
dla każdego $y \in R$ zachodzi $\lim_{x \rightarrow -\infty} F(x, y) = 0$,
oraz $\lim_{x \rightarrow +\infty, y \rightarrow +\infty} F(x, y) = 1$,

W5 dla dowolnych punktów: (x_1, y_1) i (x_2, y_2) takich, że $x_1 \leq x_2$ i $y_1 \leq y_2$, zachodzi nierówność:

$$F(x_2, y_2) - F(x_2, y_1) - F(x_1, y_2) + F(x_1, y_1) \geq 0. \quad (2.75)$$

Podobnie, jak w przypadku zmiennej losowej jednowymiarowej dowodzi się twierdzenia:

Twierdzenie 26 Jeżeli funkcja F spełnia własności W1, W2, W3, W4 i W5, to jest ona dystrybuantą pewnej dwuwymiarowej zmiennej losowej (X, Y) .

Korzystając z dystrybuanty wyznaczamy prawdopodobieństwa zdarzeń, np.

$$\begin{aligned} P(x_1 < X \leq x_2, y_1 < Y \leq y_2) &= \\ &= F(x_2, y_2) - F(x_2, y_1) - F(x_1, y_2) + F(x_1, y_1). \end{aligned} \quad (2.76)$$

Dystrybuantę zmiennej losowej (X, Y) nazywamy czasem dystrybuantą rozkładu łącznego, w przeciwieństwie do dystrybuant rozkładów brzegowych, określonych poniżej:

Twierdzenie 27 Jeżeli F jest dystrybuantą zmiennej losowej dwuwymiarowej (X, Y) , to funkcja

$$F_X(x) = F(x, \infty) \quad \text{dla } x \in R \quad (2.77)$$

jest dystrybuantą zmiennej losowej X , zaś funkcja

$$F_Y(y) = F(\infty, y) \quad \text{dla } y \in R \quad (2.78)$$

jest dystrybuantą zmiennej losowej Y .

Funkcję F_X nazywamy dystrybuantą brzegową zmiennej losowej X . Funkcję F_Y nazywamy dystrybuantą brzegową zmiennej losowej Y . Z twierdzenia wynika, że rozkład prawdopodobieństwa zmiennej losowej dwuwymiarowej (X, Y) wyznacza rozkłady prawdopodobieństwa zmiennych losowych X i Y . Rozkłady te nazywamy rozkładami brzegowymi. Rozkłady brzegowe zmiennych losowych X i Y nie wyznaczają rozkładu dwuwymiarowej zmiennej losowej (X, Y) .

Podobnie, jak w przypadku jednowymiarowych zmiennych losowych będziemy rozważać dwa podstawowe typy dwuwymiarowych zmiennych losowych: zmienne losowe typu skokowego i typu ciągłego.

Definicja 28 Zmienna losowa (X, Y) jest typu skokowego, jeżeli przyjmuje co najwyżej przeliczalną liczbę wartości (x_i, y_k) oraz:

$$P(X = x_i, Y = y_k) = p_{ik} \geq 0, \quad i, k = 1, 2, \dots \quad (2.79)$$

przy czym

$$\sum_i \sum_k p_{ik} = 1 \quad (2.80)$$

Jeżeli dwuwymiarowa zmienna (X, Y) przyjmuje skończoną liczbę wartości, to wygodnie jest rozkład zadać w tabeli dwudzielczej:

x_i					
y_k	x_1	x_2	\dots	x_m	$p_{\bullet k}$
y_1	p_{11}	p_{21}	\dots	p_{m1}	$p_{\bullet 1}$
y_2	p_{12}	p_{22}	\dots	p_{m2}	$p_{\bullet 2}$
\dots	\dots	\dots	\dots	\dots	\dots
y_s	p_{1s}	p_{2s}	\dots	p_{ms}	$p_{\bullet s}$
$p_{i\bullet}$	$p_{1\bullet}$	$p_{2\bullet}$	\dots	$p_{m\bullet}$	1

Jeżeli dany jest łączny rozkład zmiennej losowej (X, Y) to można wyznaczyć rozkłady brzegowe. Wtedy:

$$p_{i\bullet} = P(X = x_i) = \sum_k P(X = x_i, Y = y_k) = \sum_k p_{ik}, \quad (2.81)$$

$$p_{\bullet k} = P(Y = y_k) = \sum_i P(X = x_i, Y = y_k) = \sum_i p_{ik} \quad (2.82)$$

są rozkładami prawdopodobieństwa dla jednowymiarowych zmiennych losowych skokowych, odpowiednio, X i Y .

(patrz: **Przykład 2.9**)

Jako przykład wielowymiarowego rozkładu skokowego podamy rozkład wielomianowy.

Definicja 29 *Zmienna losowa (X_1, X_2, \dots, X_k) ma rozkład wielomianowy, jeżeli*

$$P(X_1 = n_1, \dots, X_k = n_k) = \frac{n!}{n_1! \dots n_k!} p_1^{n_1} \dots p_k^{n_k},$$

gdzie $p_i \in [0, 1]$, $i = 1, 2, \dots, k$, $p_1 + \dots + p_k = 1$, $n_1 + \dots + n_k = n$.

Jak widać, rozkład wielomianowy jest uogólnieniem rozkładu dwumianowego i opisuje rozkład wyników przy n -krotnym powtórzeniu doświadczenia o k możliwych rezultatach. X_i oznacza liczbę wyników i -tego typu w serii.

Przejdziemy teraz do określenia zmiennej losowej typu ciągłego.

Definicja 30 *Zmienna losowa (X, Y) jest typu ciągłego, jeżeli istnieje nieujemna funkcja f , zwana gęstością, taka, że dystrybuantę tej zmiennej można przedstawić w postaci:*

$$F(x, y) = \int_{-\infty}^x \int_{-\infty}^y f(u, t) du dt \quad \text{dla } (x, y) \in R^2. \quad (2.83)$$

Jeżeli zmienna losowa (X, Y) jest typu ciągłego o gęstości f , to:

$$\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x, y) dx dy = 1 \quad (2.84)$$

oraz w punktach ciągłości gęstości f zachodzi

$$\frac{\partial^2 F(x, y)}{\partial x \partial y} = f(x, y). \quad (2.85)$$

(patrz: **Przykład 2.10**)

Podobnie jak w przypadku zmiennej losowej typu skokowego, można wyznaczyć gęstości brzegowe dla poszczególnych współrzędnych. Funkcje:

$$f_X(x) = \int_{-\infty}^{+\infty} f(x, y) dy, \quad (2.86)$$

$$f_Y(y) = \int_{-\infty}^{+\infty} f(x, y) dx \quad (2.87)$$

są gęstościami, odpowiednio, zmiennych losowych X i Y .

(patrz: **Przykład 2.10**)

Jako przykład dwuwymiarowego rozkładu ciągłego podamy rozkład normalny.

Definicja 31 *Zmienna losowa (X, Y) ma rozkład dwuwymiarowy normalny z parametrami $\mu_X, \mu_Y, \sigma_X > 0, \sigma_Y > 0$ i $\rho \in (-1, 1)$, jeżeli jej gęstość f wyraża się wzorem*

$$f(x, y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp \left\{ -\frac{1}{2(1-\rho^2)} \left[\frac{(x-\mu_X)^2}{\sigma_X^2} - 2\rho \frac{(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \frac{(y-\mu_Y)^2}{\sigma_Y^2} \right] \right\}. \quad (2.88)$$

Można wykazać, że jeżeli dwuwymiarowa zmienna losowa (X, Y) ma rozkład normalny, to zmienne X i Y mają, odpowiednio, jednowymiarowe rozkłady normalne $N(\mu_X, \sigma_X)$ i $N(\mu_Y, \sigma_Y)$.

2.2.2 Niezależność zmiennych losowych

Jednym z ważniejszych pojęć rachunku prawdopodobieństwa jest pojęcie niezależności zmiennych losowych.

Definicja 32 Zmienne losowe X i Y nazywamy niezależnymi zmiennymi losowymi, jeżeli dla dowolnych zbiorów A i B na prostej zachodzi równość

$$P(X_1 \in A, X_2 \in B) = P(X_1 \in A)P(X_2 \in B). \quad (2.89)$$

W szczególnym przypadku, jeżeli zbiory $A = (-\infty, x)$, $B = (-\infty, x)$ to warunkiem koniecznym i wystarczającym na to, aby X i Y były niezależnymi zmiennymi losowymi jest, aby dla każdego $(x, y) \in R^2$ dystrybuanta F dwuwymiarowej zmiennej losowej (X, Y) była iloczynem dystrybuant brzegowych F_X i F_Y :

$$F(x, y) = F_X(x)F_Y(y) \quad \forall (x, y) \in R^2. \quad (2.90)$$

Warunkiem koniecznym i wystarczającym niezależności zmiennych losowych X i Y o gęstościach brzegowych f_X , f_Y jest

$$f(x, y) = f_X(x)f_Y(y) \quad \forall (x, y) \in R^2, \quad (2.91)$$

gdzie f jest gęstością dwuwymiarowej zmiennej losowej (X, Y) .

(patrz: **Przykład 2.10**)

W przypadku skokowym, zmienne losowe X i Y są niezależne wtedy i tylko wtedy, gdy

$$P(X = x_i, Y = y_k) = P(X = x_i)P(Y = y_k) \quad \forall i, k = 1, 2, \dots, \quad (2.92)$$

co zapisujemy również w postaci:

$$p_{ik} = p_{i\bullet} p_{\bullet k} \quad \forall i, k = 1, 2, \dots. \quad (2.93)$$

(patrz: **Przykład 2.9**)

2.2.3 Kowariancja i współczynnik korelacji

Niech (X, Y) będzie dwuwymiarową zmienną losową i niech $Z = g(X, Y)$ będzie funkcją tej dwuwymiarowej zmiennej losowej. Wartość oczekiwana zmiennej losowej Z określamy wzorem:

$$EZ = Eg(X, Y) = \sum_i \sum_k g(x_i, y_k) p_{ik}, \quad (2.94)$$

jeżeli (X, Y) jest zmienną losową typu skokowego o rozkładzie $P(X = x_i, Y = y_k) = p_{ik}$ oraz

$$EZ = Eg(X, Y) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} g(x, y) f(x, y) dx dy, \quad (2.95)$$

jeżeli (X, Y) jest zmienną losową ciągłą o gęstości $f(x, y)$.

(patrz: **Przykład 2.9, 2.10**)

Omawiając własności wartości oczekiwanej podaliśmy, że wartość oczekiwana sumy równa jest sumie wartości oczekiwanych (por. (2.32)). Nasuwa się pytanie, czy podobnie zachowuje się wartość oczekiwana iloczynu dwóch zmiennych losowych. Odpowiedź na to pytanie daje następujące twierdzenie:

Twierdzenie 33 Jeżeli zmienne losowe X i Y są niezależne i mają wartości oczekiwane, to

$$E(XY) = (EX)(EY). \quad (2.96)$$

Twierdzenie odwrotne nie jest prawdziwe, tzn. dla dwóch zmiennych może zachodzić warunek (2.96), a zmienne te będą mimo to zależne.

Definicja 34 Kowariancją zmiennych losowych X i Y nazywamy liczbę $Cov(X, Y)$ określoną wzorem:

$$Cov(X, Y) = E[(X - EX)(Y - EY)]. \quad (2.97)$$

Kowariancja może być traktowana jako pewna miara zgodności dwóch zmiennych losowych, będąc średnią wartością iloczynu odchyлеń obu zmiennych od ich wartości oczekiwanych. Bezpośrednio z definicji wynika, że

$$Cov(X, Y) = E(XY) - (EX)(EY) \quad (2.98)$$

gdzie

$$E(XY) = \sum_i \sum_k x_i y_k p_{ik}, \quad (2.99)$$

gdzie zmienna losowa (X, Y) jest typu skokowego oraz $P(X = x_i, Y = y_k) = p_{ik}$, bądź

$$E(XY) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} xy f(x, y) dx dy, \quad (2.100)$$

gdzie zmienna losowa (X, Y) jest typu ciągłego o gęstości f .

(patrz: **Przykład 2.9, 2.10**)

Jeżeli $Cov(X, Y) = 0$, to zmienne losowe X i Y nazywamy nieskorelowanymi. Z powyższego twierdzenia wynika więc, że zmienne losowe niezależne są jednocześnie nieskorelowane, ale zmienne nieskorelowane mogą być zależne.

Zatrzymajmy się jeszcze nad pytaniem, czy własność podobna do (2.32) zachodzi dla wariancji. Mówiąc o tym następujące twierdzenie:

Twierdzenie 35 Jeżeli $EX^2 < \infty$ i $EY^2 < \infty$ to

$$\text{Var}(X \pm Y) = \text{Var}X + \text{Var}Y \pm 2\text{Cov}(X, Y). \quad (2.101)$$

(patrz: **Przykład 2.8, 2.9**)

Ze wzoru (2.101) wynika, że jeżeli zmienne losowe X i Y są nieskorelowane (a więc, w szczególności, gdy są niezależne), to wówczas

$$\text{Var}(X \pm Y) = \text{Var}X + \text{Var}Y. \quad (2.102)$$

Wzór (2.101) można uogólnić dla sumy n zmiennych losowych

$$\text{Var}(X_1 + \dots + X_n) = \sum_{i=1}^n \text{Var}X_i + 2 \sum_{i < j} \text{Cov}(X_i, X_j). \quad (2.103)$$

Korzystając ze wzorów (2.32), (2.103), (2.30) i (2.39) otrzymamy pożyteczne wzory na wartość oczekiwana i wariancję średniej arytmetycznej. Mianowicie, jeżeli zmienne losowe $X_1 + \dots + X_n$ są niezależne i mają jednakowe rozkłady o wartości oczekiwanej $\mu = E(X_k)$ i wariancji $\sigma^2 = \text{Var}(X_k)$, $k = 1, 2, \dots, n$, to

$$E\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \mu \quad (2.104)$$

$$\text{Var}\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n} \sigma^2. \quad (2.105)$$

Można udowodnić, że

$$\text{Cov}(X, Y) \leq \sqrt{\text{Var}X} \sqrt{\text{Var}Y}, \quad (2.106)$$

przy czym równość zachodzi wtedy i tylko wtedy, gdy zmienne losowe X i Y związane są zależnością liniową. Wynika stąd, że kowariancja istnieje, gdy istnieją odpowiednie wariancje. Warto też pamiętać, że dla dowolnych liczb rzeczywistych a, b, c, d zachodzi następujący wzór

$$\text{Cov}(aX + b, cY + d) = ac\text{Cov}(X, Y). \quad (2.107)$$

Z kowariancją związany jest pewien współczynnik, o którym mowa poniżej.

Definicja 36 *Współczynnikiem korelacji zmiennych losowych X i Y nazywamy liczbę:*

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}X} \sqrt{\text{Var}Y}}. \quad (2.108)$$

Warto odnotować pewne własności współczynnika korelacji:

- $|\rho(X, Y)| \leq 1$.
- Jeżeli zmienne losowe X i Y są niezależne, to $\rho(X, Y) = 0$.
- Dla dowolnych liczb rzeczywistych a, b, c, d zachodzi

$$|\rho(aX + b, cY + d)| = |\rho(X, Y)|.$$

- Zmienne losowe X i Y są zależne liniowo wtedy i tylko wtedy, gdy $|\rho(X, Y)| = 1$.

Tak więc współczynnik korelacji jest pewną unormowaną miarą zależności liniowej zmiennych losowych.

(patrz: **Przykład 2.8, 2.9, 2.10**)

2.2.4 Pewne rozkłady sum niezależnych zmiennych losowych

W niniejszym podrozdziale podamy kilka użytecznych twierdzeń o rozkładzie sumy zmiennych losowych dla wybranych rozkładów, będących składnikami tej sumy. Będą to twierdzenia o sumie niezależnych zmiennych losowych mających rozkład dwupunktowy, dwumianowy, Poissona, normalny i gamma.

Twierdzenie 37 Jeżeli zmienne losowe X_1, X_2, \dots, X_n są niezależne i mają jednakowe rozkłady dwupunktowe $Bern(p)$, $k = 1, 2, \dots, n$, to suma $S_n = X_1 + X_2 + \dots + X_n$ ma rozkład dwumianowy $Bin(n, p)$.

Twierdzenie 38 Jeżeli zmienne losowe X_1, X_2, \dots, X_n są niezależne i mają rozkłady dwumianowe $Bin(n_k, p)$, $k = 1, 2, \dots, n$, to suma $S_n = X_1 + X_2 + \dots + X_n$ ma rozkład dwumianowy $Bin(\sum_{k=1}^n n_k, p)$.

Twierdzenie 39 Jeżeli zmienne losowe X_1, X_2, \dots, X_n są niezależne i mają rozkłady Poissona $P(\lambda_k)$, $k = 1, 2, \dots, n$, to suma $S_n = X_1 + X_2 + \dots + X_n$ ma rozkład Poissona $P(\sum_{k=1}^n \lambda_k)$.

Twierdzenie 40 Jeżeli zmienne losowe X_1, X_2, \dots, X_n są niezależne i mają rozkłady normalne $N(\mu_k, \sigma_k)$, $k = 1, 2, \dots, n$, to suma $S_n = X_1 + X_2 + \dots + X_n$ ma rozkład normalny $N\left(\sum_{k=1}^n \mu_k, \sqrt{\sum_{k=1}^n \sigma_k^2}\right)$.

Twierdzenie 41 Jeżeli zmienne losowe X_1, X_2, \dots, X_n są niezależne i mają rozkłady gamma $\Gamma(\alpha_k, \lambda)$, $k = 1, 2, \dots, n$, to suma $S_n = X_1 + X_2 + \dots + X_n$ ma rozkład gamma $\Gamma(\sum_{k=1}^n \alpha_k, \lambda)$.

2.3 Prawa wielkich liczb i twierdzenia graniczne

Rozważmy doświadczenie polegające na wielokrotnym rzucie symetryczną monetą. Można zauważać, że częstość wypadnięcia orła stabilizuje się w pobliżu $\frac{1}{2}$. Podobnie, przy wielokrotnym rzucie kostką sześcienną, częstość wypadnięcia np. szóstki będzie w przybliżeniu równa $\frac{1}{6}$. Prawidłowość tej opisuje twierdzenie:

Twierdzenie 42 Jeżeli S_n ma rozkład dwumianowy $Bin(n, p)$, to dla każdego $\varepsilon > 0$

$$\lim_{n \rightarrow \infty} P\left(\left|\frac{S_n}{n} - p\right| \leq \varepsilon\right) = 1. \quad (2.109)$$

Powysze twierdzenie (prawo wielkich liczb Bernoulliego) należy do klasy twierdzeń związanych ze słabymi prawami wielkich liczb. Niech X_1, X_2, \dots, X_n będzie ciągiem zmiennych losowych mających wartość oczekiwana i nich

$$S_n = X_1 + X_2 + \dots + X_n. \quad (2.110)$$

Definicja 43 Mówimy, że ciąg (X_n) spełnia słabe prawo wielkich liczb (SPWL), jeżeli ciąg zmiennych losowych $(\frac{1}{n}(S_n - E(S_n))$ jest zbieżny według prawdopodobieństwa do zera, tzn. dla każdego $\varepsilon > 0$

$$\lim_{n \rightarrow \infty} P\left(\left|\frac{S_n - E(S_n)}{n}\right| \leq \varepsilon\right) = 1. \quad (2.111)$$

Wymienimy tutaj podstawowe prawa wielkich liczb.

Twierdzenie 44 (Przepis Czebyszewa). Jeżeli (X_n) będzie ciągiem zmiennych losowych niezależnych o skończonych wariancji $\sigma_n^2 = Var X_n$, $n = 1, 2, \dots$. Jeżeli:

$$\lim_{n \rightarrow \infty} \frac{1}{n^2} \sum_{k=1}^n \sigma_k^2 = 0, \quad (2.112)$$

to ciąg (X_n) spełnia SPWL.

Przy dowodzie twierdzenia korzystamy z ważnej nierówności, a mianowicie nierówności Czebyszewa, która bywa przydatna również do szacowania prawdopodobieństwa niektórych zdarzeń.

Twierdzenie 45 (Nierówność Czebyszewa). Jeżeli zmienna losowa X ma wartość oczekiwana μ i wariancję σ^2 , to dla każdego $t > 0$

$$P(|X - \mu| \geq t\sigma) \leq \frac{1}{t^2}. \quad (2.113)$$

Nierówność Czebyszewa jest prawdziwa dla dowolnych zmiennych losowych mających wariancję skończoną i wskazuje, że odchylenie standardowe σ może służyć za miarę rozproszenia. Jeżeli przyjąć odchylenie standarde za jednostkę rozrzutu, to nierówność ta mówi, że prawdopodobieństwo zdarzenia polegającego na tym, iż zmienna losowa odchyli się od swojej wartości oczekiwanej o więcej niż t jednostek, jest nie większe niż $1/t^2$. Podstawiając w nierówności $t = 3$ otrzymujemy (reguła trzysigmowa)

$$P(|X - \mu| \geq 3\sigma) \leq \frac{1}{9}.$$

Chinczyn wykazał, że istnienie skończonego odchylenia standardowego zmiennych losowych X_n nie jest konieczne, aby ciąg spełniał SPWL. Oto twierdzenie Chinczyna:

Twierdzenie 46 (Przepis wielkich liczb Chinczyna). Niech (X_n) będzie ciągiem niezależnych zmiennych losowych o tym samym rozkładzie i skończonej wartości oczekiwanej μ . Wtedy ciąg (X_n) spełnia SPWL, tzn.,

$$\lim_{n \rightarrow \infty} P\left(\left|\frac{1}{n} S_n - \mu\right| \leq \varepsilon\right) = 1. \quad (2.114)$$

Słabe prawa wielkich liczb są szczególnym przypadkiem twierdzeń granicznych i dotyczą zbieżności według prawdopodobieństwa sum niezależnych zmiennych losowych do liczby. Ważną klasą twierdzeń granicznych są twierdzenia, w których występuje zbieżność według dystrybuant sum zmiennych losowych. Jednym z podstawowych twierdzeń z tego zakresu (czyli tzw. twierdzeń integralnych) jest następujące:

Twierdzenie 47 (Lindeberga-Levy'ego). Jeżeli zmienne losowe X_1, X_2, \dots są niezależne o jednakowych rozkładach z parametrami $EX_k = \mu$, $Var X_k = \sigma^2$ dla $k = 1, 2, \dots$, to

$$\lim_{n \rightarrow \infty} P\left(a < \frac{S_n - n\mu}{\sigma\sqrt{n}} \leq b\right) = \Phi(b) - \Phi(a), \quad (2.115)$$

gdzie Φ jest dystrybuantą rozkładu normalnego $N(0, 1)$.

Oznacza to, że suma S_n ma rozkład asymptotycznie normalny $N(n\mu, \sigma\sqrt{n})$. (patrz: Przykład 2.2, 2.5)

Szczególnym przypadkiem powyższego twierdzenia jest następujące

Twierdzenie 48 (Moivre'a-Laplace'a). Jeżeli zmienne losowe X_1, X_2, \dots są niezależne o jednakowych rozkładach dwupunktowych $Bern(p)$, to

$$\lim_{n \rightarrow \infty} P\left(a < \frac{S_n - np}{\sqrt{npq}} \leq b\right) = \Phi(b) - \Phi(a). \quad (2.116)$$

W tym przypadku suma S_n ma rozkład asymptotycznie normalny $N(np, \sqrt{npq})$. Twierdzenie to może być wykorzystywane do przybliżonego wyznaczania prawdopodobieństw dla rozkładów dwumianowych (suma S_n ma rozkład dwumianowy $Bin(n, p)$).

(patrz: Przykład 2.4)

Innego oszacowania dla prawdopodobieństwa danej liczby sukcesów w rozkładzie dwumianowym dostarcza twierdzenie Poissona. Przybliżenie to możemy stosować wtedy, gdy prawdopodobieństwo sukcesu p jest małe, a liczba doświadczeń n jest duża (co najmniej 100).

Twierdzenie 49 (Poissona). Jeżeli zmienne losowe X_1, X_2, \dots są niezależne o rozkładach dwumianowych $Bin(n, p_n)$ i jeśli $np_n = \lambda$ dla $n = 1, 2, \dots$ to

$$\lim_{n \rightarrow \infty} P(X_n = k) = \lim_{n \rightarrow \infty} \binom{n}{k} p_n^k (1 - p_n)^{n-k} = \frac{\lambda^k}{k!} e^{-\lambda}. \quad (2.117)$$

(patrz: Przykład 2.3)

2.4 Przykłady

Przykład 2.1

Obliczyć wartość oczekiwana i odchylenie standardowe wysokości wygranej na loterii, jeśli jej dystrybuanta jest dana wzorem (w zł):

$$F(x) = \begin{cases} 0 & \text{dla } x < 0 \\ 0.9 & \text{dla } 0 \leq x < 100 \\ 0.975 & \text{dla } 100 \leq x < 200 \\ 1 & \text{dla } x \geq 200. \end{cases}$$

Rozwiązańe

Z określenia dystrybuanty widzimy, że zmienna losowa X przyjmuje wartości 0, 100 i 200 z prawdopodobieństwami

$$\begin{aligned} P(X = 0) &= 0.9, \\ P(X = 100) &= 0.975 - 0.9 = 0.075, \\ P(X = 200) &= 1 - 0.975 = 0.025. \end{aligned}$$

Korzystając ze wzorów (2.18) i (2.41) obliczamy wartość średnią i wariancję wygranej

$$\begin{aligned} EX &= 0 \cdot 0.9 + 100 \cdot 0.075 + 200 \cdot 0.025 = 12.5 \\ VarX &= EX^2 - (EX)^2 \\ &= 0^2 \cdot 0.9 + 100^2 \cdot 0.075 + 200^2 \cdot 0.025 - 12.5^2 \\ &= 1750 - 156.25 = 1593.75. \end{aligned}$$

Przykład 2.2

Rzucamy dwiema kostkami do gry. Jeżeli suma oczek jest równa 2 to otrzymujemy 12 zł., jeżeli suma oczek jest równa 3 to otrzymujemy 6 zł., a w każdym pozostałym przypadku placimy 1 zł. Niech zmienna losowa X oznacza wygraną.

- Podać rozkład zmiennej losowej X i wyznaczyć jej dystrybuantę F .
- Wyznaczyć wartość oczekiwana $\mu = EX$ i wariancję $\sigma^2 = VarX$ wygranej.
- Założmy, że grę powtarzamy 100 razy. Korzystając z twierdzenia Lindeberga-Levy'ego, oszacować prawdopodobieństwo, że przegramy co najmniej 1 zł.

Rozwiązańe

a) Gra (doświadczenie) polega na rzucie dwiema kostkami, a więc zbiorem zdarzeń elementarnych Ω jest zbiór par:

$$\Omega = \{(i, j) : i, j = 1, 2, \dots, 6\}.$$

Wygrana X przyjmuje trzy wartości: -1, 6, 12. Określając rozkład wygranej należy podać z jakimi prawdopodobieństwami te wartości są przyjmowane. Wygranie 12 zł. realizuje się w przypadku jednej pary (na obu kostkach jedynka). Wygranie 6 zł. realizuje się w przypadku dwóch par (para (2, 1) i (1, 2)). Przegrywamy 1 zł. w 33 trzech pozostałych przypadkach. Tak więc:

$$P(X = -1) = \frac{33}{36}, \quad P(X = 6) = \frac{2}{36}, \quad P(X = 12) = \frac{1}{36}.$$

Rozkład wygranej możemy przedstawić w postaci tabelki:

x_i	-1	6	12
p_i	$\frac{33}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Korzystając z określenia (2.7) dystrybuanty zmiennej losowej skokowej, wyznaczamy:

$$F(x) = P(X \leq x) = \begin{cases} 0 & \text{dla } x < -1, \\ \frac{33}{36} & \text{dla } -1 \leq x < 6, \\ \frac{35}{36} & \text{dla } 6 \leq x < 12, \\ 1 & \text{dla } x \geq 12. \end{cases}$$

b) Korzystając ze wzoru (2.18) wyznaczamy wartość oczekiwana wygranej:

$$\mu = EX = (-1) \cdot \frac{33}{36} + 6 \cdot \frac{2}{36} + 12 \cdot \frac{1}{36} = -\frac{1}{4}.$$

Do wyznaczenia wariancji wykorzystujemy wzór (2.41):

$$(2.23) \quad \sigma^2 = VarX = \left(-1 + \frac{1}{4}\right)^2 \frac{33}{36} + \left(6 + \frac{1}{4}\right)^2 \frac{2}{36} + \left(12 + \frac{1}{4}\right)^2 \frac{1}{36} = \frac{115}{144} \cdot \frac{329}{48}$$

$$\text{Var } X = (-1)^2 \cdot \frac{33}{36} + 6^2 \cdot \frac{2}{36} + 12^2 \cdot \frac{1}{36} = (-\frac{1}{4})^2 \cdot \frac{329}{48}$$

c) W tym przypadku mamy do czynienia z ciągiem stu niezależnych zmiennych losowych X_1, X_2, \dots, X_{100} , z których każda ma ten sam rozkład co zmienna losowa X . Niech

$$S_{100} = X_1 + X_2 + \dots + X_{100}$$

będzie wygraną w 100 grach. Należy oszacować prawdopodobieństwo, że wygrana S_{100} będzie mniejsza niż -1 zł. Korzystając z twierdzenia Lindeberga-Levy'ego i wzoru (2.115) mamy:

$$P(S_{100} < -1) = P\left(\frac{S_{100} - 100\mu}{\sigma\sqrt{100}} < \frac{-1 - 100\mu}{\sigma\sqrt{100}}\right) \simeq \Phi\left(\frac{-1 - 100\mu}{\sigma\sqrt{100}}\right).$$

gdzie Φ jest dystrybuantą rozkładu normalnego $N(0, 1)$. Wykorzystując wyznaczone w punkcie b) wartości μ i σ i korzystając z tablic dystrybuanty rozkładu normalnego $N(0, 1)$ otrzymujemy:

$$P(S_{100} < -1) \simeq \Phi\left(\frac{-1 - 100(-0.25)}{2.228\sqrt{100}}\right) \simeq \Phi(1.08) = 0.8599.$$

Przykład 2.3

Centrala obsługuje 100 abonentów. Prawdopodobieństwo tego, że abonent zgłosi się do centrali w ciągu godziny jest równe 0.02.

- a) Obliczyć prawdopodobieństwo, że w ciągu godziny będą co najmniej cztery zgłoszenia oraz oszacować to prawdopodobieństwo korzystając z przybliżenia rozkładem Poissona.
- b) Jaka jest najbardziej prawdopodobna liczba zgłoszeń do centrali w ciągu godziny?

Rozwiązańe

a) Niech X będzie zmienną losową przyjmującą wartości równe liczbie abonentów zgłaszających się do centrali w ciągu godziny. Zmienna losowa X ma rozkład dwumianowy z parametrami $n = 100$ i $p = 0.02$. Należy wyznaczyć prawdopodobieństwo:

$$\begin{aligned} P(X \geq 4) &= 1 - P(X \leq 3) \\ &= 1 - P(X = 0) - P(X = 1) - P(X = 2) - P(X = 3). \end{aligned} \quad (2.118)$$

Korzystając ze wzoru (2.44):

$$P(X = 0) = \binom{100}{0} (0.02)^0 (0.98)^{100} = 0.1326,$$

$$P(X = 1) = \binom{100}{1} (0.02)^1 (0.98)^{99} = 0.2707,$$

$$P(X = 2) = \binom{100}{2} (0.02)^2 (0.98)^{98} = 0.2734,$$

$$P(X = 3) = \binom{100}{3} (0.02)^3 (0.98)^{97} = 0.1823.$$

Podstawiając wyznaczone wartości do wzoru (2.118) otrzymujemy:

$$P(X \geq 4) = 0.859.$$

Niech teraz $\lambda = np = 100 \cdot 0.02 = 2$. Korzystając z przybliżenia rozkładem Poissona (wzór (2.117)) otrzymujemy:

$$P(X = 0) \simeq \frac{2^0}{0!} e^{-2} \simeq 0.1353,$$

$$P(X = 1) \simeq \frac{2^1}{1!} e^{-2} \simeq 0.2707,$$

$$P(X = 2) \simeq \frac{2^2}{2!} e^{-2} \simeq 0.2707,$$

$$P(X = 3) \simeq \frac{2^3}{3!} e^{-2} \simeq 0.1804$$

i po podstawieniu do wzoru (2.118) mamy:

$$P(X \geq 4) \simeq 0.8571.$$

Jak widzimy dokładność przybliżenia jest duża, rzędu $2 \cdot 10^{-3}$.

b) Dla rozkładu dwumianowego, najbardziej prawdopodobna liczba sukcesów zależy od tego czy liczba $(n+1)p$ jest liczbą całkowitą czy nie. Ponieważ w naszym przypadku $(n+1)p = (100+1)0.02 = 2.02$, więc najbardziej prawdopodobna liczba zgłoszeń to $[(n+1)p] = [2.02] = 2$ zgłoszenia w ciągu godziny.

Przykład 2.4

Prawdopodobieństwo popsuca się aparatu w trakcie sprawdzania jego niezdolności jest równe $p = 0.05$.

- a) Jakie jest prawdopodobieństwo, że w trakcie sprawdzania 100 aparatów popsuje się nie mniej niż 4, ale nie więcej niż 10 aparatów?
- b) Ile aparatów należałoby sprawdzić, aby z prawdopodobieństwem większym niż 0.95 liczba aparatów zepsutych stanowiła od 3% do 7% liczby sprawdzanych aparatów?

Rozwiązańe

a) Z każdym doświadczeniem (sprawdzaniem aparatu) związana jest zmienność losowa X_k , $k = 1, 2, \dots, 100$ przyjmująca wartość 1, w przypadku gdy sprawdzany aparat uległ uszkodzeniu i wartość 0, w przeciwnym przypadku. Zmienność X_k ma rozkład dwupunktowy:

$$P(X_k = 1) = p = 0.05, \quad P(X_k = 0) = q = 0.95.$$

Rozważmy sumę:

$$S_{100} = X_1 + X_2 + \dots + X_{100},$$

ktora przyjmuje wartości równe liczbie aparatów uszkodzonych, wśród 100 sprawdzanych. Należy oszacować prawdopodobieństwo:

$$P(4 < S_{100} \leq 10).$$

Korzystając z twierdzenia Moivre'a-Laplace'a i wzoru (2.116) otrzymujemy:

$$\begin{aligned} P(4 < S_{100} \leq 10) &= P\left(\frac{4 - 100p}{\sqrt{100pq}} < \frac{S_{100} - 100p}{\sqrt{100pq}} \leq \frac{10 - 100p}{\sqrt{100pq}}\right) \\ &= P\left(\frac{-1}{\sqrt{4.75}} < \frac{S_{100} - 5}{\sqrt{4.75}} \leq \frac{5}{\sqrt{4.75}}\right) = \\ &\simeq P(-0.46 < \frac{S_{100} - 5}{\sqrt{4.75}} < 2.29) \simeq \\ &\simeq \Phi(2.29) - \Phi(-0.46) = \Phi(2.29) - (1 - \Phi(0.46)) \\ &= \Phi(2.29) + \Phi(0.46) - 1 \simeq 0.99 + 0.68 - 1 = 0.67. \end{aligned}$$

b) Niech n będzie liczbą aparatów, które należy pobrać do sprawdzenia i niech

$$S_n = X_1 + X_2 + \dots + X_n$$

będzie liczbą aparatów, które ulegną zepsuciu podczas ich sprawdzania. Należy wyznaczyć n tak, aby spełniona była nierówność:

$$P\left(\frac{3}{100}n < S_n \leq \frac{7}{100}n\right) > 0.95.$$

Korzystając ponownie z twierdzenia Moivre'a-Laplace'a mamy:

$$\begin{aligned} P\left(\frac{3}{100}n < S_n \leq \frac{7}{100}n\right) &= P\left(\frac{0.03n - np}{\sqrt{npq}} < \frac{S_n - np}{\sqrt{npq}} \leq \frac{0.07n - np}{\sqrt{npq}}\right) \\ &\simeq \Phi\left(\frac{2n}{\sqrt{475n}}\right) - \Phi\left(\frac{-2n}{\sqrt{475n}}\right) \\ &= \Phi\left(\frac{2n}{\sqrt{475n}}\right) - \left(1 - \Phi\left(\frac{2n}{\sqrt{475n}}\right)\right) \\ &= 2\Phi\left(\frac{2n}{\sqrt{475n}}\right) - 1 > 0.95. \end{aligned}$$

Z powyższego wynika, że:

$$\Phi\left(\frac{2n}{\sqrt{475n}}\right) > 0.975.$$

Ponieważ $\Phi(1.96) = 0.975$ więc

$$\frac{2n}{\sqrt{475n}} > 1.96 \Leftrightarrow n > (0.98)^2 475 = 456.19.$$

Do sprawdzenia należy pobrać co najmniej 457 aparatów.

Przykład 2.5

Zmienność losowa X ma rozkład ciągły o gęstości:

$$f(x) = \begin{cases} 0 & \text{dla } x < -1 \\ -x & \text{dla } -1 \leq x < 0 \\ Cx^2 & \text{dla } 0 \leq x < 1 \\ 0 & \text{dla } x \geq 1 \end{cases}$$

- a) Wyznaczyć stałą C .
- b) Wyznaczyć dystrybuantę $F(x) = P(X \leq x)$ i naszkicować jej wykres.
- c) Obliczyć prawdopodobieństwa $P(X \leq \frac{1}{2})$, $P(X > -\frac{1}{3})$ oraz $P(-\frac{1}{4} < X \leq \frac{1}{4})$.
- d) Obliczyć wartość oczekiwana $\mu = EX$ i wariancję $\sigma^2 = Var X$.
- e) Założymy, że zmienność losowa X_1, X_2, \dots, X_{121} są niezależne o takim samym rozkładzie co X i niech

$$S_{121} = X_1 + X_2 + \dots + X_{121}.$$

Korzystając z twierdzenia Lindeberga-Levy'ego oszacować prawdopodobieństwo: $P(2 < S_{121} \leq 16)$.

Rozwiązańe

a) Stałą C wyznaczamy z warunku (2.9). Ponieważ gęstość zeruje się poza przedziałem $(-1; 1]$ więc:

$$1 = \int_{-1}^1 f(x) dx = \int_{-1}^0 (-x) dx + \int_0^1 Cx^2 dx = -\left[\frac{x^2}{2}\right]_{-1}^0 + C\left[\frac{x^3}{3}\right]_0^1 = \frac{1}{2} + C\frac{1}{3}.$$

Rozwiązuając równanie $\frac{1}{2} + C\frac{1}{3} = 1$ wnosimy, że $C = \frac{3}{2}$.

b) Dystrybuantę F wyznaczać będziemy w przedziałach.

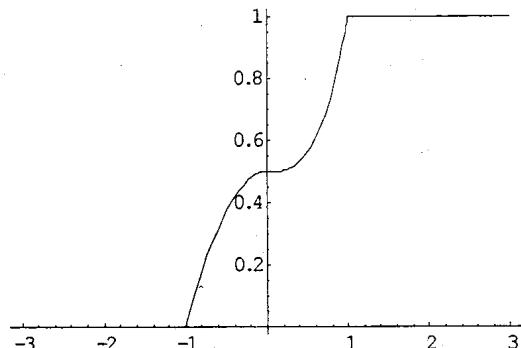
$$-\int_{-1}^{\frac{x}{2}} \frac{u^2}{2} du = -\left[\frac{u^3}{6} \right]_{-1}^{\frac{x}{2}} = \frac{1}{2} \left(\frac{x^3}{6} - \frac{-1}{6} \right) = \frac{1}{2} \cdot \frac{x^3 + 1}{6}$$

- dla $x \in (-\infty, -1)$, $F(x) = \int_{-\infty}^x 0 du = 0$,
- dla $x \in [-1, 0)$, $F(x) = \int_{-\infty}^{-1} 0 du + \int_{-1}^x (-u) du = -\left[\frac{u^2}{2} \right]_{-1}^x = \frac{1}{2}(1-x^2)$,
- dla $x \in [0, 1)$, $F(x) = \int_{-\infty}^{-1} 0 du + \int_{-1}^0 (-u) du + \int_0^x \frac{3}{2} u^2 du = -\left[\frac{u^3}{3} \right]_{-1}^0 + \frac{3}{2} \left[\frac{u^3}{3} \right]_0^x = \frac{1}{2}(1+x^3)$,
- dla $x \geq 1$, $F(x) = \int_{-\infty}^{-1} 0 du + \int_{-1}^0 (-u) du + \int_0^1 \frac{3}{2} u^2 du + \int_1^x 0 du = 1$.

Wyznaczoną dystrybuantę możemy opisać następującym wzorem:

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

Wykres dystrybuanty przedstawia rysunek (rys. 2.6).



Rys. 2.6 – Wykres dystrybuanty

b) Wyznaczamy kolejno prawdopodobieństwa wykorzystując dystrybuantę

$$P(X \leq \frac{1}{2}) = F(\frac{1}{2}) = \frac{1}{2}(1 + (\frac{1}{2})^3) = \frac{9}{16},$$

$$\begin{aligned} P(X > -\frac{1}{3}) &= 1 - P(X \leq -\frac{1}{3}) = 1 - F(-\frac{1}{3}) = 1 - \frac{1}{2}(1 - (-\frac{1}{3})^2) = \\ &= 1 - \frac{8}{18} = \frac{5}{9}, \end{aligned}$$

$$P(-\frac{1}{4} < X \leq \frac{1}{4}) = F(\frac{1}{4}) - F(-\frac{1}{4}) = \frac{1}{2}(1 + (\frac{1}{4})^3) - \frac{1}{2}(1 - (-\frac{1}{4})^2) = \frac{5}{128}.$$

c) Korzystając ze wzoru (2.18):

$$\begin{aligned} \mu &= EX = \int_{-\infty}^{+\infty} xf(x)dx = \int_{-1}^0 (-x^2)dx + \int_0^1 \frac{3}{2}x^3 dx = \\ &= -\left[\frac{x^3}{3} \right]_{-1}^0 + \frac{3}{2} \left[\frac{x^4}{4} \right]_0^1 = -\frac{1}{3} + \frac{3}{8} = \frac{1}{24} = 0.04. \end{aligned}$$

Do wyznaczenia wariancji zastosujemy wzór (2.41):

$$\begin{aligned} \sigma^2 &= EX^2 - \mu^2 = \int_{-\infty}^{+\infty} x^2 f(x)dx - \mu^2 = \\ &= \int_{-1}^0 (-x^3)dx + \int_0^1 \frac{3}{2}x^4 dx - \left(\frac{1}{24} \right)^2 = \\ &= -\left[\frac{x^4}{4} \right]_{-1}^0 + \frac{3}{2} \left[\frac{x^5}{5} \right]_0^1 - \left(\frac{1}{24} \right)^2 = \frac{11}{20} - \left(\frac{1}{24} \right)^2 = 0.55. \end{aligned}$$

e) Korzystając z twierdzenia Lindeberga-Levy'ego i wzoru (2.115) mamy:

$$\begin{aligned} P(2 < S_{121} \leq 16) &= P\left(\frac{2-121\mu}{\sigma\sqrt{121}} < \frac{S_{121}-121\mu}{\sigma\sqrt{121}} \leq \frac{16-121\mu}{\sigma\sqrt{121}}\right) = \\ &= P\left(\frac{-2.84}{8.14} < \frac{S_{121}-4.84}{8.14} \leq \frac{11.16}{8.14}\right) \simeq \Phi(1.37) - \Phi(-0.35) = \\ &= \Phi(1.37) - (1 - \Phi(0.35)) = 0.915 + 0.637 - 1 = 0.552. \end{aligned}$$

Przykład 2.6

Czas między kolejnymi zgłoszeniami do sieci informatycznej ma rozkład wykładniczy z parametrem $\lambda = 0.5 [\text{h}^{-1}]$ (oznacza to, że średni czas między kolejnymi zgłoszeniami wynosi 2[h]). Obliczyć prawdopodobieństwo tego, że w ciągu ośmiu godzin w sieci będzie pracowało co najmniej trzech użytkowników jeżeli w chwili początkowej sieć jest wolna.

Rozwiązanie

Niech X będzie zmienną losową przyjmującą wartości równe liczbie użytkowników pracujących w ciągu ośmiu określonych godzin. Ponieważ czas między kolejnymi zgłoszeniami ma rozkład wykładniczy, więc ilość zgłoszeń w systemie ma rozkład Poissona z parametrem $\lambda t = 0.5 \cdot 8 = 4$. Zgodnie ze wzorem (2.52) mamy

$$\begin{aligned} P(X \geq 3) &= 1 - P(X \leq 2) = 1 - P(X = 0) - P(X = 1) - P(X = 2) \\ &= 1 - \frac{4^0}{0!} e^{-4} - \frac{4^1}{1!} e^{-4} - \frac{4^2}{2!} e^{-4} = 1 - 13e^{-4} \simeq 0.76. \end{aligned}$$

Przykład 2.7

Wiadomo, iż przeciętny wiek kobiety w chwili urodzenia dziecka wynosi 26.9 lat, przy odchyleniu standardowym 5.5 roku. Zakładamy, że rozkład wieku kobiet w chwili urodzenia dziecka jest normalny.

- Jakie jest prawdopodobieństwo, że dziecko urodzi kobietę mającą nie więcej niż 30 lat?
- Jakie jest prawdopodobieństwo, że dziecko urodzi kobietę w wieku powyżej 40 lat?
- Jaka jest frakcja kobiet rodzących dziecko w wieku między 25 i 38 lat?
- Wyznaczyć wiek kobiet rodzących dziecko, którego nie przekracza 80% badanej populacji osób.

Rozwiążanie

Niech X będzie zmienną losową określającą wiek kobiety rodzącej dziecko. Zmienna X ma rozkład normalny $N(26.9, 5.5)$.

a) Szukane prawdopodobieństwo wyznaczamy korzystając z tablic dystrybuanty rozkładu normalnego standardowego pamiętając, że jeżeli X ma rozkład $N(26.9, 5.5)$, to zmienna losowa $\frac{X-26.9}{5.5}$ ma rozkład $N(0, 1)$.

$$P(X \leq 30) = P\left(\frac{X - 26.9}{5.5} \leq \frac{30 - 26.9}{5.5}\right) = \Phi(0.56) = 0.7123.$$

b) Należy wyznaczyć prawdopodobieństwo:

$$\begin{aligned} P(X > 40) &= 1 - P(X \leq 40) \\ &= 1 - P\left(\frac{X - 26.9}{5.5} \leq \frac{40 - 26.9}{5.5}\right) = 1 - \Phi(2.38) \\ &\simeq 1 - 0.99 = 0.01. \end{aligned}$$

c) Obliczamy najpierw prawdopodobieństwo:

$$\begin{aligned} P(25 < X \leq 38) &= P\left(\frac{25 - 26.9}{5.5} < \frac{X - 26.9}{5.5} \leq \frac{38 - 26.9}{5.5}\right) \\ &= \Phi(2.02) - \Phi(-0.35) \\ &= \Phi(2.02) - (1 - \Phi(0.35)) \simeq 0.98 + 0.64 - 1 = 0.62. \end{aligned}$$

Uzyskany wynik oznacza, że 62% kobiet rodzi dziecko między 25 a 38 rokiem życia.

d) Niech t oznacza wiek, którego nie przekracza 80% badanej populacji kobiet. Z treści zadania wynika, że:

$$P(X \leq t) = P\left(\frac{X - 26.9}{5.5} \leq \frac{t - 26.9}{5.5}\right) = \Phi\left(\frac{t - 26.9}{5.5}\right) = 0.80,$$

a więc tym razem mamy wyznaczyć kwantyl rzędu 0.8 rozkładu normalnego. Z tablic odczytujemy, że $u_{0.8} \simeq 0.84$ i otrzymujemy równanie:

$$\frac{t - 26.9}{5.5} = 0.84.$$

Zatem $t = 31.52$ czyli wiek, którego nie przekracza 80% badanej populacji kobiet to około 31 i pół roku.

Przykład 2.8

Sprzedawca zatrudniony w sklepie z komputerami dostaje miesięcznie stałą pensję w wysokości 500 zł, a do tego 10 zł za każdy sprzedany komputer i 20 zł za każdy sprzedany zestaw oprogramowania. W ciągu miesiąca udaje mu się sprzedać średnio 40 komputerów oraz 20 zestawów oprogramowania, z odchyleniami standardowymi, odpowiednio, 10 i 5. Współczynnik korelacji pomiędzy liczbą sprzedanych komputerów i liczbą sprzedanych zestawów oprogramowania wynosi 0.5. Obliczyć średnie miesięczne wynagrodzenie tego sprzedawcy oraz odchylenie standardowe miesięcznego wynagrodzenia.

Rozwiążanie

Niech X będzie zmienną losową przyjmującą wartości równe liczbie komputerów sprzedanych przez sprzedawcę w ciągu miesiąca a Y zmienną losową przyjmującą wartości równe liczbie sprzedanych zestawów oprogramowania. Średnie miesięczne wynagrodzenie sprzedawcy wynosi zatem

$$W = 10X + 20Y + 500.$$

Należy obliczyć wartość średnią EW i odchylenie standardowe $\sigma = \sqrt{VarW}$. Z treści zadania wynika, że $EX = 40$ i $EY = 20$ oraz $\sqrt{VarX} = 10$ i $\sqrt{VarY} = 5$. Korzystając ze wzorów (2.32), (2.30) i (2.39), (2.101) otrzymujemy:

$$EW = E(10X + 20Y + 500) = 10EX + 20EY + 500 = 1300$$

$$\begin{aligned} VarW &= 10^2VarX + 20^2VarY + 2 \cdot 10 \cdot 20Cov(X, Y) \\ &= 20000 + 400Cov(X, Y). \end{aligned}$$

Do wyznaczenia kowariancji wykorzystamy znajomość współczynnika korelacji. Ze wzoru (2.108) mamy

$$Cov(X, Y) = \sqrt{VarX}\sqrt{VarY}\rho(X, Y) = 10 \cdot 5 \cdot 0.5 = 25.$$

W związku z powyższym

$$\sigma = \sqrt{VarW} = \sqrt{30000} = 173.21.$$

Tak więc, średnie miesięczne wynagrodzenie sprzedawcy wynosi 1300 zł a odchylenie standardowe 173.21 zł.

Przykład 2.9

Pewna firma budowlana prowadzi działalność w Warszawie oraz poza Warszawą. Badania dotyczące działalności firmy w ciągu ostatnich kilku lat

pokały, że rozkład prawdopodobieństwa liczby inwestycji prowadzonych jednocześnie w Warszawie oraz poza Warszawą przedstawia się następująco:

		Y (poza Warszawą)		
		0	1	2
X (w Warszawie)	0	0.1	0.3	0.2
	1	0.2	0.1	0.1

- a) Czy liczby inwestycji prowadzonych jednocześnie w Warszawie i poza Warszawą są niezależne?
- b) Czy liczby inwestycji prowadzonych jednocześnie w Warszawie i poza Warszawą są skorelowane (jeżeli tak, to w jakim stopniu)?
- c) Jaka jest średnia i wariancja łącznej liczby inwestycji prowadzonych jednocześnie w Warszawie i poza Warszawą?

Rozwiązańe

a) Korzystając ze wzorów (2.81) i (2.82) wyznaczamy rozkłady brzegowe zmiennych losowych X i Y . Otrzymujemy:

$$\begin{aligned} p_{1 \bullet} &= P(X = 0) = \frac{1}{10} + \frac{3}{10} + \frac{2}{10} = \frac{3}{5}, \\ p_{2 \bullet} &= P(X = 1) = \frac{2}{10} + \frac{1}{10} + \frac{1}{10} = \frac{2}{5} \end{aligned}$$

oraz

$$\begin{aligned} p_{\bullet 1} &= P(Y = 0) = \frac{1}{10} + \frac{2}{10} = \frac{3}{10}, \\ p_{\bullet 2} &= P(Y = 1) = \frac{3}{10} + \frac{1}{10} = \frac{4}{10}, \\ p_{\bullet 3} &= P(Y = 2) = \frac{2}{10} + \frac{1}{10} = \frac{3}{10}. \end{aligned}$$

Ponieważ np. $P(X = 0, Y = 2) = 0.2 \neq P(X = 0)P(Y = 2) = 0.18$ więc zmienne X i Y nie są niezależne bo nie jest spełniony warunek (2.93).

b) Wyznaczmy teraz współczynnik korelacji $\rho(X, Y)$ między liczbą inwestycji w Warszawie i poza Warszawą. Korzystając z rozkładów brzegowych, wyznaczamy wartości średnie i wariancje zmiennych losowych X i Y

$$\begin{aligned} EX &= 0 \cdot \frac{3}{5} + 1 \cdot \frac{2}{5} = \frac{2}{5}, \\ EY &= 0 \cdot \frac{3}{10} + 1 \cdot \frac{4}{10} + 2 \cdot \frac{3}{10} = 1, \\ VarX &= EX^2 - (EX)^2 = 0^2 \cdot \frac{3}{5} + 1^2 \cdot \frac{2}{5} - \frac{4}{25} = \frac{6}{25}, \\ VarY &= EY^2 - (EY)^2 = 0^2 \cdot \frac{3}{10} + 1^2 \cdot \frac{4}{10} + 2^2 \cdot \frac{3}{10} - 1 = \frac{3}{5}, \end{aligned}$$

a następnie (wykorzystując wzór (2.99)) wartość oczekiwana iloczynu:

$$E(XY) = 1 \cdot 1 \cdot \frac{1}{10} + 1 \cdot 2 \cdot \frac{1}{10} = \frac{3}{10}.$$

Korzystając ze wzoru (2.98) otrzymujemy kowariancję

$$Cov(X, Y) = E(XY) - (EX)(EY) = \frac{3}{10} - \frac{2}{5} = -\frac{1}{10}.$$

Zatem współczynnik korelacji $\rho(X, Y)$ określony wzorem (2.108) wynosi

$$\rho(X, Y) = \frac{-\frac{1}{10}}{\sqrt{\frac{6}{25}} \sqrt{\frac{3}{5}}} = -\frac{1}{2} \sqrt{\frac{5}{2}} = -0.79.$$

Wartość wyznaczonego współczynnika korelacji wskazuje na dość wysokie skorelowanie liczby inwestycji prowadzonych jednocześnie w Warszawie i poza Warszawą.

c) Łączną liczbą inwestycji jest suma zmiennych losowych $X + Y$. Korzystając ze wzorów (2.32) i (2.101) otrzymujemy

$$\begin{aligned} E(X + Y) &= EX + EY = \frac{2}{5} + 1 = \frac{7}{5}, \\ Var(X + Y) &= VarX + VarY + 2Cov(X, Y) = \\ &= \frac{6}{25} + \frac{3}{5} + 2 \cdot \left(-\frac{1}{10}\right) = \frac{16}{25}. \end{aligned}$$

Przykład 2.10

Zmienna losowa (X, Y) ma rozkład ciągły o gęstości f określonej wzorem

$$f(x, y) = \begin{cases} Cx & \text{dla } 0 < x < y < 1, \\ 0 & \text{dla pozostałych } x \text{ i } y. \end{cases}$$

- a) Wyznaczyć stałą C .
- b) Obliczyć wartość dystrybuanty $F(\frac{1}{2}, 2)$.
- c) Wyznaczyć gęstości brzegowe f_X i f_Y i sprawdzić, czy X, Y są niezależne.
- d) Obliczyć współczynnik korelacji ρ .

Rozwiązańe

- a) Stałą C wyznaczamy z warunku (2.84)

$$\begin{aligned} 1 &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x, y) dx dy = \int_0^1 dx \int_x^1 Cxy dy = C \int_0^1 x [y]_x^1 dx = \\ &= C \int_0^1 x(1-x) dx = C \int_0^1 (x-x^2) dx = C \left[\frac{x^2}{2} - \frac{x^3}{3} \right]_0^1 = C \frac{1}{6}. \end{aligned}$$

Wynika stąd, że $C = 6$.

b) Z określenia (2.83) dystrybuanty mamy

$$\begin{aligned} F\left(\frac{1}{2}, 2\right) &= \int_{-\infty}^{1/2} \int_{-\infty}^2 f(x, y) dx dy = \int_0^{1/2} dx \int_x^1 6x dy = 6 \int_0^{1/2} x[y]_x^1 dx \\ &= 6 \int_0^{1/2} x(1-x) dx = 6 \left[\frac{x^2}{2} - \frac{x^3}{3} \right]_0^{1/2} = \frac{1}{2}. \end{aligned}$$

c) Gęstości brzegowe wyznaczamy wykorzystując wzory (2.86) i (2.87).

Jeżeli $x \in (0, 1)$, to

$$f_X(x) = \int_{-\infty}^{+\infty} f(x, y) dy = \int_x^1 6x dy = 6x(1-x)$$

i dla $x \notin (0, 1)$, $f_X(x) = 0$. Tak więc

$$f_X(x) = \begin{cases} 6x(1-x) & \text{dla } x \in (0, 1) \\ 0 & \text{dla } x \notin (0, 1). \end{cases}$$

W podobny sposób wyznaczamy gęstość brzegową dla zmiennej Y . Jeżeli $y \in (0, 1)$, to

$$f_Y(y) = \int_{-\infty}^{+\infty} f(x, y) dx = \int_0^y 6x dx = 3y^2$$

i dla $y \notin (0, 1)$, $f_Y(y) = 0$. Zatem

$$f_Y(y) = \begin{cases} 3y^2 & \text{dla } y \in (0, 1) \\ 0 & \text{dla } y \notin (0, 1). \end{cases}$$

Wyznaczone gęstości brzegowe nie spełniają warunku (2.91), a więc X i Y nie są niezależne (czyli są zależne).

d) Aby wyznaczyć współczynnik korelacji określony wzorem (2.108), należy najpierw obliczyć wartości średnie i wariancje zmiennych X i Y :

$$\begin{aligned} EX &= \int_{-\infty}^{+\infty} xf_X(x) dx = 6 \int_0^1 x^2(1-x) dx = 6 \left[\frac{x^3}{3} - \frac{x^4}{4} \right]_0^1 = \frac{1}{2}, \\ EY &= \int_{-\infty}^{+\infty} yf_Y(y) dy = 3 \int_0^1 y^3 dy = \frac{3}{4}, \\ VarX &= EX^2 - (EX)^2 = \int_{-\infty}^{+\infty} x^2 f_X(x) dx - \frac{1}{4} = \\ &= 6 \int_0^1 x^3(1-x) dx - \frac{1}{4} = 6 \left[\frac{x^4}{4} - \frac{x^5}{5} \right]_0^1 - \frac{1}{4} = \frac{1}{20}, \\ VarY &= EY^2 - (EY)^2 = \int_{-\infty}^{+\infty} y^2 f_Y(y) dy - \frac{9}{16} = \\ &= 3 \int_0^1 y^4 dy - \frac{9}{16} = \frac{3}{5} - \frac{9}{16} = \frac{3}{80}. \end{aligned}$$

Korzystając ze wzorów (2.98) i (2.100) policzymy kowariancję zmiennych X i Y

$$\begin{aligned} Cov(X, Y) &= E(XY) - (EX)(EY) = \\ &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} xyf(x, y) dx dy - \left(\frac{1}{2} \right) \cdot \left(\frac{3}{4} \right) = \\ &= 6 \int_0^1 x^2 dx \int_x^1 y dy - \frac{3}{8} = 3 \int_0^1 x^2 [y^2]_x^1 dx - \frac{3}{8} = \\ &= 3 \left[\frac{x^3}{3} - \frac{x^5}{5} \right]_0^1 - \frac{3}{8} = \frac{2}{5} - \frac{3}{8} = \frac{1}{40}. \end{aligned}$$

Wstawiając wyznaczone wartości do wzoru (2.108) otrzymujemy

$$\rho(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}} = \frac{\frac{1}{40}}{\sqrt{\frac{1}{20}}\sqrt{\frac{3}{80}}} = \frac{\sqrt{3}}{3} = 0.58.$$

2.5 Zadania

Zadanie 2.1

Na pewnej uczelni działa sieć informatyczna. Na podstawie obserwacji stwierdzono, że dziennie występują co najwyżej trzy awarie sieci. Dwie awarie występują dwa razy rzadziej niż jedna awaria, zaś trzy awarie dwa razy rzadziej niż dwie. Średnia liczba awarii wynosi dziennie $\frac{11}{16}$. Niech X będzie zmienną losową przyjmującą wartości równe liczbie awarii w ciągu jednego dnia (wartości 0, 1, 2 i 3).

- Podać rozkład zmiennej losowej X .
- Wyznaczyć dystrybuantę F .
- Obliczyć wariancję dziennej liczby awarii.
- Korzystając z twierdzenia Lindeberga-Levy'ego oszacować prawdopodobieństwo tego, że liczba awarii w ciągu 144 dni pracy sieci komputerowej będzie w granicach od 81 do 123.

Zadanie 2.2

Szacuje się, że zaledwie 10% Polaków posiada kartę kredytową. Jakie jest prawdopodobieństwo, że w losowej grupie 8 osób co najmniej dwie będą posiadały kartę kredytową?

Zadanie 2.3

Według ostatnich badań 90% młodych ludzi, w wieku od 21 do 25 lat mieszka z rodzicami. Jakie jest prawdopodobieństwo, że w losowej grupie 7 osób w tym wieku co najwyżej jedna z nich będzie mieszkała samodzielnie?

Zadanie 2.4

Obliczyć wartość oczekiwana i odchylenie standardowe wysokości premii wypłacanej w pewnym przedsiębiorstwie, jeżeli wiadomo, że dystrybuanta rozkładu wysokości premii dana jest wzorem (w tys. zł):

$$F(x) = \begin{cases} 0 & \text{dla } x < 0.5 \\ 0.5 & \text{dla } 0.5 \leq x < 1 \\ 0.9 & \text{dla } 1 \leq x < 2 \\ 1 & \text{dla } x \geq 2. \end{cases}$$

Zadanie 2.5

Prawdopodobieństwo tego, że w ciągu czasu T przestanie działać jeden kondensator, jest równe 0.2. Wyznaczyć prawdopodobieństwo, że spośród 100 kondensatorów w ciągu czasu T przestanie działać:

- nie mniej niż 20 kondensatorów,
- nie więcej niż 28 kondensatorów,

- co najmniej 14 ale nie więcej niż 26 kondensatorów.

Zadanie 2.6

Co trzeci detal wykonany w pewnej fabryce jest wadliwy. Obliczyć prawdopodobieństwo, że w losowo wybranej próbie 1200 detali liczba tych, które będą wadliwe, będzie większa od 375 i mniejsza od 425.

Zadanie 2.7

Daltoniści stanowią około 6% populacji mężczyzn. Jakie jest prawdopodobieństwo, że w losowej grupie 250 mężczyzn znajdzie się co najmniej 10, ale nie więcej niż 20 daltonistów?

Zadanie 2.8

Przeprowadzone badania wykazały, że około 34% pracujących dojeżdża do miejsca pracy własnym pojazdem, 56% dojeżdża korzystając z publicznych środków transportu, zaś pozostałe 10% osób dociera do miejsca pracy piechotą, bądź pracuje w miejscu zamieszkania. Obliczyć prawdopodobieństwo tego, że co najmniej 170 spośród 200 losowo wybranych osób dojeżdża do swego miejsca pracy.

Zadanie 2.9

W związku z organizacją konferencji planuje się wynająć salę konferencyjną. Sala ta może pomieścić co najwyżej 100 osób. Aby móc pokryć koszty wynajęcia sali, w konferencji powinno wziąć udział nie mniej niż 70 osób. Organizatorzy postanowili wysłać zaproszenia na konferencję do 120 osób. Szacuje się, że prawdopodobieństwo tego, iż poszczególne osoby zechcą wziąć udział w tej konferencji wynosi 70%.

- Jakie jest prawdopodobieństwo tego, że konferencja będzie mogła się odbyć we wspomnianej sali?
- Czy liczba wystosowanych zaproszeń jest odpowiednia?

Zadanie 2.10

W celu oceny jakości wyprodukowanej partii przedmiotów dokonuje się ich sprawdzenia. Prawdopodobieństwo, że losowo wybrany przedmiot okaże się wadliwy, jest stałe i równe 0.1. Partię przedmiotów odrzuca się jako niedobra przy znalezieniu co najmniej 10 przedmiotów wadliwych. Ile przedmiotów należy sprawdzić, aby z prawdopodobieństwem 0.6 można było twierdzić, że partia zawierająca 10% przedmiotów wadliwych nie będzie przyjęta?

Zadanie 2.11

Po mieście jeździ 10000 samochodów. Prawdopodobieństwo wezwania pogotowia technicznego w ciągu doby przez samochód wynosi 0.0001. Obliczyć prawdopodobieństwo tego, że w ciągu losowo wybranej doby pogotowie techniczne będzie wzywane co najmniej trzy razy.

Zadanie 2.12

Oddział banku otrzymuje w ciągu tygodnia około 300 wniosków o wydanie karty kredytowej. Stwierdzono, że około 1% wniosków nie zostaje rozpatrzonych pozytywnie. Obliczyć prawdopodobieństwo odrzucenia w danym tygodniu trzech lub więcej wniosków o wydanie karty kredytowej. Porównać dokładne rozwiązanie z wynikiem uzyskanym za pomocą twierdzenia Poissona.

Zadanie 2.13

Urządzenie składa się z wielu niezależnie pracujących elementów, z jednakowymi (bardzo małymi) prawdopodobieństwami awarii w czasie T . Znaleźć średnią liczbę nie działających elementów w czasie T , jeśli prawdopodobieństwo tego, że uległ awarii co najmniej jeden element w tym czasie, jest równe 0.98.

Zadanie 2.14

Czas (w miesiącach) pomiędzy awariami kopiarki ma rozkład wykładniczy z parametrem 0.5. W przypadku wystąpienia awarii kopiarka ta jest natychmiast naprawiana.

- Określić rozkład liczby awarii kopiarki w ciągu miesiąca.
- Ilu awarii możemy spodziewać się, średnio, w ciągu miesiąca?
- Ilu awarii możemy spodziewać się, średnio, w ciągu roku?

Zadanie 2.15

Z dotyczących obserwacji wynika, że liczba klientów przybywających w ciągu godziny do oddziału banku ma rozkład Poissona o średniej 4 (klientów na godzinę).

- Jaki jest rozkład prawdopodobieństwa czasu między przyjściem kolejnych klientów?
- Jaki jest średni czas oraz odchylenie standardowe czasu pomiędzy chwilami przybyć kolejnych klientów?
- Jeżeli w danej chwili do oddziału wszedł klient, to jakie jest prawdopodobieństwo, że w ciągu najbliższych 30 minut kolejny klient przybędzie do oddziału?
- Jakie jest prawdopodobieństwo, że w ciągu godziny do oddziału banku nie przyjdzie ani jeden klient?

Zadanie 2.16

Wzrost pewnej grupy osób opisany jest rozkładem normalnym o wartości oczekiwanej 173 cm i odchyleniu standardowym 6 cm.

a) Jakie jest prawdopodobieństwo, że losowo wybrana osoba ma więcej niż 181 cm wzrostu?

b) Jakie jest prawdopodobieństwo, że losowo wybrana osoba ma nie więcej niż 179 cm wzrostu?

c) Jaka jest frakcja osób mających wzrost pomiędzy 167 i 180 cm?

d) Wyznaczyć wartość wzrostu, którego nie przekracza 60% badanej populacji osób.

Zadanie 2.17

Obliczyć wartość oczekiwana i odchylenie standardowe czasu oczekiwania na dowiezienie pizzy, jeżeli wiadomo, że dystrybuanta czasu oczekiwania dana jest wzorem (w min.):

$$F(x) = \begin{cases} 0 & \text{dla } x < 10, \\ \frac{x-10}{10} & \text{dla } 10 \leq x < 20, \\ 1 & \text{dla } x \geq 20. \end{cases}$$

Zadanie 2.18

Zmienna losowa X ma rozkład ciągły o gęstości

$$f(x) = \begin{cases} 0 & \text{dla } x < 0, \\ C(1+x)^{-4} & \text{dla } x \geq 0. \end{cases}$$

- Wyznaczyć stałą C .
- Wyznaczyć dystrybuantę F .
- Obliczyć prawdopodobieństwa $P(X > 2)$ i $P(\frac{1}{4} < X \leq 1)$.
- Obliczyć wartość oczekiwana i odchylenie standardowe.
- Załóżmy, że zmienne losowe X_1, X_2, \dots, X_{300} są niezależne o takim samym rozkładzie co X . Korzystając z twierdzenia Lindeberga - Levy'ego oszacować prawdopodobieństwo

$$P\left(\left|\frac{1}{300} \sum_{k=1}^{300} X_k - \frac{1}{2}\right| < \frac{1}{20}\right).$$

Zadanie 2.19

Dealer samochodowy prowadzi sprzedaż samochodów we wszystkie dni tygodnia, z czego w soboty udaje mu się sprzedawać nie więcej niż dwa auta, a w

niedziele co najwyżej jeden samochód, z następującymi prawdopodobieństwami:

		Y (niedziela)	
		0	1
X (sobota)	0	$\frac{1}{10}$	$\frac{1}{10}$
	1	$\frac{3}{10}$	$\frac{2}{10}$
	2	$\frac{1}{10}$	$\frac{2}{10}$

- a) Czy liczby samochodów sprzedawanych, odpowiednio, w sobotę i w niedziele, są niezależne?
- b) Czy liczby samochodów sprzedawanych, odpowiednio, w sobotę i w niedziele, są skorelowane (jeżeli tak, to w jakim stopniu)?
- c) Jaka jest średnia i wariancja liczby samochodów sprzedawanych przez tego dealera w ciągu całego weekendu?

Zadanie 2.20

Firma ubezpieczeniowa wypłaca agentom zajmującym się ubezpieczeniami komunikacyjnymi stałą pensję w wysokości 600 zł miesięcznie, a ponadto 10 zł za każdą zawartą umowę OC i 20 zł za każdą zawartą umowę AC. Pewien agent zawiera w miesiącu średnio 60 umów OC i 40 umów AC z odchyleniem standardowym, odpowiednio, 15 i 10. Współczynnik korelacji między liczbą zawieranych umów OC i AC wynosi 0.7. Obliczyć średni miesięczny dochód agenta oraz odchylenie standardowe miesięcznego dochodu.

Zadanie 2.21

Zmienna losowa (X, Y) ma rozkład ciągły o gęstości f określonej wzorem

$$f(x, y) = \begin{cases} Cxy & \text{dla } 0 < y < x < 1 \\ 0 & \text{dla pozostałych } x \text{ i } y. \end{cases}$$

- a) Wyznaczyć stałą C .
- b) Obliczyć wartość dystrybuanty $F(2, \frac{1}{2})$.
- c) Wyznaczyć gęstości brzegowe f_X i f_Y i sprawdzić, czy X, Y są niezależne.
- d) Obliczyć współczynnik korelacji ρ .

ODPOWIEDZI

Zadanie 2.1

a) $P(X = 0) = \frac{9}{16}$, $P(X = 1) = \frac{4}{16}$, $P(X = 2) = \frac{2}{16}$, $P(X = 3) = \frac{1}{16}$.

b) $F(x) = \begin{cases} 0 & \text{dla } x < 0 \\ \frac{9}{16} & \text{dla } 0 \leq x < 1 \\ \frac{13}{16} & \text{dla } 1 \leq x < 2 \\ \frac{15}{16} & \text{dla } 2 \leq x < 3 \\ 1 & \text{dla } x \geq 3. \end{cases}$

c) $Var X = 75/128$,

d) 0.97.

Zadanie 2.2

0.19

Zadanie 2.3

0.19

Zadanie 2.4

$EX = 0.85$, $Var X = 0.2025$

Zadanie 2.5

- a) 0.5,
- b) 0.977,
- c) 0.866

Zadanie 2.6

0.87

Zadanie 2.7

0.82

Zadanie 2.8

0.99

Zadanie 2.9

- a) 0.997,
- b) tak

Zadanie 2.10

195

Zadanie 2.11

0.08

Zadanie 2.12

0.577

Zadanie 2.13

$-\ln 0.02$

Zadanie 2.14

- a) rozkład Poissona z parametrem $\lambda = 0.5$,
- b) 0.5,
- c) 6

Zadanie 2.15

- a) rozkład wykładniczy z parametrem $\lambda = 4$,
- b) 0.25 godz.,
- c) 0.86,

d) 0.018

Zadanie 2.16

- a) 0.09,
 b) 0.84,
 c) 72,
 d) 174.56 cm

Zadanie 2.17

$$EX = 15, \text{Var}X = 8.33$$

Zadanie 2.18

- a) $C = 3$,
 b) $F(x) = \begin{cases} 0 & \text{dla } x < 0 \\ 1 - (1+x)^{-3} & \text{dla } x \geq 0 \end{cases}$,
 c) $P(X > 2) = 0.96, P(\frac{1}{4} < X \leq 1) = 0.387$,
 d) $EX = \frac{1}{2}, \text{Var}X = \frac{\sqrt{3}}{2}$
 e) 0.68

Zadanie 2.19

- a) są zależne,
 b) tak, współczynnik korelacji $\rho(X, Y) = \frac{1}{7}$ (słaba korelacja),
 c) $E(X + Y) = 1.6, \text{Var}(X + Y) = 0.84$

Zadanie 2.20

Średni dochód wynosi 2000 zł a odchylenie standardowe wynosi 323.26 zł

Zadanie 2.21

- a) $C = 8$,
 b) $F(2, \frac{1}{2}) = \frac{7}{16}$,
 c) $f_X(x) = \begin{cases} 4x^3 & \text{dla } x \in (0, 1) \\ 0 & \text{dla } x \notin (0, 1) \end{cases}$,
 $f_Y(y) = \begin{cases} 4y(1-y^2) & \text{dla } y \in (0, 1) \\ 0 & \text{dla } y \notin (0, 1) \end{cases}$,
 d) $\rho(X, Y) = \frac{2\sqrt{66}}{33} \simeq 0.49$.

3

Statystyka opisowa

3.1 Wprowadzenie

W potocznym rozumieniu **statystyka** jest zbiorem metod służących gromadzeniu, prezentacji, analizie i interpretacji danych, w celu podejmowania decyzji. Termin statystyka obejmuje w równej mierze **statystykę teoretyczną**, zajmującą się dowodzeniem twierdeń i badaniem własności określonych obiektów matematycznych, jak i **statystykę stosowaną**, która pokazuje, w jaki sposób owe pojęcia i twierdzenia mogą być wykorzystane do rozwiązywania praktycznych problemów. Z kolei w ramach statystyki stosowanej możemy wyodrębnić **statystykę opisową**, zajmującą się wstępnią analizą danych oraz **wnioskowanie statystyczne**, czyli metodologię wyciągania wniosków na temat pewnych właściwości badanej zbiorowości na podstawie dostępnych danych.

W niniejszym rozdziale przedstawione zostaną podstawowe pojęcia i metody używane w statystyce opisowej. W kolejnych rozdziałach omówione zostaną dwa zasadnicze filary wnioskowania statystycznego: teoria estymacji (rozdział 4) i teoria weryfikacji hipotez (rozdział 5).

3.2 Podstawowe pojęcia

Badanie statystyczne dotyczy zawsze jakiegoś zbioru elementów (osób, przedmiotów, zjawisk), nazywanego w statystyce **populacją** (populacją generalną). Celem każdego badania statystycznego jest poznanie określonych

właściwości elementów rozważanej populacji. Ową właściwość, ze względu na którą prowadzi się badanie, nazywamy **cechą** (statystyczną). W praktyce możemy mieć do czynienia z dwojakiego rodzaju cechami: jakościowymi i ilościowymi. **Cecha jakościowa**, to właściwość niemierzalna (np. kolor oczu, marka samochodu, stan cywilny), natomiast **cecha ilościowa**, to właściwość mierzalna (jak waga, wzrost, czas działania, itp.). Charakter badanej cechy decyduje o wyborze narzędzi statystycznych, które należy użyć do opisu i dalszego wnioskowania statystycznego.

Jeżeli elementy populacji różnią się między sobą wartościami badanej cechy, wówczas można mówić o **rozkładzie cechy** w populacji. Wartość cechy, wyznaczoną dla danego elementu próby, nazywamy **obserwacją** lub **pomiarem**.

W celu uzyskania informacji o rozkładzie interesującej nas cechy można prowadzić badanie wszystkich elementów populacji lub tylko pewnego podzbiuru populacji. W pierwszym przypadku mamy do czynienia z tzw. **badaniem pełnym** (kompletnym, całkowitym, stuprocentowym). Mimo oczywistych zalet, badania pełne prowadzi się w praktyce bardzo rzadko. Spowodowane jest to głównie względami ekonomicznymi (duży koszt badania) oraz czasochłonnością. Nieraz badanie pełne jest wręcz niewykonalne (gdy populacja jest nieskończona lub o bardzo dużej liczności), bądź bezsensowne (np. gdy badanie ma charakter niszczący). Stąd też zazwyczaj mamy do czynienia z **badaniem częściowym**, w którym bezpośredniemu badaniu podlega jedynie podzbiór populacji zwany **próbą** (próbką). **Statystyka matematyczna** zajmuje się wyłącznie badaniami częściowymi i to takimi, w których dobór próby podlega pewnym regułom. Skoro bowiem badanie próby stanowić ma podstawę do wnioskowania o rozkładzie cechy w całej populacji, naturalne jest wymaganie, aby próba była **reprezentatywna**. Oznacza to, że rozkład cechy w próbie nie powinien różnić się istotnie od rozkładu cechy w całej populacji. Tak więc próba reprezentatywna ma stanowić swoją "miniaturę" populacji. Aby to osiągnąć elementy próbki zwykle losuje się spośród elementów populacji. Tak utworzona próba nazywana jest **próbą losową**. Jeżeli wszystkie elementy populacji mają jednakowe szanse (prawdopodobieństwo) dostania się do próby, wówczas mamy do czynienia z **próbą losową prostą**. W praktyce stosuje się także inne schematy losowania – zajmuje się tym szczegółowo dziedzina statystyki zwana **metodą reprezentacyjną**. Jednakże w dalszej części książki mówiąc o próbie losowej będziemy zawsze mieć na myśli właśnie próbę losową prostą.

Zbiór obserwacji jeszcze nieprzetworzonych w żaden sposób, nieogrupowanych i nieorganizowanych nazywamy **danymi surowymi**. Statystyka opisowa zajmuje się wstępna analizą próbki – a więc właśnie danych surowych – odwołującą się w większym stopniu do intuicji, niż do rachunku prawdopodobieństwa. Jej celem jest uporządkowanie danych, wyznaczenie pewnych charakterystyk liczbowych opisujących badaną cechę oraz graficzne przedstawienie wyników obserwacji. Czynności te, zazwyczaj, poprzedzają

wnioskowanie statystyczne będące przedmiotem właściwej statystyki matematycznej.

3.3 Rozkład empiryczny cechy

3.3.1 Uwagi wstępne

Punktem wyjścia każdego wnioskowania statystycznego i podstawą wszelkich analiz statystycznych dotyczących badanej cechy jest **rozkład empiryczny** tej cechy. Pojęciem tym określamy przyporządkowanie poszczególnym wartościom cechy, obserwowanym w próbie, liczności bądź częstości ich występowania. W zależności od tego, czy mamy do czynienia z danymi jakościowymi, czy ilościowymi, inaczej przebiega konstrukcja rozkładu empirycznego i inna jest jego reprezentacja graficzna.

3.3.2 Metody opisu danych jakościowych

Opis danych jakościowych nie przedstawia większych trudności. Wszak już z samej definicji, dane tego typu utożsamiamy z listą kategorii (kompletną i rozłączną, co oznacza, że każda z obserwacji występujących w próbie należy do jednej i tylko do jednej kategorii). Jedyne co pozostaje do zrobienia z surowymi danymi, to policzyć, ile razy w próbie wystąpiła obserwacja z każdej kategorii. I takie właśnie zestawienie kategorii wraz z licznosciami obserwacji dla poszczególnych kategorii tworzy **rozkład empiryczny dla danych jakościowych**. Wspomnianą najprostszą postać rozkładu można, jeśli zachodzi taka potrzeba, uzupełnić o częstości wystąpienia obserwacji dla poszczególnych kategorii, tzn. liczności kategorii podzielonych przez łączną liczbę obserwacji. Mamy więc w istocie dwa rodzaje rozkładów empirycznych dla danych jakościowych: **rozkład empiryczny licznosci i rozkład empiryczny częstości**.

Bardziej formalnie

Definicja 50 *Rozkładem empirycznym licznosci dla danych jakościowych nazywamy zbiór par uporządkowanych*

$$\{(C_i, n_i) : i = 1, \dots, k\}, \quad (3.1)$$

gdzie C_1, \dots, C_k oznacza rozłączną i kompletną listę kategorii odpowiadającą możliwym wynikom obserwacji, natomiast n_1, \dots, n_k są licznosciami obserwacji dla odpowiednich kategorii.

Definicja 51 *Rozkładem empirycznym częstości dla danych jakościowych nazywamy zbiór par uporządkowanych*

$$\{(C_i, f_i) : i = 1, \dots, k\}, \quad (3.2)$$

gdzie C_1, \dots, C_k oznacza rozłączną i kompletną listę kategorii odpowiadających możliwym wynikom obserwacji, natomiast f_1, \dots, f_k są częstotliwościami obserwacji dla odpowiednich kategorii, tzn.

$$f_i = \frac{n_i}{n}, \quad (3.3)$$

przy czym $n = n_1 + \dots + n_k$ jest licznością próby.

W praktyce stosuje się dwie ilustracje graficzne rozkładu empirycznego dla danych jakościowych: wykres kołowy oraz wykres słupkowy.

Wykres kołowy, to nic innego jak koło (reprezentujące ogół obserwacji w próbie) z wyodrębnionymi wycinkami (reprezentującymi poszczególne kategorie), których pole jest proporcjonalne do częstości odpowiadających im kategorii. Przykład wykresu kołowego przedstawia rysunek 3.4 na str. 98.

Na wykresie słupkowym kategorie reprezentowane są przez rysowane pionowo albo poziomo prostokąty (rysunek 3.5 na str. 98), których długość jest proporcjonalna do liczności lub częstości odpowiadających im kategorii. (patrz: **Przykład 3.3**)

3.3.3 Metody opisu danych ilościowych

Omawiając dane typu ilościowego rozpatrzymy oddzielnie tzw. dane indywidualne oraz tzw. dane pogrupowane.

Dane indywidualne, to po prostu nieprzetworzony pełen zbiór surowych obserwacji. Założmy, że nasza próba liczy n obserwacji, które mogą się różnić, albo nie. Przyjmijmy, że w próbie występuje k różnych wartości w_1, \dots, w_k , przy czym $1 \leq k \leq n$. Założmy dodatkowo, że wartość w_i występuje w próbie dokładnie n_i razy, przy czym, oczywiście, $n_1 + \dots + n_k = n$. Wprowadzone oznaczenia pozwalały zdefiniować podstawowe narzędzie opisu rozkładu empirycznego indywidualnych danych ilościowych, jakim jest dystrybuanta empiryczna.

Definicja 52 Dystrybuantą empiryczną nazywamy funkcję

$$F_n(x) = \begin{cases} 0 & \text{dla } x < w_1 \\ \frac{1}{n} \sum_{j=1}^i n_j & \text{dla } w_i \leq x < w_{i+1} \\ 1 & \text{dla } x \geq w_k, \end{cases} \quad (3.4)$$

gdzie $i = 1, \dots, k - 1$.

Ilustracją graficzną dystrybuanty empirycznej jest krzywa schodkowa niemalejąca, o skokach w punktach w_i , $i = 1, \dots, k$, o wielkościach opisanych wzorem (3.4). Wspomnijmy w tym miejscu, że pojęcie dystrybuanty

empirycznej – bardzo ważne w statystyce – wystąpi jeszcze wielokrotnie w niniejszej książce, przy czym dalej podana zostanie jeszcze inna, równoważna, definicja tego pojęcia.

W praktyce dystrybuantę empiryczną postaci (3.4) wyznacza się zwykle jedynie dla prób o stosunkowo małej liczności lub dla prób składających się z obserwacji przyjmujących niewielką liczbę różnych wartości. Przy dużej liczności próbki, w celu ułatwienia analizy, dane grupuje się w przedziały klasowe (klasy), tworząc tzw. szereg rozdzielczy. Najczęściej (choć nie jest to konieczne) tworzy się szereg rozdzielczy z klasami o jednakowej długości. Liczbę klas k dobiera się w zależności od liczności próbki n . Zazwyczaj zaleca się, aby

$$\frac{3}{4}\sqrt{n} \leq k \leq \sqrt{n}. \quad (3.5)$$

Właściwy dobór liczby klas jest niezmiernie ważny, bowiem ustalenie zbyt małej liczby powoduje nadmierną utratę informacji, zaś zbyt duża liczba klas sprawia, że dane stają się mało przejrzyste.

Jeżeli przyjmuje się, że klasy będą miały jednakową długość, wówczas ową długość klasy b wyznacza się ze wzoru

$$b \simeq \frac{X_{n:n} - X_{1:n}}{k}, \quad (3.6)$$

gdzie $X_{1:n}$ i $X_{n:n}$ oznaczają, odpowiednio, najmniejszą i największą obserwację w próbie.

Tradycyjny szereg rozdzielczy jest tablicą zawierającą trzy kolumny: w pierwszej kolumnie umieszcza się granice klas, w drugiej środki przedziałów klasowych x_i^0 , a w trzeciej liczności n_i obserwacji należących do kolejnych klas, przy czym x_i^0 wyznacza się ze wzoru

$$x_i^0 = \frac{\xi_i^- + \xi_i^+}{2}, \quad (3.7)$$

gdzie ξ_i^- i ξ_i^+ oznaczają, odpowiednio, dolną i górną granicę i -tego przedziału klasowego.

Czasem oprócz liczności wyznacza się również tzw. liczności skumulowane. Licznością skumulowaną i -tej klasy nazywamy łączną licznosć danej klasy oraz klas poprzedzających daną klasę, czyli sumę

$$cn_i = \sum_{j=1}^i n_j. \quad (3.8)$$

Dzieląc licznosć lub licznosć skumulowaną i -tej klasy przez liczbę wszystkich obserwacji otrzymujemy, odpowiednio, częstość lub częstość skumulowaną i -tej klasy szeregu rozdzielczego.

Pełen szereg rozdzielczy zapisuje się w postaci tablicy zawierającej końce przedziałów klasowych, środki przedziałów klasowych, liczności, liczności skumulowane, częstotliwości oraz częstotliwości skumulowane poszczególnych klas.

Do wyjaśnienia pozostaje jeszcze kwestia sposobu postępowania z obserwacjami leżącymi na granicy klas, a dokładniej odpowiedź na pytanie, do której klasy je zakwalifikować. Trzeba po prostu zdecydować się, czy klasy traktujemy jako przedziały lewostronnie domknięte i prawostronnie otwarte, tzn. $[., \cdot)$, czy też na odwrót – lewostronnie otwarte i prawostronnie domknięte tzn. $(\cdot, \cdot]$. By uniknąć tego typu dylematów można też wyznaczać granice klas w ten sposób, aby były one liczbami zawierającymi o jedną cyfrę znaczącą więcej niż obserwacje (wówczas żadna z obserwacji nie znajdzie się na granicy klas). Jest to, w istocie, sprawa drugorzędna, a najważniejsze to pamiętać, że każda z obserwacji musi znaleźć się w jednej i tylko w jednej klasie.

Graficzną ilustracją szeregu rozdzielczego jest **histogram**. Jest to wykres słupkowy, którego podstawą stanowią przedziały klasowe, natomiast wysokości słupków są proporcjonalne do liczności n_i poszczególnych klas. Taki histogram zwany jest **histogramem liczności**. Przykładowy histogram liczności zamieszczono na rysunku 3.1 na str. 94. Biorąc zaś wysokości słupków proporcjonalne do częstotliwości ($f_i = \frac{n_i}{n}$), liczności skumulowanych (cn_i) lub częstotliwości skumulowanych ($cf_i = \sum_{j=1}^i f_j$) otrzymamy, odpowiednio, **histogram częstotliwości**, **histogram liczności skumulowanych** (rys. 3.2 na str. 94) oraz **histogram częstotliwości skumulowanych**.

Inną formą graficznej prezentacji danych pogrupowanych są łamane (wieloboki) liczności bądź częstotliwości. Łącząc punkty o współrzędnych (x_i^0, n_i) , gdzie x_i^0 oznacza środek i -tego przedziału klasowego, otrzymujemy tzw. **łamana liczności** (rys. 3.1 na str. 94). **Łamana częstotliwości** otrzymuje się przez połączenie punktów o współrzędnych (x_i^0, f_i) . Z kolei **łamana liczności skumulowanych** (rys. 3.2 na str. 94) oraz **łamana częstotliwości skumulowanych** uzyskujemy łącząc punkty o współrzędnych, odpowiednio, (ξ_i^+, cn_i) bądź (ξ_i^+, cf_i) , gdzie ξ_i^+ oznacza górny kraniec i -tego przedziału klasowego.

(patrz: **Przykład 3.1**)

3.4 Syntetyczne charakterystyki próby

3.4.1 Uwagi wstępne

Choć rozkład empiryczny próby jest dobrze opisany przez szereg rozdzielczy, często jednak bardziej użyteczne w praktyce są pewne wskaźniki liczbowe, pozwalające w syntetyczny sposób scharakteryzować próbę i ocenić jej własności. Owa syntetyczna ocena może dotyczyć różnych aspektów, a w szczególności: poziomu cechy, jej zróżnicowania oraz kształtu rozkładu cechy. Stąd też wśród znanych charakterystyk wyróżniamy: **miary położenia**,

miary rozproszenia i charakterystyki kształtu. Przedstawione poniżej syntetyczne charakterystyki próby nazywa się czasem **statystykami opisowymi**. Na zakończenie niniejszego podrozdziału przedstawimy jeszcze pewną miarę korelacji dla dwóch prób stosowaną powszechnie w przypadku badania związku między dwiema cechami.

3.4.2 Miary położenia

Wśród miar położenia wyodrębnia się **miary tendencji centralnej**, wskazujące wartości "typowe" badanej cechy oraz **miary pozycji**, określające położenie wybranych obserwacji względem innych obserwacji w próbie. Najczęściej używanymi miarami tendencji centralnej są: średnia arytmetyczna, mediana i moda.

Średnia arytmetyczna (średnia) z próby losowej X_1, \dots, X_n nazywamy liczbę określona wzorem

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i. \quad (3.9)$$

Jeżeli nie dysponujemy danymi surowymi, a jedynie pogrupowanymi w szeregu rozdzielczy, wówczas średnią wylicza się ze wzoru

$$\bar{X} \simeq \frac{1}{n} \sum_{i=1}^k n_i x_i^0, \quad (3.10)$$

gdzie k oznacza liczbę klas w szeregu rozdzielczym, n_i jest licznością klas, natomiast x_i^0 jest środkiem klas. W przypadku, kiedy dostępne są dane surowe, jak i pogrupowane, zaleca się korzystać ze wzoru (3.9), bowiem wzór (3.10) jest jedynie wzorem przybliżonym.

Średnią arytmetyczną można interpretować jako współrzędną środkową masy układu punktów materialnych o masach jednostkowych umieszczonej w punktach o współrzędnych X_i .

Medianą z próby nazywamy taką wartość cechy, że co najmniej 50% obserwacji przyjmuje wartość nie większą od niej i jednocześnie co najmniej 50% obserwacji ma wartość nie mniejszą od tej wartości. Jeżeli liczność próby n jest liczbą nieparzystą, wówczas mediana jest równa środkowej obserwacji w uporządkowanej niemalejąco próbie, natomiast gdy n jest liczbą parzystą, wtedy za medianę przyjmuje się średnią arytmetyczną z dwóch środkowych obserwacji. Formalnie, można to ująć wzorem

$$Med = \begin{cases} X_{\frac{n+1}{2}:n} & \text{gdy } n \text{ jest nieparzyste} \\ \frac{1}{2}(X_{\frac{n}{2}:n} + X_{\frac{n}{2}+1:n}) & \text{gdy } n \text{ jest parzyste}, \end{cases} \quad (3.11)$$

gdzie $X_{k:n}$ oznacza k -tą statystykę pozycyjną, czyli k -tą obserwację w uporządkowanej niemalejąco próbie

$$X_{1:n} \leq X_{2:n} \leq \dots \leq X_{n:n}. \quad (3.12)$$

Dla danych pogrupowanych medianę wyznaczamy w sposób przybliżony ze wzoru

$$Med \approx x_L + \frac{b}{n_M} \left(\frac{n}{2} - \sum_{i=1}^{M-1} n_i \right), \quad (3.13)$$

gdzie x_L jest dolną granicą klasy, w której znajduje się mediana (klasy medianowej), b jest szerokością klasy medianowej, n_M jest licznością klasy medianowej, zaś M jest numerem klasy, w której znajduje się mediana.

Modą (wartością modalną, dominantą) nazywamy wartość najczęściej powtarzającą się w próbie. Często zakłada się dodatkowo, że nie może to być wartość najmniejsza ani największa.

Porównując ze sobą trzy wymienione powyżej miary tendencji centralnej nie można jednoznacznie orzec, która z nich jest najlepsza. Każda z nich ma swoje zalety i wady i decyza, którą z nich najlepiej zastosować zależy od konkretnej sytuacji. Zaletą średniej arytmetycznej jest to, że przy wyznaczaniu typowej wartości cechy bierze pod uwagę całą próbę. To samo może być jednak czasem i wadą, bowiem dzięki tej właściwości średnia jest bardzo wrażliwa na wpływ obserwacji odstających (tzw. outlierów). Obserwacje odstające, to te wartości w próbie, które są bardzo małe albo bardzo duże w stosunku do ogółu obserwacji. Najczęściej, choć nie zawsze, outliersy są rezultatem błędów grubych. Dlatego też należy być bardzo ostrożnym przy wyciąganiu wniosków na podstawie próby zawierającej obserwacje odstające.

W przeciwnieństwie do średniej, mediana jest całkowicie odporna na wpływ obserwacji odstających, bowiem wyznaczana jest na podstawie środkowych wartości w próbie i ignoruje całkowicie wartości ekstremalne.

W porównaniu ze średnią i medianą, moda używana jest w praktyce stosunkowo rzadko. Warto zaznaczyć, że o ile średnia i mediana z próby może być wyliczona zawsze i w sposób jednoznaczny, to moda nie musi istnieć, bądź też może być więcej niż jedna moda w próbie. Z kolei modę można wyznaczyć zarówno dla danych ilościowych, jak i jakościowych, podczas gdy średnią i medianę daje się wyliczyć tylko dla danych ilościowych.

Wśród innych miar położenia warto jeszcze wspomnieć średnią geometryczną

$$\bar{X}_G = \sqrt[n]{X_1 \cdot \dots \cdot X_n} \quad (3.14)$$

oraz średnią harmoniczną

$$\bar{X}_H = \frac{n}{\sum_{i=1}^n \frac{1}{X_i}}. \quad (3.15)$$

Nietrudno zauważyć, że nie dla każdej próby da się wyliczyć \bar{X}_G i \bar{X}_H . Ogólnie, przy założeniu, że owe średnie istnieją, można udowodnić iż

$$\bar{X}_H \leq \bar{X}_G \leq \bar{X}. \quad (3.16)$$

Najczęściej stosowanymi miarami pozycji są kwartyły. Mówiąc nieformalnie, kwartyły, to trzy wartości dzielące próbę na cztery równe części. A dokładniej, pierwszy kwartyl (kwartyl dolny) Q_1 , to taka wartość cechy, że co najmniej 25% obserwacji przyjmuje wartość nie większą od niej i jednocześnie co najmniej 75% obserwacji ma wartość nie mniejszą od tej wartości. Drugi kwartyl, jak nie trudno się domyślić, jest równy medianie. Natomiast trzeci kwartyl (górnego kwartylu) Q_3 , to taka wartość cechy, że co najmniej 75% obserwacji przyjmuje wartość nie większą od niej i jednocześnie co najmniej 25% obserwacji ma wartość nie mniejszą od tej wartości. W praktyce dolny kwartyl wyznacza się jako medianę z podpróbki złożonej z obserwacji leżących w uporządkowanym niemalejąco ciągu obserwacji na lewo od mediany z całej próby. Z kolei górny kwartyl wyznacza się jako medianę z podpróbki złożonej z obserwacji leżących w uporządkowanym niemalejąco ciągu obserwacji na prawo od mediany z całej próby.

Oprócz kwartyli w statystyce opisowej stosuje się także decyle, dzielące uporządkowaną niemalejąco próbę na 10 równych części oraz percentile (centyle), dzielące uporządkowaną niemalejąco próbę na 100 równych części.

(patrz: Przykład 3.1, 3.2)

3.4.3 Miary rozproszenia

Informacja o "typowej" wartości badanej cechy bywa często niewystarczająca. Przykładowo może się zdarzyć, że średnie w dwóch próbach są sobie równe, gdy tymczasem próbki te różnią się istotnie rozrzedzeniem obserwacji. Stąd potrzeba znajomości miary rozproszenia (rozrzutu) obserwacji w próbie.

Najprostszą miarą rozproszenia jest rozstęp, czyli odległość między najmniejszą i największą obserwacją w próbie. Korzystając z wprowadzonych powyżej oznaczeń możemy rozstęp zapisać wzorem

$$R = X_{n:n} - X_{1:n}, \quad (3.17)$$

gdzie $X_{n:n}$ i $X_{1:n}$ oznaczają, odpowiednio, największą i najmniejszą obserwację w próbie. Tak więc rozstęp podaje długość przedziału, na którym rozproszone są obserwacje występujące w próbie.

Zaletą rozstępu, jako miary rozproszenia, jest prostota obliczeniowa. Rozstęp ma jednak wiele mankamentów wynikających z tego, że jest on funkcją wyłącznie wartości ekstremalnych. W szczególności, jest on nieodporny na wpływ obserwacji odstających. Wady tej pozbawiony jest rozstęp międzykwartylowy dany wzorem

$$IQR = Q_3 - Q_1, \quad (3.18)$$

gdzie Q_3 i Q_1 oznaczają, odpowiednio, trzeci i pierwszy kwartyl. Jak widać ze wzoru, rozstęp międzykwartylowy podaje długość odcinka, na którym

leży 50% środkowych wartości w uporządkowanej niemalejąco próbie. Czasem jako miarę rozrzutu używa się połowę rozstępu międzykwartylowego, zwana odchyleniem ćwiartkowym

$$Q = \frac{Q_3 - Q_1}{2}. \quad (3.19)$$

Wymienione dotąd miary rozproszenia nie są w pełni zadowalające, gdyż zbudowane są wyłącznie na dwóch miarach pozycyjnych, a więc ignorują informacje o większości obserwacji. Tymczasem pożądana byłaby znajomość takiej miary rozrzutu, która brałaby pod uwagę wszystkie obserwacje rozważanej próby. Taką miarą rozproszenia jest wariancja z próby oraz odchylenie standardowe.

Wariancja z próby zdefiniowana jest wzorem

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2. \quad (3.20)$$

A zatem wariancja równa jest średniokwadratowemu odchyleniu obserwacji od średniej. W niektórych publikacjach wariancję definiuje się jako sumę kwadratów odchyleń od średniej dzieloną przez n , a nie przez $n-1$, jak to ma miejsce we wzorze (3.20). Powód, dla którego posłużyliśmy się dzielnikiem $n-1$, a nie n , wyjaśniony zostanie w podrozdz. 4.4. A mówiąc nieformalnie, przyjmując we wzorze (3.20) dzielnik n uzyskiwalibyśmy na podstawie badań częstociowych systematycznie zaniżone oszacowania wariancji dla całej populacji.

W praktyce użyteczny bywa równoważny wzór na wariancję

$$S^2 = \frac{1}{n-1} \left[\sum_{i=1}^n X_i^2 - \frac{1}{n} \left(\sum_{i=1}^n X_i \right)^2 \right]. \quad (3.21)$$

Dla danych pogrupowanych otrzymujemy wzór przybliżony postaci

$$S^2 \simeq \frac{1}{n-1} \sum_{i=1}^k n_i (x_i^0 - \bar{X})^2, \quad (3.22)$$

gdzie k oznacza liczbę klas w szeregu rozdzielczym, n_i jest licznością klasy, natomiast x_i^0 jest środkiem klasy, bądź równoważny wzór

$$S^2 \simeq \frac{1}{n-1} \left[\sum_{i=1}^k n_i (x_i^0)^2 - \frac{1}{n} \left(\sum_{i=1}^k n_i x_i^0 \right)^2 \right]. \quad (3.23)$$

Wariancja podaje miarę rozproszenia w jednostkach będących kwadratem jednostek, w których dokonano pomiaru badanej cechy. Ponieważ w wielu przypadkach wygodnie jest wyrażać rozrzut w tych samych jednostkach, co

obserwacje, często korzysta się z miary rozproszenia, zwanej **odchyleniem standardowym**, która jest pierwiastkiem kwadratowym z wariancji. Tak więc odchylenie standardowe S równe jest

$$S = \sqrt{S^2} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2}. \quad (3.24)$$

W sytuacji gdy chcemy porównać rozrzut dwóch lub więcej prób, nierzaz wygodniej jest skorzystać ze względnej miary rozproszenia. Taką miarą jest współczynnik zmienności

$$V = \frac{S}{\bar{X}} (100\%). \quad (3.25)$$

(patrz: Przykład 3.1, 3.2)

3.4.4 Charakterystyki kształtu

Oprócz wspomnianych miar położenia i rozrzutu warto jeszcze wspomnieć o tzw. *miarach kształtu*, spośród których najważniejsze znaczenie ma skośność oraz kurtoza.

O rozkładzie empirycznym powiemy, że jest **symetryczny**, jeżeli dla każdej wartości cechy $X_i < \bar{X}$ istnieje w próbie wartość $X_j > \bar{X}$ taka, że $\bar{X} - X_i = X_j - \bar{X}$. Rozkład, który nie jest symetryczny nazywamy **asymetrycznym** lub **skośnym**. W przypadku rozkładu skośnego możemy mieć do czynienia z asymetrią dodatnią (prawostawną) albo ujemną (lewostronną), w zależności od tego, czy bardziej wydłużone jest prawe, czy lewe ramię wykresu rozkładu.

Skośność, zwana także **współczynnikiem asymetrii**, jest wielkością niemianowaną charakteryzującą stopień i kierunek asymetrii rozkładu empirycznego badanej cechy. Współczynnik ten wyliczamy ze wzoru

$$A = \frac{n \sum_{i=1}^n (X_i - \bar{X})^3}{(n-1)(n-2)S^3}, \quad (3.26)$$

przy czym $A = 0$ oznacza, że obserwacje są symetrycznie rozłożone względem średniej, $A > 0$ mówi o dodatniej asymetrii, natomiast $A < 0$ wskazuje na asymetrię ujemną. Trzeba zaznaczyć, że współczynnik asymetrii powinno badać się jedynie dla rozkładów jednomodalnych.

W tym miejscu warto wspomnieć o relacji między średnią \bar{X} , medianą Med i modą Mo w kontekście asymetrii. Otóż przy asymetrii dodatniej mamy $Mo < Med < \bar{X}$ zaś przy asymetrii ujemnej $\bar{X} < Med < Mo$. W przypadku rozkładu symetrycznego wspomniane trzy parametry są sobie równe, tzn. $\bar{X} = Med = Mo$.

W pewnych zastosowaniach korzysta się ze standaryzowanego współczynnika asymetrii danego wzorem

$$A_s = \frac{A}{\sqrt{\frac{6}{n}}}, \quad (3.27)$$

który dla dużego n ($n > 150$) ma rozkład asymptotyczny normalny.

Rzadziej stosowaną miarą asymetrii jest pozycyjny współczynnik asymetrii, zdefiniowany za pomocą kwartyli Q_1 , Med , Q_3 oraz odchylenia ćwiartkowego Q :

$$A_{poz} = \frac{(Q_3 - Med) - (Med - Q_1)}{2Q}. \quad (3.28)$$

Inną ważną charakterystyką kształtu rozkładu jest kurtoza, zwana także współczynnikiem spłaszczenia, zdefiniowana wzorem

$$K = \frac{n(n+1) \sum_{i=1}^n (X_i - \bar{X})^4}{(n-1)(n-2)(n-3)S^4} - \frac{3(n-1)^2}{(n-2)(n-3)}. \quad (3.29)$$

Współczynnik ten wskazuje na ile wykres rozkładu empirycznego badanej cechy jest płaski (bądź stromy) względem wykresu rozkładu normalnego. Kurtoza rozkładu normalnego wynosi zero. Ujemna wartość kurtozy oznacza, że wykres rozkładu badanej cechy jest bardziej płaski niż wykres rozkładu normalnego i ma krótsze ramiona (ogony). Gdy $K > 0$ wtedy wykres rozkładu jest albo bardziej stromy niż wykres rozkładu normalnego (duże skupienie obserwacji wokół mediany) albo ma tzw. "ciężkie ogony". Podobnie, jak w przypadku współczynnika asymetrii, badanie kurtozy ma sens wyłącznie w przypadku rozkładów jednomodalnych.

Standaryzowana kurtoza

$$K_s = \frac{K}{\sqrt{\frac{24}{n}}}, \quad (3.30)$$

stosowana w pewnych sytuacjach, ma – podobnie jak standaryzowany współczynnik asymetrii – rozkład asymptotycznie normalny.

(patrz: Przykład 3.2)

3.4.5 Miary korelacji

Jak wiadomo, zależność między dwiema zmiennymi losowymi można scharakteryzować za pomocą współczynnika korelacji (2.108). Dla dwuwymiarowego rozkładu empirycznego możemy natomiast wyliczyć tzw. współczynnik korelacji z próby (współczynnik korelacji Pearsona). W przypadku jednoczesnego badania dwóch cech pewnej populacji naszą próbą jest

ciąg par $(X_1, Y_1), \dots, (X_n, Y_n)$, gdzie X_i oraz Y_i oznaczają, odpowiednio, wartości pierwszej i drugiej cechy przyjmowane przez i -ty element próby. Dla takiej próby empiryczny współczynnik korelacji r dany jest wzorem

$$r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 \sum_{i=1}^n (Y_i - \bar{Y})^2}}, \quad (3.31)$$

gdzie $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ i $\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$.

Współczynnik (3.31) przyjmuje wartości z przedziału $[-1, 1]$. Jeżeli $r = 1$ lub $r = -1$, oznacza to, że punkty $(X_1, Y_1), \dots, (X_n, Y_n)$ leżą na prostej, odpowiednio, rosnącej bądź malejącej. Wartości współczynnika korelacji bliskie 1 bądź -1 wskazują na silną korelację liniową. Współczynnik (3.31) jest nie tylko miarą siły korelacji, ale też wskazuje jej kierunek: dodatnie wartości r (bliskie 1) mówią o korelacji dodatniej, co oznacza tendencję występowania dużych wartości X z dużymi wartościami Y oraz występowania małych wartości X z małymi wartościami Y , natomiast ujemne wartości r (bliskie -1) mówią o korelacji ujemnej, tzn. że duże wartości X wykazują tendencję do występowania z małymi wartościami Y zaś małe wartości X z dużymi wartościami Y . Jeżeli wartość współczynnika r wynosi零, oznacza to brak skorelowania liniowego, co nie jest zasadniczo jednoznaczne z brakiem istnienia jakiegokolwiek związku między badanymi cechami (inny sposób $r = 0$ oznacza, że nie ma związku między cechami, lub też jest, ale na pewno nie jest to związek liniowy).

(patrz: Przykład 5.15)

3.5 Użyteczne wykresy statystyki opisowej

3.5.1 Uwagi wstępne

Na wstępnie tego rozdziału omówiono podstawowe metody graficznej prezentacji danych, jakimi są – w przypadku danych jakościowych – wykres kołowy i słupkowy, natomiast – w przypadku danych ilościowych – histogramy i łamane liczności (częstości). Z uwagi na to, że metody graficzne są niezwykle użyteczne w praktyce i niejednokrotnie pozwalały szybciej zinterpretować uzyskane informacje niż zestawy liczb, poniżej zaprezentujemy jeszcze dwa inne wykresy stosowane w statystyce opisowej: wykres skrzynkowy i wykres łodygowo-ilościowy.

3.5.2 Wykres skrzynkowy

Wykres skrzynkowy (ang. box-plot) lub wykres typu "skrzynka z wąsami" (ang. box-and-whisker-plot) jest uniwersalnym narzędziem pozwalającym

ująć na jednym rysunku wiadomości dotyczące położenia, rozproszenia i kształtu rozkładu empirycznego badanej cechy. Przykładowy wykres skrzynkowy zamieszczono na rys. 3.3.

Centrum wykresu zajmuje prostokąt ("skrzynka"), którego lewy bok wyznaczony jest przez dolny kwartyl, a prawy bok przez góry kwartyl. Tak więc wewnątrz owego prostokąta znajduje się 50% środkowych obserwacji, a zatem długość prostokąta odpowiada rozstępowi międzykwartylowemu. "Skrzynka" przedzielona jest pionową linią odpowiadającą medianie. Krzyżek leżący wewnątrz "skrzynki" oznacza średnią arytmetyczną. Na zewnątrz skrzynki znajdują się dwa odcinki ("wąsy") poprowadzone na lewo od dolnego kwartyla i na prawo od górnego kwartyla do punktów, odpowiednio, x_* i x^* . W tradycyjnym wykresie skrzynkowym punkty te odpowiadają najmniejszej i największej obserwacji w próbie, tzn. $x_* = X_{1:n}$, oraz $x^* = X_{n:n}$.

Stosuje się też bardziej wyrafinowane wykresy skrzynkowe, pozwalające wykrywać obserwacje odstające. Na takich wykresach wąsy sięgają, jak wcześniej, od x_* do Q_1 oraz od Q_3 do x^* , tyle, że teraz

$$\begin{aligned} x_* &= \min \left\{ X_i : X_i \in [Q_1 - \frac{3}{2}(Q_3 - Q_1), Q_1] \right\} \\ x^* &= \max \left\{ X_i : X_i \in [Q_3, Q_3 + \frac{3}{2}(Q_3 - Q_1)] \right\}. \end{aligned} \quad (3.32)$$

Obserwacje, które nie mieścią się w odcinku $[x_*, x^*]$ oznaczane są kropkami i są to właśnie tzw. obserwacje odstające.

Wykresy skrzynkowe są szczególnie użyteczne przy porównywaniu kilku prób. Dzięki nim można szybko zorientować się, w której z prób badana cecha osiąga najwyższy poziom, jak lokuje się typowy poziom cechy w każdej z prób, które pomiary wykazują największy, bądź najmniejszy rozrzut, jaki jest kształtem (skośność) rozkładów empirycznych w poszczególnych próbach itd.

(patrz: Przykład 3.2)

3.5.3 Wykres łodygowo-liściowy

Innym sposobem prezentacji wyników pomiarów w syntetyczny sposób graficzny jest **wykres łodygowo-liściowy** (ang. stem-and-leaf display). Zasada tworzenia tego wykresu jest następująca: początкова cyfra (cyfry) każdej z liczb odpowiadających obserwacjom, tworzy tzw. "łodygę". Wartości te wypisywane są pionowo, w rosnącym porządku, jedna pod drugą. Poostała część każdej z liczb, czyli tzw. "liść" umieszczany jest po prawej stronie odpowiadającej mu "łodygi". W ten sposób otrzymujemy wykres przypominający histogram zbudowany z cyfr i obrócony w prawo o 90°, przy czym wartości tworzące "łodygę" są odpowiednikami klas.

Wykres łodygowo-liściowy, podobnie jak histogram, pozwala w łatwy sposób określić zakres zmienności, skoncentrowanie i asymetrię rozkładu badanej cechy, a także wykryć ewentualne obserwacje odstające. (patrz: Przykład 3.1)

3.6 Przykłady

Przykład 3.1

Firma budowlana zainteresowana jest jakością betonu. Jedną z pożądanych cech jest jego odpowiednia wytrzymałość na ściskanie. Aby sprawdzić, czy beton proponowany przez dostawcę spełnia normy jakości, przebadano 40 próbek i otrzymano następujące wyniki wytrzymałości na ściskanie: 20.3, 23.4, 21.5, 21.8, 22.0, 24.5, 23.4, 22.7, 24.1, 22.5, 19.6, 21.0, 23.8, 22.3, 21.4, 20.7, 20.1, 24.9, 23.5, 20.1, 23.3, 25.0, 22.3, 19.5, 22.2, 23.4, 24.3, 22.2, 23.7, 21.4, 21.1, 22.2, 23.0, 19.8, 22.1, 23.8, 22.4, 22.1, 23.2, 20.8.

- Zbudować szereg rozdzielczy.
- Narysować histogram liczności oraz łamana liczności.
- Narysować histogram skumulowanych liczności oraz łamana skumulowanych liczności.
- Sporządzić wykres łodygowo-liściowy.

Rozwiążanie

a) Aby sporządzić szereg rozdzielczy musimy najpierw wybrać stosowną liczbę klas i ustalić długość każdej klasy. Najczęściej przyjmuje się liczbę klas z przedziału $\frac{3}{4}\sqrt{n} \leq k \leq \sqrt{n}$ (por. (3.5)), gdzie n oznacza liczbę wszystkich obserwacji. W naszym zadaniu $n = 40$, stąd $4.7434 \leq k \leq 6.3246$. Zatem możemy взять liczbę klas k równą 5 lub 6. Wybierzmy liczbę klas równą 5.

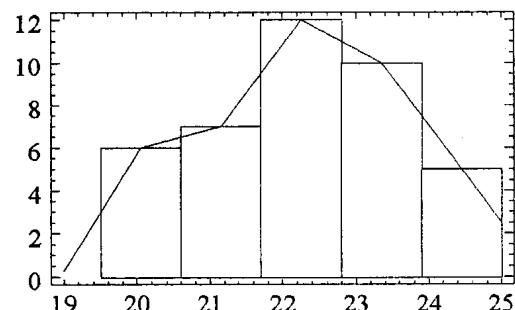
Teraz możemy wyliczyć długość każdej z klas (zazwyczaj przyjmuje się, że mają one taką samą długość). W tym celu znajdujemy obserwację najmniejszą $X_{1:n} = 19.3$ i największą $X_{n:n} = 25.1$ oraz rozstęp (czyli odległość między najmniejszą i największą obserwacją) $R = 25.0 - 19.5 = 5.5$. Długość klasy znajdujemy ze wzoru (3.6) $\frac{R}{k} = \frac{5.5}{5} = 1.1$. Możemy więc już przystąpić do konstrukcji szeregu rozdzielczego:

numer klasy	granice klas	x_i^0	n_i	f_i	cn_i	$c f_i$
1	[19.5, 20.6)	20.05	6	0.15	6	0.15
2	[20.6, 21.7)	21.15	7	0.175	13	0.325
3	[21.7, 22.8)	22.25	12	0.3	25	0.625
4	[22.8, 23.9)	23.35	10	0.25	35	0.875
5	[23.9, 25.0]	24.45	5	0.125	40	1.0

Sposób wyznaczania częstości f_i , liczności skumulowanych cn_i i częstości skumulowanych cf_i , zamieszczonych w powyższej tabeli, omówimy na przykładzie czwartej klasy: $n_4 = 10$, stąd częstość $f_4 = \frac{n_4}{n} = \frac{10}{40} = 0.25$; liczność skumulowana tej klasy jest równa $cn_4 = n_1 + n_2 + n_3 + n_4 = 6 + 7 + 12 + 10 = 35$, natomiast częstość skumulowana $cf_4 = 0.15 + 0.175 + 0.3 + 0.25 = 0.875$.

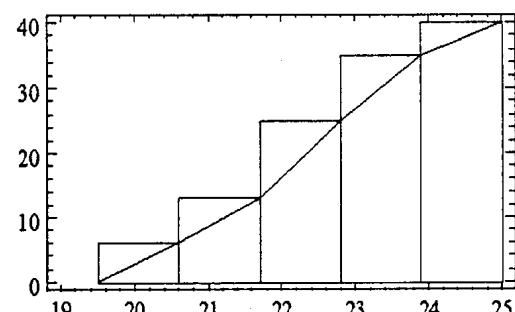
Poniżej zamieszczone są dwa wykresy przedstawiające histogram licznosci z nałożoną na niego łamana licznosci (rys. 3.1) oraz histogram licznosci skumulowanych z nałożoną na niego łamana licznosci skumulowanych (rys. 3.2).

b)



Rys. 3.1 – Histogram i łamana licznosci

c)



Rys. 3.2 – Histogram i łamana licznosci skumulowanych

d) Wykres łodygowo-liściowy wygląda zaś następująco

19	5 6 8
20	1 1 3 7 8
21	0 1 4 4 5 8
22	0 1 1 2 2 2 3 3 4 5 7
23	0 2 3 4 4 4 5 7 8 8
24	1 3 5 9
25	0

Przykład 3.2

18 studentów drugiego roku zapytano na ilu wykładach z RPiS byli w ciągu semestru. Uzyskano następujące odpowiedzi:

12, 15, 9, 13, 15, 13, 14, 10, 13, 1, 12, 14, 10, 6, 14, 12, 11, 13.

- a) Wyznaczyć podstawowe statystyki próbkkowe.
- b) Sporządzić i opisać wykres skrzynkowy.

Rozwiążanie

a) Zaczniemy od wyznaczenia podstawowych miar położenia. Średnia:

$$\begin{aligned}\bar{X} &= \frac{1}{n} \sum_{i=1}^n X_i = \frac{1}{18} (12 + 15 + 9 + 13 + 15 + 13 + 14 + \\ &+ 10 + 13 + 1 + 12 + 14 + 10 + 6 + 14 + 12 + 11 + 13) = \\ &= 11.5.\end{aligned}$$

Moda (najczęściej powtarzająca się liczba wy słuchanych wykładów):

$$MoX = 13.$$

Aby wyznaczyć medianę oraz kwartyle, należy najpierw niemalejaco uporządkować dane:

1, 6, 9, 10, 10, 11, 12, 12, 12, 13, 13, 13, 14, 14, 14, 15, 15.

Nasza próbka składa się z 18 elementów, czyli ma parzystą licznosć, zatem mediana to średnia z dwóch środkowych obserwacji:

$$MedX = \frac{12 + 13}{2} = 12.5.$$

Oznacza to, że co najmniej 50% studentów było na nie mniej niż 12.5 wykładach, zaś co najmniej 50% na co najwyżej 12.5 wykładach.

Kwartyl dolny to mediana z pierwszej połowy uporządkowanej rosnaco próbki, czyli z pierwszych 9 obserwacji: 1, 6, 9, 10, 10, 11, 12, 12, 12.

Ponieważ obecnie mamy nieparzystą liczbę elementów, to jako medianę wybieramy środkową obserwację, czyli 10. Zatem dolny kwartył

$$Q_1 = 10.$$

Oznacza to, że co najmniej 75% studentów było na nie mniej niż 10 wykłach, zaś co najmniej 25% na co najwyżej 10 wykładach.

Kwartył gorny to mediana z drugiej połowy uporządkowanej rosnąco próbki, czyli z ostatnich 9 obserwacji: 13, 13, 13, 13, 14, 14, 14, 15, 15. Znów mamy nieparzystą liczbę elementów, więc mediana to środkowa obserwacja, czyli 14. Zatem gorny kwartył

$$Q_3 = 14.$$

Oznacza to, że co najmniej 25% studentów było na nie mniej niż 14 wykłach, zaś co najmniej 75% studentów było na co najwyżej 14 wykładach.

Przejdźmy teraz do miar rozproszenia. Najprostsze z nich – rozstęp i rozstęp międzykwartylowy – wynoszą:

$$R = 15 - 1 = 14$$

$$IQR = 14 - 10 = 4.$$

Obliczamy wariancję:

$$\begin{aligned} S^2 &= \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 = \\ &= \frac{1}{17} [(12 - 11.5)^2 + (15 - 11.5)^2 + (9 - 11.5)^2 + (13 - 11.5)^2 + \\ &+ (15 - 11.5)^2 + (13 - 11.5)^2 + (14 - 11.5)^2 + (10 - 11.5)^2 + \\ &+ (13 - 11.5)^2 + (1 - 11.5)^2 + (12 - 11.5)^2 + (14 - 11.5)^2 + \\ &+ (10 - 11.5)^2 + (6 - 11.5)^2 + (14 - 11.5)^2 + (12 - 11.5)^2 + \\ &+ (11 - 11.5)^2 + (13 - 11.5)^2] = 12.0294. \end{aligned}$$

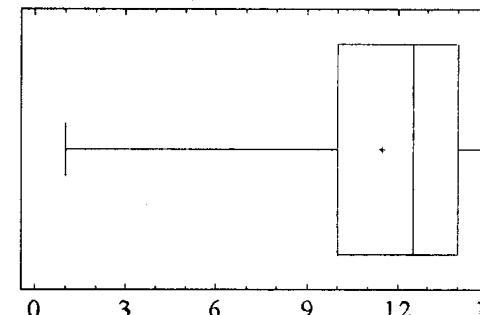
Stąd odchylenie standardowe wynosi

$$S = \sqrt{S^2} = \sqrt{12.0294} = 3.46834.$$

Współczynnik asymetrii, obliczony na podstawie wzoru (3.26), wynosi -1.90335 , co oznacza, że rozkład empiryczny badanej cechy jest asymetryczny i jest to asymetria ujemna (rozkład skośny w lewo). Świadczy też o tym relacja między średnią, medianą i modeą, a mianowicie:

$$\bar{X} < \text{Med}X < \text{Mo}X.$$

b) Wykres skrzynkowy przedstawiono na rys. 3.3.



Rys. 3.3 Wykres skrzynkowy

Przykład 3.3

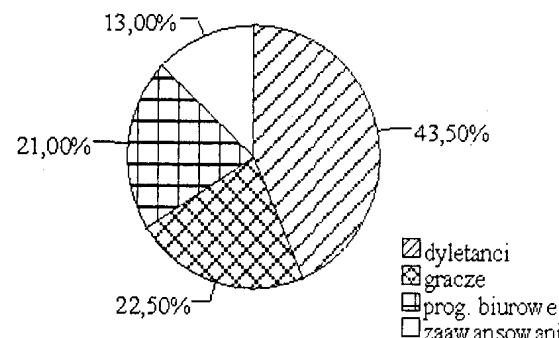
Badania ankietowe przeprowadzone na 200 osobowej próbie losowej pokazały, że 87 osób nie umie posługiwać się komputerem, 45 osób wykorzystuje komputer wyłącznie do gier, 42 osoby umieją posługiwać się jedynie podstawowymi programami biurowymi, zaś pozostałe osoby dysponują bardziej zaawansowanymi umiejętnościami i wiedzą w zakresie wykorzystania komputera. Zilustrować powyższe wyniki za pomocą wykresu kołowego i słupkowego.

Rozwiążanie

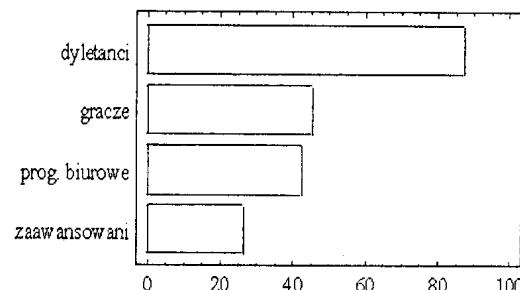
Zacznijmy od przedstawienia rozkładu empirycznego badanej przez nas cechy, którą możemy określić jako "umiejętności w zakresie obsługi komputera". W badaniu wyodrębniono cztery kategorie, którym nadaliśmy, odpowiednio, etykiety: dyletanci, gracze, programy biurowe, zaawansowani. Rozkład empiryczny przedstawia poniższa tabela

kategorie	n_i	f_i
dyletanci	87	0.435
gracze	45	0.225
prog. biurowe	42	0.21
zaawansowani	26	0.13

Ilustracją graficzną tego rozkładu empirycznego jest następujący wykres kołowy (rys. 3.4) i wykres słupkowy (rys. 3.5).



Rys. 3.4 – Wykres kołowy



Rys. 3.5 – Wykres słupkowy

3.7 Zadania

Zadanie 3.1

Zamieszczone poniżej dane to wyniki pomiarów lepkości pewnego produktu chemicznego, uzyskane w odstępach jednogodzinnych:
14.1, 13.3, 14.3, 13.4, 12.7, 11.3, 13.3, 12.2, 14.8, 14.7, 13.1, 11.3, 12.9, 12.0, 11.9, 11.7, 12.8, 12.6, 13.1, 13.1, 12.6, 12.9, 13.8, 11.4, 11.1, 14.2, 13.8, 14.1, 13.6.

Utworzyć dla tych danych:

- szereg rozdzielczy,
- histogram,
- lamaną liczności,
- wykres łodygowo-liściowy.

Zadanie 3.2

Utworzyć i opisać wykres skrzynkowy dla następujących danych:

8, 5, 17, 18, 9, 4, 17, 16, 12, 14.

Zadanie 3.3

W pewnym doświadczeniu farmakologicznym bada się utlenianie tkankowe wątroby. Dokonano 38 pomiarów tego utleniania i otrzymano następujące wyniki (ilość μl tlenu zużytego w ciągu 1 godziny przez 100 mg tkanki):

ilość zużytego tlenu	liczba pomiarów
15 - 25	4
25 - 35	6
35 - 45	12
45 - 55	7
55 - 65	6
65 - 75	3

Utworzyć dla tych danych histogram liczności i lamaną liczności.

4

Estymacja

4.1 Wprowadzenie

Teoria estymacji jest działem statystyki poświęconym szacowaniu wartości parametrów (bądź ich funkcji) rozkładu badanej cechy lub, ewentualnie, postaci rozkładu cechy. Jeżeli przedmiotem zainteresowania jest szacowanie jedynie parametrów rozkładu, wówczas mówimy o **estymacji parametrycznej**. Jeśli natomiast postępowanie dotyczy szacowania postaci rozkładu, to mamy wtedy do czynienia z **estymacją nieparametryczną**. Bieżący rozdział poświęcony jest zasadniczo estymacji parametrycznej. Kilka uwag dotyczących estymacji nieparametrycznej znajdzie Czytelnik w ostatnim punkcie tego rozdziału.

W zależności od sposobu, w jaki dokonuje się oszacowania interesującej nas wielkości można mówić o **estymacji punktowej** oraz o **estymacji przedziałowej**. Estymacja punktowa dostarcza ocenę liczbową nieznanego parametru w postaci jednej, konkretnej wartości. Niestety, szacując wartość parametru za pomocą estymacji punktowej nie mamy możliwości oceny dokładności szacowania. Możliwość tę daje estymacja przedziałowa, prowadząca do oceny parametru za pomocą pewnego przedziału liczbowego, tzw. **przedziału ufności**, zawierającego prawdziwą wartość poszukiwanego parametru na z góry zadany poziomie ufności.

W podrozdz. 4.2 omówiono niektóre własności estymatorów punktowych. Metody wyznaczania estymatorów przedstawiono w podrozdz. 4.3. W podrozdz. 4.4 zamieszczono listę podstawowych estymatorów. Podrozdz. 4.5 poświęcony jest wyznaczaniu przedziałów ufności oraz estymacji przedzia-

owej o zadanej precyzyji. Podrozdz. 4.6 dotyczy natomiast estymacji nieparametrycznej.

4.2 Podstawowe własności estymatorów

Przyjmijmy, że badana cecha ma rozkład F_θ , gdzie θ jest nieznanym parametrem tego rozkładu. Wartość parametru θ będziemy szacować na podstawie próby losowej X_1, \dots, X_n pochodzącej z badanej populacji. Każde narzędzie, za pomocą którego będziemy starali się dokonać oceny owego parametru nazywamy estymatorem. Formalnie estymatorem parametru θ rozkładu nazywamy dowolną statystykę z próby $\hat{\theta}_n = \hat{\theta}_n(X_1, \dots, X_n)$.

Z powyższej definicji wynika, że estymator jest zmienną losową, której rozkład w oczywisty sposób zależy od rozkładu badanej cechy, a w szczególności od parametru θ .

Choć dla danego parametru θ można utworzyć wiele estymatorów (w definicji podano, że estymatorem jest dowolna funkcja próby), to jednak – z oczywistych przyczyn – interesować się będziemy wyłącznie takimi estymatorami, które "dobrze" szacują θ . Ponieważ szacowania parametru dokonuje się na podstawie próby losowej, istnieje więc możliwość popełnienia błędu. Powstaje więc pytanie, w jaki sposób oceniać jakość estymatora. Najczęściej stosowanym kryterium oceny estymacji jest tzw. błąd średniokwadratowy:

$$R(\hat{\theta}_n, \theta) = E(\hat{\theta}_n - \theta)^2. \quad (4.1)$$

Byłoby rzeczą wielce pożądaną, gdyby istniał taki estymator, który minimalizowałby błąd średniokwadratowy dla dowolnej wartości parametru θ . Niestety, za wyjątkiem trywialnych sytuacji, estymator taki nie istnieje (por. Zieliński [18], Silvey [16]). Pozostaje więc sformułować jeszcze inne kryteria, które pozwolą uzyskać estymatory o możliwie "optimalnych" własnościach. Takimi własnościami są przede wszystkim: zgodność, nieobciążoność i efektywność.

Definicja 53 Mówimy, że estymator $\hat{\theta}_n$ parametru θ jest zgodny, jeżeli

$$\lim_{n \rightarrow \infty} P\left\{|\hat{\theta}_n - \theta| < \varepsilon\right\} = 1 \quad \forall \varepsilon > 0. \quad (4.2)$$

Zgodność estymatora odpowiada intuicyjnemu i naturalnemu postulatowi, aby przy dostatecznie dużej liczności próby estymator $\hat{\theta}_n$ przyjmował z dużym prawdopodobieństwem wartość bliskie estymowanemu parametrowi θ (a więc przy dostatecznie licznej próbie ryzyko popełnienia dużego błędu w szacowaniu było niewielkie).

Przykładowo, z prawa wielkich liczb Czebyszewa wynika natychmiast, że średnia arytmetyczna z próby \bar{X}_n jest estymatorem zgodnym wartości

oczekiwanej. Z kolei, z prawa wielkich liczb Bernoulliego wnioskujemy, że częstość jest estymatorem zgodnym prawdopodobieństwa sukcesu w rozkładzie dwupunktowym.

Inną ważną własnością jest nieobciążoność.

Definicja 54 Mówimy, że estymator $\hat{\theta}_n$ parametru θ jest nieobciążony, jeżeli

$$E(\hat{\theta}_n) = \theta. \quad (4.3)$$

W przeciwnym przypadku, tzn. gdy $E(\hat{\theta}_n) \neq \theta$, estymator $\hat{\theta}_n$ nazywamy obciążonym, a wielkość

$$b_n(\hat{\theta}_n) = E(\hat{\theta}_n) - \theta \quad (4.4)$$

nazywamy obciążeniem estymatora.

Nieobciążoność estymatora oznacza, że uzyskiwane dzięki niemu oceny parametru nie są obciążone błędem systematycznym, tzn., że stosując go nie będziemy z zasady ani przeszacowywać ani nie niedoszacowywać θ , ale średnio rzecz biorąc otrzymamy tyle, ile trzeba. W przypadku estymatora obciążonego miarą wspomnianego błędu systematycznego jest właśnie obciążenie estymatora.

Warto zaznaczyć, że estymator nieobciążony pozostanie dalej nieobciążony przy zmianie liczności próbki. Jednakże w przypadku estymatora obciążonego może się zdarzyć, że zwiększenie liczności próby pociągnie za sobą zmniejszenie obciążenia.

Definicja 55 Mówimy, że estymator $\hat{\theta}_n$ parametru θ jest asymptotycznie nieobciążony, jeżeli

$$\lim_{n \rightarrow \infty} b_n(\hat{\theta}_n) = \lim_{n \rightarrow \infty} E(\hat{\theta}_n) - \theta = 0. \quad (4.5)$$

Oznacza to, że dla dostatecznie dużej próby obciążenie estymatora asymptotycznie nieobciążonego jest pomijalne.

Przykładowo, średnia arytmetyczna z próby \bar{X}_n jest estymatorem nieobciążonym wartości oczekiwanej. Podobnie, wariancja z próby dana wzorem $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$ jest estymatorem nieobciążonym wariancji rozkładu. Natomiast inny estymator wariancji: $S_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2$ jest asymptotycznie nieobciążony.

Warto w tym miejscu wspomnieć, że

Twierdzenie 56 Jeżeli estymator jest zgodny, to jest asymptotycznie nieobciążony.

Twierdzenie przeciwe nie jest prawdziwe, ale zachodzi

Twierdzenie 57 Jeżeli estymator $\hat{\theta}_n$ jest asymptotycznie nieobciążony oraz jeżeli jego wariancja spełnia warunek $\lim_{n \rightarrow \infty} \text{Var}(\hat{\theta}_n) = 0$, to $\hat{\theta}_n$ jest zgodny.

Dla danego parametru θ może istnieć wiele estymatorów nieobciążonych. Powstaje więc kwestia wyboru najlepszego z nich. Jeżeli więc $\hat{\theta}_n^*$ i $\hat{\theta}_n^{**}$ są dwoma estymatorami nieobciążonymi parametru θ , to powiemy, że $\hat{\theta}_n^*$ jest **estymatorem efektywniejszym**, niż $\hat{\theta}_n^{**}$, wtedy gdy

$$\text{Var}(\hat{\theta}_n^*) < \text{Var}(\hat{\theta}_n^{**}). \quad (4.6)$$

Oznacza to, że ten estymator jest efektywniejszy, którego wartości są bardziej skupione wokół θ , lub innymi słowy, który ma mniejszy rozrzut. Tak jest dla dowolnej pary estymatorów, a ogólniej

Definicja 58 Estymator nieobciążony parametru θ , który ma najmniejszą wariancję spośród wszystkich nieobciążonych estymatorów danego parametru, nazywamy **estymatorem efektywnym** (najefektywniejszym).

Pozostaje jednak pytanie jak znaleźć taki estymator i w jaki sposób zmierzyć efektywność danego estymatora. Można wykazać, że przy dość ogólnych założeniach, wariancja dowolnego nieobciążonego estymatora $\hat{\theta}_n$ spełnia następującą nierówność

$$\text{Var}(\hat{\theta}_n) \geq \frac{1}{nE\left(\frac{\partial}{\partial\theta} \ln f(x, \theta)\right)^2}, \quad (4.7)$$

gdzie f oznacza gęstość rozkładu badanej cechy w przypadku rozkładu ciągłego albo funkcję prawdopodobieństwa w przypadku rozkładu dyskretnego.

Nierówność (4.7) zwana jest **nierównością Rao-Cramera** lub **nierównością informacyjną**, a wyrażenie występujące w mianowniku prawej strony tej nierówności nosi nazwę **informacji Fishera** zawartej w próbce. To ostatnie pojęcie ma dość naturalną interpretację. Otóż od "dobrego" estymatora można oczekwać, że będzie on charakteryzował się małą wariancją, co w świetle nierówności informacyjnej równoważne jest faktowi wykorzystania dużej ilości informacji zawartej w próbie.

Nierówność Rao-Cramera i pojęcie informacji Fishera wykorzystano przy konstrukcji miary efektywności estymatora.

Definicja 59 Miarą efektywności estymatora nieobciążonego $\hat{\theta}_n$ jest liczba

$$\text{eff}(\hat{\theta}_n) = \frac{1}{\text{Var}(\hat{\theta}_n) n E\left(\frac{\partial}{\partial\theta} \ln f(x, \theta)\right)^2}, \quad (4.8)$$

zwana **efektywnością estymatora $\hat{\theta}_n$** .

Nietrudno zauważyć, że dla dowolnego estymatora nieobciążonego $\hat{\theta}_n$ zachodzi

$$0 < \hat{\theta}_n \leq 1, \quad (4.9)$$

przy czym jeżeli $\text{eff}(\hat{\theta}_n) = 1$, oznacza to, że $\hat{\theta}_n$ jest estymatorem efektywnym.

Estymator, który staje się efektywny gdy liczność próby dąży do nieskończoności, tzn. gdy zachodzi

$$\lim_{n \rightarrow \infty} \text{eff}(\hat{\theta}_n) = 1 \quad (4.10)$$

nazywamy estymatorem **asymptotycznie efektywnym**.

Można pokazać, że jeżeli interesująca nas cecha ma rozkład normalny to średnia arytmetyczna z próby \bar{X}_n jest estymatorem efektywnym wartości oczekiwanej.

4.3 Metody wyznaczania estymatorów

4.3.1 Wstęp

Znanych jest wiele metod wyznaczania estymatorów. Wśród nich najbardziej znanymi są: **metoda momentów** oraz **metoda największej wiarygodności**. Obie te metody naszkicujemy pobiennie poniżej. Warto jednak pamiętać, że obok wymienionych metod istnieją i inne, w szczególności **metoda kwantylów**, czy **metoda najmniejszych kwadratów**, wykorzystywana na przykład w analizie regresji. W literaturze przedmiotu mówi się również o **estymacji bayesowskiej**, o **estymatorach minimaksowych**, o **L-estymatorach**, **M-estymatorach**, itp. Obecnie, mając do dyspozycji wielkie moce obliczeniowe, konstruuje się nowe klasy estymatorów, wykorzystujących możliwości stwarzane przez komputer. Warto w tym miejscu wspomnieć o takich metodach jak **bootstrap**, **Jackknife**, czy **cross-validation**.

4.3.2 Metoda momentów

Przypuśćmy, że nieznany parametr θ można wyrazić za pomocą funkcji kilku momentów rozkładu badanej cechy, tzn.

$$\theta = g(EX, EX^2, \dots, EX^r) \quad (4.11)$$

(momenty rozkładu można przedstawić jako pewne funkcje parametrów rozpatrywanego rozkładu, a stąd da się wyprowadzić wzór postaci (4.11)).

Oznaczmy przez M_k k -ty moment empiryczny z próby X_1, \dots, X_n , gdzie

$$M_k = \frac{1}{n} \sum_{i=1}^n X_i^k. \quad (4.12)$$

Idea metody momentów polega na tym, że za estymator poszukiwanego parametru θ przyjmuje się wspomnianą w (4.11) funkcję, tyle że momentów empirycznych, a nie teoretycznych. Tak więc estymatorem $\hat{\theta}_n$ parametru θ , wyznaczonym metodą momentów, jest wielkość

$$\hat{\theta}_n = g(M_1, M_2, \dots, M_r). \quad (4.13)$$

(patrz: Przykład 4.1)

Podstawową zaletą metody momentów jest jej prostota. Estymatory otrzymane tą metodą są zazwyczaj zgodne. Niestety, nie posiadają one na ogół zbyt dobrych innych własności statystycznych – są zazwyczaj obciążone i nieefektywne. Dlatego też metodę tę stosuje się raczej na etapie badań wstępnych, bądź też gdy nie możemy skonstruować lepszego estymatora.

4.3.3 Metoda największej wiarogodności

W przeciwnieństwie do metody momentów, metoda największej wiarogodności prowadzi do estymatorów charakteryzujących się pożądanymi własnościami. Idea metody największej wiarogodności sprowadza się do wyboru takiej wartości $\hat{\theta}_n$, jako estymatora parametru θ , która maksymalizuje prawdopodobieństwo (lub gęstość rozkładu cechy) otrzymania takiej realizacji próby, jaką właśnie otrzymano. W celu bardziej ścisłego przedstawienia tej metody musimy najpierw zdefiniować tzw. funkcję wiarogodności.

Definicja 60 Niech x_1, \dots, x_n będzie realizacją próby X_1, \dots, X_n . Jeżeli rozkład badanej cechy jest dyskretny, wówczas funkcją wiarogodności dla realizacji próby nazywamy wyrażenie

$$L(x_1, \dots, x_n; \theta) = p(x_1; \theta) \cdot \dots \cdot p(x_n; \theta), \quad (4.14)$$

gdzie $p(x_i; \theta)$ oznacza prawdopodobieństwo przyjęcia przez zmienną losową X wartość x_i , natomiast w przypadku gdy rozkład badanej cechy jest ciągły, funkcja wiarogodności przyjmuje postać

$$L(x_1, \dots, x_n; \theta) = f(x_1; \theta) \cdot \dots \cdot f(x_n; \theta), \quad (4.15)$$

gdzie $f(x; \theta)$ oznacza gęstość rozkładu.

Definicja 61 $\hat{\theta}_n$ jest estymatorem największej wiarogodności parametru θ , jeżeli maksymalizuje on wartość funkcji wiarogodności $L(x_1, \dots, x_n; \theta)$.

Warto w tym miejscu nadmienić, że ze względów rachunkowych maksymalizuje się zwykle nie funkcję L , ale jej logarytm $\ln L$, bowiem w przeważającej liczbie przypadków ułatwia to obliczenia, a funkcja $\ln L$ osiąga maksimum w tym samym punkcie co funkcja L .

W przypadku, gdy funkcja $\ln L$ jest choćby dwukrotnie różniczkowalna względem zmiennej θ , algorytm wyznaczania estymatora metodą największej wiarogodności przedstawia się następująco:

- znaleźć funkcję wiarogodności L ;
- znaleźć $\ln L$;
- obliczyć pochodną $\frac{\partial}{\partial \theta}(\ln L)$;
- znaleźć rozwiązanie θ_0 równania $\frac{\partial}{\partial \theta}(\ln L) = 0$;
- sprawdzić, czy w θ_0 funkcja $\ln L$ osiąga maksimum, tzn. czy

$$\left. \frac{\partial^2}{\partial \theta^2} (\ln L) \right|_{\theta=\theta_0} < 0. \quad (4.16)$$

Jeżeli spełniony jest warunek (4.16), oznacza to, że w punkcie θ_0 funkcja $\ln L$, a także funkcja L osiąga maksimum, a więc $\hat{\theta}_n = \theta_0$ jest estymatorem największej wiarogodności parametru θ .

(patrz: Przykład 4.2, 4.3)

Estymatory wyznaczone metodą największej wiarogodności mają bardzo dobre własności: są zgodne, co najmniej asymptotycznie nieobciążone oraz co najmniej asymptotycznie efektywne. Ponadto wiadomo, że jeśli w danym przypadku istnieje estymator efektywny, to można go uzyskać metodą największej wiarogodności. Inną cenną właściwość estymatorów największej wiarogodności ujmuje następujące twierdzenie.

Twierdzenie 62 Jeżeli $\hat{\theta}_n$ jest estymatorem największej wiarogodności parametru θ to dowolna funkcja tego estymatora $h(\hat{\theta}_n)$ jest estymatorem największej wiarogodności funkcji parametru $h(\theta)$.

Reasumując, metoda największej wiarogodności jest zalecaną metodą wyznaczania estymatorów. Niestety zdarzają się sytuacje, w których estymator największej wiarogodności nie istnieje.

4.4 Przegląd estymatorów

4.4.1 Uwagi wstępne

W niniejszym podrozdziale zamieszczono syntetyczny przegląd najczęściej wykorzystywanych w praktyce estymatorów. Oprócz podania stosownych wzorów wymieniono ich podstawowe własności statystyczne.

Czytelnik z pewnością zauważy, że podane poniżej estymatory znane są mu już z rozdziału poświęconego statystyce opisowej. Nie jest to bynajmniej przypadek. Co więcej, można rzec iż niniejszy podrozdział zawiera

matematyczne motywacje do stosowania tych właśnie statystyk we wstępnej analizie danych.

4.4.2 Estymatory wartości oczekiwanej

Podstawowym estymatorem wartości oczekiwanej jest średnia arytmetyczna z próby

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i. \quad (4.17)$$

Jest to estymator zgodny i nieobciążony, a ponadto, jeżeli badana cecha ma rozkład normalny, jest również estymatorem efektywnym.

Wartość oczekiwana można również estymować za pomocą mediany z próby. Estymator ten jest zgodny i asymptotycznie nieobciążony.

Jak już to wspomniano w rozdziale 3, minkiemem średniej jest duża wrażliwość na obecność obserwacji odstających (outlierów) w próbie. Zaletą mediany jest jej odporność na występowanie obserwacji odstających. W tym miejscu pojawia się naturalne pytanie: czy można skonstruować estymator wartości oczekiwanej, który łączyłby w sobie zalety obu wspomnianych powyżej estymatorów? Przykładem estymatora, który jest zbliżony do średniej, a jednocześnie pozwala wyeliminować (lub chociaż zmniejszyć) wpływ obserwacji odstających na wynik szacowania jest tzw. **średnia ucięta** (ang. α -trimmed mean)

$$U_n = \frac{1}{n - 2[n\alpha]} \sum_{k=[n\alpha]+1}^{n-[n\alpha]} X_{k:n}, \quad (4.18)$$

gdzie α ($0 < \alpha < \frac{1}{2}$) jest parametrem mówiącym o stopniu "ucięcia" próby, $X_{k:n}$ jest k -tą statystyką pozycyjną w uporządkowanej próbie X_1, \dots, X_n (tzn. $X_{1:n} \leq X_{2:n} \leq \dots \leq X_{n:n}$), natomiast $[n\alpha]$ oznacza część całkowitą liczby $n\alpha$. Zatem średnia ucięta, to nic innego jak średnia arytmetyczna policzona dla środkowych $(1 - 2\alpha)100\%$ obserwacji, a więc średnia arytmetyczna z próbki powstałej po odrzuceniu frakcji α obserwacji najmniejszych i frakcji α obserwacji największych.

Innym estymatorem, podobnym w budowie do średniej uciętej jest następujący estymator (ang. α -windsorized mean)

$$V_n = \frac{1}{n} \left([n\alpha]X_{[n\alpha]+1:n} + \sum_{k=[n\alpha]+1}^{n-[n\alpha]} X_{k:n} + [n\alpha]X_{n-[n\alpha]:n} \right). \quad (4.19)$$

Różnica pomiędzy dwoma powyższymi estymatorami polega na tym, że w (4.19), w przeciwieństwie do (4.18), frakcja α najmniejszych oraz największych obserwacji nie jest pomijana lecz zastępowana obserwacjami

równymi percentylami z próby rzędu, odpowiednio, α i $1 - \alpha$. Dzięki temu zabiegowi eliminuje się potencjalne obserwacje odstające, a średnia jest nadal liczona na podstawie n obserwacji.

4.4.3 Estymatory wariancji

W przypadku, gdy znana jest wartość oczekiwana μ rozkładu badanej cechy zalecanym estymatorem wariancji jest

$$\tilde{S}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2. \quad (4.20)$$

Estymator ten jest zgodny i nieobciążony, a jeżeli badana cecha ma rozkład normalny, jest również estymatorem efektywnym.

Gdy wartość oczekiwana μ rozkładu badanej cechy nie jest znana, to do szacowania wariancji używa się jednego z poniższych estymatorów

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2 \quad (4.21)$$

lub

$$S_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2. \quad (4.22)$$

Estymator (4.21) jest zgodny, nieobciążony i w przypadku rozkładu normalnego cechy asymptotycznie efektywny. Z kolei estymator (4.22) jest zgodny i asymptotycznie nieobciążony. Co ciekawe, jeżeli badana cecha ma rozkład normalny to estymatorem wariancji uzyskanym metodą największej wiarygodności jest właśnie (4.22).

4.4.4 Estymatory odchylenia standardowego

Odchylenie standardowe można estymować za pomocą statystyk \tilde{S} , S oraz S_n , otrzymanych po obliczeniu pierwiastka, odpowiednio, z wyrażeń (4.20), (4.21) i (4.22). Estymatory te są zgodne, ale nie mają innych dobrych właściwości statystycznych. Stąd też w praktyce stosuje się estymatory będące funkcjami wymienionych statystyk, zawierającymi pewne poprawki (np. likwidujące obciążenie). Przykładowo, jeżeli badana cecha ma rozkład normalny, wówczas zgodnymi, nieobciążonymi i asymptotycznie efektywnymi estymatorami odchylenia standardowego są

$$\frac{\Gamma(\frac{n-1}{n})}{\Gamma(n)} \sqrt{\frac{n-1}{2}} S \quad (4.23)$$

lub

$$\frac{\Gamma(\frac{n-1}{n})}{\Gamma(n)} \sqrt{\frac{n}{2}} S_n, \quad (4.24)$$

gdzie $\Gamma(x)$ oznacza wartość funkcji gamma w punkcie x .

Ponadto, odchylenie standardowe można szacować wykorzystując rozstęp. Mianowicie, jeżeli badana cecha ma rozkład normalny, to dla prób o malej liczności zgodnym, nieobciążonym i asymptotycznie efektywnym estymatorem odchylenia standardowego jest

$$Rd_n, \quad (4.25)$$

gdzie d_n jest współczynnikiem zależnym od liczności próby. Wartości współczynników można znaleźć np. w pracy [11].

4.4.5 Estymator wskaźnika struktury

Naturalnym estymatorem wskaźnika struktury (prawdopodobieństwa sukcesu w rozkładzie dwupunktowym) jest

$$\hat{p} = \frac{k}{n}, \quad (4.26)$$

gdzie k oznacza liczbę elementów wyróżnionych w próbie (liczbę sukcesów) o liczności n . Estymator ten jest zgodny, nieobciążony i efektywny.

4.5 Przedziały ufności

4.5.1 Pojęcie przedziału ufności

Metody estymacji omawiane powyżej pozwalają uzyskiwać punktowe oszacowania nieznanego parametru rozkładu badanej cechy. Niestety nie dają one odpowiedzi na pytanie o dokładność uzyskanej oceny. Istnieje jednakże i inne podejście do zagadnienia estymacji – estymacja przedziałowa. W podejściu tym zyskujemy informację o dokładności estymacji kosztem rezygnacji z oceny punktowej, na rzecz oceny przedziałowej. Estymacja przedziałowa sprowadza się zatem do wyznaczenia pewnego przedziału, zwanego przedziałem ufności, który z określonym z góry prawdopodobieństwem pokrywa nieznaną wartość szacowanego parametru θ .

Definicja 63 Przedział losowy ($\underline{\theta}, \bar{\theta}$), którego końcami są statystyki $\underline{\theta} = \underline{\theta}(X_1, \dots, X_n)$ oraz $\bar{\theta} = \bar{\theta}(X_1, \dots, X_n)$, gdzie $\underline{\theta} < \bar{\theta}$, nazywamy przedziałem ufności dla parametru θ na poziomie ufności $1 - \alpha$ ($0 < \alpha < 1$), jeżeli

$$P\{\underline{\theta}(X_1, \dots, X_n) < \theta < \bar{\theta}(X_1, \dots, X_n)\} \geq 1 - \alpha. \quad (4.27)$$

Dla danego poziomu ufności można znaleźć wiele przedziałów ufności spełniających relację (4.27). W praktyce interesować nas będą przedziały ufności o jak najmniejszej długości, bowiem owa długość przedziału

$$l_n = \bar{\theta}(X_1, \dots, X_n) - \underline{\theta}(X_1, \dots, X_n) \quad (4.28)$$

jest miarą precyzji estymacji.

Byłoby naturalne chcieć estymować nieznany parametr przy możliwie wysokim współczynnikiem ufności. Niestety nie jest to zbyt korzystne, bowiem wraz ze wzrostem poziomu ufności rośnie też długość przedziału ufności, a więc zmniejsza się precyzja estymacji.

W kolejnym podrozdziale podano wzory na najczęściej stosowane w praktyce przedziały ufności

4.5.2 Przedziały ufności dla wartości oczekiwanej

Model 1

Niech X_1, X_2, \dots, X_n będzie próbą prostą z populacji o rozkładzie normalnym $N(\mu, \sigma^2)$ o znanej wariancji σ^2 . Wtedy dla ustalonego poziomu ufności $1 - \alpha$ najkrótszy przedział ufności dla wartości oczekiwanej ma postać

$$\left(\bar{X} - u_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{X} + u_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right), \quad (4.29)$$

gdzie $u_{1-\frac{\alpha}{2}}$ oznacza kwantyl rozkładu normalnego standardowego rzędu $1 - \frac{\alpha}{2}$.

Model 2

Załóżmy, jak poprzednio, że badana cecha ma rozkład normalny $N(\mu, \sigma^2)$, tyle, że tym razem wariancja σ^2 jest nieznana. Można wykazać, że statystyka

$$t = \frac{\bar{X} - \mu}{S} \sqrt{n}, \quad (4.30)$$

ma rozkład t-Studenta o $n-1$ stopniach swobody. Korzystając z tego otrzymujemy przedział ufności dla wartości oczekiwanej μ na poziomie ufności $1 - \alpha$

$$\left(\bar{X} - t_{1-\frac{\alpha}{2}}^{[n-1]} \frac{S}{\sqrt{n}}, \bar{X} + t_{1-\frac{\alpha}{2}}^{[n-1]} \frac{S}{\sqrt{n}} \right), \quad (4.31)$$

gdzie $t_{1-\frac{\alpha}{2}}^{[n-1]}$ jest kwantylem rzędu $1 - \frac{\alpha}{2}$ rozkładu t-Studenta o $n-1$ stopniach swobody.

(patrz: **Przykład 4.4**)

Model 3

Załóżmy, że cecha X rozkładu populacji ma rozkład dowolny o nieznanej, ale skończonej wartości oczekiwanej i wariancji. Założymy ponadto, że dysponujemy dużą próbą (tzn. o liczności $n \geq 100$). Wówczas korzystając z twierdzenia Lindeberga-Levy'ego można pokazać, że statystyka

$$U = \frac{\bar{X} - \mu}{S} \sqrt{n}$$

ma rozkład asymptotyczny $N(0, 1)$. Stąd też, ze względu na dużą próbę, otrzymujemy następujący wzór na przedział ufności dla wartości oczekiwanej

$$\left(\bar{X} - u_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}}, \bar{X} + u_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}} \right). \quad (4.32)$$

(patrz: **Przykład 4.5**)

4.5.3 Przedziały ufności dla wariancji i odchylenia standardowego

Model 1

Załóżmy, że cecha X ma rozkład normalny $N(\mu, \sigma)$ o nieznanych parametrach μ i σ . Założymy, że dysponujemy próbą o liczności $n \leq 50$. Dowodzi się, że statystyka

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} \quad (4.33)$$

ma rozkład chi-kwadrat o $n-1$ stopniach swobody. Stąd przedział ufności dla wariancji σ^2 dany jest wzorem

$$\left(\frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2}, n-1}}, \frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2}, n-1}} \right), \quad (4.34)$$

gdzie $\chi^2_{1-\frac{\alpha}{2}, n-1}$ i $\chi^2_{\frac{\alpha}{2}, n-1}$ są kwantylami rzędu, odpowiednio, $1 - \frac{\alpha}{2}$ i $\frac{\alpha}{2}$ rozkładu chi-kwadrat o $n-1$ stopniach swobody.

Przy tych samych założeniach przedział ufności dla odchylenia standar-dowego σ ma postać

$$\left(\sqrt{\frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2}, n-1}}}, \sqrt{\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2}, n-1}}} \right). \quad (4.35)$$

Model 2

Załóżmy teraz, że cecha X ma rozkład normalny $N(\mu, \sigma)$ o nieznanych parametrach μ i σ , a próba ma liczbność $n \geq 50$. Korzystając z faktu, że statystyka $\frac{S}{\sigma} \sqrt{2(n-1)}$ ma asymptotycznie rozkład normalny $N(1 - \frac{2}{9(n-1)}, \frac{1}{3} \sqrt{\frac{2}{n-1}})$ otrzymujemy następujący przedział ufności dla wariancji σ^2

$$\left(\frac{2nS^2}{(\sqrt{2n-3} + u_{1-\frac{\alpha}{2}})^2}, \frac{2nS^2}{(\sqrt{2n-3} - u_{1-\frac{\alpha}{2}})^2} \right), \quad (4.36)$$

oraz dla odchylenia standardowego

$$\left(\frac{S\sqrt{2n}}{\sqrt{2n-3} + u_{1-\frac{\alpha}{2}}}, \frac{S\sqrt{2n}}{\sqrt{2n-3} - u_{1-\frac{\alpha}{2}}} \right). \quad (4.37)$$

(patrz: **Przykład 4.5**)

4.5.4 Przedział ufności dla wskaźnika struktury

Załóżmy, że badana cecha ma rozkład dwupunktowy z nieznanym parametrem p , a liczność próby jest duża ($n \geq 100$). Z twierdzenia Moivre'a - Laplace'a wynika, że statystyka $\frac{k}{n}$, gdzie k oznacza liczbę elementów wyróżnionych w próbie (liczbę zaobserwowanych sukcesów), ma w przybliżeniu rozkład $N(p, \sqrt{\frac{p(1-p)}{n}})$. Stąd przedział ufności dla wskaźnika struktury p przyjmuje postać

$$\left(\frac{k}{n} - u_{1-\frac{\alpha}{2}} \sqrt{\frac{\frac{k}{n}(1-\frac{k}{n})}{n}}, \frac{k}{n} + u_{1-\frac{\alpha}{2}} \sqrt{\frac{\frac{k}{n}(1-\frac{k}{n})}{n}} \right). \quad (4.38)$$

(patrz: **Przykład 4.7**)

4.5.5 Przedziały ufności dla średniego czasu zdatności

Jednym z podstawowych parametrów niezawodnościowych obiektu jest wartość oczekiwana czasu zdatności (czasu bezawaryjnej pracy) tego obiektu. Założymy, że czas zdatności T obiektu nienaprawialnego ma rozkład wykładniczy o średniej θ . W praktyce parametr θ szacuje się na podstawie różnych planów badania. Poniżej przedstawimy przedziały ufności dla czasu zdatności obiektu dla trzech najczęściej stosowanych planów badania.

Model 1

W planie (n, B, n) badaniu podlega n obiektów. Obiekty uszkodzone w czasie badań nie są wymieniane na nowe ani nie są naprawiane. Badanie

kończy się z chwilą uszkodzenia ostatniego n -tego obiektu. Jest to plan badań zapewniający uzyskanie n -elementowej próby prostej, umożliwiający bezpośrednio stosowanie klasycznych metod statystyki matematycznej. Jednak plan ten ma poważną wadę, ponieważ czas badania może być bardzo długi. Czas badania jest bowiem zmienną losową $T_B = \max\{T_1, T_2, \dots, T_n\}$, gdzie T_1, T_2, \dots, T_n oznaczają długości odcinków czasu zdatności kolejnych elementów poddanych badaniu. W przypadku stosowania tego planu otrzymujemy następujący przedział ufności dla średniego czasu zdatności θ na poziomie ufności $1 - \alpha$

$$\left(\frac{2n\theta_1^*}{\chi_{1-\frac{\alpha}{2}, 2n}^2}, \frac{2n\theta_1^*}{\chi_{\frac{\alpha}{2}, 2n}^2} \right), \quad (4.39)$$

gdzie

$$\theta_1^* = \frac{1}{n} \sum_{i=1}^n T_i \quad (4.40)$$

jest nieobciążonym estymatorem parametru θ .

(patrz: **Przykład 4.9**)

Model 2

Plan (n, B, r) - różni się od planu (n, B, n) tym, że badanie kończy się z chwilą uszkodzenia się r -tego obiektu ($r < n$). Taka reguła końca badań skraca czas badania, lecz zmienia rozkład z próby. W tym przypadku przedział ufności dla średniego czasu zdatności ma postać

$$\left(\frac{2r\theta_2^*}{\chi_{1-\frac{\alpha}{2}, 2r}^2}, \frac{2r\theta_2^*}{\chi_{\frac{\alpha}{2}, 2r}^2} \right), \quad (4.41)$$

gdzie

$$\theta_2^* = \frac{1}{r} \left[\sum_{i=1}^r T_i + (n-r)T_r \right], \quad (4.42)$$

natomiast T_r jest chwilą wystąpienia ostatniego, r -tego, uszkodzenia.

(patrz: **Przykład 4.10**)

Model 3

Plan (n, B, t) - różni się od planu (n, B, r) regułą końca badania, która stanowi limitowany czas badania. Można w tym miejscu zwrócić uwagę na fakt, że plany badań z limitowanym czasem badania mogą prowadzić do uzyskania próbek o bardzo małej liczności. W tym przypadku przedział ufności dla średniego czasu zdatności przyjmuje postać

$$\left(\frac{2m\theta_3^*}{\chi_{1-\frac{\alpha}{2}, 2(m+1)}^2}, \frac{2m\theta_3^*}{\chi_{\frac{\alpha}{2}, 2m}^2} \right), \quad (4.43)$$

gdzie

$$\theta_3^* = \frac{1}{m} \left[\sum_{i=1}^m T_i + (n-m)t \right], \quad (4.44)$$

natomiast m jest liczbą zarejestrowanych uszkodzeń do chwili t , kończącej badanie (wzór ma sens jedynie gdy $m > 0$).

(patrz: **Przykład 4.11**)

4.5.6 Estymacja przedziałowa o zadanej precyzyji

Jak to już wspomniano w podrozdz. 4.4, przy ustalonej wielkości próby zwiększenie poziomu ufności pociąga za sobą powiększenie długości przedziału ufności, a co za tym idzie, zmniejszenie dokładności szacowania nieznanego parametru. Dlatego też najbardziej uzasadnionym sposobem postępowania przy wyznaczaniu przedziałów ufności jest ustalenie z góry poziomu ufności $1 - \alpha$ oraz pożądanej precyzyji estymacji, tzn. długości l przedziału ufności, a następnie wyznaczenie takiej liczności próby n , przy której przyjęte postulaty odnośnie warunków estymacji będą spełnione. Postępowanie takie nie zawsze jest łatwe lub możliwe do przeprowadzenia, bowiem wyznaczenie n jako funkcji $1 - \alpha$ i l wymaga często dodatkowo znajomości niektórych innych parametrów rozkładu badanej cechy. Poniżej przedstawimy kilka modeli, w których zadanie estymacji o zadanej precyjacji można przeprowadzić w zadowalający sposób.

Model 1

Załóżmy, że badana cecha ma rozkład normalny $N(\mu, \sigma)$ o znanej wariancji σ^2 . Dla zadanego poziomu ufności $1 - \alpha$ długość najkrótszego przedziału ufności (por. (4.29)) jest równa $2u_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$. A zatem, skoro długość przedziału ufności ma być nie większa niż $l = 2d$, więc wymagana liczność próby musi spełniać nierówność

$$n \geq \left(u_{1-\frac{\alpha}{2}} \frac{\sigma}{d} \right)^2. \quad (4.45)$$

Ponieważ prawa strona powyższej nierówności nie jest zazwyczaj liczbą całkowitą, minimalna liczność próby równa jest najmniejszej liczbie naturalnej większej równej od $(u_{1-\frac{\alpha}{2}} \frac{\sigma}{d})^2$.

(patrz: **Przykład 4.6**)

Model 2

W przypadku, gdy badana cecha ma rozkład normalny $N(\mu, \sigma)$ o nieznanej wariancji σ^2 , aby otrzymać przedział ufności o zadanej długości $l = 2d$, należy zastosować tzw. dwuetapową procedurę Steina.

W pierwszym etapie z populacji pobiera się próbę wstępna o liczności n_0 (małej), dla której obliczamy średnią \bar{X}_0 oraz wariancję z próby S_0^2 , gdzie

$$\bar{X}_0 = \frac{1}{n_0} \sum_{i=1}^{n_0} X_i, \quad S_0^2 = \frac{1}{n_0 - 1} \sum_{i=1}^{n_0} (X_i - \bar{X}_0)^2. \quad (4.46)$$

Następnie, na mocy wzoru (4.31), wyliczamy wartość

$$k = \left(t_{1-\frac{\alpha}{2}}^{[n_0-1]} \frac{S_0}{d} \right)^2. \quad (4.47)$$

Jeżeli $k \leq n_0$, to liczebność n_0 próby wstępnej jest wystarczająca. Natomiast jeżeli $k > n_0$, to do próbki wstępnej należy jeszcze doliczyć próbę o liczności n_1 równej najmniejszej liczbie naturalnej większej od $k - n_0$, co oznacza, że do wyznaczenia przedziału ufności o zadanej precyzyji wymagana jest próbka o liczności $n_0 + n_1$.
(patrz: **Przykład 4.4**)

Model 3

Aby oszacować wartość wskaźnika struktury p zadaną precyją $l = 2d$, wymaganą liczbę próbki wyznacza się ze wzoru (4.38). Po odpowiednich przekształceniach otrzymujemy

$$n \geq u_{1-\frac{\alpha}{2}}^2 \frac{\hat{p}(1-\hat{p})}{d^2}. \quad (4.48)$$

Jeżeli znany jest rzad wielkości parametru p , to po podstawieniu odpowiedniej wartości w miejsce \hat{p} we wzorze (4.48), otrzymujemy natychmiast wymaganą minimalną liczbę próbki.

(patrz: **Przykład 4.7, 4.8**)

Jeżeli natomiast nie mamy żadnych informacji o rzędzie wielkości parametru p , to we wzorze (4.48) podstawiamy $\hat{p} = \frac{1}{2}$, czyli wartość, dla której iloczyn $\hat{p}(1-\hat{p})$ jest największy. Wówczas wymaganą liczbę próbki spełniać musi nierówność

$$n \geq u_{1-\frac{\alpha}{2}}^2 \frac{1}{4d^2}. \quad (4.49)$$

(patrz: **Przykład 4.8**)

W obu przypadkach jako minimalną liczbę próbki potrzebną do konstrukcji przedziału ufności o zadanej długości, należy oczywiście przyjąć najmniejszą liczbę naturalną spełniającą, odpowiednio (4.48) bądź (4.49).

4.6 Estymacja nieparametryczna

4.6.1 Uwagi wstępne

Jak to już wspomniano we wprowadzeniu do niniejszego rozdziału, czasem szacowanie wybranych parametrów rozkładu badanej cechy, co jest domeną estymacji parametrycznej, okazuje się niewystarczające. Niekiedy bowiem ważna jest znajomość postaci samego rozkładu. Przykładowo, przedmiotem estymacji może być gęstość rozkładu prawdopodobieństwa, dystrybuanta czy funkcja niezawodności (funkcja przeżycia). Są to typowe zadania estymacji nieparametrycznej.

4.6.2 Estymacja gęstości rozkładu

Mówiąc o estymacji parametrycznej rozważaliśmy różne kryteria jakości estymatorów. W przypadku estymacji gęstości powszechnie stosowanym kryterium jest scałkowany błąd średniokwadratowy określony wzorem

$$R(\hat{f}_n; f) = E \left[\int_{-\infty}^{+\infty} (\hat{f}_n(x) - f(x))^2 dx \right], \quad (4.50)$$

gdzie f oznacza estymowaną gęstość, natomiast \hat{f}_n - badany estymator. Jak widać ze wzoru (4.50), $R(\hat{f}_n; f)$ jest miarą przeciętnego globalnego dopasowania estymatora do estymowanej funkcji.

Najprostszym estymatorem gęstości rozkładu jest **histogram**, znany już z rozdziału 3 poświęconego statystyce opisowej. Jeśli X_1, \dots, X_n jest próbą losową, to estymator ów możemy zapisać następującym wzorem

$$\hat{f}_n(x) = \frac{\text{liczba tych } X_i, \text{ które należą do tej samej klasy co } x}{nh}, \quad (4.51)$$

gdzie h oznacza szerokość klasy. Właściwy wybór szerokości klasy ma zasadnicze znaczenie dla właściwości rozważanego estymatora. W praktyce dość dobre rezultaty dla szerokiej klasy funkcji f uzyskuje się dla

$$h \approx \frac{3.486 S_n}{\sqrt[3]{n}}, \quad (4.52)$$

gdzie S_n jest estymatorem odchylenia standardowego (por. wzór (4.22)).

Zasadniczym mankamentem histogramu, jako estymatora gęstości, jest to, iż jest on zawsze funkcja nieciągła (bez względu na to, czy estymowana funkcja jest ciągła, czy nie). Niedogodność tę można wyeliminować stosując tzw. **estymatory jądrowe** postaci

$$\hat{f}_n(x) = \frac{1}{nh} \sum_{i=1}^n K \left(\frac{x - X_i}{h} \right), \quad (4.53)$$

gdzie h ($h > 0$) jest zadaną liczbą, zwaną szerokością pasma (okna), natomiast K jest pewną funkcją spełniającą warunek

$$\int_{-\infty}^{+\infty} K(x)dx = 1, \quad (4.54)$$

zwaną jądrem.

W zastosowaniach jako jądro przyjmuje się często gęstość standardowego rozkładu normalnego (mówimy wtedy o tzw. jądrze gaussowskim). W tym przypadku dla szerokiej klasy funkcji f optymalna szerokość pasma wynosi

$$h = \frac{1.06 S_n}{\sqrt[5]{n}}. \quad (4.55)$$

Innym, często przyjmowanym jądrem jest tzw. jądro Epanecznikowa dane wzorem

$$K_e(t) = \begin{cases} \frac{3}{4\sqrt{5}} & \text{dla } -\sqrt{5} \leq t \leq \sqrt{5} \\ 0 & \text{w przeciwnym razie.} \end{cases} \quad (4.56)$$

Wówczas optymalna szerokość pasma wynosi

$$h = \frac{1.05 S_n}{\sqrt[5]{n}}. \quad (4.57)$$

4.6.3 Estymacja dystrybuanty

Naturalnym estymatorem nieznanej dystrybuanty F jest tzw. dystrybuanta empiryczna \hat{F}_n , wspominana już w rozdziale 3, dana wzorem

$$\hat{F}_n(x) = \frac{\#\{X_i : X_i \leq x\}}{n}. \quad (4.58)$$

Na mocy twierdzenia Gliwenki-Cantellego (por. np. [18]) jest to estymator zgodny i nieobciążony.

4.7 Przykłady

Przykład 4.1

Niech X_1, \dots, X_n będzie próbą losową z rozkładu gamma $\Gamma(\alpha, \beta)$, gdzie $\alpha, \beta > 0$. Stosując metodę momentów wyznaczyć estymatory parametrów α, β .

Rozwiążanie

Przypomnijmy, że rozkład gamma $\Gamma(\alpha, \beta)$ to rozkład ciągły o gęstości:

$$f(x) = \begin{cases} 0 & \text{dla } x \leq 0, \\ \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} & \text{dla } x > 0. \end{cases}$$

Mamy znaleźć estymatory dwóch parametrów: α i β , więc potrzebne nam będą dwa pierwsze momenty zwykłe: EX i EX^2 . Ze wzorów (2.67) wiadomo, że dla rozkładu gamma, $EX = \frac{\alpha}{\beta}$, $VarX = \frac{\alpha}{\beta^2}$. Ponieważ $VarX = EX^2 - (EX)^2$, stąd

$$EX^2 = VarX + (EX)^2 = \frac{\alpha}{\beta^2} + \left(\frac{\alpha}{\beta}\right)^2 = \frac{\alpha(1+\alpha)}{\beta^2}.$$

Z układu równań:

$$\begin{cases} EX = \frac{\alpha}{\beta}, \\ EX^2 = \frac{\alpha(1+\alpha)}{\beta^2} \end{cases}$$

wyznaczamy α i β :

$$\begin{aligned} \alpha &= \frac{(EX)^2}{EX^2 - (EX)^2}, \\ \beta &= \frac{EX}{EX^2 - (EX)^2}. \end{aligned}$$

Aby otrzymać estymatory α i β wystarczy do powyższych wzorów w miejsce momentów teoretycznych EX i EX^2 , wstawić momenty empiryczne M_1 i M_2 , gdzie:

$$\begin{aligned} M_1 &= \frac{1}{n} \sum_{i=1}^n X_i = \bar{X}, \\ M_2 &= \frac{1}{n} \sum_{i=1}^n X_i^2. \end{aligned}$$

Stąd:

$$\begin{aligned} \hat{\alpha} &= \frac{(M_1)^2}{M_2 - (M_1)^2} = \frac{(\bar{X})^2}{\frac{1}{n} \sum_{i=1}^n X_i^2 - (\bar{X})^2}, \\ \hat{\beta} &= \frac{M_1}{M_2 - (M_1)^2} = \frac{\bar{X}}{\frac{1}{n} \sum_{i=1}^n X_i^2 - (\bar{X})^2}. \end{aligned}$$

Zauważmy, że mianowniki można zapisać w prostszej postaci:
 $\frac{1}{n} \sum_{i=1}^n X_i^2 - (\bar{X})^2 = S_n^2$. Stąd:

$$\hat{\alpha} = \frac{(\bar{X})^2}{S_n^2},$$

$$\hat{\beta} = \frac{\bar{X}}{S_n^2}.$$

Przykład 4.2

Niech X_1, \dots, X_n będzie próbą losową z rozkładu Poissona $P(\lambda)$, gdzie $\lambda > 0$. Stosując metodę największej wiarogodności wyznaczyć estymator parametru λ .

Rozwiążanie

Wyznaczamy najpierw funkcję wiarogodności

$$L(x_1, \dots, x_n; \lambda) = p(x_1; \lambda) \dots p(x_n; \lambda),$$

gdzie $p(x_i; \lambda)$ oznacza prawdopodobieństwo przyjęcia przez zmienną losową X wartości x_i . Ponieważ w naszym zadaniu X ma rozkład Poissona $P(\lambda)$, stąd

$$p(x_i; \lambda) = P(X = x_i) = e^{-\lambda} \frac{\lambda^{x_i}}{x_i!} \quad \text{dla } x_i = 0, 1, 2, \dots$$

Zatem

$$\begin{aligned} L(x_1, \dots, x_n; \lambda) &= e^{-\lambda} \frac{\lambda^{x_1}}{x_1!} \dots e^{-\lambda} \frac{\lambda^{x_n}}{x_n!} = \\ &= \prod_{i=1}^n e^{-\lambda} \frac{\lambda^{x_i}}{x_i!} = e^{-n\lambda} \frac{\lambda^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!}. \end{aligned}$$

Znajdujemy teraz logarytm funkcji wiarogodności:

$$\ln L(x_1, \dots, x_n; \lambda) = \ln \left(e^{-n\lambda} \frac{\lambda^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!} \right) = -n\lambda + \sum_{i=1}^n x_i \ln \lambda - \ln \left(\prod_{i=1}^n x_i! \right),$$

obliczamy pochodną

$$\frac{\partial}{\partial \lambda} (\ln L(x_1, \dots, x_n; \lambda)) = -n + \sum_{i=1}^n x_i \frac{1}{\lambda},$$

po czym szukamy λ będącego rozwiązaniem równania

$$\frac{\partial}{\partial \lambda} (\ln L(x_1, \dots, x_n; \lambda)) = 0.$$

Mamy

$$-n + \sum_{i=1}^n x_i \frac{1}{\lambda} = 0$$

czyli

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

Sprawdźmy, czy dla tej wartości λ funkcja $\ln L(x_1, \dots, x_n; \lambda)$ osiąga maksimum. W tym celu policzmy drugą pochodną:

$$\frac{\partial^2}{\partial \lambda^2} (\ln L(x_1, \dots, x_n; \lambda)) = - \sum_{i=1}^n x_i \frac{1}{\lambda^2} < 0.$$

Dla wyliczonej przez nas λ pochodna ta jest ujemna, więc funkcja $\ln L$ osiąga w tym punkcie maksimum. Zatem

$$\hat{\lambda} = \bar{x}$$

jest estymatorem największej wiarogodności parametru λ .

Przykład 4.3

Niech X_1, \dots, X_n będzie próbą losową z rozkładu wykładniczego $Exp(\lambda)$, gdzie $\lambda > 0$. Stosując metodę największej wiarogodności wyznaczyć estymator parametru λ .

Rozwiążanie

Przypomnijmy, że rozkład wykładniczy ma następującą gęstość:

$$f(x; \lambda) = \begin{cases} \lambda e^{-\lambda} & \text{dla } x > 0 \\ 0 & \text{dla } x \leq 0. \end{cases}$$

Bez straty ogólności będziemy rozpatrywać wyłącznie $x > 0$. Zatem funkcja wiarogodności L przyjmuje postać:

$$\begin{aligned} L(x_1, \dots, x_n; \lambda) &= f(x_1; \lambda) \dots f(x_n; \lambda) = \prod_{i=1}^n \lambda e^{-\lambda x_i} \\ &= \lambda^n e^{-\lambda \sum_{i=1}^n x_i}. \end{aligned}$$

Stąd

$$\ln L(x_1, \dots, x_n; \lambda) = n \ln \lambda - \lambda \sum_{i=1}^n x_i.$$

Obliczamy teraz pochodną funkcji $\ln L$

$$\frac{\partial}{\partial \lambda} (\ln L(x_1, \dots, x_n; \lambda)) = \frac{n}{\lambda} - \sum_{i=1}^n x_i$$

i wyznaczamy λ będącą rozwiązaniem równania

$$\frac{\partial}{\partial \lambda} (\ln L(x_1, \dots, x_n; \lambda)) = 0$$

tzn.

$$\frac{n}{\lambda} - \sum_{i=1}^n x_i = 0.$$

Otrzymujemy

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

Pozostaje sprawdzić, czy funkcja $\ln L$ przyjmuje maksimum dla wyznaczonej przez nas λ . W tym celu sprawdzany znak drugiej pochodnej

$$\frac{\partial^2}{\partial \lambda^2} (\ln L(x_1, \dots, x_n; \lambda)) = -\frac{n}{\lambda^2} < 0.$$

Ponieważ jest on ujemny, zatem

$$\hat{\lambda} = \frac{1}{\bar{x}}.$$

Przykład 4.4

Czas montowania bębna w pralce automatycznej jest zmienną losową o rozkładzie normalnym. Zmierzono czas montowania bębna przez 6 losowo wybranych robotników i otrzymano następujące wyniki (w minutach): 6.2, 7.1, 6.3, 6.9, 7.5, 7.0.

- Na podstawie otrzymanych danych wyznaczyć przedział ufności dla średniego czasu montażu bębna w pralce. Przyjąć poziom ufności 0.95.
- Ile pomiarów czasu montażu bębna w pralce należałoby wykonać dla oszacowania średniego czasu montażu z maksymalnym błędem wynoszącym ± 0.1 minuty, na poziomie ufności 0.95?

Rozwiązanie

a) Badany czas montowania bębna ma rozkład normalny z nieznanym odchyleniem standardowym. Zatem przedział ufności dla średniego czasu montażu bębna w pralce ma postać (4.31):

$$\left(\bar{X} - t_{1-\frac{\alpha}{2}}^{[n-1]} \frac{S}{\sqrt{n}}, \bar{X} + t_{1-\frac{\alpha}{2}}^{[n-1]} \frac{S}{\sqrt{n}} \right),$$

gdzie

$$\begin{aligned} \bar{X} &= \frac{1}{n} \sum_{i=1}^n X_i = \frac{1}{6} (6.2 + 7.1 + 6.3 + 6.9 + 7.5 + 7.0) \simeq 6.833, \\ S^2 &= \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{1}{5} [(6.2 - 6.833)^2 + (7.1 - 6.833)^2 + \\ &\quad + (6.3 - 6.833)^2 + (6.9 - 6.833)^2 + (7.5 - 6.833)^2 + \\ &\quad + (7.0 - 6.833)^2] \simeq 0.24667, \\ S &= \sqrt{S^2} = \sqrt{0.24667} \simeq 0.49666. \end{aligned}$$

Poziom ufności $1 - \alpha = 0.95$, zatem $1 - \frac{\alpha}{2} = 0.975$ i z tablic kwantyli rozkładu t-Studenta odczytujemy

$$t_{1-\frac{\alpha}{2}}^{[n-1]} = t_{0.975}^{[5]} = 2.5706.$$

Stąd szukany przedział ufności jest postaci:

$$\left(6.833 - 2.5706 \frac{0.49666}{\sqrt{6}}, 6.833 + 2.5706 \frac{0.49666}{\sqrt{6}} \right) = (6.3118, 7.3542).$$

b) W celu określenia liczby pomiarów dla oszacowania średniego czasu montażu bębna w pralce z maksymalnym błędem wynoszącym ± 0.1 minuty, skorzystamy z dwuetapowej procedury Steina. Sześć pomiarów czasu montażu bębna w pralce, podane w zadaniu, traktujemy jako próbę wstępnią ($n_0 = 6$) i stąd po podstawieniu do wzoru (4.47) otrzymamy

$$k = \left(t_{1-\frac{\alpha}{2}}^{[n_0-1]} \frac{S_0}{d} \right)^2 = \left(2.5706 \frac{0.49666}{0.1} \right)^2 = 162.99$$

A zatem w celu oszacowania czasu montażu bębna w pralce zadaną precyzją należy wykonać aż 163 pomiary (ponieważ jednak 6 już mamy, to trzeba wykonać jeszcze 157 pomiarów).

Przykład 4.5

Wysokość zarobków losowej grupy pracowników pewnego przedsiębiorstwa przedstawia się następująco:

zarobki (w tys. złotych)	liczba osób
0.6 – 1.0	22
1.0 – 1.4	115
1.4 – 1.8	48
1.8 – 2.2	15

- a) Znaleźć przedział ufności dla wysokości średniej pensji w tym przedsiębiorstwie. Przyjąć poziom ufności 0.95.
- b) Na poziomie ufności 0.95 wyznaczyć przedział ufności dla odchylenia standardowego zarobków w tym przedsiębiorstwie.

Rozwiążanie

a) Nie znamy rozkładu wysokości zarobków, ale mamy próbę o dużej liczności ($n = 200$). Aby wyznaczyć przedział ufności dla średniej korzystamy ze wzoru (4.32)

$$\left(\bar{X} - u_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}}, \bar{X} + u_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}} \right).$$

Ponieważ dysponujemy wyłącznie danymi pogrupowanymi, to średnią \bar{X} i wariancję S^2 wyliczamy z następujących wzorów:

$$\begin{aligned}\bar{X} &= \frac{1}{n} \sum_{i=1}^r n_i x_i^0, \\ S^2 &= \frac{1}{n-1} \sum_{i=1}^r n_i (x_i^0 - \bar{X})^2.\end{aligned}$$

Obliczenia najwygodniej przeprowadzić w tabeli:

klasy	n_i	x_i^0	$n_i x_i^0$	$(x_i^0 - \bar{x})^2$	$n_i (x_i^0 - \bar{x})^2$
0.6 – 1.0	22	0.8	17.6	0.262144	5.76717
1.0 – 1.4	115	1.2	138.0	0.012544	1.44256
1.4 – 1.8	48	1.6	76.8	0.082944	3.98131
1.8 – 2.2	15	2.0	30.0	0.473344	7.10016
suma	200		262.4		18.29120

Stąd

$$\bar{X} = \frac{1}{200} 262.4 = 1.312,$$

$$S^2 = \frac{1}{199} 18.2912 \simeq 0.091916,$$

$$S = \sqrt{0.091916} \simeq 0.30318.$$

Zadany poziom ufności wynosi $1 - \alpha = 0.95$, stąd $1 - \frac{\alpha}{2} = 0.975$. Z tablic kwantylów rozkładu normalnego standardowego odczytujemy:

$$u_{1-\frac{\alpha}{2}} = u_{0.975} = 1.95996.$$

Zatem szukany przedział ufności jest postaci:

$$\left(1.312 - 1.95996 \frac{0.30318}{\sqrt{200}}, 1.312 + 1.95996 \frac{0.30318}{\sqrt{200}} \right) = (1.27, 1.354).$$

b) Aby wyznaczyć przedział ufności dla odchylenia standardowego używamy wzoru (4.37)

$$\left(\frac{S\sqrt{2n}}{\sqrt{2n-3+u_{1-\frac{\alpha}{2}}}}, \frac{S\sqrt{2n}}{\sqrt{2n-3-u_{1-\frac{\alpha}{2}}}} \right).$$

Ponieważ poziom ufności jest taki sam jak w pkt. a) więc mamy:

$$\left(\frac{0.30318\sqrt{400}}{\sqrt{397}+1.95996}, \frac{0.30318\sqrt{400}}{\sqrt{397}-1.95996} \right) = (0.277, 0.3375).$$

Przykład 4.6

Ustalić, jak liczna powinna być próba, aby na jej podstawie można było oszacować wzrost noworodków, jeżeli wiadomo, że ma on rozkład normalny o odchyleniu standardowym 1.5 cm. Przyjąć, że maksymalny błąd oszacowania średniego wzrostu na poziomie ufności 0.99 ma wynosić 0.5 cm.

Rozwiążanie

Badany wzrost noworodków ma rozkład normalny ze znany odchyleniem standardowym. Korzystając ze wzoru (4.45) na minimalną liczbę próby niezbędną do konstrukcji przedziału ufności dla średniej otrzymujemy:

$$n \geq \left(u_{1-\frac{\alpha}{2}} \frac{\sigma}{d} \right)^2,$$

gdzie $\sigma = 1.5$, $d = 0.5$ i $1 - \alpha = 0.99$, czyli $\alpha = 0.01$, stąd $1 - \frac{\alpha}{2} = 0.995$. Z tablic kwantylów rozkładu normalnego odczytujemy, że $u_{1-\frac{\alpha}{2}} = u_{0.995} = 2.57582$. A zatem

$$n \geq \left(\frac{2.57582 \cdot 1.5}{0.5} \right)^2 \simeq 59.7136,$$

z czego wynika, że trzeba zmierzyć wzrost 60 noworodków.

Przykład 4.7

Fabryka zakupiła nowy agregat. Wylosowano 500 wyprodukowanych przez ten agregat detali. Okazało się, że 20 z nich nie spełnia normy jakości.

- Podać 95% przedział ufności dla wadliwości.
- Jak liczną próbę należałoby pobrać, aby móc oszacować przedziałowo wadliwość nowego agregatu z dokładnością $\pm 1\%$ na poziomie ufności 95%?

Rozwiązańie

a) Mamy tu rozkład dwupunktowy (detal jest wadliwy albo dobry). Stosujemy wzór (4.38) na przedział ufności dla wskaźnika struktury

$$\left(\frac{k}{n} - u_{1-\frac{\alpha}{2}} \sqrt{\frac{\frac{k}{n}(1-\frac{k}{n})}{n}}, \frac{k}{n} + u_{1-\frac{\alpha}{2}} \sqrt{\frac{\frac{k}{n}(1-\frac{k}{n})}{n}} \right),$$

gdzie $n = 500$, $\frac{k}{n} = \frac{20}{500} = 0.04$, a ponieważ $1 - \alpha = 0.95$, stąd $1 - \frac{\alpha}{2} = 0.975$ i odczytany z tablic rozkładu normalnego standardowego kwantyl $u_{1-\frac{\alpha}{2}} = u_{0.975} = 1.95996$. Zatem szukany przedział ufności to

$$\left(0.04 - 1.95996 \sqrt{\frac{0.04(1-0.04)}{500}}, 0.04 + 1.95996 \sqrt{\frac{0.04(1-0.04)}{500}} \right) = \\ = (0.023, 0.057).$$

b) Wyznaczmy teraz licznosć próby niezbędną do oszacowania wadliwości nowego agregatu z dokładnością $\pm 1\%$. Ponieważ wiemy, że na 500 badanych detali 20 okazało się wadliwych, to znamy rzad wielkości szacowanego wskaźnika struktury $\hat{p} = \frac{20}{500} = 0.04$ i wybieramy wzór (4.48), gdzie $d = 0.01$:

$$n \geq u_{1-\frac{\alpha}{2}}^2 \frac{\hat{p}(1-\hat{p})}{d^2} = (1.95996)^2 \frac{0.04(1-0.04)}{(0.01)^2} \simeq 1475.1$$

Zatem minimalna licznosć próby wynosi 1476.

Przykład 4.8

Ile osób należałoby wylosować niezależnie do próby, aby z maksymalnym błędem 1.5% oszacować na poziomie ufności 0.98 procent osób, które oglądają codziennie telewizję, jeśli:

- ze wstępnych badań wynika, że spodziewany rzad wielkości szacowanego procentu wynosi 65%,

- nie robiliśmy żadnych wstępnych sondaży.

Rozwiązańie

Zadanie dotyczy ponownie modelu dwupunktowego (mamy dwie możliwości: albo oglądamy codziennie telewizję, albo nie). Stosujemy więc wzór na minimalną licznosć próby niezbędną do oszacowania wartości wskaźnika struktury.

a) Jeżeli znamy rzad wielkości szacowanego wskaźnika struktury (procentu), to korzystamy ze wzoru (4.48). Ponieważ $1 - \alpha = 0.98$, stąd $1 - \frac{\alpha}{2} = 0.99$ i odczytany z tablic rozkładu normalnego standardowego kwantyl $u_{1-\frac{\alpha}{2}} = u_{0.99} = 2.32634$. Zatem mamy

$$n \geq u_{1-\frac{\alpha}{2}}^2 \frac{\hat{p}(1-\hat{p})}{d^2} = (2.32634)^2 \frac{0.65(1-0.65)}{0.015^2} \simeq 5471.99,$$

czyli należy wylosować 5472 osoby.

b) Jeśli nie znamy rzadu wielkości szacowanego wskaźnika struktury, to musimy posłużyć się wzorem (4.49). Mamy więc

$$n \geq u_{1-\frac{\alpha}{2}}^2 \frac{1}{4d^2} = (2.32634)^2 \frac{1}{4 \cdot 0.015^2} \simeq 6013.2,$$

czyli w tym przypadku należy wylosować aż 6014 osób.

Przykład 4.9

Podczas badania 20 sztuk obiektów planem (n, B, n) zaobserwowano następujące chwile uszkodzeń (w godzinach):

30, 40, 54, 62, 63, 67, 72, 75, 79, 81, 81, 82, 84, 86, 87, 89, 90, 95, 100, 110.

Wyznaczyć przedział ufności dla wartości średniej θ czasu zdatności obiektu przyjmując poziom ufności $1 - \alpha = 0.95$. Zakładamy, że czas zdatności ma rozkład wykładniczy.

Rozwiązańie

Zadanie dotyczy planu badań (n, B, n), więc do wyznaczenia przedziału ufności dla średniej wartości czasu zdatności interesujących nas obiektów skorzystamy ze wzorów (4.39) i (4.40).

$$\begin{aligned} \theta_1^* &= \frac{1}{n} \sum_{i=1}^n T_i \\ &= \frac{1}{20} (30 + 40 + 54 + 62 + 63 + 67 + 72 + 75 + 79 + 81 \\ &\quad + 81 + 82 + 84 + 86 + 87 + 89 + 90 + 95 + 100 + 110) \\ &= 76.35. \end{aligned}$$

Ponieważ $1 - \alpha = 0.95$, więc $1 - \frac{\alpha}{2} = 0.975$ i odpowiednie kwantyle rozkładu chi-kwadrat, odczytane z tablic, wynoszą

$$\begin{aligned}\chi^2_{1-\frac{\alpha}{2}, 2n} &= \chi^2_{0.975, 40} = 59.3420, \\ \chi^2_{\frac{\alpha}{2}, 2n} &= \chi^2_{0.025, 40} = 24.4330.\end{aligned}$$

Stąd szukany przedział ufności dla średniej wartości czasu zdatności θ ma postać

$$\begin{aligned}\left(\frac{2n\theta_1^*}{\chi^2_{1-\frac{\alpha}{2}, 2n}}, \frac{2n\theta_1^*}{\chi^2_{\frac{\alpha}{2}, 2n}} \right) &= \left(\frac{2 \cdot 20 \cdot 76.35}{59.3420}, \frac{2 \cdot 20 \cdot 76.35}{24.4330} \right) = \\ &= (51.464, 124.99).\end{aligned}$$

Przykład 4.10

Podczas badania 20 sztuk obiektów planem (n, B, r) rejestrano chwile uszkodzeń pierwszych 10-ciu z nich. Chwile zarejestrowane są następujące (w godzinach):

6.35, 9.62, 18.99, 19.56, 24.64, 34.21, 34.44, 74.95, 63.31, 64.07.

Zakładając, że czas zdatności ma rozkład wykładniczy, wyznaczyć przedział ufności dla wartości oczekiwanej θ czasu zdatności, przyjmując poziom ufności $1 - \alpha = 0.9$.

Rozwiązańie:

Badamy obiekty planem (n, B, r), więc do wyznaczenia przedziału ufności dla wartości oczekiwanej czasu zdatności zastosujemy wzory (4.41) i (4.42).

$$\begin{aligned}\theta_2^* &= \frac{1}{r} \left[\sum_{i=1}^r T_i + (n-r)T_r \right] \\ &= \frac{1}{10} [6.35 + 9.62 + 18.99 + 19.56 + 24.64 + 34.21 + 34.44 \\ &\quad + 74.95 + 63.31 + 64.07 + (20-10)64.07] \\ &= 99.084.\end{aligned}$$

Ponieważ $1 - \alpha = 0.9$, więc $1 - \frac{\alpha}{2} = 0.95$ i odpowiednie kwantyle rozkładu chi-kwadrat, odczytane z tablic, wynoszą

$$\begin{aligned}\chi^2_{1-\frac{\alpha}{2}, 2r} &= \chi^2_{0.95, 20} = 31.4104, \\ \chi^2_{\frac{\alpha}{2}, 2r} &= \chi^2_{0.05, 20} = 10.8508.\end{aligned}$$

Stąd szukany przedział ufności dla średniej wartości czasu zdatności θ ma postać

$$\begin{aligned}\left(\frac{2r\theta_2^*}{\chi^2_{1-\frac{\alpha}{2}, 2r}}, \frac{2r\theta_2^*}{\chi^2_{\frac{\alpha}{2}, 2r}} \right) &= \left(\frac{2 \cdot 10 \cdot 99.084}{31.4104}, \frac{2 \cdot 10 \cdot 99.084}{10.8508} \right) = \\ &= (63.09, 182.63).\end{aligned}$$

Przykład 4.11

Przy badaniu 15 sztuk obiektów planem (n, B, t) zaobserwowano, że do chwili $t = 80$ [h] uszkodziło się 8 z nich w chwilach:

7.7, 8.6, 16.56, 22.12, 31.52, 39.12, 73.09, 75.92 [h].

Wyznaczyć przedział ufności dla średniej wartości czasu zdatności θ przyjmując poziom ufności $1 - \alpha = 0.9$. Zakładamy, że czas zdatności ma rozkład wykładniczy.

Rozwiązańie:

Tym razem obiekty badaliśmy planem (n, B, t). Stąd przedział ufności dla średniej wartości czasu zdatności wyliczać będziemy według wzorów (4.43) i (4.44).

$$\begin{aligned}\theta_3^* &= \frac{1}{m} \left[\sum_{i=1}^m T_i + (n-m)t \right] = \\ &= \frac{1}{8} [7.7 + 8.6 + 16.56 + 22.12 + 31.52 + 39.12 + \\ &\quad + 73.09 + 75.92 + (15-8)80] = 104.33.\end{aligned}$$

Ponieważ $1 - \alpha = 0.9$, więc $1 - \frac{\alpha}{2} = 0.95$ i odpowiednie kwantyle rozkładu chi-kwadrat, odczytane z tablic, wynoszą

$$\begin{aligned}\chi^2_{1-\frac{\alpha}{2}, 2(m+1)} &= \chi^2_{0.95, 18} = 28.8693, \\ \chi^2_{\frac{\alpha}{2}, 2m} &= \chi^2_{0.05, 16} = 7.9616.\end{aligned}$$

Stąd szukany przedział ufności dla średniej wartości czasu zdatności θ ma postać

$$\begin{aligned}\left(\frac{2m\theta_3^*}{\chi^2_{1-\frac{\alpha}{2}, 2(m+1)}}, \frac{2m\theta_3^*}{\chi^2_{\frac{\alpha}{2}, 2m}} \right) &= \left(\frac{2 \cdot 8 \cdot 104.33}{28.8693}, \frac{2 \cdot 8 \cdot 104.33}{7.9616} \right) = \\ &= (57.822, 209.67).\end{aligned}$$

4.8 Zadania

Zadanie 4.1

Niech X_1, \dots, X_n będzie próbą losową z rozkładu dyskretnego o prawdopodobieństwach:

$$\begin{aligned} P(X_i = -1) &= P(X_i = 2) = P(X_i = 4) = \theta, \\ P(X_i = 1) &= 1 - 3\theta, \end{aligned}$$

dla $i = 1, \dots, n$ oraz dla $\theta \in (0, 1)$. Metodą momentów wyznaczyć estymator parametru θ .

Zadanie 4.2

Niech X_1, \dots, X_n będzie próbą losową z rozkładu wykładniczego o gęstości:

$$f(x, \theta) = \begin{cases} \theta e^{-\theta x} & \text{dla } x > 0 \\ 0 & \text{dla } x \leq 0 \end{cases}$$

Metodą momentów wyznaczyć estymator parametru θ .

Zadanie 4.3

Niech X_1, \dots, X_n będzie próbą losową z rozkładu jednostajnego o gęstości:

$$f(x, \theta) = \begin{cases} \frac{1}{2\theta} & \text{dla } x \in [-\theta, \theta] \\ 0 & \text{dla } x \notin [-\theta, \theta] \end{cases}$$

Metodą momentów wyznaczyć estymator parametru θ .

Zadanie 4.4

Ile przebiegów pociągów pasażerskich należałoby wylosować niezależnie do próby, aby z maksymalnym błędem 5% oszacować na poziomie ufności 0.90 nieznany procent opóźnionych pojazdów na stację docelową?

Zadanie 4.5

Ilu studentów należałoby wylosować niezależnie do próby, aby z maksymalnym błędem 2% oszacować na poziomie ufności 0.95 procent studentów, którzy byli podczas wakacji za granicą, jeśli spodziewany rząd wielkości szacowanego procentu jest 25%?

Zadanie 4.6

Telewizja podała, że pewien program cieszy się zainteresowaniem aż 75% telewidzów. Na 2200 losowo wybranych telewidzów 1386 potwierdziło zainteresowanie owym programem.

- Oszacować przedziałowo procent telewidzów zainteresowanych wspomnianym programem. Przyjąć poziom ufności 0.95.
- Ilu telewidzów należałoby wylosować do próby, aby w świetle przedstawionych wyżej danych móc oszacować oglądalność owego programu z maksymalnym błędem $\pm 1\%$ (na tym samym, co poprzednio, poziomie ufności)?

Zadanie 4.7

Na 200 połączeń telefonicznych w pewnej centrali 14 okazało się błędnych.

- Na poziomie ufności 0.95 zbudować przedział ufności dla frakcji błędnych połączeń.
- Jak liczną próbę połączeń należy zbadać, aby zbudować ów przedział ufności z maksymalnym błędem 1.5%?

Zadanie 4.8

Dział kontroli jakości w zakładach chemicznych chce oszacować średnią wagę proszku do prania sprzedawanego w pudełkach o nominalnej wadze 3 kg. Pobrano w tym celu próbki losową 7 pudełek proszku do prania. Każde pudełko zważono i otrzymano następujące wyniki (w kilogramach): 2.93, 2.97, 3.05, 2.91, 3.02, 2.87, 2.92. Wiadomo, że rozkład wagi pudełka proszku do prania jest normalny.

- Na poziomie ufności 0.99 zbudować przedział ufności dla średniej wagi pudełka proszku do prania.
- Jak liczną próbkę pudełek proszku należy pobrać aby z maksymalnym błędem 50 g wyznaczyć 99% przedział ufności dla średniej wagi pudełka proszku do prania?

Zadanie 4.9

Zmierzono czas świecenia 39 żarówek i stwierdzono, że dla 4 żarówek był on krótszy niż 1000 godzin, w przypadku 18 żarówek zawierał się w przedziale 2000 - 3000 godzin, a w przypadku pozostałych 3 żarówek czas świecenia zawierał się w przedziale 3000 - 4000 godzin. Stwierdzono również, że rozkład prawdopodobieństwa opisujący czas świecenia wspomnianych żarówek jest rozkładem normalnym. Wyznaczyć przedział ufności dla odchylenia standardowego czasu świecenia żarówek (przyjąć poziom ufności 0.95).

Zadanie 4.10

Z partii kondensatorów wybrano losowo 6 sztuk i zmierzono ich pojemność, otrzymując wyniki: 4.3, 4.4, 4.3, 4.2, 4.3, 4.4 (pF). Zakładamy, że pojemność ma rozkład normalny.

- Wyznaczyć 95% przedział ufności dla odchylenia standardowego pojemności tych kondensatorów.
- Ile kondensatorów należałoby wylosować niezależnie do próby, aby na poziomie ufności 95% otrzymać przedział ufności dla średniej pojemności, z maksymalnym błędem 0.05 (pF)?

Zadanie 4.11

Wysokość zaröbków losowej próby pracowników pewnego przedsiębiorstwa przedstawia się następująco:

zarobki (w tys. zł)	liczba osób
0.6 - 1.0	3
1.0 - 1.4	10
1.4 - 1.8	12
1.8 - 2.2	5

Zakładając, że rozkład zarobków jest normalny, znaleźć przedział ufności dla wariancji wysokości zarobków w tym przedsiębiorstwie. Przyjąć poziom ufności 0.95.

ODPOWIEDZI

Zadanie 4.1

$$\hat{\theta} = 2\bar{X}.$$

Zadanie 4.2

$$\hat{\theta} = \frac{1}{\bar{X}}.$$

Zadanie 4.3

$$\hat{\theta} = \frac{3}{n} \sum_{i=1}^n X_i^2 \quad (\text{Wskazówka: skorzystać z } EX^2).$$

Zadanie 4.4

$$n = 271.$$

Zadanie 4.5

$$n = 1801.$$

Zadanie 4.6

- $p \in (0.6098, 0.6502)$

- $n = 8955.$

Zadanie 4.7

- $p \in (0.034639, 0.10536)$

- $n = 1112.$

Zadanie 4.8

- $\mu \in (2.86, 3.04)$

- $n = 24.$

Zadanie 4.9

$$\sigma \in (641.841, 1012.171).$$

Zadanie 4.10

- $\sigma \in (0.047, 0.1848444)$

- $n = 17.$

Zadanie 4.11

$$\sigma^2 \in (0.08036, 0.22897).$$

5

Weryfikacja hipotez

5.1 Wprowadzenie

Drugim podstawowym działem statystyki, obok teorii estymacji, jest teoria weryfikacji hipotez. **Hipotezą statystyczną** nazywamy dowolne przypuszczenie dotyczące rozkładu badanej cechy. Weryfikacji takiej hipotezy dokonuje się na podstawie pobranej próby losowej. Jej zaś celem jest odpowiedź na pytanie, czy postawiona hipoteza jest prawdziwa czy też fałszywa.

Narzędzia służące do weryfikacji hipotez nazywamy **testami statystycznymi**. W zależności od tego, czy rozpatrujemy model parametryczny, czy nieparametryczny, mamy do czynienia z hipotezami parametrycznymi bądź nieparametrycznymi, a w konsekwencji z testami, odpowiednio, parametrycznymi bądź nieparametrycznymi.

W kolejnym podrozdziale omówimy podstawowe pojęcia teorii weryfikacji hipotez. Następnie przedstawimy podstawowe testy parametryczne i nieparametryczne, a w szczególności testy dotyczące wartości oczekiwanej, wariancji, wskaźnika struktury, testy zgodności oraz metody testowania hipotez dotyczących niezależności lub korelacji cech.

5.2 Pojęcia podstawowe

Niech $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ będzie przestrzenią statystyczną, gdzie \mathcal{X} oznacza przestrzeń prób (zbior możliwych wyników obserwacji), \mathcal{A} jest σ -cięciem podzbiorów zbioru \mathcal{X} , a $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ jest rodziną rozkładów praw-

dopodobieństwa na \mathcal{A} . Rozpatrujemy pewną hipotezę H dotyczącą parametru θ . Gdyby wartość parametru θ była znana, było by wiadomo również, czy owa hipoteza jest prawdziwa. A zatem rodzinę rozkładów \mathcal{P} można podzielić na dwie rozłączne podrodziny: podrodzinę $\{P_\theta : \theta \in \Theta_H\}$ zawierającą rozkłady, dla których rozważana hipoteza jest prawdziwa i podrodzinę $\{P_\theta : \theta \in \Theta_K\}$, gdzie $\Theta_H, \Theta_K \subset \Theta$ i $\Theta_H \cap \Theta_K = \emptyset$, zawierającą rozkłady, dla których rozważana hipoteza jest fałszywa. Postawiona hipoteza, zwana często **hipotezą zerową**, jest więc formalnie równoważna orzeczeniu $H : \theta \in \Theta_H$. Natomiast orzeczenie $K : \theta \in \Theta_K$ nazywamy **hipotezą alternatywną**.

Hipotezę H (K) nazywamy **prostą**, jeżeli odpowiadający jej zbiór Θ_H (Θ_K) jest jednoelementowy (wówczas hipoteza jednoznacznie określa rozkład badanej cechy); w przeciwnym razie mówimy o hipotezie **złożonej**.

Na podstawie zaobserwowanej próby losowej X_1, \dots, X_n możemy podjąć jedną z dwóch decyzji: przyjąć H (i odrzucić K), bądź odrzucić H (i przyjąć K).

Definicja 64 *Testem statystycznym nazywamy regułę decyzyjną, przypisującą możliwym realizacjom próby losowej X_1, X_2, \dots, X_n decyzję odrzucenia albo przyjęcia weryfikowanej hipotezy.*

Tak więc test hipotezy H będziemy utożsamiali z funkcją $\varphi : \mathcal{X} \rightarrow \{0, 1\}$, gdzie 0 odpowiada przyjęciu hipotezy zerowej, natomiast 1 jej odrzuceniu.

Każdy test statystyczny rozbija przestrzeń prób \mathcal{X} na dwa rozłączne podzbiory: $\{(x_1, \dots, x_n) \in \mathcal{X} : \varphi(x_1, \dots, x_n) = 0\}$ – zbiór przyjęć hipotezy H i $W_\alpha = \{(x_1, \dots, x_n) \in \mathcal{X} : \varphi(x_1, \dots, x_n) = 1\}$ – zbiór odrzuceń hipotezy H (i przyjęć hipotezy K), nazywany też **zbiorem krytycznym** lub **obszarem krytycznym**.

W większości przypadków spotykanych w praktyce, test statystyczny ma następującą postać

$$\varphi(X_1, \dots, X_n) = \begin{cases} 1 & \text{gdy } T(X_1, \dots, X_n) \in W_\alpha \\ 0 & \text{gdy } T(X_1, \dots, X_n) \notin W_\alpha \end{cases} \quad (5.1)$$

gdzie $T = T(X_1, \dots, X_n)$ jest pewną funkcją próby zwaną **statystyką testową**.

W literaturze przedmiotu rozważa się również tzw. testy randomizowane $\tilde{\varphi} : \mathcal{X} \rightarrow [0, 1]$, w których do podjęcia decyzji uruchamia się dodatkowy mechanizm losowy, niezależny od \mathcal{X} , a hipotezę zerową odrzuca się, bądź przyjmuje, odpowiednio, z prawdopodobieństwem $\tilde{\varphi}(x)$ i $1 - \tilde{\varphi}(x)$. Ta ogólniejsza koncepcja testu statystycznego powstała w celu uzyskania pewnej elegancji teorii, same zaś testy randomizowane nie są w zasadzie stosowane w praktyce.

W wyniku testowania hipotezy H możemy podjąć trafną decyzję, bądź też popełnić jeden z dwóch błędów: odrzucić H , gdy jest ona prawdziwa (tzw. **błąd pierwszego rodzaju**) albo też przyjąć H , gdy jest ona fałszywa

(tzw. **błąd drugiego rodzaju**). Prawdopodobieństwo popełnienia błędu pierwszego rodzaju wyraża się wzorem

$$\alpha(\varphi) = P\{\varphi(X_1, \dots, X_n) = 1 \mid H\} = P\{T(X_1, \dots, X_n) \in W_\alpha \mid H\}, \quad (5.2)$$

natomiast prawdopodobieństwo popełnienia błędu drugiego rodzaju wynosi

$$\beta(\varphi) = P\{\varphi(X_1, \dots, X_n) = 0 \mid K\} = P\{T(X_1, \dots, X_n) \notin W_\alpha \mid K\}. \quad (5.3)$$

Najlepszym testem byłby oczywiście taki test, który minimalizowałby prawdopodobieństwa popełnienia obu błędów jednocześnie. Niestety taki test nie istnieje, bowiem przy ustalonej liczności próby zmniejszanie prawdopodobieństwa popełnienia błędu pierwszego rodzaju powoduje wzrost prawdopodobieństwa popełnienia błędu drugiego rodzaju i na odwrót.

Dlatego też w klasycznej teorii weryfikacji hipotez testy konstruuje się w ten sposób, że przyjmuje się górne ograniczenie na prawdopodobieństwo popełnienia błędu pierwszego rodzaju (tzw. **poziom istotności testu**), a następnie poszukuje się takiego testu, który – przy ograniczeniu na błąd pierwszego rodzaju – minimalizuje prawdopodobieństwo popełnienia błędu drugiego rodzaju. Mamy więc do czynienia z zadaniem postaci:

$$\beta(\varphi) \rightarrow \min, \quad (5.4)$$

przy warunku

$$\alpha(\varphi) \leq \alpha, \quad (5.5)$$

gdzie α jest przyjętym poziomem istotności (zwykle jest to mała liczba, np. 0.05 albo 0.01).

Zadanie to jest równoważne maksymalizacji **mocy testu** M (czyli prawdopodobieństwu odrzucenia hipotezy H , gdy jest ona fałszywa) przy zadanym poziomie istotności, tzn.

$$M(\varphi) \rightarrow \max, \quad (5.6)$$

przy warunku (5.5). Nietrudno zauważyć, że $M(\varphi) = 1 - \beta(\varphi)$.

Wspomniane powyżej pojęcie mocy testu znajduje zastosowanie przy porządkowaniu testów, np. w celu wyznaczenia testu optymalnego. I tak powiemy, że test φ jest mocniejszy od testu ψ , jeżeli

$$M(\varphi) \geq M(\psi). \quad (5.7)$$

Test, który jest mocniejszy od wszystkich innych testów danej klasy zwany jest **testem jednostajnie najmocniejszym**.

Problem konstrukcji testu jednostajnie najmocniejszego może być efektywnie rozwiązany niestety tylko w niewielu sytuacjach. Metodę konstrukcji

takiego testu w przypadku, gdy H i K są hipotezami prostymi, podaje podstawowy lemat Neymana–Pearsona (por. np. [1], [13], [18]). Stosunkowo łatwo można też skonstruować test jednostajnie najmocniejszy dla tzw. hipotez jednostronnych (np. $H : \theta > \theta_0$). W przypadku gdy nie istnieje test jednostajnie najmocniejszy w klasie wszystkich testów na poziomie istotności α , może się zdarzyć, że istnieje test jednostajnie najmocniejszy w danej podklasie tych testów, określonej w pewien naturalny sposób. Takim naturalnym warunkiem ograniczającym klasę testów jest warunek **nieobciążoności**, mówiący, że minimalna moc testu (względem wszystkich alternatyw) powinna być nie mniejsza od poziomu istotności testu. Innym naturalnym warunkiem zawężającym klasę rozpatrywanych testów jest warunek **niezmienneości**: test φ jest niezmienniczy względem danej grupy przekształceń \mathcal{G} , jeżeli $\varphi(x) = \varphi(gx)$ dla każdego $x \in \mathcal{X}$ i dla każdego $g \in \mathcal{G}$. Teoria testów jednostajnie najmocniejszych została szczegółowo omówiona np. w pracach [1], [13].

Warto na zakończenie wspomnieć, że poza omówionym powyżej klasycznym podejściem do konstrukcji testów optymalnych, pochodzących od Neymana i Pearsona, istnieją i inne podejścia, np. bayesowskie, czy minimaksowe.

5.3 Algorytmy testowania hipotez

Bez względu na postać testowanych hipotez, sam przebieg ich weryfikacji można opisać następującym algorytmem:

1. postawić hipotezę zerową H i hipotezę alternatywną K ;
2. wyspecyfikować model matematyczny (np. zakładamy, że próba losowa X_1, \dots, X_n pochodzi z rozkładu normalnego o nieznanej wariancji);
3. przyjąć poziom istotności α ;
4. obliczyć wartość statystyki testowej $T = T(X_1, \dots, X_n)$;
5. wyznaczyć obszar krytyczny W_α (w zależności od przyjętego poziomu istotności oraz hipotezy alternatywnej);
6. podjąć decyzję:
 - jeśli $T(X_1, \dots, X_n) \in W_\alpha$, wówczas odrzucić hipotezę H ,
 - jeśli $T(X_1, \dots, X_n) \notin W_\alpha$, wówczas nie ma podstaw do odrzucenia hipotezy H .

Uwaga!

Sytuacja, w której wartość statystyki testowej nie wpada do obszaru krytycznego, nie jest równoznaczna z potwierdzeniem prawdziwości hipotezy H . Oznacza to wyłącznie tyle, że przedstawione dane nie dają wystarczających argumentów na rzecz odrzucenia H . Stąd takie ostrożne sformułowanie, że jeżeli $T(X_1, \dots, X_n) \notin W_\alpha$, wówczas mówimy jedynie o braku podstaw do odrzucenia hipotezy zerowej. Ten stan rzeczy jest następstwem nieuwzględnienia ewentualnych konsekwencji błędu drugiego rodzaju, a zabezpieczaniu się jedynie przed możliwością popełnienia błędu pierwszego rodzaju. Jednakże trzeba zaznaczyć, że w praktyce zdarzenie $T(X_1, \dots, X_n) \notin W_\alpha$ na ogół jest interpretowane jako sugestia decyzji o akceptacji hipotezy H .

Przedstawiony powyżej algorytm będącym nazywany klasycznym algorytmem weryfikacji hipotez. W tym miejscu warto wspomnieć o jeszcze innym algorytmie, w którym podstawową rolę odgrywa pojęcie istotności testu (**poziomu krytycznego** lub ang. **p-value**).

Definicja 65 Istotnością testu nazywamy najmniejszy poziom istotności, przy którym odrzucamy rozważaną hipotezę.

Stąd też, jeśli obliczona dla danej próbki istotność testu p jest nie większa od α , to rozważaną hipotezę należy odrzucić. W przeciwnym przypadku, tzn. gdy $p > \alpha$, nie ma podstaw do odrzucenia danej hipotezy. To ostatnie podejście jest obecnie coraz częściej stosowane, bowiem popularne programy statystyczne nastawione są właśnie na obliczanie istotności testu.

A zatem, w przypadku posługiwania się pojęciem istotności testu, algorytm testowania hipotez wygląda następująco:

1. postawić hipotezę zerową H i hipotezę alternatywną K ;
2. wyspecyfikować model matematyczny;
3. przyjąć poziom istotności α ;
4. obliczyć wartość istotności testu p ;
5. podjąć decyzję:
 - jeśli $p \leq \alpha$, wówczas odrzucić hipotezę H ,
 - jeśli $p > \alpha$, wówczas nie ma podstaw do odrzucenia hipotezy H .

5.4 Testy dla wartości oczekiwanej

5.4.1 Testy dla pojedynczej próby

Założymy, że jesteśmy zainteresowani weryfikacją hipotezy dotyczącej wartości oczekiwanej μ , a konkretnie

$$H : \mu = \mu_0, \quad (5.8)$$

wobec jednej z trzech hipotez alternatywnych:

$$K : \mu \neq \mu_0 \quad (5.9)$$

$$K' : \mu < \mu_0$$

$$K'' : \mu > \mu_0.$$

W zależności od posiadanych informacji o rozkładzie badanej cechy wyróżniamy trzy modele.

Model 1

Załóżmy, że badana cecha X ma rozkład normalny $N(\mu, \sigma)$ o znanym odchyleniu standardowym σ . Statystyka testowa przyjmuje w tym przypadku postać

$$T = \frac{\bar{X} - \mu_0}{\sigma} \sqrt{n}. \quad (5.10)$$

Przy założeniu prawdziwości hipotezy H statystyka ta ma rozkład normalny $N(0, 1)$, w związku z czym obszar krytyczny – w zależności od przyjętej hipotezy alternatywnej (K , K' albo K'') – ma postać

$$\begin{aligned} W_\alpha &= (-\infty, -u_{1-\frac{\alpha}{2}}] \cup [u_{1-\frac{\alpha}{2}}, +\infty), \\ W'_\alpha &= (-\infty, -u_{1-\alpha}], \\ W''_\alpha &= [u_{1-\alpha}, +\infty), \end{aligned} \quad (5.11)$$

gdzie $u_{1-\frac{\alpha}{2}}$ i $u_{1-\alpha}$ są, odpowiednio, kwantylami rozkładu normalnego $N(0, 1)$ rzędów $1 - \frac{\alpha}{2}$ i $1 - \alpha$.

(patrz: Przykład 5.1)

Model 2

Jeżeli cecha X ma rozkład normalny $N(\mu, \sigma)$ o nieznanym odchyleniu standardowym σ , to do weryfikacji hipotezy H wykorzystujemy test zbudowany na statystyce

$$T = \frac{\bar{X} - \mu_0}{S} \sqrt{n}, \quad (5.12)$$

która przy założeniu prawdziwości hipotezy H ma rozkład t-Studenta o $n - 1$ stopniach swobody. W zależności od przyjętej hipotezy alternatywnej obszar krytyczny przybiera postać

$$\begin{aligned} W_\alpha &= (-\infty, -t_{1-\frac{\alpha}{2}}^{[n-1]}] \cup [t_{1-\frac{\alpha}{2}}^{[n-1]}, +\infty), \\ W'_\alpha &= (-\infty, -t_{1-\alpha}^{[n-1]}], \\ W''_\alpha &= [t_{1-\alpha}^{[n-1]}, +\infty), \end{aligned} \quad (5.13)$$

gdzie $t_{1-\frac{\alpha}{2}}^{[n-1]}$ i $t_{1-\alpha}^{[n-1]}$ są, odpowiednio, kwantylami rozkładu t-Studenta o $n - 1$ stopniach swobody rzędów $1 - \frac{\alpha}{2}$ i $1 - \alpha$.
(patrz: Przykład 5.2)

Model 3

Jeżeli próba pochodzi z dowolnego rozkładu (posiadającego jednakże skończoną wariancję), ale jest wystarczająco duża ($n \geq 100$), wówczas statystyka testowa przyjmuje postać

$$T = \frac{\bar{X} - \mu_0}{S} \sqrt{n}. \quad (5.14)$$

Przy założeniu prawdziwości hipotezy H i dla dostatecznie dużej próby statystyka ta ma w przybliżeniu rozkład normalny $N(0, 1)$, w związku z czym obszar krytyczny – w zależności od przyjętej hipotezy alternatywnej – ma postać

$$\begin{aligned} W_\alpha &= (-\infty, -u_{1-\frac{\alpha}{2}}] \cup [u_{1-\frac{\alpha}{2}}, +\infty), \\ W'_\alpha &= (-\infty, -u_{1-\alpha}], \\ W''_\alpha &= [u_{1-\alpha}, +\infty). \end{aligned} \quad (5.15)$$

5.4.2 Testy dla dwóch prób niezależnych

W praktyce istotną rolę odgrywają testy statystyczne, za pomocą których można porównywać wartości oczekiwane badanej cechy w dwóch zbiorowościach statystycznych. W szczególności interesująca jest weryfikacja hipotezy, że obie porównywane średnie są jednakowe, tzn.

$$H : \mu_1 = \mu_2, \quad (5.16)$$

przy jednej z trzech hipotez alternatywnych:

$$\begin{aligned} K &: \mu_1 \neq \mu_2 \\ K' &: \mu_1 < \mu_2 \\ K'' &: \mu_1 > \mu_2. \end{aligned} \quad (5.17)$$

Podstawą testu są dwie próbki X_1, \dots, X_{n_1} i Y_1, \dots, Y_{n_2} o licznosciami, odpowiednio n_1 i n_2 . Podobnie jak to miało miejsce przy weryfikacji hipotez dotyczących wartości oczekiwanej pojedynczej populacji, w zależności od posiadanych informacji o porównywanych populacjach, rozważamy różne modele.

Model 1

Załóżmy, że próbki X_1, \dots, X_{n_1} i Y_1, \dots, Y_{n_2} są niezależne i pochodzą z populacji o rozkładach normalnych, odpowiednio, $N(\mu_1, \sigma_1)$ i $N(\mu_2, \sigma_2)$

oraz że odchylenia standardowe σ_1 i σ_2 są znane. Wówczas statystyka testowa ma postać

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}, \quad (5.18)$$

gdzie \bar{X} i \bar{Y} są, odpowiednio, średnimi arytmetycznymi z pobranych próbek. Statystyka (5.18) przy założeniu prawdziwości hipotezy zerowej $H : \mu_1 = \mu_2$ ma rozkład normalny $N(0, 1)$. Obszar krytyczny – w zależności od przyjętej hipotezy alternatywnej K , K' i K'' – będzie miał postać

$$\begin{aligned} W_\alpha &= (-\infty, -u_{1-\frac{\alpha}{2}}] \cup [u_{1-\frac{\alpha}{2}}, +\infty), \\ W'_\alpha &= (-\infty, -u_{1-\alpha}], \\ W''_\alpha &= [u_{1-\alpha}, +\infty), \end{aligned} \quad (5.19)$$

gdzie liczby $u_{1-\frac{\alpha}{2}}$ i $u_{1-\alpha}$ są, odpowiednio, kwantylami rozkładu $N(0, 1)$ rzędów $1 - \frac{\alpha}{2}$ i $1 - \alpha$.

Model 2

Załóżmy teraz, że próby X_1, \dots, X_{n_1} i Y_1, \dots, Y_{n_2} są niezależne i pochodzą z populacji o rozkładach normalnych, odpowiednio, $N(\mu_1, \sigma_1)$ i $N(\mu_2, \sigma_2)$ o nieznanych odchyleniach standardowych σ_1 i σ_2 , ale równych, tzn. $\sigma_1 = \sigma_2$. Ponieważ najczęściej nie wiemy, czy założenie to jest spełnione, wobec tego należy najpierw zweryfikować hipotezę o równości wariancji (patrz poniżej) i dopiero wtedy, gdy hipoteza o równości wariancji $\sigma_1^2 = \sigma_2^2$ (przy ustalonym poziomie istotności α) nie będzie odrzucona – wówczas stosować poniższy test.

Do weryfikacji hipotezy o równości wartości oczekiwanych wykorzystujemy statystykę

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \sqrt{n_1 + n_2 - 2}, \quad (5.20)$$

gdzie \bar{X} i S_1^2 oraz \bar{Y} i S_2^2 są, odpowiednio, średnimi arytmetycznymi i wariancjami z pobranych próbek. Przy prawdziwości hipotezy H statystyka (5.20) ma rozkład t-Studenta o $n_1 + n_2 - 2$ stopniach swobody. W zależności od postaci hipotezy alternatywnej obszar krytyczny wygląda następująco:

$$\begin{aligned} W_\alpha &= (-\infty, -t_{1-\frac{\alpha}{2}}^{[n_1+n_2-2]}] \cup [t_{1-\frac{\alpha}{2}}^{[n_1+n_2-2]}, +\infty), \\ W'_\alpha &= (-\infty, -t_{1-\alpha}^{[n_1+n_2-2]}], \\ W''_\alpha &= [t_{1-\alpha}^{[n_1+n_2-2]}, +\infty), \end{aligned} \quad (5.21)$$

gdzie liczby $t_{1-\frac{\alpha}{2}}^{[n_1+n_2-2]}$ i $t_{1-\alpha}^{[n_1+n_2-2]}$ są, odpowiednio, kwantylami rzędu $1 - \frac{\alpha}{2}$ i $1 - \alpha$ rozkładu t-Studenta o $n_1 + n_2 - 2$ stopniach swobody.

(patrz: Przykład 5.4)

Model 3

Załóżmy teraz, że próby X_1, \dots, X_{n_1} i Y_1, \dots, Y_{n_2} są niezależne i pochodzą z populacji o rozkładach normalnych, odpowiednio, $N(\mu_1, \sigma_1)$ i $N(\mu_2, \sigma_2)$ o nieznanych i różnych odchyleniach standardowych σ_1 i σ_2 , tzn. $\sigma_1 \neq \sigma_2$. Odpowiada to sytuacji, gdy hipoteza o równości wariancji $\sigma_1^2 = \sigma_2^2$ (przy ustalonym poziomie istotności α) zostaje odrzucona. Wówczas statystyka testowa będzie mieć postać

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}. \quad (5.22)$$

Rozkład tej statystyki (zwanej statystyką Cochran-Coxa) zależy od liczności próbek i nieznanego stosunku wariancji σ_1/σ_2 . Można jednak znaleźć przybliżone wartości kwantylów tego rozkładu. Mianowicie, dla prób o liczebnościach, odpowiednio, n_1 i n_2 kwantyl rzędu ξ jest dany wzorem

$$c(\xi, n_1, n_2) \simeq \frac{\frac{S_1^2}{n_1} t_\xi^{[n_1-1]} + \frac{S_2^2}{n_2} t_\xi^{[n_2-1]}}{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}, \quad (5.23)$$

gdzie $t_\xi^{[n_1-1]}$ i $t_\xi^{[n_2-1]}$ są kwantylami rozkładu t-Studenta rzędu ξ , odpowiednio, o $n_1 - 1$ i $n_2 - 1$ stopniach swobody.

Obszar krytyczny – w zależności od przyjętej hipotezy alternatywnej K , K' i K'' – będzie miał postać

$$\begin{aligned} W_\alpha &= (-\infty, -c(1 - \frac{\alpha}{2}, n_1, n_2)] \cup [c(1 - \frac{\alpha}{2}, n_1, n_2), +\infty), \\ W'_\alpha &= (-\infty, -c(1 - \alpha, n_1, n_2)], \\ W''_\alpha &= [c(1 - \alpha, n_1, n_2), +\infty). \end{aligned} \quad (5.24)$$

Model 4

Jeżeli próby X_1, \dots, X_{n_1} i Y_1, \dots, Y_{n_2} są niezależne i pochodzą z populacji o nieznanych rozkładach, mających jednak skończone wariancje σ_1^2 i σ_2^2 , oraz jeżeli dysponujemy próbami o dużych liczebnościach n_1 i n_2 ($n_1, n_2 \geq 100$), wówczas statystyka testowa jest dana wzorem

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}. \quad (5.25)$$

Przy założeniu prawdziwości hipotezy H i dla dostatecznie dużej próby statystyka ta ma w przybliżeniu rozkład normalny $N(0, 1)$, w związku

z czym obszar krytyczny – w zależności od przyjętej hipotezy alternatywnej – ma postać

$$\begin{aligned} W_\alpha &= (-\infty, -u_{1-\frac{\alpha}{2}}] \cup [u_{1-\frac{\alpha}{2}}, +\infty), \\ W'_\alpha &= (-\infty, -u_{1-\alpha}], \\ W''_\alpha &= [u_{1-\alpha}, +\infty). \end{aligned} \quad (5.26)$$

(patrz: Przykład 5.3)

5.4.3 Test dla obserwacji parami zależnych

Rozważmy obecnie sytuację, w której próby X_1, \dots, X_n i Y_1, \dots, Y_n są parami zależne. Innymi słowy mamy próbę losową utworzoną przez uporządkowane pary zmiennych losowych $(X_1, Y_1), \dots, (X_n, Y_n)$, przy czym obserwacje w parach są zależne, podczas gdy pary są wzajemnie niezależne. Z taką sytuacją mamy do czynienia na przykład wtedy, gdy badania dotyczą tych samych jednostek. Przyjmijmy, dodatkowo, założenie o normalności rozkładu. Statystyką testową będzie w tym wypadku

$$T = \frac{\bar{Z}}{S_Z} \sqrt{n}, \quad (5.27)$$

gdzie \bar{Z} oraz S_Z jest, odpowiednio, średnią i odchyleniem standardowym z próby Z_1, \dots, Z_n otrzymanej poprzez odejmowanie w parach, tzn. $Z_i = X_i - Y_i$, dla $j = 1, \dots, n$. Przy prawdziwości hipotezy H statystyka (5.27) ma rozkład t-Studenta o $n - 1$ stopniach swobody. Stąd – w zależności od przyjętej hipotezy alternatywnej – obszar krytyczny przyjmuje postać

$$\begin{aligned} W_\alpha &= (-\infty, -t_{1-\frac{\alpha}{2}}^{[n-1]}] \cup [t_{1-\frac{\alpha}{2}}^{[n-1]}, +\infty), \\ W'_\alpha &= (-\infty, -t_{1-\alpha}^{[n-1]}], \\ W''_\alpha &= [t_{1-\alpha}^{[n-1]}, +\infty). \end{aligned} \quad (5.28)$$

(patrz: Przykład 5.5)

5.5 Testy dla mediany

5.5.1 Testy dla pojedynczej próby

W tym punkcie podamy dwa testy służące do weryfikacji hipotezy dotyczącej wartości mediana med . Weryfikujemy hipotezę zerową

$$H : med = m_0, \quad (5.29)$$

przy jednej z trzech hipotez alternatywnych

$$\begin{aligned} K &: med \neq m_0, \\ K' &: med < m_0, \\ K'' &: med > m_0, \end{aligned} \quad (5.30)$$

na zadany poziomie istotności α . O rozkładzie badanej cechy X zakładamy jedynie tyle, że jest on ciągły w otoczeniu mediany.

Test znaków

Nazwa testu pochodzi stąd, że statystyką testową jest liczba obserwacji przewyższających m_0 (czyli liczba znaków "+" w różnicach $X_i - m_0$, $i = 1, 2, \dots, n$), tzn.

$$T = \sum_{i=1}^n I(X_i - m_0 > 0), \quad (5.31)$$

gdzie $I(\cdot)$ jest funkcją indykatorową, tzn.

$$I(A) = \begin{cases} 1 & \text{gdy zdanie } A \text{ jest prawdziwe} \\ 0 & \text{w przeciwnym przypadku.} \end{cases} \quad (5.32)$$

Przy założeniu prawdziwości hipotezy zerowej H statystyka ta ma rozkład dwumianowy

$$P(R = r) = \binom{n}{r} \left(\frac{1}{2}\right)^n, \quad r = 0, 1, \dots, n. \quad (5.33)$$

Obszar krytyczny, w zależności od przyjętej hipotezy alternatywnej, ma postać:

$$\begin{aligned} W_\alpha &= [0, r_{\frac{\alpha}{2}}] \cup [n - r_{\frac{\alpha}{2}}, n] \\ W'_\alpha &= [0, r_\alpha] \\ W''_\alpha &= [n - r_\alpha, n], \end{aligned} \quad (5.34)$$

przy czym r_α jest wartością krytyczną rozkładu statystyki T dla testu na poziomie istotności α . Tablice z wartościami krytycznymi znajdują się w podrozdz. 6.6.

(patrz: Przykład 5.6)

Uwaga!

Cecha X ma rozkład ciągły a więc wszystkie różnice $X_i - m_0$, $i = 1, 2, \dots, n$ są różne od zera. W praktyce jednak, ze względu na stosowane zaokrąglenia pomiarów zmiennej X , może się okazać, że niektóre z różnic są zerami. W takich przypadkach istnieją dwie drogi postępowania:

1. jeżeli liczba zer jest mała w stosunku do n , to elementy X_i dla których różnice są zerami, można pominąć, korygując wartość n do liczby obserwacji, które zostały,
2. jeżeli liczba zer nie jest mała w stosunku do n , to tym elementom, dla których różnica jest zero przyporządkowujemy znak "+" z prawdopodobieństwem $1/2$ i znak "-" z tym samym prawdopodobieństwem.

Test rangowanych znaków

Omówiony wcześniej test znaków nie jest testem o dużej mocy, ponieważ z informacji zawartych w próbie wykorzystuje jedynie znaki różnic $X_i - m_0$, ($i = 1, 2, \dots, n$). Mocniejszym testem jest test rangowanych znaków, który uwzględnia również wielkości tych różnic. Test ten wymaga jednak większej liczby założeń, niż test znaków. Aby zastosować test rangowanych znaków rozkład badanej cechy musi być ciągły i symetryczny.

Postępowanie testujące można opisać następująco. Obliczamy różnice $X_i - m_0$, ($i = 1, 2, \dots, n$) i nadajemy numery (rang) wartościom bezwzględnych tych różnic poczynając od 1 dla najmniejszej wartości bezwzględnej. Wyznaczone rangi grupujemy oddzielnie dla różnic dodatnich i ujemnych i obliczamy sumy tych rang T^+ i T^- . Oznaczmy przez $R(|X_i - m_0|)$ rangę różnic $X_i - m_0$. Wówczas

$$\begin{aligned} T^+ &= \sum_{i=1}^n I(X_i - m_0 > 0) R(|X_i - m_0|), \\ T^- &= \sum_{i=1}^n I(X_i - m_0 < 0) R(|X_i - m_0|). \end{aligned} \quad (5.35)$$

Przy założeniu prawdziwości hipotezy zerowej H i przy założeniu, że cecha X ma rozkład ciągły i symetryczny, statystyki T^+ i T^- mają ten sam rozkład, który wspólnie oznaczamy przez T . Jeżeli wyznaczone wartości T^+ i T^- spełniają (w zależności od hipotezy alternatywnej) nierówności

$$\begin{aligned} T^+ &\leq q_{\frac{\alpha}{2}} \text{ lub } T^- \leq q_{\frac{\alpha}{2}} \quad (\text{dla hipotezy alternatywnej } K) \\ T^+ &\leq q_{\alpha} \quad (\text{dla hipotezy alternatywnej } K') \\ T^- &\leq q_{\alpha} \quad (\text{dla hipotezy alternatywnej } K''), \end{aligned} \quad (5.36)$$

gdzie q_{α} jest wartością krytyczną rozkładu statystyki T dla testu na poziomie istotności α , to hipotezę zerową odrzucamy na korzyść odpowiedniej hipotezy alternatywnej. W przeciwnym przypadku nie ma podstaw do odrzucenia hipotezy zerowej. Tablice wartości krytycznych rozkładu zmiennej losowej T można znaleźć w podrozdz. 6.7.

Statystyka testu rangowanych znaków ma dla dużych wartości n , rozkład asymptotycznie normalny $N\left(\frac{1}{4}n(n+1), \sqrt{\frac{1}{24}n(n+1)(2n+1)}\right)$. Wtedy

wartość krytyczną t_{α} wyznaczamy ze wzoru

$$t_{\alpha} = \frac{1}{4}n(n+1) + u_{\alpha} \sqrt{\frac{1}{24}n(n+1)(2n+1)}, \quad (5.37)$$

gdzie u_{α} jest kwantylem rzędu α rozkładu normalnego $N(0, 1)$.

5.5.2 Test dla dwóch prób niezależnych

Załóżmy, że próby X_1, X_2, \dots, X_n i Y_1, Y_2, \dots, Y_m pochodzą z rozkładu ciągłego o medianach, odpowiednio, med_X i med_Y . Bez straty ogólności możemy przyjąć, że $n \leq m$. Do weryfikacji hipotezy o równości median obu populacji, tzn.

$$H : med_X = med_Y, \quad (5.38)$$

przy jednej z trzech hipotez alternatywnych:

$$\begin{aligned} K &: med_X \neq med_Y, \\ K' &: med_X < med_Y, \\ K'' &: med_X > med_Y, \end{aligned} \quad (5.39)$$

służy test Wilcoxona, zwany również testem Manna-Whitneya-Wilcoxon.

Obserwacje z obu prób ustawiamy w jeden ciąg niemalejący i elementem tego ciągu przyporządkowujemy rangi. Niech R_1, \dots, R_n będą rangami elementów z pierwszej próby. Statystyka testowa dana jest wzorem

$$T = \sum_{i=1}^n R_i. \quad (5.40)$$

W zależności od przyjętej alternatywy obszar krytyczny ma postać:

$$\begin{aligned} W_{\alpha} &= [0, w_{n,m}(\alpha)] \cup [w'_{n,m}(\alpha), +\infty) \\ W'_{\alpha} &= [0, w_{n,m}(2\alpha)] \\ W''_{\alpha} &= [w'_{n,m}(2\alpha), +\infty), \end{aligned} \quad (5.41)$$

przy czym $w_{n,m}(\alpha)$ i $w'_{n,m}(\alpha)$ są wartościami krytycznymi rozkładu statystyki T dla poziomu istotności α . W podrozdz. 6.8 zamieszczono wartości krytyczne $w_{n,m}(\alpha)$. Wartości $w'_{n,m}(\alpha)$ oblicza się ze wzoru

$$w'_{n,m}(\alpha) = n(n+m+1) - w_{n,m}(\alpha). \quad (5.42)$$

(patrz: Przykład 5.7)

5.5.3 Testy dla obserwacji parami zależnych

Rozważmy ponownie sytuację, w której próby X_1, \dots, X_n i Y_1, \dots, Y_n są parami zależne, czyli gdy mamy próbę losową utworzoną przez uporządkowane pary zmiennych losowych $(X_1, Y_1), \dots, (X_n, Y_n)$, przy czym obserwacje w parach są zależne, a pary są wzajemnie niezależne. Tym razem jednak zajmiemy się nie wartością oczekiwana, ale medianą. A konkretnie, naszym celem będzie weryfikacja hipotezy

$$H : \text{med}_{X-Y} = 0, \quad (5.43)$$

wobec jednej z hipotez alternatywnych

$$K : \text{med}_{X-Y} \neq 0 \quad (5.44)$$

$$K' : \text{med}_{X-Y} < 0$$

$$K'' : \text{med}_{X-Y} > 0.$$

Do weryfikacji hipotezy (5.43) można wykorzystać omówiony powyżej test znaków lub test rangowanych znaków. Dokładniej, niech $Z_i = X_i - Y_i$ dla $i = 1, \dots, n$. Jeżeli ciąg różnic Z_1, \dots, Z_n ma rozkład ciągły w otoczeniu mediany, wówczas hipotezę (5.43) weryfikujemy za pomocą testu znaków, natomiast jeżeli rozkład ciągu różnic jest ciągły i symetryczny, to możemy użyć testu rangowanych znaków. W obu przypadkach statystyki testowe (5.31) oraz (5.35) należy zmodyfikować zastępując X zmienną Z .

(patrz: Przykład 5.8)

Uwaga!

W ogólności, mediana różnicy nie jest równa różnicy median. Jeżeli jednak rozkłady zmiennych losowych X i Y są symetryczne i zachodzi $\text{med}_X = \text{med}_Y$ oraz gdy zmienna losowa $Z = X - Y$ ma rozkład symetryczny, wówczas $\text{med}_Z = \text{med}_X - \text{med}_Y$ i wtedy hipoteza (5.43) jest równoważna hipotezie $H : \text{med}_X = \text{med}_Y$.

5.6 Testy dla wariancji

5.6.1 Testy dla pojedynczej próby

Załóżmy, że badana cecha X ma rozkład normalny $N(\mu, \sigma)$ o nieznanych parametrach μ i σ . Testujemy hipotezę zerową

$$H : \sigma^2 = \sigma_0^2, \quad (5.45)$$

przy jednej z trzech możliwych hipotez alternatywnych:

$$K : \sigma^2 \neq \sigma_0^2 \quad (5.46)$$

$$K' : \sigma^2 < \sigma_0^2$$

$$K'' : \sigma^2 > \sigma_0^2.$$

Model 1

Podstawą testu jest statystyka

$$T = \frac{(n-1)S^2}{\sigma_0^2}, \quad (5.47)$$

która przy założeniu prawdziwości hipotezy zerowej H ma rozkład chi-kwadrat o $n-1$ stopniach swobody. Przy ustalonym poziomie istotności α mamy, odpowiednio dla K , K' i K'' , następujące obszary krytyczne

$$W_\alpha = (0, \chi_{\frac{\alpha}{2}, n-1}^2] \cup [\chi_{1-\frac{\alpha}{2}, n-1}^2, +\infty) \quad (5.48)$$

$$W'_\alpha = (0, \chi_{\alpha, n-1}^2]$$

$$W''_\alpha = [\chi_{1-\alpha, n-1}^2, +\infty),$$

przy czym $\chi_{\beta, n-1}^2$ jest kwantylem rzędu β rozkładu chi-kwadrat o $n-1$ stopniach swobody.

(patrz: Przykład 5.2)

Model 2

W przypadku, gdy liczność próby $n \geq 50$, to do weryfikacji hipotezy H wykorzystujemy statystykę

$$T = \sqrt{\frac{2nS^2}{\sigma_0^2}} - \sqrt{2n-3}, \quad (5.49)$$

która ma asymptotycznie rozkład normalny $N(0, 1)$. W tym przypadku obszar krytyczny – stosownie do przyjętej hipotezy alternatywnej – ma postać

$$W_\alpha = (-\infty, -u_{1-\frac{\alpha}{2}}] \cup [u_{1-\frac{\alpha}{2}}, +\infty) \quad (5.50)$$

$$W'_\alpha = (-\infty, -u_{1-\alpha}]$$

$$W''_\alpha = [u_{1-\alpha}, +\infty).$$

5.6.2 Testy dla dwóch prób niezależnych

Przyjmijmy, że próby X_1, \dots, X_{n_1} i Y_1, \dots, Y_{n_2} są niezależne i pochodzą z populacji o rozkładach normalnych, odpowiednio, $N(\mu_1, \sigma_1)$ i $N(\mu_2, \sigma_2)$ oraz że parametry rozkładów μ_1 , μ_2 , σ_1 i σ_2 są nieznane. Rozważamy hipotezę zerową o równości dwóch wariancji

$$H : \sigma_1^2 = \sigma_2^2, \quad (5.51)$$

wobec hipotezy alternatywnej

$$K'' : \sigma_1^2 > \sigma_2^2. \quad (5.52)$$

Statystyką testową będzie w tym przypadku

$$T'' = \frac{S_2^2}{S_1^2} \quad (5.53)$$

Jeżeli hipoteza zerowa jest prawdziwa, to statystyka ta ma rozkład F-Snedecora o $(n_1 - 1, n_2 - 1)$ stopniach swobody. Przy ustalonym poziomie istotności α mamy następujący obszar krytyczny

$$W_\alpha = [F_{1-\alpha}^{[n_1-1, n_2-1]}, +\infty), \quad (5.54)$$

gdzie $F_{1-\alpha}^{[n_1-1, n_2-1]}$ jest kwantylem rzędu $1 - \alpha$ rozkładu F-Snedecora o $(n_1 - 1, n_2 - 1)$ stopniach swobody.

Jeżeli natomiast jesteśmy zainteresowani weryfikacją hipotezy H wobec hipotezy alternatywnej

$$K' : \sigma_1^2 < \sigma_2^2, \quad (5.55)$$

wówczas korzystamy ze statystyki testowej

$$T' = \frac{1}{T''} = \frac{S_2^2}{S_1^2}, \quad (5.56)$$

która przy prawdziwości hipotezy H ma rozkład F-Snedecora o $(n_2 - 1, n_1 - 1)$ stopniach swobody. Zatem w tym przypadku mamy następujący obszar krytyczny

$$W_\alpha = [F_{1-\alpha}^{[n_2-1, n_1-1]}, +\infty), \quad (5.57)$$

gdzie $F_{1-\alpha}^{[n_2-1, n_1-1]}$ jest kwantylem rzędu $1 - \alpha$ rozkładu F-Snedecora o $(n_2 - 1, n_1 - 1)$ stopniach swobody.

Wreszcie, gdy weryfikujemy hipotezę H względem dwustronnej hipotezy alternatywnej

$$K : \sigma_1^2 \neq \sigma_2^2, \quad (5.58)$$

wtedy korzystamy ze statystyki testowej

$$T = \begin{cases} T' & \text{gdy } S_1^2 > S_2^2, \\ T'' & \text{gdy } S_2^2 > S_1^2, \end{cases} \quad (5.59)$$

i obszaru krytycznego

$$W_\alpha = \begin{cases} [F_{1-\alpha}^{[n_1-1, n_2-1]}, +\infty) & \text{gdy } S_1^2 > S_2^2, \\ [F_{1-\alpha}^{[n_2-1, n_1-1]}, +\infty) & \text{gdy } S_2^2 > S_1^2. \end{cases} \quad (5.60)$$

5.7 Testy dla wskaźnika struktury

5.7.1 Testy dla pojedynczej próby

Zakładamy, że próba pochodzi z rozkładu dwupunktowego. Weryfikowana hipoteza dotyczy nieznanego parametru p (prawdopodobieństwa sukcesu)

$$H : p = p_0, \quad (5.61)$$

wobec jednej z trzech hipotez alternatywnych:

$$\begin{aligned} K &: p = p_0, \\ K' &: p < p_0, \\ K'' &: p > p_0. \end{aligned} \quad (5.62)$$

Do weryfikacji hipotezy H wykorzystujemy wskaźnik struktury z próby

$$\hat{p} = \frac{k}{n}, \quad (5.63)$$

gdzie k jest liczbą elementów wyróżnionych (sukcesów) w próbie o liczności n . W zależności od liczności próby wyróżniamy dwa modele.

Model 1

Jeżeli dysponujemy liczną próbą, tzn. $n \geq 100$, wówczas statystyka testowa ma postać

$$T = \frac{k - np_0}{\sqrt{np_0(1 - p_0)}}. \quad (5.64a)$$

Na podstawie centralnego twierdzenia granicznego Moivre'a - Laplace'a wiemy, że statystyka ta ma w przybliżeniu rozkład $N(0, 1)$. A zatem – w zależności od przyjętej hipotezy alternatywnej (K , K' albo K'') – obszar krytyczny ma postać

$$\begin{aligned} W_\alpha &= (-\infty, -u_{1-\frac{\alpha}{2}}] \cup [u_{1-\frac{\alpha}{2}}, +\infty), \\ W'_\alpha &= (-\infty, -u_{1-\alpha}], \\ W''_\alpha &= [u_{1-\alpha}, +\infty). \end{aligned} \quad (5.65)$$

(patrz: **Przykład 5.9**)

Model 2

Jeżeli próba nie jest dostatecznie duża korzystamy ze statystyki testowej

$$T = 2 \left(\arcsin \sqrt{\frac{k}{n}} - \arcsin \sqrt{p_0} \right) \sqrt{n}, \quad (5.66)$$

mającej w przybliżeniu rozkład $N(0, 1)$. A zatem obszar krytyczny – stosownie do przyjętej hipotezy alternatywnej – ma postać

$$\begin{aligned} W_\alpha &= (-\infty, -u_{1-\frac{\alpha}{2}}] \cup [u_{1-\frac{\alpha}{2}}, +\infty) \\ W'_\alpha &= (-\infty, -u_{1-\alpha}] \\ W''_\alpha &= [u_{1-\alpha}, +\infty). \end{aligned} \quad (5.67)$$

5.7.2 Testy dla dwóch niezależnych prób

Rozważmy obecnie sytuację, w której próby X_1, \dots, X_n i Y_1, \dots, Y_n pochodzą z rozkładów dwupunktowych o nieznanych parametrach, odpowiednio p_1 i p_2 . Weryfikujemy hipotezę o równości obu wskaźników struktury (prawdopodobieństw sukcesu)

$$H : p_1 = p_2, \quad (5.68)$$

wobec jednej z hipotez alternatywnych

$$\begin{aligned} K &: p_1 \neq p_2, \\ K' &: p_1 < p_2, \\ K'' &: p_1 > p_2. \end{aligned} \quad (5.69)$$

Podobnie jak w przypadku weryfikacji hipotezy dla jednej próby podamy dwa modele, zależne od liczności próby.

Model 1

Jeżeli obie próby mają dużą licznosć, tzn. $n_1, n_2 \geq 100$, wówczas statystyka testowa jest dana wzorem

$$T = \frac{\frac{k_1}{n_1} - \frac{k_2}{n_2}}{\sqrt{\frac{k}{n}(1 - \frac{k}{n})(\frac{1}{n_1} + \frac{1}{n_2})}}, \quad (5.70)$$

gdzie k_1 i k_2 oznaczają liczbę elementów wyróżnionych (sukcesów), odpowiednio, w pierwszej i drugiej próbie, n jest łączną liczbą obserwacji, tzn. $n = n_1 + n_2$, natomiast k oznacza łączną liczbę elementów wyróżnionych w obu próbach, czyli $k = k_1 + k_2$.

Przy założeniu prawdziwości hipotezy H statystyka ta ma rozkład asymptotycznie normalny, w związku z czym obszar krytyczny – w zależności od przyjętej alternatywy – ma postać

$$\begin{aligned} W_\alpha &= (-\infty, -u_{1-\frac{\alpha}{2}}] \cup [u_{1-\frac{\alpha}{2}}, +\infty) \\ W'_\alpha &= (-\infty, -u_{1-\alpha}] \\ W''_\alpha &= [u_{1-\alpha}, +\infty). \end{aligned} \quad (5.71)$$

(patrz: Przykład 5.10)

Model 2

Jeżeli natomiast liczności prób nie są wystarczająco duże, korzystamy ze statystyki testowej

$$T = 2 \left(\arcsin \sqrt{\frac{k_1}{n_1}} - \arcsin \sqrt{\frac{k_2}{n_2}} \right) \sqrt{\frac{n_1 n_2}{n_1 + n_2}}, \quad (5.72)$$

która, zakładając że hipoteza H jest prawdziwa, ma w przybliżeniu rozkład $N(0, 1)$. Stąd ponownie, obszar krytyczny – stosownie do przyjętej alternatywy – wygląda następująco

$$\begin{aligned} W_\alpha &= (-\infty, -u_{1-\frac{\alpha}{2}}] \cup [u_{1-\frac{\alpha}{2}}, +\infty) \\ W'_\alpha &= (-\infty, -u_{1-\alpha}] \\ W''_\alpha &= [u_{1-\alpha}, +\infty). \end{aligned} \quad (5.73)$$

5.8 Testy zgodności

5.8.1 Uwagi wstępne

Omawiając w poprzednim rozdziale podstawowe testy parametryczne spotkaliśmy się często z założeniem normalności rozkładu badanej cechy. Można jednak zapytać: a skąd wiemy, że ów rozkład jest normalny? Jak się o tym przekonać? Podobne pytania można oczywiście zadawać odnośnie i innych rozkładów.

Postawiony powyżej problem można sprowadzić do zadania weryfikacji hipotezy o postaci rozkładu (w szczególności, o normalności rozkładu). Do weryfikacji takiej hipotezy stosuje się tzw. *testy zgodności*.

Załóżmy, że interesuje nas postać rozkładu badanej cechy X . Dystrybuantę owego nieznanego rozkładu oznaczymy przez F . Hipoteza dotycząca postaci F może być dwojakiego rodzaju: może to być hipoteza prosta

$$H : F = F_0, \quad (5.74)$$

gdzie F_0 jest określona dystrybuantą, bądź też hipoteza złożona

$$H : F \in \mathcal{F}, \quad (5.75)$$

gdzie \mathcal{F} oznacza pewną rodzinę dystrybuant (rozkładów). Przykładem hipotezy typu (5.74) było by: "X ma rozkład wykładniczy o wartości oczekiwanej 100", podczas gdy w drugim przypadku mielibyśmy do czynienia z bardziej ogólną hipotezą – "X ma rozkład wykładniczy". W obu przypadkach hipoteza alternatywna K jest po prostu zaprzeczeniem hipotezy zerowej H , tzn. dla hipotezy (5.74) mamy

$$K : F \neq F_0, \quad (5.76)$$

natomiast dla (5.75)

$$H : F \notin \mathcal{F}. \quad (5.77)$$

Poniżej przedstawione zostaną najbardziej znane i uniwersalne testy zgodności, jakimi są: test zgodności chi-kwadrat oraz test Kolmogorowa. Ponadto istnieją testy ukierunkowane na weryfikację zgodności z określonym rozkładem. W podrozdziale 5.8.4 omówimy pokrótko metody testowania normalności.

Czasem przedmiotem zainteresowania jest nie tyle znajomość rozkładu cechy lecz to, czy dwie (lub więcej niż dwie) próbki pochodzą z tej samej populacji (mają ten sam rozkład). Do rozstrzygnięcia tej kwestii służą również odpowiednie testy zgodności (por. podrozdz. 5.8.5 i 5.8.6).

5.8.2 Test zgodności chi-kwadrat

Test ten wymaga dużej próby (z reguły $n \geq 100$), ale może być stosowany zarówno wobec rozkładów dyskretnych jak i ciągłych. Przeprowadza się go na podstawie danych pogrupowanych w szereg rozdzielczy według zasad przedstawionych w rozdziale 3 (podrozdz. 3.3.3). Dodatkowo wymaga się, aby liczności poszczególnych klas nie były mniejsze niż 5 (tzn. $n_i \geq 5$).

Hipotezę prostą (5.74), w której dystrybuanta F_0 jest w pełni określona, testujemy porównując zaobserwowane liczności n_i z licznosciami hipotetycznymi np_i odpowiadającymi oczekiwany licznosciom poszczególnych klas przy założeniu prawdziwości hipotezy zerowej. Wielkości p_i obliczamy ze wzoru

$$p_i = F_0(\zeta_i) - F_0(\zeta_{i-1}), \quad (5.78)$$

gdzie ζ_0, \dots, ζ_k oznaczają krańce przedziałów klasowych.

Statystyka testowa

$$T = \sum_{i=1}^k \frac{(n_i - np_i)^2}{np_i} \quad (5.79)$$

ma przy prawdziwości hipotezy asymptotyczny rozkład chi-kwadrat o $k-1$ stopniach swobody.

Zbyt duże wartości tej statystyki będą świadczyć przeciwko postawionej hipotezie, stąd obszar krytyczny testu na przyjętym poziomie istotności α będzie miał postać

$$W_\alpha = [\chi^2_{1-\alpha, k-1}, +\infty), \quad (5.80)$$

gdzie $\chi^2_{1-\alpha, k-1}$ oznacza kwantyl rzędu $1-\alpha$ rozkładu chi-kwadrat o $k-1$ stopniach swobody. Jeżeli obliczona wartość statystyki (5.79) należy do obszaru krytycznego (5.80), to odrzucamy hipotezę rozważaną H na rzecz

hipotezy alternatywnej (w tym przypadku odpowiada to stwierdzeniu, że F_0 nie jest rozkładem badanej cechy). W przeciwnym przypadku stwierdzamy brak podstaw do odrzucenia postawionej hipotezy o postaci rozkładu na zadany poziomie istotności.

(patrz: Przykład 5.11)

W praktycznych zastosowaniach na ogół nie znamy rozkładu F_0 . Wówczas testujemy hipotezę złożoną (5.75), w której \mathcal{F} jest pewną rodziną rozkładów zależących od r nieznanych parametrów $\theta_1, \dots, \theta_r$, tzn.

$$H : F \in \mathcal{F} = \{F : F = F_{\theta_1, \dots, \theta_r}\}. \quad (5.81)$$

W tym przypadku szacujemy najpierw nieznane parametry (najlepiej metodą największej wiarodnościi) otrzymując $\hat{\theta}_1, \dots, \hat{\theta}_r$, po czym rozważaną hipotezę złożoną (5.81) zastępujemy hipotezą prostą

$$H : F = F_{\hat{\theta}_1, \dots, \hat{\theta}_r}, \quad (5.82)$$

wyrażającą przypuszczenie, że badana cecha ma rozkład należący do rozpatrywanej rodziny rozkładów, a dokładniej jest to rozkład $F_{\hat{\theta}_1, \dots, \hat{\theta}_r}$.

Statystyką testową jest tu również statystyka (5.79), tyle że wielkości p_i oblicza się teraz ze wzoru

$$p_i = F_{\hat{\theta}_1, \dots, \hat{\theta}_r}(\zeta_i) - F_{\hat{\theta}_1, \dots, \hat{\theta}_r}(\zeta_{i-1}). \quad (5.83)$$

Przy prawdziwości hipotezy (5.82) statystyka ta ma asymptotyczny rozkład chi-kwadrat o $k-r-1$ stopniach swobody, gdzie r jest liczbą estymowanych parametrów, a obszar krytyczny testu na przyjętym poziomie istotności α ma postać

$$W_\alpha = [\chi^2_{1-\alpha, k-r-1}, +\infty). \quad (5.84)$$

(patrz: Przykład 5.12)

5.8.3 Test Kołmogorowa

Test zgodności Kołmogorowa można stosować nawet dla małych prób, ale tylko dla rozkładów ciągłych. Istotą tego testu jest porównanie dystrybuanty empirycznej \hat{F}_n , zbudowanej na podstawie próby, z dystrybuantą teoretyczną F_0 .

Za statystykę testową do weryfikacji hipotezy prostej (5.74) Kołmogorow przyjął następującą miarę odległości dystrybuanty empirycznej od dystrybuanty teoretycznej

$$D_n = \sup_x |\hat{F}_n(x) - F_0(x)|. \quad (5.85)$$

Przy założeniu prawdziwości hipotezy H statystyka D_n ma rozkład niezależny od F_0 . Tablice wartości krytycznych $D_n(\alpha)$ tego rozkładu znaleźć można w podrozdz. 6.11. Obszar krytyczny rozważanego testu ma postać

$$W_\alpha = [D_n(\alpha), 1]. \quad (5.86)$$

W praktycznych zastosowaniach wartość statystyki (5.85) wyznacza się ze wzoru

$$D_n = \max\{D_n^+, D_n^-\}, \quad (5.87)$$

gdzie

$$D_n^+ = \max_{1 \leq i \leq n} \left| \frac{i}{n} - F_0(x_{i:n}) \right|, \quad (5.88)$$

$$D_n^- = \max_{1 \leq i \leq n} \left| F_0(x_{i:n}) - \frac{i-1}{n} \right|, \quad (5.89)$$

natomiast $x_{1:n} \leq \dots \leq x_{n:n}$ są uporządkowanymi rosnąco wartościami próbki.

Przy testowaniu hipotezy złożonej (5.81) postępujemy podobnie jak z testem zgodności chi-kwadrat, a więc estymujemy nieznane parametry rozkładu, po czym problem sprowadzamy do weryfikacji hipotezy (5.82). Niestety w tym przypadku rozkład statystyki D_n zależy od hipotetycznej dystrybuanty (a więc w szczególności od prawdziwych, ale nieznanych, wartości parametrów), co sprawia, że bezpośrednie zastosowanie testu zgodności Kołmogorowa nie jest właściwe. Jeżeli jednak liczność próbki jest dostatecznie duża, można posługiwać się granicznym rozkładem statystyki D_n , tzw. rozkładem λ -Kołmogorowa. W tym przypadku korzystamy ze statystyki testowej

$$T = \sqrt{n} D_n, \quad (5.90)$$

gdzie D_n oblicza się ze wzoru (5.87), z tą różnicą, że we wzorach (5.88) i (5.89) zamiast F_0 należy wstawić $F_{\hat{\theta}_1, \dots, \hat{\theta}_r}$. Obszar krytyczny jest postaci

$$W_\alpha = [\lambda_\alpha, +\infty), \quad (5.91)$$

gdzie λ_α jest wartością krytyczną testu na poziomie istotności α . Dla wybranych trzech poziomów istotności wartości krytyczne wynoszą

α	0.01	0.05	0.10
λ_α	1.628	1.354	1.224

5.8.4 Testy normalności

Przedstawione powyżej testy zgodności: chi-kwadrat i Kołmogorowa mogą być stosowane do weryfikacji szerokiej gamy hipotez o postaci rozkładu, w szczególności do testowania hipotez o normalności rozkładu. Z uwagi na fundamentalne znaczenie rozkładu normalnego, testowanie hipotez o normalności jest zagadniением bardzo ważnym w praktyce. Oprócz wspomnianych powyżej testów znane są i inne testy zgodności ukierunkowane właśnie na weryfikację hipotez o normalności. Do takich specjalistycznych testów zgodności należą m.in. test Shapiro - Wilka, test wykorzystujący współczynnik asymetrii oraz test wykorzystujący kurtozę.

Statystyka testowa testu Shapiro - Wilka dana jest wzorem

$$T = \frac{\left(\sum_{i=1}^{[n/2]} a_i(n) (X_{n-i+1:n} - X_{i:n}) \right)^2}{\sum_{i=1}^n (X_i - \bar{X})^2}, \quad (5.92)$$

gdzie $a_i(n)$ są pewnymi stałymi zależnymi od liczności próby (wartości tych stałych można znaleźć np. w podrozdz. 6.9), natomiast $[n/2]$ oznacza część całkowitą wyrażenia $n/2$. Obszar krytyczny ma postać

$$W_\alpha = (-\infty, w(\frac{\alpha}{2}, n)] \cup [w(1 - \frac{\alpha}{2}, n), +\infty), \quad (5.93)$$

gdzie $w(\frac{\alpha}{2}, n)$ i $w(1 - \frac{\alpha}{2}, n)$ oznaczają kwantyle rzędu, odpowiednio, $\frac{\alpha}{2}$ i $1 - \frac{\alpha}{2}$ rozkładu statystyki (5.92), podane np. w podrozdz. 6.10.

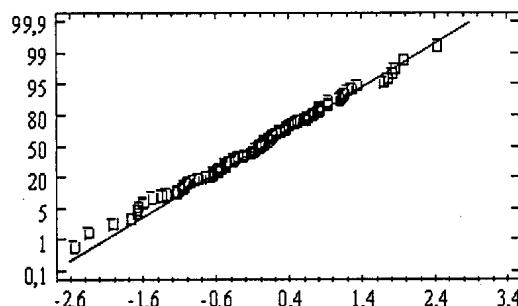
Do testowania normalności wykorzystuje się również standaryzowany współczynnik asymetrii oraz standaryzowaną kurtozę. Współczynniki te, mają w przybliżeniu (przy pewnych założeniach) rozkład normalny. Stąd, jeżeli obliczona wartość standaryzowanego współczynnika asymetrii wykracza poza przedział $(-2, 2)$, oznacza to, iż na poziomie istotności 0.05 możemy odrzucić hipotezę zerową o normalności rozkładu badanej cechy, gdyż rozkład ów nie jest symetryczny. Z kolei gdy obliczona wartość standaryzowanej kurtozy wykracza poza przedział $(-2, 2)$, oznacza to, iż na poziomie istotności 0.05 możemy odrzucić hipotezę zerową o normalności rozkładu badanej cechy, gdyż rozkład ów jest albo zbyt płaski, albo zbyt stromy, albo też ma zbyt "ciękie ogony" (tzn., że jego gęstość zbiega do osi odciętych wolniej niż gęstość rozkładu normalnego).

W praktyce często korzysta się również z tzw. wykresu normalności, pozwalającego na wizualną ocenę danych pod kątem zgodności z rozkładem normalnym. Wykres ten tworzy zbiór punktów $\{(X_{i:n}, Y_i) : i = 1, \dots, n\}$, gdzie $X_{1:n}, \dots, X_{n:n}$ oznacza uporządkowaną rosnącą próbę, natomiast $Y_i = u_{p(i)}$ jest kwantylem rzędu $p(i)$ rozkładu normalnego standardowego,

przy czym

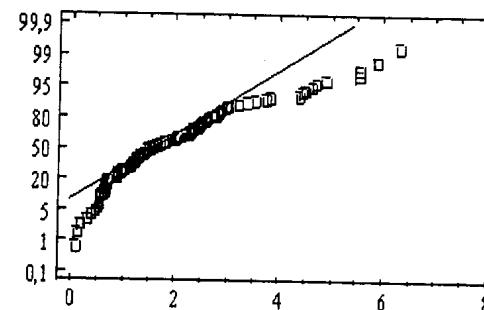
$$p(i) = \frac{i - \frac{3}{8}}{n + \frac{1}{4}}. \quad (5.94)$$

Ten sposób skalowania osi rzędnych sprawia, że wykres dystrybuanty rozkładu normalnego jest linią prostą. Jeżeli więc układ punktów tworzących wykres, reprezentujących jednocześnie dystrybuantę empiryczną badanej cechy, zbliżony jest do linii prostej, świadczy to o zgodności rozkładu badanej cechy z rozkładem normalnym (por. rys. 5.1). I vice versa, jeżeli otrzymany wykres daleki jest od prostej, oznacza to, że rozkład badanej cechy nie jest z pewnością rozkładem normalnym.

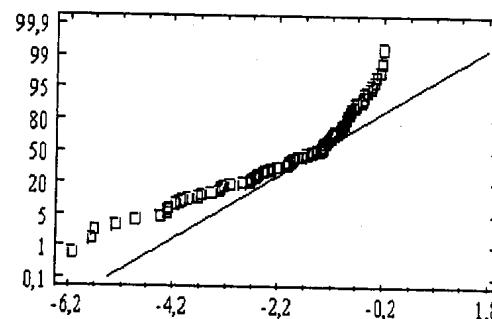


Rys. 5.1 Wykres normalności

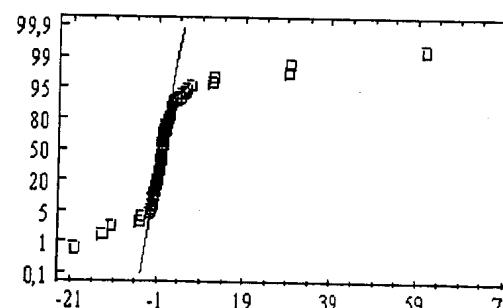
Warto zaznaczyć, że obserwacja wykresu normalności w przypadku wskaźania na brak zgodności z rozkładem normalnym, może doświadczonemu badaczowi dostarczyć i innych cennych informacji o rozkładzie badanej cechy: czy jest to rozkład symetryczny, czy raczej skośny (i o jakiej asymetrii), czy "ogony" rozkładu są "cięzsze" niż w rozkładzie normalnym, itd. Odpowiednie przykłady pokazano na rysunkach: rys. 5.2 – rozkład o skośności dodatniej na wykresie normalności, rys. 5.3 – rozkład o skośności ujemnej, rys. 5.4 – rozkład symetryczny o "ciężkich ogonach" (cięzszych, niż dla rozkładu normalnego), rys. 5.5 – rozkład symetryczny o ograniczonym nośniku.



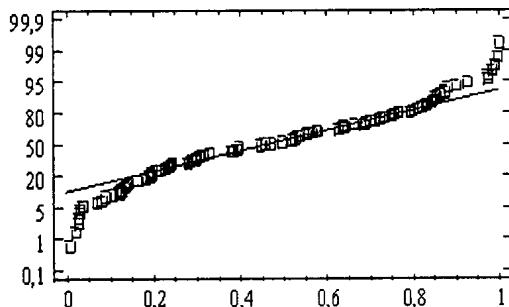
Rys. 5.2 – Rozkład o skośności dodatniej



Rys. 5.3 – Rozkład o skośności ujemnej



Rys. 5.4 – Rozkład o "ciężkich ogonach"



Rys. 5.5 – Rozkład o ograniczonym nośniku

5.8.5 Test Kolmogorowa-Smirnowa

Test Kolmogorowa-Smirnowa stosowany jest do weryfikacji hipotezy

$$H : F_1 = F_2, \quad (5.95)$$

o identyczności rozkładów badanej cechy dla dwóch populacji, wobec hipotezy alternatywnej orzekającej, że rozkłady te istotnie się różnią, tzn.

$$K : F_1 \neq F_2. \quad (5.96)$$

Podobnie jak test Kolmogorowa, wymaga on spełnienia założenia o ciągłości rozkładów.

Niech X_1, \dots, X_{n_1} będzie próbą losową pochodzącą z pierwszej populacji, natomiast Y_1, \dots, Y_{n_2} próbą losową pochodzącą z drugiej populacji. Statystyką testową jest

$$D_{n,m} = \sup_x \left| \hat{F}_{n_1}(x) - \hat{F}_{n_2}(x) \right|, \quad (5.97)$$

gdzie \hat{F}_{n_1} i \hat{F}_{n_2} oznaczają, odpowiednio, dystrybuanty empiryczne wyznaczone na podstawie pierwszej i drugiej próbki. Zbyt duże wartości tej statystyki świadczą przeciw hipotezie H , stąd też obszar krytyczny testu ma postać

$$W_\alpha = [d(\alpha, n_1, n_2), 1], \quad (5.98)$$

gdzie $d(\alpha, n_1, n_2)$ jest wartością krytyczną rozkładu statystyki (5.97). W tablicy zamieszczonej w podrozdz. 6.12 podano wartości krytyczne tego testu pomnożone przez liczności obu prób, tzn. $n_1 n_2 d(\alpha, n_1, n_2)$.

5.8.6 Test Kruskala-Wallisa

Uogólnieniem rozważanego powyżej przypadku testowania hipotezy o identyczności rozkładów badanej cechy dla dwóch populacji jest weryfikacja

hipotezy o identyczności rozkładów k populacji, gdzie $k > 2$, tzn.

$$H : F_1 = F_2 = \dots = F_k, \quad (5.99)$$

wobec hipotezy alternatywnej, że rozkład badanej cechy nie we wszystkich populacjach jest taki sam. Do weryfikacji tej hipotezy stosuje się test Kruskala-Wallisa. Wymaga on spełnienia założenia o ciągłości rozkładów.

Testowanie przebiega następująco. Założymy, że mamy k próbek o liczbach n_1, \dots, n_k , przy czym $\sum_{i=1}^k n_i = n$. Obserwacje pochodzące ze wszystkich k próbek ustawiamy w porządku rosnącym, po czym numerujemy kolejnymi liczbami naturalnymi (nadajemy tzw. rangi). Jeżeli kilka kolejnych wyników ma tę samą wartość, to każdemu z nich przypisujemy rangę będącą średnią arytmetyczną przypisanych im liczb naturalnych. Następnie dla każdej próbki oddzielnie wyznaczamy sumę rang R_i , po czym obliczamy wartość statystyki testowej

$$T = \sum_{i=1}^k \frac{12}{n_i(n-n_i)(n+1)} \left(R_i - \frac{n_i(n+1)}{2} \right)^2. \quad (5.100)$$

Przy założeniu prawdziwości hipotezy (5.99) statystyka (5.100) ma asymptotyczny rozkład chi-kwadrat o $k-1$ stopniach swobody.

Obszar krytyczny ma postać

$$W_\alpha = [\chi_{1-\alpha, k-1}^2, +\infty), \quad (5.101)$$

gdzie $\chi_{1-\alpha, k-1}^2$ oznacza kwantyl rzędu $1-\alpha$ rozkładu chi-kwadrat o $k-1$ stopniach swobody. Tak więc duże wartości statystyki (5.100) świadczą przeciwko hipotezie (5.99).

5.9 Testowanie niezależności

5.9.1 Test niezależności chi-kwadrat

Niejednokrotnie badając pewną populację mamy informacje dotyczące dwóch cech owej populacji i w związku z tym interesuje nas, czy owe cechy są niezależne, czy też występuje między nimi jakaś zależność. Mozemy więc założyć, że naszą próbą jest ciąg par $(X_1, Y_1), \dots, (X_n, Y_n)$, gdzie X_i oraz Y_i oznaczają, odpowiednio, wartości pierwszej i drugiej cechy przyjmowane przez i -ty element próby. Najdogodniejszą formą zapisu wyników takiego badania jest tablica korelacyjna (tablica dwudzielcza) zawierająca tyle wierszy, ile wyróżniamy poziomów pierwszej cechy i tyle kolumn, ile wyróżniamy poziomów drugiej cechy. Założymy, że wyróżniliśmy r różnych poziomów pierwszej cechy i c różnych poziomów drugiej cechy. Nasza tabela korelacyjna ma więc wymiar $r \times c$. Wewnątrz każdej komórki tabeli korelacyjnej, a więc na przecięciu pewnego wiersza i pewnej kolumny, wpisuje się

liczbę tych elementów próby, dla których zaobserwowano wartość poziomu pierwszej cechy odpowiadającą poziomowi tego wiersza i jednocześnie wartość drugiej cechy równą poziomowi danej kolumny.

Po utworzeniu tabeli korelacyjnej możemy przejść do testowania niezależności cech za pomocą tzw. **testu niezależności chi-kwadrat**. Weryfikować będziemy hipotezę

$$H: \text{cechy są niezależne} \quad (5.102)$$

względem hipotezy alternatywnej

$$K: \text{cechy są zależne.} \quad (5.103)$$

Statystyka testowa jest dana wzorem

$$T = \sum_{j=1}^{rc} \frac{(O_j - E_j)^2}{E_j}, \quad (5.104)$$

gdzie O_j oznacza liczbę obserwacji w j -tej komórce tabeli korelacyjnej (sposób numerowania komórek jest dowolny), natomiast E_j jest tzw. oczekiwana liczbą obserwacji, która powinna znaleźć się w j -tej komórce, jeżeli rozpatrywane cechy są w istocie niezależne. Oczekiwana liczbę obserwacji wylicza się dla każdej komórki ze wzoru

$$E_j = \frac{\sum_j^r \cdot \sum_j^c}{n}, \quad (5.105)$$

gdzie \sum_j^r oznacza sumę obserwacji w wierszu, w którym położona jest j -ta komórka, \sum_j^c jest sumą obserwacji w kolumnie, do której należy j -ta komórka, zaś n jest licznością próby.

Przy założeniu prawdziwości hipotezy zerowej oraz dla licznej próby ($n \geq 100$), statystyka (5.104) ma w przybliżeniu rozkład chi-kwadrat o liczbie stopni swobody df

$$df = (r-1)(c-1), \quad (5.106)$$

gdzie r i c oznaczają liczbę wierszy i kolumn w tabeli korelacyjnej.

Ze wzoru (5.104) wynika, że duże wartości statystyki testowej przemawiają przeciwko hipotezie zerowej. Stąd obszar krytyczny ma postać

$$W_\alpha = [\chi_{df, 1-\alpha}^2, +\infty), \quad (5.107)$$

przy czym $\chi_{df, 1-\alpha}^2$ jest kwantylem rozkładu chi-kwadrat rzędu $1-\alpha$ o df stopniach swobody.

(patrz: **Przykład 5.13**)

Warto zauważyć, że testu niezależności chi-kwadrat można używać bez względu na typ badanych cech, tzn. zarówno dla cech jakościowych jak i ilościowych.

Na zakończenie tego podrozdziału rozpatrzmy jeszcze krótko przypadek odrzucenia hipotezy zerowej. Sytuacja ta oznacza, że badane cechy są zależne. Nasuwa się w tym miejscu naturalne pytanie o siłę związku obu cech. W literaturze przedmiotu można znaleźć różne miary siły związku cech zależnych. Są to często miary wykorzystujące wartość statystyki (5.104). Przykładem takiej miary jest współczynnik Cramera.

5.9.2 Test dla współczynnika korelacji rangowej

Czasem mamy do czynienia z danymi niemierzalnymi (a więc jakościowymi), dla których jednak można wprowadzić pewien porządek, np. odpowiadający preferencjom. Gdy obserwacje zostaną uszeregowane zgodnie z przyjętym kryterium porządkowym, możemy im przypisać rangi, czyli numery miejsc zajmowanych w uporządkowanym ciągu. W ten sposób badanie zależności między dwiema cechami można sprowadzić do badania korelacji między rangami.

Załóżmy, że dysponujemy n elementową próbą. Niech R_i oznacza rangę pierwszej cechy przyporządkowaną i -temu elementowi próby, natomiast S_i – rangę drugiej cechy przyporządkowaną i -temu elementowi próby, przy czym $i = 1, \dots, n$. Oznaczmy przez d_i różnicę rang dla i -tego elementu próby, tzn.

$$d_i = R_i - S_i. \quad (5.108)$$

Definicja 66 *Współczynnikiem korelacji rangowej Spearmana nazywamy współczynnik dany wzorem*

$$r_S = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)}. \quad (5.109)$$

Można wykazać, że współczynnik ten przyjmuje wartości z przedziału $[-1, 1]$, przy czym $r_S = 1$ odpowiada idealnej zgodności rang w obu ciągach, natomiast $r_S = -1$ oznacza maksymalną niezgodność rang (tzn. kolejność ustawienia względem jednej cechy jest dokładnie odwrotna niż względem drugiej cechy). Przypadek $r_S = 0$ oznacza brak korelacji rang, a więc w szczególności niezależność cech.
(patrz: **Przykład 5.14**)

Współczynnik korelacji rangowej Spearmana służy do konstrukcji następującego testu niezależności. Podobnie, jak w przypadku testu niezależności chi-kwadrat, testujemy hipotezę zerową

$$H: \text{cechy są niezależne} \quad (5.110)$$

względem hipotezy alternatywnej

$$K : \text{cechy są zależne.} \quad (5.111)$$

Model 1

W przypadku próby o małej liczności, tzn. $8 < n < 30$, korzystamy ze statystyki testowej

$$T = \frac{r_s}{\sqrt{1 - r_s^2}} \sqrt{n - 2}, \quad (5.112)$$

która przy założeniu prawdziwości hipotezy zerowej ma rozkład t-Studenta o $n - 2$ stopniach swobody. Obszar krytyczny ma postać

$$W_\alpha = (-\infty, -t_{1-\frac{\alpha}{2}}^{[n-2]}] \cup [t_{1-\frac{\alpha}{2}}^{[n-2]}, +\infty) \quad (5.113)$$

gdzie $t_{1-\frac{\alpha}{2}}^{[n-2]}$ jest kwantylem rozkładu t-Studenta o $n - 2$ stopniach swobody rzędu $1 - \frac{\alpha}{2}$.

Model 2

Dla prób o większej liczności, tzn. dla $n \geq 30$, jako statystykę testową przyjmujemy po prostu (5.109), czyli

$$T = r_s, \quad (5.114)$$

która przy prawdziwości H ma w przybliżeniu rozkład $N(0, \frac{1}{\sqrt{n-1}})$, a zatem obszar krytyczny dany jest wzorem

$$W_\alpha = (-\infty, -\frac{u_{1-\frac{\alpha}{2}}}{\sqrt{n-1}}] \cup [\frac{u_{1-\frac{\alpha}{2}}}{\sqrt{n-1}}, +\infty), \quad (5.115)$$

gdzie $u_{1-\frac{\alpha}{2}}$ jest kwantylem rozkładu $N(0, 1)$ rzędu $1 - \frac{\alpha}{2}$.

5.10 Testy dla współczynnika korelacji

5.10.1 Test hipotezy o braku korelacji liniowej

Jak już wcześniej wspomniano, miarą zależności liniowej cech ilościowych jest współczynnik korelacji Pearsona. Założmy, że badane cechy mają pewien rozkład dwuwymiarowy o nieznanym współczynniku korelacji ρ . Na podstawie n -elementowej próby losowej $(X_1, Y_1), \dots, (X_n, Y_n)$ chcemy stwierdzić, czy prawdziwe jest przypuszczenie, że badane cechy są nieskorelowane. A zatem weryfikujemy hipotezę zerową

$$H : \rho = 0, \quad (5.116)$$

wobec jednej z alternatyw

$$\begin{aligned} K &: \rho \neq 0, \\ K' &: \rho < 0, \\ K'' &: \rho > 0. \end{aligned} \quad (5.117)$$

Model 1

Założmy, że badane cechy mają dwuwymiarowy rozkład normalny o nieznanym współczynniku korelacji ρ , a liczność próby $n \geq 3$. Jeżeli spełnione są powyższe założenia oraz gdy hipoteza zerowa jest prawdziwa, wtedy statystyka testowa

$$T = \frac{r}{\sqrt{1 - r^2}} \sqrt{n - 2} \quad (5.118)$$

ma rozkład t-Studenta o $n - 2$ stopniach swobody. W zależności od przyjętej hipotezy alternatywnej obszar krytyczny przybiera postać

$$\begin{aligned} W_\alpha &= (-\infty, -t_{1-\frac{\alpha}{2}}^{[n-2]}] \cup [t_{1-\frac{\alpha}{2}}^{[n-2]}, +\infty), \\ W'_\alpha &= (-\infty, -t_{1-\alpha}^{[n-2]}], \\ W''_\alpha &= [t_{1-\alpha}^{[n-2]}, +\infty), \end{aligned} \quad (5.119)$$

gdzie $t_{1-\frac{\alpha}{2}}^{[n-2]}$ i $t_{1-\alpha}^{[n-2]}$ są, odpowiednio, kwantylami rozkładu t-Studenta o $n - 2$ stopniach swobody rzędów $1 - \frac{\alpha}{2}$ i $1 - \alpha$.
(patrz: Przykład 5.15)

Model 2

Założmy, że badane cechy mają dwuwymiarowy rozkład normalny o nieznanym współczynniku korelacji ρ , a liczność próby jest duża, tzn. $n \geq 100$. Można wówczas posłużyć się testem o statystyce testowej

$$T = \frac{r}{1 - r^2} \sqrt{n}, \quad (5.120)$$

mającej, przy prawdziwości hipotezy H , rozkład asymptotyczny normalny standardowy, w związku z czym obszar krytyczny – w zależności od przyjętej hipotezy alternatywnej (K , K' albo K'') – ma postać

$$\begin{aligned} W_\alpha &= (-\infty, -u_{1-\frac{\alpha}{2}}] \cup [u_{1-\frac{\alpha}{2}}, +\infty) \\ W'_\alpha &= (-\infty, -u_{1-\alpha}], \\ W''_\alpha &= [u_{1-\alpha}, +\infty), \end{aligned} \quad (5.121)$$

gdzie $u_{1-\frac{\alpha}{2}}$ i $u_{1-\alpha}$ są, odpowiednio, kwantylami rozkładu normalnego $N(0, 1)$ rzędów $1 - \frac{\alpha}{2}$ i $1 - \alpha$.

Model 3

W przypadku gdy rozkład dwuwymiarowy nie jest znany, ale za to dysponujemy próbą o liczności $n \geq 50$, należy użyć testu o statystyce testowej

$$T = nr^2, \quad (5.122)$$

mającej, przy prawdziwości H , rozkład asymptotyczny chi-kwadrat o jednym stopniu swobody. Jest to jednakże test do weryfikacji wyłącznie hipotezy zerowej H wobec hipotezy alternatywnej K . Obszar krytyczny jest dany wzorem

$$W_\alpha = [\chi_{1,1-\alpha}^2, +\infty), \quad (5.123)$$

gdzie $\chi_{1,1-\alpha}^2$ jest kwantylem rzędu $1 - \alpha$ rozkładu chi-kwadrat o jednym stopniu swobody.

5.10.2 Test dla współczynnika korelacji liniowej

Załóżmy, że badane cechy mają pewien rozkład dwuwymiarowy o nieznanym współczynniku korelacji ρ . Na podstawie n -elementowej próby losowej $(X_1, Y_1), \dots, (X_n, Y_n)$ weryfikujemy hipotezę zerową, że prawdziwa wartość współczynnika korelacji wynosi ρ_0 , tzn.

$$H : \rho = \rho_0, \quad (5.124)$$

wobec jednej z alternatyw

$$\begin{aligned} K &: \rho \neq \rho_0 \\ K' &: \rho < \rho_0 \\ K'' &: \rho > \rho_0. \end{aligned} \quad (5.125)$$

Dodatkowo zakładamy, że wartość estymatora współczynnika korelacji obliczona dla danej próby spełnia warunek $|r| \neq 0$ oraz że liczność próby $n \geq 10$. Jeżeli spełnione są podane założenia oraz gdy hipoteza zerowa jest prawdziwa, wtedy statystyka testowa

$$T = \frac{1}{2} \left(\ln \frac{1+r}{1-r} - \ln \frac{1+\rho_0}{1-\rho_0} \right) \sqrt{n-3}, \quad (5.126)$$

ma w przybliżeniu rozkład normalny $N(0, 1)$. Stąd obszar krytyczny – w zależności od przyjętej hipotezy alternatywnej – ma postać

$$\begin{aligned} W_\alpha &= (-\infty, -u_{1-\frac{\alpha}{2}}] \cup [u_{1-\frac{\alpha}{2}}, +\infty) \\ W'_\alpha &= (-\infty, -u_{1-\alpha}] \\ W''_\alpha &= [u_{1-\alpha}, +\infty). \end{aligned} \quad (5.127)$$

5.11 Przykłady**Przykład 5.1**

Czas rozwiązywania jednego zadania na egzaminie z RPiS jest zmienną losową o rozkładzie normalnym, z odchyleniem standardowym 5 minut. Wykładowca przewiduje na tę czynność 10 minut. Wśród studentów panuje jednak przekonanie, że taki czas jest zbyt krótki. Zmierzono czas rozwiązywania zadania przez wybranych losowo 6 studentów i otrzymano następujące wyniki (w minutach): 17, 8.5, 20, 10.5, 11, 15.5. Na poziomie istotności 0.05 stwierdzić, czy przekonanie studentów jest słuszne.

Rozwiązańe

Czas rozwiązywania zadania ma rozkład normalny ze znanym odchyleniem standardowym, więc korzystać będziemy z modelu 1 opisanego w podrozdziale 5.4.1. Będziemy testować hipotezę zerową $H : \mu = 10$ (średni czas rozwiązywania zadania wynosi 10 minut) względem hipotezy alternatywnej $K'' : \mu > 10$ (średni czas rozwiązywania zadania jest dłuższy niż 10 minut). Statystyka testowa przyjmuje postać (5.10):

$$T = \frac{\bar{X} - \mu_0}{\sigma} \sqrt{n},$$

gdzie $n = 6$, $\mu_0 = 10$, $\sigma = 5$,

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i = \frac{1}{6} (17 + 8.5 + 20 + 10.5 + 11 + 15.5) = 13.75.$$

Stąd

$$T = \frac{13.75 - 10}{5} \sqrt{6} = 1.8371.$$

Obszar krytyczny W''_α odpowiadający hipotezie K'' ma, zgodnie z (5.11), postać

$$W''_\alpha = [u_{1-\alpha}, +\infty).$$

Ponieważ poziom istotności $\alpha = 0.05$, to $1 - \alpha = 0.95$. Z tablic kwantyle rozkładu normalnego odczytujemy: $u_{1-\alpha} = u_{0.95} = 1.64485$. Stąd

$$W''_\alpha = [1.64485, +\infty).$$

Widzimy, że nasza statystyka testowa $T = 1.8371 \in W''_\alpha$, a zatem odrzucamy hipotezę zerową na korzyść hipotezy alternatywnej. Stwierdzamy więc, że przekonanie studentów jest słuszne, tzn. czas podany przez wykładowcę jest zbyt krótki.

Przykład 5.2

Liczba punktów uzyskanych na egzaminie z RPiS przez losowo wybranych studentów przedstawia się następująco:

punkty	liczba osób
0 – 10	6
10 – 20	12
20 – 30	10
30 – 40	3

Zakładamy, że rozkład liczb zdobytych punktów jest normalny.

- a) Czy na podstawie powyższych danych można uznać, że średnia liczba punktów zdobytych przez studentów drugiego roku na egzaminie z RPiS jest mniejsza niż 19 ? Przyjąć poziom istotności 0.01.
- b) Na poziomie istotności 0.01 sprawdzić, czy odchylenie standardowe liczby punktów z egzaminu z RPiS wynosi 10.

Rozwiążanie

Badana liczba punktów z egzaminu z RPiS ma rozkład normalny o nieznanym odchyleniu standardowym, więc będziemy używać modelu 2 opisanego w podręczniku 5.4.1.

a) Będziemy testować hipotezę zerową $H : \mu = 19$ (średnia liczba zdobytych punktów wynosi 19) względem hipotezy alternatywnej $K' : \mu < 19$ (średnia liczba zdobytych punktów jest mniejsza niż 19). Statystykę testową obliczamy ze wzoru (5.12)

$$T = \frac{\bar{X} - \mu_0}{S} \sqrt{n},$$

gdzie $n = 6 + 12 + 10 + 3 = 31$, $\mu_0 = 19$, natomiast \bar{X} oraz S najwygodniej jest policzyć w tabelce:

klasy	n_i	x_i^0	$n_i x_i^0$	$(x_i^0 - \bar{X})^2$	$n_i (x_i^0 - \bar{X})^2$
0 – 10	6	5	30	174.24	1045.44
10 – 20	12	15	180	10.24	122.88
20 – 30	10	25	250	46.24	462.40
30 – 40	3	35	105	282.24	846.72
suma	31		565		2477.44

Stąd

$$\begin{aligned}\bar{X} &= \frac{1}{n} \sum_{i=1}^r n_i x_i^0 = \frac{565}{31} \simeq 18.2, \\ S^2 &= \frac{1}{n-1} \sum_{i=1}^r n_i (x_i^0 - \bar{X})^2 = \frac{2477.44}{30} \simeq 82.58, \\ S &= \sqrt{S^2} = \sqrt{82.58} \simeq 9.0874.\end{aligned}$$

Statystyka testowa jest zatem równa:

$$T = \frac{18.2 - 19}{9.0874} \sqrt{31} = -0.49015,$$

a obszar krytyczny odpowiadający hipotezie alternatywnej K' jest postaci (5.13):

$$W'_\alpha = (-\infty, -t_{1-\alpha}^{[n-1]}].$$

W naszym zadaniu $\alpha = 0.01$, czyli $1 - \alpha = 0.99$ i odczytujemy z tablic kwantylu rozkładu t-Studenta, że $t_{1-\alpha}^{[n-1]} = t_{0.99}^{[30]} = 2.4573$. Stąd

$$W'_\alpha = (-\infty, -2.4573].$$

Ponieważ statystyka testowa $T \notin W'_\alpha$, zatem nie ma podstaw do odrzucenia hipotezy zerowej. W rezultacie stwierdzamy, że średnia liczba punktów uzyskiwanych z egzaminu z RPiS nie jest mniejsza niż 19.

b) Interesuje nas test dla wariancji w celu weryfikacji hipotezy zerowej $H : \sigma^2 = 100$ (odchylenie standardowe σ liczby punktów z egzaminu z RPiS wynosi 10) względem hipotezy alternatywnej $K : \sigma^2 \neq 100$ (odchylenie standardowe liczby punktów z egzaminu z RPiS jest rózne od 10). Statystyka testowa jest dana wzorem (5.47)

$$T = \frac{(n-1) S^2}{\sigma_0^2},$$

gdzie $n = 31$, $\sigma_0^2 = 100$, $S^2 = 82.58$. Stąd

$$T = \frac{(31-1) 82.58}{100} = 24.774.$$

Obszar krytyczny odpowiadający hipotezie alternatywnej K ma postać (5.48):

$$W_\alpha = (0, \chi_{\frac{\alpha}{2}, n-1}^2] \cup [\chi_{1-\frac{\alpha}{2}, n-1}^2, +\infty).$$

U nas $\alpha = 0.01$, a stąd $\frac{\alpha}{2} = 0.005$ i $1 - \frac{\alpha}{2} = 0.995$. Z tablic kwantyli rozkładu χ^2 odczytujemy:

$$\begin{aligned}\chi_{\frac{\alpha}{2}, n-1}^2 &= \chi_{0.005, 30}^2 = 13.7867, \\ \chi_{1-\frac{\alpha}{2}, n-1}^2 &= \chi_{0.995, 30}^2 = 53.6720.\end{aligned}$$

Stąd

$$W_\alpha = (0, 13.7867] \cup [53.6720, +\infty).$$

Ponieważ statystyka testowa $T = 24.774 \notin W_\alpha$, więc nie ma podstaw do odrzucenia hipotezy zerowej. A zatem przyjmujemy, że odchylenie standarde liczby punktów z egzaminu z RPiS wynosi 10.

Przykład 5.3

Porównano średnie ocen uzyskanych w ciągu ostatniego roku przez studentów dwóch równoległych lat informatyki i zarządzania. Dla Wydziału Informatyki średnia wyniosła 4.1 z wariancją 1.8, natomiast dla Wydziału Zarządzania średnia wyniosła 3.6 z wariancją 2. Powyższe wyniki obliczono na podstawie 250 ocen uzyskanych przez studentów informatyki i 200 ocen studentów zarządzania. Na poziomie istotności 0.05 zweryfikować hipotezę, że przeciętne wyniki osiągane przez studentów informatyki są lepsze od przeciętnych wyników studentów zarządzania.

Rozwiązańe

Nie mamy informacji na temat typu rozkładu ocen uzyskiwanych przez studentów w obu populacjach (czyli studentów informatyki i zarządzania). Dysponujemy jednak próbami o dużych liczebnościach ($n_1 = 250$, $n_2 = 200$). Do porównania średnich w tych dwóch populacjach użyjemy zatem modelu 4 opisanego w podrozdziale 5.4.2. Testować będziemy hipotezę zerową $H : \mu_1 = \mu_2$ (przeciętne wyniki osiągane przez studentów obu wydziałów są jednakowe) wobec hipotezy alternatywnej $K'': \mu_1 > \mu_2$ (przeciętne wyniki osiągane przez studentów informatyki są lepsze od wyników studentów zarządzania). Do testowania wykorzystujemy statystykę (5.25):

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}.$$

W naszym przypadku $\bar{X} = 4.1$, $S_1^2 = 1.8$, $\bar{Y} = 3.6$, $S_2^2 = 2$. Stąd

$$T = \frac{4.1 - 3.6}{\sqrt{\frac{1.8}{250} + \frac{2}{200}}} = 3.8125.$$

Obszar krytyczny odpowiadający naszej hipotezie alternatywnej K'' wygląda następująco (5.26):

$$W_\alpha'' = [u_{1-\alpha}, +\infty).$$

U nas $\alpha = 0.05$, a więc $1 - \alpha = 0.95$ i odczytany z tablic rozkładu normalnego kwantyl $u_{1-\alpha} = u_{0.95} = 1.64485$. Zatem

$$W_\alpha'' = [1.64485, +\infty).$$

Ponieważ statystyka testowa $T = 3.8125 \in W_\alpha''$, więc odrzucamy hipotezę zerową na korzyść alternatywnej i stwierdzamy, że studenci informatyki przeciętnie osiągają lepsze wyniki niż studenci zarządzania.

Przykład 5.4

Badano grubość płyt metalowych przed i po obróbce chemicznej. Dla 10 losowo wybranych płyt przed obróbką otrzymano średnią z próby 0.451 mm i wariancję z próby 0.02. Natomiast dla 15 losowo wybranych płyt po obróbce chemicznej otrzymano średnią z próby 0.550 mm z wariancją 0.017. Sprawdzić, czy grubość płyt zmienia się podczas obróbki. Założyć, że grubość płyt przed i po obróbce ma rozkład normalny o tej samej wariancji. Przyjmując poziom istotności 0.05.

Rozwiązańe

W zadaniu tym porównujemy dwie populacje płyt pochodzących z rozkładów normalnych o tej samej wariancji ($\sigma_1^2 = \sigma_2^2$). Interesuje nas porównanie średnich, więc będziemy stosować model 2 opisany w podrozdz. 5.4.2. Testować będziemy hipotezę zerową $H : \mu_1 = \mu_2$ (grubość płyt nie zmienia się podczas obróbki) wobec hipotezy alternatywnej $K : \mu_1 \neq \mu_2$ (grubość płyt zmienia się podczas obróbki). Do testowania wykorzystamy statystykę (5.20)

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \sqrt{n_1 + n_2 - 2},$$

gdzie $n_1 = 10$, $n_2 = 15$, $\bar{X} = 0.451$, $S_1^2 = 0.02$, $\bar{Y} = 0.550$, $S_2^2 = 0.017$. Wstawiając do wzoru otrzymujemy:

$$T = \frac{0.451 - 0.550}{\sqrt{\frac{(10-1)0.02 + (15-1)0.017}{10+15-2} \left(\frac{1}{10} + \frac{1}{15} \right)}} \sqrt{10+15-2} = -1.7988.$$

Obszar krytyczny odpowiadający hipotezie K jest następujący (5.21):

$$W_\alpha = (-\infty, -t_{1-\frac{\alpha}{2}}^{[n_1+n_2-2]}] \cup [t_{1-\frac{\alpha}{2}}^{[n_1+n_2-2]}, +\infty).$$

W naszym zadaniu $\alpha = 0.05$, a stąd $1 - \frac{\alpha}{2} = 0.975$ i odczytany z tablic kwantyl rozkładu t-Studenta $t_{1-\frac{\alpha}{2}}^{[n_1+n_2-2]} = t_{0.975}^{[23]} = 2.0687$. Zatem obszar krytyczny ma postać

$$W_\alpha = (-\infty, -2.0687] \cup [2.0687, +\infty).$$

Widzimy, że statystyka $T = -1.7988 \notin W_\alpha$, a zatem nie ma podstaw do odrzucenia hipotezy zerowej, czyli stwierdzamy, że grubość płyt metalowych nie zmienia się podczas obróbki chemicznej.

Przykład 5.5

Podczas sprawdzianu z ortografią 8 losowo wybranych dzieci popełniło: 1, 3, 2, 7, 6, 5, 4, 8 błędów. Przez miesiąc, drogą licznych dyktandów, ćwiczono ortografię, po czym powtórzono sprawdzian na tej samej grupie dzieci. Tym razem dzieci te popełniły, odpowiednio, następującą liczbę błędów: 0, 1, 3, 5, 5, 3, 2, 4. Zakładamy, że rozkład liczb popełnionych błędów jest normalny. Czy na podstawie przedstawionych danych można stwierdzić, że dyktanda wpływają na poprawę ortografii? Przyjąć poziom istotności 0.01.

Rozwiązańe

W zadaniu tym występują dwie zależne próbki. Możemy na naszą próbę losową patrzeć jak na zbiór uporządkowanych par $(X_1, Y_1), \dots, (X_8, Y_8)$, gdzie X_j oznacza liczbę błędów popełnionych przez j -te dziecko przed serią dyktandów, zaś Y_j - liczbę błędów popełnionych przez to samo dziecko po serii dyktandów, $j = 1, \dots, 8$. Ponieważ wiemy, że liczby błędów mają rozkłady normalne, więc w celu porównania średnich stosujemy model opisany w podrozdziale 5.4.3.

Będziemy testować hipotezę zerową $H : \mu_1 = \mu_2$ (dyktanda nie wpływają na poprawę ortografii) wobec hipotezy alternatywnej $K'' : \mu_1 > \mu_2$ (dyktanda wpływają na poprawę ortografii, tzn. średnia liczba błędów popełnionych przez dzieci przed serią dyktandów jest większa niż po dyktandach). Statystyka testowa jest dana wzorem (5.27)

$$T = \frac{\bar{Z}}{S_Z} \sqrt{n},$$

gdzie \bar{Z} oraz S_Z jest, odpowiednio, średnią i odchyleniem standardowym z próbki Z_1, \dots, Z_n otrzymanej poprzez odejmowanie w parach, tzn. $Z_i = X_i - Y_i$, dla $i = 1, \dots, n$. W naszym przypadku otrzymamy

$$\begin{aligned} Z_1 &= 1 - 0 = 1, \quad Z_2 = 3 - 1 = 2, \quad Z_3 = 2 - 3 = -1, \quad Z_4 = 7 - 5 = 2, \\ Z_5 &= 6 - 5 = 1, \quad Z_6 = 5 - 3 = 2, \quad Z_7 = 4 - 2 = 2, \quad Z_8 = 8 - 4 = 4. \end{aligned}$$

Stąd

$$\begin{aligned} \bar{Z} &= \frac{1}{n} \sum_{i=1}^n Z_i = \frac{1}{8} (1 + 2 - 1 + 2 + 1 + 2 + 2 + 4) = 1.625, \\ S_Z^2 &= \frac{1}{n-1} \sum_{i=1}^n (Z_i - \bar{Z})^2 = \frac{1}{8-1} [(1 - 1.625)^2 + (2 - 1.625)^2 + \\ &\quad + (-1 - 1.625)^2 + (2 - 1.625)^2 + (1 - 1.625)^2 + \\ &\quad + (2 - 1.625)^2 + (2 - 1.625)^2 + (4 - 1.625)^2] = 1.9821, \\ S_Z &= \sqrt{S_Z^2} = \sqrt{1.9821} = 1.4079, \end{aligned}$$

a więc

$$T = \frac{1.625}{1.4079} \sqrt{8} = 3.2646.$$

Obszar krytyczny jest postaci (5.28)

$$W_\alpha'' = [t_{1-\alpha}^{[n-1]}, +\infty).$$

W naszym zadaniu $\alpha = 0.01$, a więc $1 - \alpha = 0.99$ i odczytany z tablic kwantyle rozkładu t-Studenta o $n-1 = 8-1 = 7$ stopniach swobody wynosi $t_{0.99}^{[7]} = 2.998$. Zatem

$$W_\alpha'' = [2.998, +\infty).$$

Jak widać statystyka testowa $T = 3.2646 \in W_\alpha''$, a zatem odrzucamy hipotezę zerową na rzecz hipotezy alternatywnej. Oznacza to, że dyktanda istotnie wpływają na poprawę ortografii.

Przykład 5.6

Pewien metalurg testował rozciągliwość nowego stopu i otrzymał następujące wyniki: 122.66, 119.97, 119.36, 120.19, 120.02, 121.14, 119.33, 119.13, 121.35, 119.48, 119.78, 123.95, 119.51, 125.96, 121.32.

Spodziewana, na podstawie rozważań teoretycznych, rozciągliwość tego stopu wynosi 120. Czy otrzymane rezultaty potwierdzają te oczekiwania? Przyjąć poziom istotności $\alpha = 0.05$.

Rozwiązańe

W zadaniu tym nie mamy informacji, że próbka pochodzi z rozkładu normalnego. Ponadto mamy małą liczbę obserwacji (nie jest spełniony warunek $n \geq 100$). Zatem do sprawdzenia, czy rozciągliwość nowego stopu rzeczywiście wynosi 120 nie możemy zastosować testu dla średniej. Użyjemy zatem testu znaków, w którym zamiast średniej bada się medianę.

Stawiamy następujące hipotezy:

$$\begin{aligned} H &: \text{med} = 120 \\ K &: \text{med} \neq 120. \end{aligned}$$

Statystykę testową wyliczamy ze wzoru (5.31):

$$T = \sum_{i=1}^n I(X_i - m_0 > 0)$$

U nas $m_0 = 120$, w związku z czym otrzymujemy

$$\begin{aligned} x_1 - m_0 &= 122.66 - 120 = 2.66 > 0 \Rightarrow I(x_1 - m_0 > 0) = 1, \\ x_2 - m_0 &= 119.97 - 120 = -0.03 < 0 \Rightarrow I(x_2 - m_0 > 0) = 0, \\ x_3 - m_0 &= 119.36 - 120 = -0.64 < 0 \Rightarrow I(x_3 - m_0 > 0) = 0, \\ x_4 - m_0 &= 120.19 - 120 = 0.19 > 0 \Rightarrow I(x_4 - m_0 > 0) = 1 \\ &\dots \end{aligned}$$

W rezultacie

$$T = 1 + 0 + 0 + 1 + 1 + 0 + 0 + 1 + 0 + 0 + 1 + 0 + 1 + 1 = 8.$$

Obszar krytyczny ma postać (wzór (5.34)):

$$W_\alpha = [0, r_{\frac{\alpha}{2}}] \cup [n - r_{\frac{\alpha}{2}}, n].$$

Wartość krytyczną $r_{\frac{\alpha}{2}}$ wyznaczamy z tablicy zamieszczonej w podrozdz. 6.6. Skoro $\alpha = 0.05$, to $\frac{\alpha}{2} = 0.025$ i odczytana wartość dla $n = 12$ wynosi $r_{\frac{\alpha}{2}} = 2$. Stąd

$$\begin{aligned} W_\alpha &= [0, 2] \cup [12 - 2, 12] \\ &= [0, 2] \cup [10, 12]. \end{aligned}$$

Ponieważ $T = 8 \notin W_\alpha$, to przyjmujemy hipotezę H i stwierdzamy, że rozciągliwość nowego stopu faktycznie wynosi 120.

Przykład 5.7

Asystent porównuje szybkość pisania programu przez chłopców i dziewczęta. Napisanie programu zajęło 8 losowo wybranym chłopcom, odpowiednio, 13.5, 22, 29, 16.5, 21, 17, 20, 12.5 minut, podczas gdy 7 losowo wybranym dziewczętom zajęło, odpowiednio, 15, 28, 27.5, 23, 22.5, 26, 19 minut. Czy prawdziwe jest przypuszczenie, że chłopcom napisanie tego typu programu zajmuje mniej czasu, aniżeli dziewczętom? Zweryfikować stosowną hipotezę na poziomie istotności $\alpha = 5\%$.

Rozwiązanie

Ponieważ nie mamy informacji o rozkładzie czasu i nasze próbki nie są duże, zamiast średnich porównywać będziemy mediany. Z uwagi na to, że próby X_1, \dots, X_n oraz Y_1, \dots, Y_m są niezależne zastosujemy test Wilcoxon'a. Ponieważ w podrozdz. 5.5.2 założono, że $n \leq m$, więc X -y odpowiadają będą pomiarom czasu w grupie dziewcząt, natomiast Y -ki – w grupie chłopców. Weryfikować będziemy hipotezę zerową $H : \text{med}_X = \text{med}_Y$, mówiącą, że chłopcom i dziewczętom napisanie tego typu programu zajmuje przeciętnie tyle samo czasu, wobec hipotezy alternatywnej $K' : \text{med}_X > \text{med}_Y$, że chłopcom napisanie programu zajmuje przeciętnie mniej czasu niż dziewczętom.

Statystyka testowa dana jest wzorem (5.40)

$$T = \sum_{i=1}^n R_i.$$

Aby wyznaczyć rangi R_i obserwacje z obu prób ustawiamy w jeden ciąg niemalejący, pamiętając, z której próbki pochodzą:

numer pozycji	1	2	3	4	5	6	7	8
wartość	12.5	13.5	15	16.5	17	19	20	21
obserwacja	Y	Y	X	Y	Y	X	Y	Y

numer pozycji	9	10	11	12	13	14	15
wartość	22	22.5	23	26	27.5	28	29
obserwacja	Y	X	X	X	X	X	Y

Patrząc teraz na numery porządkowe poszczególnych obserwacji z pierwszej próby (dla dziewcząt) odczytujemy odpowiadające im rangi, tzn. 3, 6, 10, 11, 12, 13, 14. Mamy zatem:

$$T = 3 + 6 + 10 + 11 + 12 + 13 + 14 = 69.$$

Obszar krytyczny ma postać (wzór (5.41)):

$$W''_\alpha = [w'_{n,m}(2\alpha), +\infty).$$

Wartość krytyczną $w_{2\alpha}$ rozkładu statystyki T odczytujemy z tablic zamieszczonych w podrozdz. 6.8. A konkretnie, dla $n = 7$ i $m = 8$ otrzymujemy $w_{7,8}(2\alpha) = w_{7,8}(0.1) = 44$. Stąd

$$w'_{n,m}(2\alpha) = n(n+m+1) - w_{n,m}(2\alpha) = 112 - 44 = 68.$$

Mamy więc

$$W''_\alpha = [68, +\infty),$$

a zatem $T = 69 \in W''_\alpha$, co oznacza, że odrzucamy hipotezę H , czyli można twierdzić, że napisanie tego typu programu zajmuje chłopcom istotnie mniej czasu niż dziewczętom.

Przykład 5.8

Losową grupę 15 kobiet poddano 6 tygodniowej diecie odchudzającej. Uzyskano następujące wyniki (waga przed i po kuracji w kilogramach):

przed kuracją	88	69.1	86.5	59.3	57.7	82.1	94.9	92.8
po kuracji	74.9	68.1	75.9	55	52.8	83.3	85.1	85.7
przed kuracją	64	91.5	86	59.2	91.3	60.4	58	
po kuracji	65	85.5	77.7	57.9	90.1	58.3	59.3	

Czy wyniki te potwierdzają skuteczność diety? Przyjąć poziom istotności 0.05.

Rozwiązaanie

W zadaniu tym mamy obserwacje parami zależne. Nie mamy informacji, że rozkład wagi jest rozkładem normalnym i próbki są tylko 15 elementowe. Weryfikować będziemy hipotezę zerową $H : \text{med}_{X-Y} = 0$ (brak efektów diety) wobec hipotezy alternatywnej $K' : \text{med}_{X-Y} > 0$ (waga przed kuracją była większa niż po kuracji odchudzającej).

Do weryfikacji tak postawionej hipotezy zastosujemy test znaków o statystyce testowej

$$T = \sum_{i=1}^n I(X_i - Y_i > 0).$$

W naszym przypadku

$$X_1 - Y_1 = 88 - 74.9 = 13.1 > 0 \Rightarrow I(X_1 - Y_1 > 0) = 1$$

$$X_2 - Y_2 = 69.1 - 68.1 = 1 > 0 \Rightarrow I(X_2 - Y_2 > 0) = 1$$

.....

$$X_6 - Y_6 = 82.1 - 83.3 = -1.2 < 0 \Rightarrow I(X_6 - Y_6 > 0) = 0$$

.....

W rezultacie otrzymamy

$$T = 1 + 1 + 1 + 1 + 1 + 0 + 1 + 1 + 0 + 1 + 1 + 1 + 1 + 1 + 0 = 12.$$

Obszar krytyczny ma postać:

$$W''_\alpha = [n - r_\alpha, n].$$

Wartość krytyczną r_α odczytujemy z tablicy (podrozdz. 6.6). Dla próbki o liczności $n = 15$ oraz poziomu istotności $\alpha = 0.05$ otrzymujemy $r_\alpha = 3$. Stąd

$$W''_\alpha = [15 - 3, 15] = [12, 15].$$

Ponieważ $T = 12 \in W''_\alpha$, to odrzucamy hipotezę H i stwierdzamy, że dieta jest skuteczna.

Przykład 5.9

Fabryka zakupiła nowy agregat. Producent zapewnia, że przeciętnie tylko 1 na 100 wyprodukowanych przez ten agregat detali jest wadliwy. Aby to sprawdzić, wylosowano 500 detali i okazało się, że 20 z nich nie spełnia normy jakości. Czy na podstawie takiego wyniku badań można obalić zapewnienie producenta aggregatu? Przyjąć poziom istotności 0.05.

Rozwiązaanie

W zadaniu tym mamy model dwupunktowy: detal wyprodukowany przez agregat jest albo wadliwy, albo dobry. Będziemy zatem weryfikować hipotezę o wskaźniku struktury dla jednej populacji. Stawiamy następującą hipotezę zerową $H : p = 0.01$ (tylko jeden na sto wyprodukowanych przez agregat detali jest wadliwy) przeciw hipotezie alternatywnej $K' : p > 0.01$ (więcej niż jeden na sto detali jest wadliwy). Ponieważ mamy próbke o dużej liczności to statystykę testową wyliczamy ze wzoru (5.64a)

$$T = \frac{k - np_0}{\sqrt{np_0(1 - p_0)}},$$

gdzie $k = 20$, $n = 500$, $p_0 = 0.01$. Stąd

$$T = \frac{20 - 500 \cdot 0.01}{\sqrt{500 \cdot 0.01 \cdot (1 - 0.01)}} = 6.742.$$

Obszar krytyczny odpowiadający hipotezie alternatywnej K' jest postaci (5.65):

$$W''_\alpha = [u_{1-\alpha}, +\infty).$$

Poziom istotności $\alpha = 0.05$, a więc $1 - \alpha = 0.95$ i z tablic kwatylów rozkładu normalnego odczytujemy, że $u_{1-\alpha} = u_{0.95} = 1.64485$. Zatem

$$W''_\alpha = [1.64485, +\infty).$$

Ponieważ nasza statystyka testowa $T = 6.742 \in W''_\alpha$, więc odrzucamy hipotezę zerową na rzecz alternatywnej, co oznacza, że zapewnienia producenta aggregatu nie są słusze i rzeczywista wadliwość aggregatu jest większa od zapowiadanej.

Przykład 5.10

20 spośród 100 losowo wybranych studentów studiów dziennych i 40 spośród 120 losowo wybranych studentów studiów zaocznych zdało egzamin z RPiS w pierwszym terminie. Czy na podstawie powyższych danych można stwierdzić, że studenci studiów zaocznych poważniej potraktowali ten ciekawy przedmiot? Przyjąć poziom istotności 0.1.

Rozwiążanie

W zadaniu tym mamy dwie populacje: populację słuchaczy studiów dziennych i studiów zaocznych. Każdy student ma dwie możliwości: może zdać egzamin z RPiS w pierwszym terminie lub nie, czyli mamy do czynienia z rozkładem dwupunktowym. Interesuje nas zatem weryfikacja hipotezy o równości wskaźników struktury dwóch populacji. Mamy duże liczności obu prób ($n_1 = 100$, $n_2 = 120$). Korzystać będziemy więc z modelu opisanego w podrozdziale 5.7.2.

Oznaczmy przez p_1 i p_2 odsetek studentów studiów dziennych i zaocznych, odpowiednio, zdających egzamin z RPiS w pierwszym terminie. Będziemy weryfikować hipotezę zerową $H : p_1 = p_2$ (czyli studenci studiów dziennych i zaocznych traktują egzamin tak samo poważnie, bo procent zdających go w pierwszym terminie jest w obu populacjach taki sam) wobec hipotezy alternatywnej $K' : p_1 < p_2$ (czyli studenci zaoczni potraktowali egzamin poważniej niż dzienni). Statystyka testowa jest postaci (5.70)

$$T = \frac{\frac{k_1}{n_1} - \frac{k_2}{n_2}}{\sqrt{\frac{k}{n}(1 - \frac{k}{n})(\frac{1}{n_1} + \frac{1}{n_2})}},$$

gdzie $k_1 = 20$, $n_1 = 100$, $k_2 = 40$, $n_2 = 120$, $k = k_1 + k_2 = 20 + 40 = 60$, $n = n_1 + n_2 = 100 + 120 = 220$. Zatem

$$T = \frac{\frac{20}{100} - \frac{40}{120}}{\sqrt{\frac{60}{220}(1 - \frac{60}{220})(\frac{1}{100} + \frac{1}{120})}} = -2.2049.$$

Obszar krytyczny odpowiadający hipotezie alternatywnej K' jest postaci (5.71)

$$W'_\alpha = (-\infty, -u_{1-\alpha}].$$

Ponieważ $\alpha = 0.1$, to $1 - \alpha = 0.9$ i z tablic kwantylami rozkładu normalnego odczytujemy, że $u_{1-\alpha} = u_{0.9} = 1.28155$. Stąd

$$W'_\alpha = (-\infty, -1.28155].$$

Nasza statystyka testowa $T = -2.2049 \in W'_\alpha$, więc odrzucamy hipotezę zerową na rzecz alternatywnej. Czyli stwierdzamy, że studenci studiów zaocznych poważniej potraktowali egzamin z RPiS.

Przykład 5.11

Badania grupy krwi 200 osób dały następujące wyniki: grupę A miało 37 osób, grupę B - 52 osoby, grupę AB - 66 osób, natomiast grupę 0 miało 45 osób. Czy na podstawie tych wyników można przyjąć hipotezę o równomiernym rozkładzie wszystkich grup krwi? Przyjąć poziom istotności 0.05.

Rozwiążanie

Ponieważ chcemy sprawdzić, czy badany rozkład grup krwi jest zgodny z rozkładem równomiernym, będziemy stosować test zgodności chi-kwadrat opisany w podrozdziale 5.8.2. Możemy skorzystać z tego testu, bowiem dysponujemy dużą próbą ($n = 200$). Sformułujmy hipotezę zerową i alternatywną:

H : rozkład grup krwi jest równomierny,

K : rozkład grup krwi nie jest równomierny.

Statystykę testową wyliczamy ze wzoru (5.79)

$$T = \sum_{i=1}^k \frac{(n_i - np_i)^2}{np_i}.$$

Dla uproszczenia rachunków zróbcmy tabelkę:

grupa krwi	n_i	p_i	np_i	$(n_i - np_i)^2$	$\frac{(n_i - np_i)^2}{np_i}$
A	37	$\frac{1}{4}$	50	169	3.38
B	52	$\frac{1}{4}$	50	4	0.08
AB	66	$\frac{1}{4}$	50	256	5.12
0	45	$\frac{1}{4}$	50	25	0.5
suma	200				9.08

Przyjęliśmy, że $p_i = \frac{1}{4}$ dla $i = 1, 2, 3, 4$, ponieważ przy założeniu, że rozkład grup krwi jest równomierny, prawdopodobieństwo natrafienia na osobę z określona grupą krwi jest równe $\frac{1}{4}$ (mamy cztery grupy krwi). W rezultacie otrzymujemy

$$T = 9.08.$$

Obszar krytyczny testu ma postać (5.80):

$$W_\alpha = [\chi^2_{1-\alpha, k-1}, +\infty).$$

U nas $1 - \alpha = 1 - 0.05 = 0.95$, natomiast liczba stopni swobody wynosi $k - 1 = 4 - 1 = 3$. Z tablic kwantylami rozkładu χ^2 odczytujemy, że $\chi^2_{1-\alpha, k-1} = \chi^2_{0.95, 3} = 7.8147$. Zatem mamy następujący obszar krytyczny

$$W_\alpha = [7.8147, +\infty).$$

Ponieważ nasza statystyka testowa $T = 9.08 \in W_\alpha$, to odrzucamy hipotezę zerową H na rzecz alternatywnej K . Zatem stwierdzamy, że rozkład grup krwi nie jest rozkładem równomiernym.

Przykład 5.12

W pewnym zakładzie przeprowadzono badanie absencji pracowników. Przez kolejnych 300 dni zapisywano liczbę osób, które nie przyszły danego dnia do pracy. Otrzymano następujące dane:

liczba opuszczonych dni	liczba osób
0	50
1	100
2	80
3	40
4	20
5	10

Na poziomie istotności 0.05 zweryfikować hipotezę, że rozkład liczby nieobecnych osób jest rozkładem Poissona.

Rozwiązańe

Przez X oznaczmy liczbę osób nieobecnych w pracy danego dnia. Chcemy sprawdzić, czy zmienna losowa X ma rozkład Poissona, czyli musimy zastosować test zgodności. Ponieważ rozkład Poissona jest rozkładem dyskretnym, a my dysponujemy liczną próbą ($n = 300$), dlatego też posłużymy się ponownie testem zgodności chi-kwadrat.

Weryfikować będziemy hipotezę zerową H wobec hipotezy alternatywnej K , gdzie

H : rozkład liczby nieobecnych osób jest rozkładem Poissona,

K : rozkład liczby nieobecnych osób nie jest rozkładem Poissona.

Jeśli zmienna losowa X ma rozkład Poissona z parametrem λ to

$$p_i = P(X = i) = \frac{\lambda^i}{i!} e^{-\lambda}, \quad i = 0, 1, 2, \dots$$

W naszym przypadku nie znamy, niestety, parametru λ . Wiadomo jednak, że dla rozkładu Poissona wartość oczekiwana równa jest właśnie parametrowi rozkładu, tzn. $EX = \lambda$. A zatem oszacujmy λ za pomocą średniej z próby \bar{X} (która jest estymatorem największej wiarogodności wartości oczekiwanej w rozkładzie Poissona). Oznaczmy przez x_i liczbę osób nieobecnych danego dnia w pracy, zaś przez n_i liczbę dni, w których do pracy nie przyszło x_i osób. Stąd

$$\begin{aligned}\hat{\lambda} &= \bar{X} = \frac{1}{n} \sum_{i=1}^r n_i x_i = \\ &= \frac{1}{300} (50 \cdot 0 + 100 \cdot 1 + 80 \cdot 2 + 40 \cdot 3 + 20 \cdot 4 + 10 \cdot 5) = \\ &= 1.7\end{aligned}$$

Do obliczenia statystyki testowej (5.79):

$$T = \sum_{i=1}^k \frac{(n_i - np_i)^2}{np_i}$$

wykorzystamy tabelkę:

x_i	n_i	p_i	np_i	$(n_i - np_i)^2$	$\frac{(n_i - np_i)^2}{np_i}$
0	50	0.18	54	16	0.30
1	100	0.31	93	49	0.53
2	80	0.26	78	4	0.05
3	40	0.15	45	25	0.56
4	20	0.06	18	4	0.22
≥ 5	10	0.04	12	4	0.33
suma	300				1.99

gdzie kolejne wartości p_i obliczyliśmy następująco:

$$p_1 = P(X = 0) = \frac{1.7^0}{0!} e^{-1.7} \simeq 0.18,$$

$$p_2 = P(X = 1) = \frac{1.7^1}{1!} e^{-1.7} \simeq 0.31,$$

$$p_3 = P(X = 2) = \frac{1.7^2}{2!} e^{-1.7} \simeq 0.26,$$

$$p_4 = P(X = 3) = \frac{1.7^3}{3!} e^{-1.7} \simeq 0.15,$$

$$p_5 = P(X = 4) = \frac{1.7^4}{4!} e^{-1.7} \simeq 0.06,$$

$$p_6 = P(X \geq 5) = 1 - P(X < 5) \simeq 0.04.$$

Stąd

$$T = \sum_{i=1}^6 \frac{(n_i - np_i)^2}{np_i} \simeq 1.99.$$

Obszar krytyczny wyznaczamy ze wzoru (5.84)

$$W_\alpha = [\chi^2_{1-\alpha, r-1-k}, +\infty).$$

W naszym przypadku: $1 - \alpha = 1 - 0.05 = 0.95$, $r = 6$ (liczba klas), $k = 1$ (liczba estymowanych parametrów – my wyliczaliśmy tylko λ), liczba stopni swobody $r-1-k = 6-1-1 = 4$. Z tablic kwantyle rozkładu χ^2 odczytujemy, że $\chi^2_{1-\alpha, r-1-k} = \chi^2_{0.95, 4} = 9.4877$. Zatem obszar krytyczny jest postaci

$$W_\alpha = [9.4877, +\infty).$$

Ponieważ nasza statystyka testowa $T = 1.99 \notin W_\alpha$, to nie mamy podstaw do odrzucenia hipotezy zerowej H . Zatem stwierdzamy, że liczbę osób nieobecnych danego dnia w pracy można modelować rozkładem Poissona.

Przykład 5.13

W celu zbadania, czy istnieje związek między kolorem oczu i kolorem włosów przeprowadzono badanie na losowej grupie osób i otrzymano następujące wyniki:

	niebieski kolor oczu	inny kolor oczu
włosy jasne	67	32
włosy ciemne	53	68

Zweryfikować odpowiednią hipotezę na poziomie istotności 0.01.

Rozwiązanie

Interesuje nas, czy istnieje związek między dwiema cechami: kolorem włosów i kolorem oczu. W celu rozstrzygnięcia dylematu zweryfikujemy hipotezę zerową

H : kolor oczu i kolor włosów są cechami niezależnymi,

wobec hipotezy alternatywnej

K : kolor włosów i kolor oczu są cechami zależnymi.

Do weryfikacji hipotezy o niezależności posłużymy się testem niezależności chi-kwadrat, opisanym w podrozdziale 5.9.1.

Nasze dane tworzą tablicę korelacyjną, w której – dodatkowo – zamieściłyśmy liczności brzegowe (\sum_j^r oznacza sumę obserwacji w wierszu, w którym położona jest j -ta komórka, \sum_j^c jest sumą obserwacji w kolumnie, do której należy j -ta komórka):

	niebieski kolor oczu	inny kolor oczu	\sum_j^r
włosy jasne	67	32	99
włosy ciemne	53	68	121
\sum_j^c	120	100	220

Statystyka testowa dana jest wzorem (5.104)

$$T = \sum_{j=1}^{rc} \frac{(O_j - E_j)^2}{E_j},$$

gdzie O_j oznacza liczbę obserwacji w j -tej komórce tabeli korelacyjnej, natomiast oczekiwane liczby obserwacji E_j wyliczamy dla każdej komórki ze wzoru (5.105)

$$E_j = \frac{\sum_j^r \cdot \sum_j^c}{n}.$$

W celu obliczenia wartości statystyki testowej wygodnie posłużyć się tabelką

O_j	E_j	$(O_j - E_j)^2$	$\frac{(O_j - E_j)^2}{E_j}$
67	54	169	3.13
32	45	169	3.76
53	66	169	2.56
68	55	169	3.07
		suma	12.52

W rezultacie otrzymujemy

$$T = 12.52.$$

Obszar krytyczny ma postać (5.107)

$$W_\alpha = [\chi_{df, 1-\alpha}^2, +\infty),$$

przy czym $\chi_{df, 1-\alpha}^2$ jest kwantylem rozkładu chi-kwadrat rzędu $1 - \alpha$ o $df = (r - 1)(c - 1)$ stopniach swobody. W naszym przypadku $1 - \alpha = 1 - 0.01 = 0.99$, natomiast liczba stopni swobody wynosi $df = (2 - 1)(2 - 1) = 1$. Z tablic kwantylem rozkładu χ^2 odczytujemy, że $\chi_{0.99, 1}^2 = 6.6349$ i stąd

$$W_\alpha = [6.6349, +\infty).$$

Ponieważ statystyka testowa $T = 12.519 \in W_\alpha$, więc odrzucamy hipotezę zerową H na korzyść hipotezy K . A zatem stwierdzamy, że istnieje związek między kolorem włosów i kolorem oczu.

Przykład 5.14

Interesuje nas, czy istnieje zależność pomiędzy preferencjami odnośnie przedmiotów wykładanych na II roku studiów informatycznych. Bolek i Lolek poproszeni o uporządkowanie przedmiotów – zaczynając od najbardziej przez siebie ulubionego, a kończąc na tym, który najmniej ich interesuje – podali co następuje:

Bolek	RPS	MD	WSO	BD	C++	A
Lolek	RPS	MD	C++	WSO	BD	A

przy czym RPS oznacza Rachunek Prawdopodobieństwa i Statystykę, MD – Matematykę Dyskretną, WSO – Wielodostępne Systemy Operacyjne, BD – Bazy Danych, zaś A – Język angielski.

Czy istnieje zależność między preferencjami obu chłopców? Jeżeli tak, to podać wartość miary tej współzależności.

Rozwiązanie

Niech R_i oznacza rangi nadane poszczególnym przedmiotom przez Bolka, natomiast S_i – rangi nadane przedmiotom przez Lolkę. Ponadto, niech $d_i = R_i - S_i$ oznacza różnicę rang.

przedmioty	RPS	MD	WSO	BD	C++	A
R_i	1	2	3	4	5	6
S_i	1	2	4	5	3	6
d_i	0	0	-1	-1	2	0

Współczynnik korelacji rangowej Spearmana wyliczamy ze wzoru (5.109)

$$\begin{aligned} r_s &= 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)} = \\ &= 1 - \frac{6[0^2 + 0^2 + (-1)^2 + (-1)^2 + 2^2 + 0^2]}{6(6^2 - 1)} = \\ &\simeq 1 - 0.17143 = 0.82857. \end{aligned}$$

Widzimy więc, że współczynnik korelacji jest dość bliski 1. Zatem istnieje całkiem silna zależność między preferencjami Bolka i Lolka odnośnie przedmiotów wykładanych na II roku studiów informatycznych.

Przykład 5.15

Wylosowano 10 par zawierających związek małżeński i otrzymano następujące dane o wieku (w latach):

wiek kobiety	28	24	29	27	33	29	19	22	21	23
wiek mężczyzny	33	28	30	30	35	41	22	25	26	26

- a) Wyestymować wartość współczynnika korelacji liniowej.
- b) Czy na poziomie istotności 2% można twierdzić, że wiek kobiety i wiek mężczyzny zawierających małżeństwo są skorelowane?

Rozwiązańe

a) Niech X oznacza wiek kobiety, zaś Y wiek mężczyzny. Empiryczny współczynnik korelacji liniowej między wiekiem kobiety i wiekiem mężczyznę zawierających małżeństwo obliczamy według wzoru (3.31)

$$\begin{aligned} r &= \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 \sum_{i=1}^n (Y_i - \bar{Y})^2}} \\ &= \frac{\frac{1}{n} \sum_{i=1}^n X_i Y_i - \bar{X} \bar{Y}}{\sqrt{\left(\frac{1}{n} \sum_{i=1}^n X_i^2 - \bar{X}^2\right) \left(\frac{1}{n} \sum_{i=1}^n Y_i^2 - \bar{Y}^2\right)}}, \end{aligned}$$

gdzie $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ i $\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$. W naszym przypadku otrzymujemy

$$\begin{aligned} \bar{X} &= \frac{1}{n} \sum_{i=1}^n X_i = \frac{1}{10} (28 + 24 + 29 + 27 \\ &\quad + 33 + 29 + 19 + 22 + 21 + 23) \\ &= 25.5, \\ \bar{Y} &= \frac{1}{n} \sum_{i=1}^n Y_i = \frac{1}{10} (33 + 28 + 30 + 30 \\ &\quad + 35 + 41 + 22 + 25 + 26 + 26) \\ &= 29.6, \\ \frac{1}{n} \sum_{i=1}^n X_i Y_i &= \frac{1}{10} (28 \cdot 33 + 24 \cdot 28 + 29 \cdot 30 + 27 \cdot 30 + 33 \cdot 35 \\ &\quad + 29 \cdot 41 + 19 \cdot 22 + 22 \cdot 25 + 21 \cdot 26 + 23 \cdot 26) \\ &= 773.2, \\ \frac{1}{n} \sum_{i=1}^n X_i^2 &= \frac{1}{10} (28^2 + 24^2 + 29^2 + 27^2 \\ &\quad + 33^2 + 29^2 + 19^2 + 22^2 + 21^2 + 23^2) \\ &= 667.5, \\ \frac{1}{n} \sum_{i=1}^n Y_i^2 &= \frac{1}{10} (33^2 + 28^2 + 30^2 + 30^2 \\ &\quad + 35^2 + 41^2 + 22^2 + 25^2 + 26^2 + 26^2) \\ &= 904.0. \end{aligned}$$

Stąd

$$r = \frac{773.2 - 25.5 \cdot 29.6}{\sqrt{(667.5 - 25.5^2)(904 - 29.6^2)}} \simeq 0.83963.$$

b) Aby sprawdzić czy wiek kobiety i mężczyzny zawierających małżeństwo są w istotny sposób skorelowane, stawiamy następujące hipotezy: $H: \rho = 0$ (brak korelacji liniowej) przy hipotezie alternatywnej $K: \rho \neq 0$ (istnieje liniowa zależność). Statystykę testową obliczamy ze wzoru (5.118)

$$T = \frac{r}{\sqrt{1 - r^2}} \sqrt{n - 2}.$$

Zatem

$$T = \frac{0.83963}{\sqrt{1 - 0.83963^2}} \sqrt{10 - 2} = 4.3723.$$

Zbiór krytyczny jest postaci (5.119):

$$W_\alpha = (-\infty, -t_{1-\frac{\alpha}{2}}^{[n-2]}] \cup [t_{1-\frac{\alpha}{2}}^{[n-2]}, +\infty),$$

gdzie $\alpha = 0.02$, stąd $1 - \frac{\alpha}{2} = 0.99$ i odczytany z tablic rozkładu t-Studenta kwantyl $t_{1-\frac{\alpha}{2}}^{[n-2]} = t_{0.99}^{[8]} = 2.8965$. Stąd

$$W_\alpha = (-\infty, -2.8965] \cup [2.8965, +\infty).$$

Ponieważ $T = 4.3723 \in W_\alpha$, więc odrzucamy hipotezę zerową na rzecz alternatywnej, a zatem stwierdzamy, że istnieje istotna korelacja liniowa między wiekiem kobiety i wiekiem mężczyzny zawierających małżeństwo. Co więcej, patrząc na empiryczny współczynnik korelacji ($r \approx 0.83963$) możemy powiedzieć, że korelacja ta jest stosunkowo silna.

5.12 Zadania

Zadanie 5.1

Telewizja podała, że pewien program cieszy się zainteresowaniem aż 75% telewidzów. Na 2200 losowo wybranych telewidzów 1386 potwierdziło zainteresowanie owym programem. Na poziomie istotności 0.05 stwierdzić, czy podana ocena oglądalności owego programu jest wiarygodna.

Zadanie 5.2

Na 200 połączeń telefonicznych w pewnej centrali 14 okazało się błędnych. Na poziomie istotności 0.05 zweryfikować hipotezę, że 6% połączeń w tej centrali jest błędnych.

Zadanie 5.3

Dział kontroli jakości w zakładach chemicznych chce oszacować średnią wagę proszku do prania sprzedawanego w pudełkach o nominalnej wadze 3 kg. Pobrano w tym celu próbki losową 7 pudełek proszku do prania. Każde pudełko zważono i otrzymano następujące wyniki (w kilogramach): 2.93, 2.97, 3.05, 2.91, 3.02, 2.87, 2.92. Wiadomo, że rozkład wagi pudełka proszku do prania jest normalny. Zweryfikować przypuszczenie, że średnia waga pudełka proszku do prania jest mniejsza niż 3 kg.

Zadanie 5.4

Pobrano 2 losowe próbki ziaren dwóch gatunków fasoli i zmierzono ich długość. Dla 450 ziaren pierwszego gatunku otrzymano średnią długość ziarna 12.3 mm oraz odchylenie standardowe 1.8 mm. Natomiast dla 500 elementowej próbki ziaren drugiego gatunku otrzymano średnią długość ziarna 11.9 mm i odchylenie standardowe 2.1 mm. Na poziomie istotności 0.05 zweryfikować hipotezę o równej długości ziaren obu badanych gatunków fasoli.

Zadanie 5.5

Zgodnie z założeniami dotyczącymi procesu technologicznego w pewnej fabryce, dziennie zużycie wody jest zmienną losową o rozkładzie normalnym, o wartości oczekiwanej 1000 m^3 i odchyleniu standardowym 20 m^3 . Na podstawie obserwacji prowadzonych przez 30 dni roboczych obliczono, że średnie zużycie wody wyniosło 1025 m^3 . Na poziomie istotności 0.01 stwierdzić, czy dziennie zużycie wody w tej fabryce jest istotnie większe od założonego.

Zadanie 5.6

W teście badającym pamięć 8 losowo wybranych uczniów zapamiętało następujące liczby elementów: 16, 13, 14, 21, 19, 18, 26, 17. Po specjalnym treningu pamięci grupa ta wykazała następujące wyniki: 21, 17, 20, 26, 23, 22, 21, 16. Przyjmując poziom istotności 0.05 i zakładając, że liczba zapamiętywanych elementów ma rozkład normalny, zweryfikować hipotezę, że prowadzony trening istotnie zwiększa liczbę zapamiętywanych elementów.

Zadanie 5.7

Wylosowana próba 120 rodzin zamieszkałych w Warszawie dała średnią 450 zł miesięcznych opłat za mieszkanie, z odchyleniem standardowym 150 zł. Natomiast próba losowa 100 rodzin zamieszkałych w Łodzi dała średnią 420 zł i odchylenie standardowe 120 zł miesięcznych opłat za mieszkanie. Czy na poziomie istotności 0.05 można twierdzić, że opłaty mieszkaniowe w Warszawie są średnio wyższe niż w Łodzi?

Zadanie 5.8

Plony żyta w gospodarstwach indywidualnych pewnego województwa mają rozkład normalny. Przypuszcza się, że plony są rzędu 30 q/ha. Czy przypuszczenie to jest słuszne, jeżeli dla próby 25 losowo wybranych gospodarstw otrzymano średnią wysokość plonów 28 q/ha z odchyleniem standardowym 4 q/ha? Przyjąć poziom istotności 0.05.

Zadanie 5.9

Zmierzono czas świecenia 39 żarówek i stwierdzono, że dla 4 żarówek był on krótszy niż 1000 godzin, w przypadku 18 żarówek zawierał się w przedziale 2000 - 3000 godzin, a w przypadku pozostałych 3 żarówek czas świecenia zawierał się w przedziale 3000 - 4000 godzin. Stwierdzono również, że rozkład prawdopodobieństwa opisujący czas świecenia wspomnianych żarówek jest rozkładem normalnym.

- a) Na poziomie istotności 0.05 zweryfikować hipotezę, że 50% żarówek świeciło się dłużej niż 2000 godzin.
- b) Czy na poziomie istotności 0.05 możemy stwierdzić, że średni czas świecenia żarówek jest dłuższy niż 1900 godzin?

Zadanie 5.10

Z partii kondensatorów wybrano losowo 6 sztuk i zmierzono ich pojemność, otrzymując wyniki: 4.3, 4.4, 4.3, 4.2, 4.3, 4.4 (pF). Zakładamy, że pojemność ma rozkład normalny. Czy na poziomie istotności 0.05 prawdziwe jest przypuszczenie, że średnia pojemność kondensatora przekracza 4.3 (pF)?

Zadanie 5.11

Na 160 losowo wybranych studentów Uniwersytetu Warszawskiego 34 otrzymało w ubiegłym roku stypendium za dobre wyniki w nauce. Spośród 180 losowo wybranych studentów Politechniki Warszawskiej 25 otrzymało takie stypendium. Czy na podstawie tych danych można stwierdzić, że udział studentów osiągających dobre wyniki w nauce jest wyższy na Uniwersytecie Warszawskim niż na Politechnice Warszawskiej? Przyjąć $\alpha = 0.05$.

Zadanie 5.12

Podczas kontroli pracy dwóch centrali telefonicznych stwierdzono, że w pierwszej centrali na 200 połączeń 16 było pomyłkowych, natomiast w drugiej centrali na 102 połączenia pomyłkowych było 10. Zweryfikować hipotezę, że w obu centralach jest jednakowy procent pomyłkowych połączeń. Przyjąć poziom istotności 0.01.

Zadanie 5.13

Wysunięto hipotezę, że ceny artykułów żywnościowych uległy w pewnym okresie czasu podwyżce. W celu sprawdzenia tej hipotezy wylosowano 12 rodzajów artykułów i stwierdzono, że ich ceny na początku i na końcu badanego okresu były następujące:

ceny na początku	2.5	8	1.2	3.6	2.0	4.5
ceny na końcu	2.2	10	1.5	3.8	2.2	4.0
ceny na początku	3.0	9.0	6.8	10	14	13
ceny na końcu	3.2	8.6	7.0	12	12	11.4

Na poziomie istotności $\alpha = 0.05$ zweryfikować hipotezę, że ceny artykułów żywnościowych nie uległy zmianie w ciągu badanego okresu.

Zadanie 5.14

Spośród studentów pewnego wydziału uczelni wylosowano niezależnie 14 studentów IV roku i otrzymano dla nich następujące średnie oceny uzyskane w sesji egzaminacyjnej na I i IV roku:

I rok	3.5	4.0	3.7	4.6	3.9	3.0	3.5
IV rok	4.2	3.9	3.8	4.5	4.2	3.4	3.8
I rok	3.9	4.5	4.1	4.9	3.9	4.3	3.6
IV rok	4.0	4.6	4.2	5.0	4.1	4.2	3.9

Czy te rezultaty potwierdzają hipotezę, że średnie wyniki po IV roku są lepsze niż po I roku? (Przyjmujemy poziom istotności $\alpha = 0.05$).

Zadanie 5.15

Wykonano 24 pomiary stężenia pewnego roztworu i otrzymano następujące wyniki:

19.4, 23.4, 19.4, 16.2, 22.3, 21.9, 14.8, 19.7, 17.6, 16.8, 21.2, 21.0, 21.1, 15.8, 18.2, 18.6, 22.2, 17.5, 19.7, 20.0, 21.6, 18.5, 20.6, 20.8.

Czy prawdziwe jest przypuszczenie, że stężenie tego roztworu jest większe od 19? Zweryfikować stosowną hipotezę na poziomie istotności 0.01.

Zadanie 5.16

Zmierzono czas trwania 14 losowo wybranych rozmów telefonicznych i otrzymano (w minutach): 2.2, 3.4, 12.5, 23.0, 3.8, 3.2, 2.9, 15.3, 1.5, 14.2, 13.4, 4.5, 6.0, 10.4. Zweryfikować przypuszczenie, że czas trwania rozmowy przekracza 10 minut. Przyjąć poziom istotności 0.05.

Zadanie 5.17

W pewnym doświadczeniu chemicznym bada się grubość powłoki niklowej, uzyskiwanej dla dwóch rodzajów kąpieli galwanicznych. Niezależne pomiary grubości powłoki uzyskiwanej w określonym czasie dla obu rodzajów kąpieli były następujące (w mikronach):

I kąpiel: 4.3, 3.7, 11.2, 8.7, 7.7, 11.3

II kąpiel: 8.1, 9.3, 14.7, 5.3, 7.6, 10.1, 11.1.

Na poziomie istotności 0.05 zweryfikować hipotezę, że grubość powłoki niklowej uzyskiwana w II kąpieli jest większa niż uzyskiwana w I kąpieli.

Zadanie 5.18

Dwie formacje geologiczne porównano pod względem zawartości pewnego minerału. Uzyskano następujące dane:

formacja I	7.6, 11.1, 6.8, 9.8, 4.9, 6.1, 15.1
formacja II	4.7, 6.4, 4.1, 3.7, 3.9

Czy można stwierdzić, że zawartość owego minerału w pierwszej formacji jest istotnie większa od zawartości w drugiej formacji? Przyjąć poziom istotności 0.05.

Zadanie 5.19

Wykonano 12 niezależnych pomiarów przyspieszenia ziemskiego i otrzymano następujące wyniki (w cm/s^2):

976.9, 978.2, 978.5, 977.6, 979.2, 980.4, 980.2, 978.8, 987.7, 981.0, 979.9, 978.1.

Na poziomie istotności $\alpha = 0.05$ zweryfikować hipotezę, że wartość przyspieszenia ziemskiego wynosi 980 cm/s^2 .

Zadanie 5.20

Wysokość zarobków losowej próby pracowników pewnego przedsiębiorstwa przedstawia się następująco:

zarobki (w tys. zł)	liczba osób
0.6 - 1.0	3
1.0 - 1.4	10
1.4 - 1.8	12
1.8 - 2.2	5

Czy na podstawie powyższych danych można uznać, że wariancja zarobków w tym przedsiębiorstwie wynosi 150 zł²? Zakładamy, że rozkład zarobków jest normalny. Przyjąć poziom istotności 0.05.

Zadanie 5.21

W celu zbadania poprawności działania generatora liczb pseudolosowych wygenerowano 320 cyfr (od 0 do 9) i otrzymano:

i	0	1	2	3	4	5	6	7	8	9
n_i	33	32	29	31	37	34	34	26	32	32

Sprawdzić, czy wygenerowane cyfry mają rozkład równomierny. Przyjąć poziom istotności 0.01.

Zadanie 5.22

Losową próbę studentów sptytano o ich ulubiony przedmiot. Otrzymano następujące odpowiedzi:

przedmiot	Bazy Danych	Matematyka Dyskretna	C++	RPiS
liczba studentów	190	198	187	225

Czy można przyjąć, że rozkład preferencji jest równomierny? Przyjąć poziom istotności 0.05.

Zadanie 5.23

W celu zbadania, czy istnieje związek pomiędzy wynikami egzaminów z Rachunu Prawdopodobieństwa i Statystyki (RPiS) oraz Matematyki Dyskretnej (MD), przeprowadzono badanie na 100 osobowej próbie otrzymując wyniki:

	zdany egz. z MD	nie zdany egz. z MD
zdany egz. z RPiS	23	9
nie zdany egz. z RPiS	11	57

Zweryfikować odpowiednią hipotezę, na poziomie istotności 1%.

Zadanie 5.24

Badano, czy istnieje związek między wykształceniem a tolerancją. W tym celu przeprowadzono badanie na 220 osobach i otrzymano następujące wyniki:

	tolerancja	brak tolerancji
wykształcenie wyższe	71	29
brak wyższego wykształcenia	57	63

Na poziomie istotności 0.01 zweryfikować hipotezę, że istnieje związek między wykształceniem a tolerancją.

Zadanie 5.25

Badano reakcję pacjentów na wielkość dawki pewnego leku przeciwbolesowego i otrzymano następujące dane dotyczące średniego czasu działania leku:

wielkość dawki (w mg)	0.5	1	1.5	2	3	4	8
czas działania leku (w min)	28	67	80	109	120	154	176

Wyestymować wartość współczynnika korelacji liniowej między wielkością dawki a czasem działania leku. Czy na poziomie istotności 5% można twierdzić, że wielkość dawki i średni czas działania leku są skorelowane?

Zadanie 5.26

Podczas badania zmęczenia pracowników nieprzerwaną pracą trwającą od 1 do 5 godzin, otrzymano następujące dane dotyczące średniej liczby błędów popełnionych w teście jakiemu poddano pracowników:

czas pracy (w godzinach)	1	2	3	4	5
średnia liczba błędów	2	3	6	11	20

Wyestymowac wartość współczynnika korelacji liniowej między czasem nieprzerwanej pracy a średnią liczbą popełnianych błędów. Czy na poziomie istotności 0.01 można twierdzić, że czas nieprzerwanej pracy i średnia liczba błędów są skorelowane?

Zadanie 5.27

Interesuje nas, czy istnieje zależność pomiędzy wynikami uzyskiwanymi podczas egzaminu ustnego z RPiS oraz z Matematyki Dyskretnej. Wykładowcy wypisali imiona 7 studentów w kolejności odpowiadającej ocenie zasobu wiadomości posiadanego przez danego studenta (tzn. od "najlepszego" studenta do "najgorszego" studenta):

RPiS	Jaś	Kazio	Ula	Zenek	Stefcia	Franek	Edek
Mat. Dyskr.	Kazio	Ula	Jaś	Zenek	Franek	Stefcia	Edek

Podać wartość miary współzależności pomiędzy wynikami uzyskiwanymi podczas egzaminu ustnego z RPiS oraz z Matematyki Dyskretnej i skomentować uzyskany wynik.

Zadanie 5.28

Dwaj profesorowie X i Y postanowili dokonać oceny zdolności 12 studentów. W tym celu każdy z profesorów uszeregował wspomnianych studentów od najzdolniejszego do najmniej zdolnego:

student	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII
prof. X	1	7	8	3	6	10	9	2	11	4	5	12
prof. Y	4	8	10	1	5	9	11	3	7	2	6	12

Czy opinie obu profesorów są zbieżne (przyjmując $\alpha = 0.05$)?

ODPOWIEDZI

Zadanie 5.1

$T = -12.998 \in W'_\alpha = (-\infty, -1.64]$, zatem oglądalność jest niższa niż 75%.

Zadanie 5.2

$T = 0.5955 \notin W_\alpha = (-\infty, -1.95996] \cup [1.95996)$, zatem 6% połączeń jest błędnych.

Zadanie 5.3

$T = -2.067 \in W'_\alpha = (-\infty, -1.9432]$, więc waga pudełka proszku jest mniejsza niż 3 kg.

Zadanie 5.4

$T = 3.16 \in W_\alpha = (-\infty, -1.95996] \cup [1.95996)$, zatem ziarna dwóch gatunków fasoli nie są równej długości.

Zadanie 5.5

$T = 6.85 \in W''_\alpha = [2.32634, +\infty)$, zatem zużycie wody jest istotnie większe od założonego.

Zadanie 5.6

$T = -3.5553 \in W'_\alpha = (-\infty, -1.64485]$, stąd stwierdzamy, że trening istotnie zwiększa liczbę zapamiętywanych elementów.

Zadanie 5.7

$T = 1.6477 \in W''_\alpha = [1.64485, +\infty)$, zatem opłaty za mieszkanie w Warszawie są średnio wyższe niż w Łodzi.

Zadanie 5.8

$T = -2.5 \in W_\alpha = (-\infty, -2.0639] \cup [2.0639, +\infty)$, zatem przypuszczenie, że plony są rzędu 30 q/ha, nie jest słuszne.

Zadanie 5.9

- $T = -0.803 \notin W_\alpha = (-\infty, -1.95996] \cup [1.95996, +\infty)$, zatem 50% żarówek świeciło dłużej niż 2000 godzin (uwaga: należało zastosować model dla rozkładu dwupunktowego i weryfikować hipotezę $H : p = 0.5$ przeciwko hipotezie $K : p \neq 0.5$, gdzie p to frakcja żarówek świeczących dłużej niż 2000 godzin)

- $T = 0.0795 \notin W''_\alpha = [1.686, +\infty)$, więc średni czas świecenia żarówek nie jest dłuższy niż 1900 godzin.

Zadanie 5.10

$T = 0.64999 \notin W''_\alpha = [2.0151, +\infty)$, zatem średnia pojemność kondensatora nie przekracza 4.3 (pF).

Zadanie 5.11

$T = 1.84 \in W''_\alpha = [1.644, +\infty)$, zatem na Uniwersytecie Warszawskim więcej studentów osiąga dobre wyniki w nauce.

Zadanie 5.12

$T = -0.5277 \notin W_\alpha = (-\infty, -2.57582] \cup [2.57582, +\infty)$, zatem w obu centralach jest jednakowy procent pomyłkowych połączeń.

Zadanie 5.13

$T = 5 \notin W'_\alpha = [0, 2]$, zatem ceny nie uległy zmianie.

Zadanie 5.14

$T = 3 \in W'_\alpha = [0, 3]$, więc przyjmujemy hipotezę, że średnie wyniki po IV roku są lepsze niż po I roku.

Zadanie 5.15

$T = 15 \notin W''_\alpha = [19, 24]$, zatem stwierdzamy, że stężenie tego roztworu nie jest większe od 19.

Zadanie 5.16

$T = 6 \notin W''_\alpha = [11, 14]$, więc przyjmujemy, że czas trwania rozmowy nie przekracza 10 minut.

Zadanie 5.17

$T = 38 \notin W'_\alpha = [0, 32]$, czyli grubość powłoki niklowej uzyskiwanej w II kąpieli nie jest większa niż grubość powłoki niklowej uzyskiwanej w I kąpieli.

Zadanie 5.18

$T = 17 \in W'_\alpha = [0, 23]$, a zatem można stwierdzić, że zawartość owego minerału w pierwszej formacji jest istotnie większa niż w drugiej formacji.

Zadanie 5.19

$T = 4 \notin W_\alpha = [0, 2] \cup [10, 12]$, zatem przyjmujemy, że wartość przyspieszenia ziemskiego wynosi 980 cm/s^2 .

Zadanie 5.20

$T = 24.4953 \notin W_\alpha = (0, 16.0471] \cup [45.7224, +\infty)$, zatem wariancja zarobków w tym przedsiębiorstwie wynosi 150 zł^2 .

Zadanie 5.21

$T = 2.5 \notin W_\alpha = [21.6660, +\infty)$, zatem wygenerowane cyfry mają rozkład równomierny.

Zadanie 5.22

$T = 4.49 \notin W_\alpha = [7.8147, +\infty)$, zatem rozkład preferencji jest równomierny.

Zadanie 5.23

$T = 30.08 \in W_\alpha = [6.6349, +\infty)$, więc istnieje związek pomiędzy wynikami egzaminów z RPiS i Matematyki Dyskretnej.

Zadanie 5.24

$T = 12.38 \in W_\alpha = [6.6349, +\infty)$, więc istnieje związek między wykształceniem a tolerancją.

Zadanie 5.25

$r \simeq 0.8985$, $T = 4.5768 \in W_\alpha = (-\infty, -2.5706] \cup [2.5706, +\infty)$, zatem istnieje korelacja liniowa między wielkością dawki a czasem działania leku.

Zadanie 5.26

$r \simeq 0.9442$, $T = 4.96546 \in W_\alpha = (-\infty, -4.0329] \cup [4.0321, +\infty)$, więc można twierdzić, że czas nieprzerwanej pracy i średnia liczba błędów są skorelowane.

Zadanie 5.27

$r_s \simeq 0.8571429$, zatem istnieje dość silna współzależność pomiędzy wynikami uzyskiwanymi podczas egzaminu ustnego z RPiS oraz z Matematyki Dyskretnej.

Zadanie 5.28

$T = 4.8791 \in W_\alpha = (-\infty, -2.2281] \cup [2.2281, +\infty)$, więc przyjmujemy, że opinie obu profesorów są zbieżne.

6

Tablice statystyczne

6.1 Dystrybuanta rozkładu normalnego

Tablica zawiera wartości dystrybuanty $\Phi(x)$ rozkładu normalnego standar-dowego $N(0, 1)$ dla $x \geq 0$. Dla $x < 0$ mamy

$$\Phi(x) = 1 - \Phi(-x).$$

x	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	5000	5040	5080	5120	5160	5199	5239	5279	5319	5359
0,1	5398	5438	5478	5517	5557	5596	5636	5675	5714	5753
0,2	5793	5832	5871	5910	5948	5987	6026	6064	6103	6141
0,3	6179	6217	6255	6293	6331	6368	6406	6443	6480	6517
0,4	6554	6591	6628	6664	6700	6736	6772	6808	6844	6879
0,5	6915	6950	6985	7019	7054	7088	7123	7157	7190	7224
0,6	7257	7291	7324	7357	7389	7422	7454	7486	7517	7549
0,7	7580	7611	7642	7673	7704	7734	7764	7794	7823	7852
0,8	7881	7910	7939	7967	7995	8023	8051	8078	8106	8133
0,9	8159	8186	8212	8238	8264	8289	8315	8340	8365	8389
1,0	8413	8438	8461	8485	8508	8531	8554	8577	8599	8621
1,1	8643	8665	8686	8708	8729	8749	8770	8790	8810	8830
1,2	8849	8869	8888	8907	8925	8944	8962	8980	8997	9015
1,3	9032	9049	9066	9082	9099	9115	9131	9147	9162	9177
1,4	9192	9207	9222	9236	9251	9265	9279	9292	9306	9319
1,5	9332	9345	9357	9370	9382	9394	9406	9418	9429	9441
1,6	9452	9463	9474	9484	9495	9505	9515	9525	9535	9545
1,7	9554	9564	9573	9582	9591	9599	9608	9616	9625	9633
1,8	9641	9649	9656	9664	9671	9678	9686	9693	9699	9706
1,9	9713	9719	9726	9732	9738	9744	9750	9756	9761	9767
2,0	9772	9778	9783	9788	9793	9798	9803	9808	9812	9817
2,1	9821	9826	9830	9834	9838	9842	9846	9850	9854	9857
2,2	9861	9864	9868	9871	9875	9878	9881	9884	9887	9890
2,3	9893	9896	9898	9901	9904	9906	9909	9911	9913	9916
2,4	9918	9920	9922	9925	9927	9929	9931	9932	9934	9936
2,5	9938	9940	9941	9943	9945	9946	9948	9949	9951	9952
2,6	9953	9955	9956	9957	9959	9960	9961	9962	9963	9964
2,7	9965	9966	9967	9968	9969	9970	9971	9972	9973	9974
2,8	9974	9975	9976	9977	9977	9978	9979	9979	9980	9981
2,9	9981	9982	9982	9983	9984	9984	9985	9985	9986	9986
3,0	9987	9987	9987	9988	9988	9989	9989	9989	9990	9990
3,1	9990	9991	9991	9991	9992	9992	9992	9992	9993	9993
3,2	9993	9993	9994	9994	9994	9994	9994	9995	9995	9995
3,3	9995	9995	9995	9996	9996	9996	9996	9996	9996	9997
3,4	9997	9997	9997	9997	9997	9997	9997	9997	9997	9998

6.2 Kwantyle rozkładu normalnego

Tablica zawiera wartości kwantylów u_p rozkładu normalnego standardowego $N(0, 1)$ rzędu $p \geq 0.5$. Dla $p < 0.5$ mamy

$$u_p = -u_{1-p}.$$

p	0.000	0.010	0.020	0.030	0.040	0.050
0.5	0.00000	0.02507	0.05015	0.07527	0.10043	0.12566
0.6	0.25335	0.27932	0.30548	0.33185	0.35846	0.38532
0.7	0.52440	0.55338	0.58284	0.61281	0.64335	0.67449
0.8	0.84162	0.87790	0.91537	0.95417	0.99446	1.03643
0.9	1.28155	1.34075	1.40507	1.47579	1.55477	1.64485

p	0.060	0.070	0.075	0.080	0.090	0.095
0.5	0.15097	0.17637	0.18912	0.20189	0.22754	0.24043
0.6	0.41246	0.43991	0.45376	0.46770	0.49585	0.51007
0.7	0.70630	0.73885	0.75541	0.77219	0.80642	0.82389
0.8	1.08032	1.12639	1.15035	1.17499	1.22653	1.25356
0.9	1.75069	1.88079	1.95996	2.05375	2.32634	2.57582

6.3 Kwantyle rozkładu chi-kwadrat

Tablica zawiera wartości kwantyle $\chi_{p,n}^2$ rozkładu chi-kwadrat o n stopniach swobody, rzędu p .

$n \backslash p$	0,005	0,010	0,025	0,050	0,100	0,900	0,950	0,975	0,990	0,995
1	0,000	0,000	0,001	0,004	0,016	2,706	3,841	5,024	6,635	7,879
2	0,010	0,020	0,051	0,103	0,211	4,605	5,991	7,378	9,210	10,60
3	0,072	0,115	0,216	0,352	0,584	6,251	7,815	9,348	11,34	12,84
4	0,207	0,297	0,484	0,711	1,064	7,779	9,488	11,14	13,28	14,86
5	0,412	0,554	0,831	1,145	1,610	9,236	11,07	12,83	15,09	16,75
6	0,676	0,872	1,237	1,635	2,204	10,64	12,59	14,45	16,81	18,55
7	0,989	1,239	1,690	2,167	2,833	12,02	14,07	16,01	18,48	20,28
8	1,344	1,647	2,180	2,733	3,490	13,36	15,51	17,53	20,09	21,95
9	1,735	2,088	2,700	3,325	4,168	14,68	16,92	19,02	21,67	23,59
10	2,156	2,558	3,247	3,940	4,865	15,99	18,31	20,48	23,21	25,19
11	2,603	3,053	3,816	4,575	5,578	17,28	19,68	21,92	24,73	26,76
12	3,074	3,571	4,404	5,226	6,304	18,55	21,03	23,34	26,22	28,30
13	3,565	4,107	5,009	5,892	7,041	19,81	22,36	24,74	27,69	29,82
14	4,075	4,660	5,629	6,571	7,790	21,06	23,68	26,12	29,14	31,32
15	4,601	5,229	6,262	7,261	8,547	22,31	25,00	27,49	30,58	32,80
16	5,142	5,812	6,908	7,962	9,312	23,54	26,30	28,85	32,00	34,27
17	5,697	6,408	7,564	8,672	10,09	24,77	27,59	30,19	33,41	35,72
18	6,265	7,015	8,231	9,390	10,86	25,99	28,87	31,53	34,81	37,16
19	6,844	7,633	8,907	10,12	11,65	27,20	30,14	32,85	36,19	38,58
20	7,434	8,260	9,591	10,85	12,44	28,41	31,41	34,17	37,57	40,00
21	8,034	8,897	10,28	11,59	13,24	29,62	32,67	35,48	38,93	41,40
22	8,643	9,542	10,98	12,34	14,04	30,81	33,92	36,78	40,29	42,80
23	9,260	10,20	11,69	13,09	14,85	32,01	35,17	38,08	41,64	44,18
24	9,886	10,86	12,40	13,85	15,66	33,20	36,42	39,36	42,98	45,56
25	10,52	11,52	13,12	14,61	16,47	34,38	37,65	40,65	44,31	46,93
26	11,16	12,20	13,84	15,38	17,29	35,56	38,89	41,92	45,64	48,29
27	11,81	12,88	14,57	16,15	18,11	36,74	40,11	43,19	46,96	49,65
28	12,46	13,56	15,31	16,93	18,94	37,92	41,34	44,46	48,28	50,99
29	13,12	14,26	16,05	17,71	19,77	39,09	42,56	45,72	49,59	52,34
30	13,79	14,95	16,79	18,49	20,60	40,26	43,77	46,98	50,89	53,67
31	14,46	15,66	17,54	19,28	21,43	41,42	44,99	48,23	52,19	55,00
32	15,13	16,36	18,29	20,07	22,27	42,58	46,19	49,48	53,49	56,33
33	15,82	17,07	19,05	20,87	23,11	43,75	47,40	50,73	54,78	57,65
34	16,50	17,79	19,81	21,66	23,95	44,90	48,60	51,97	56,06	58,96
35	17,19	18,51	20,57	22,47	24,80	46,06	49,80	53,20	57,34	60,27
36	17,89	19,23	21,34	23,27	25,64	47,21	51,00	54,44	58,62	61,58
37	18,59	19,96	22,11	24,07	26,49	48,36	52,19	55,67	59,89	62,88
38	19,29	20,69	22,88	24,88	27,34	49,51	53,38	56,90	61,16	64,18
39	20,00	21,43	23,65	25,70	28,20	50,66	54,57	58,12	62,43	65,48
40	20,71	22,16	24,43	26,51	29,05	51,81	55,76	59,34	63,69	66,77

6.4 Kwantyle rozkładu t-Studenta

Tablica zawiera wartości kwantyle $t_p^{[n]}$ rozkładu t-Studenta o n stopniach swobody, rzędu p .

$n \backslash p$	0,9	0,95	0,975	0,99	0,995
1	3,0777	6,3137	12,7062	31,8210	63,6559
2	1,8856	2,9200	4,3027	6,9645	9,9250
3	1,6377	2,3534	3,1824	4,5407	5,8408
4	1,5332	2,1318	2,7765	3,7469	4,6041
5	1,4759	2,0150	2,5706	3,3649	4,0321
6	1,4398	1,9432	2,4469	3,1427	3,7074
7	1,4149	1,8946	2,3646	2,9979	3,4995
8	1,3968	1,8595	2,3060	2,8965	3,3554
9	1,3830	1,8331	2,2622	2,8214	3,2498
10	1,3722	1,8125	2,2281	2,7638	3,1693
11	1,3634	1,7959	2,2010	2,7181	3,1058
12	1,3562	1,7823	2,1788	2,6810	3,0545
13	1,3502	1,7709	2,1604	2,6503	3,0123
14	1,3450	1,7613	2,1448	2,6245	2,9768
15	1,3406	1,7531	2,1315	2,6025	2,9467
16	1,3368	1,7459	2,1199	2,5835	2,9208
17	1,3334	1,7396	2,1098	2,5669	2,8982
18	1,3304	1,7341	2,1009	2,5524	2,8784
19	1,3277	1,7291	2,0930	2,5395	2,8609
20	1,3253	1,7247	2,0860	2,5280	2,8453
21	1,3232	1,7207	2,0796	2,5176	2,8314
22	1,3212	1,7171	2,0739	2,5083	2,8188
23	1,3195	1,7139	2,0687	2,4999	2,8073
24	1,3178	1,7109	2,0639	2,4922	2,7970
25	1,3163	1,7081	2,0595	2,4851	2,7874
26	1,3150	1,7056	2,0555	2,4786	2,7787
27	1,3137	1,7033	2,0518	2,4727	2,7707
28	1,3125	1,7011	2,0484	2,4671	2,7633
29	1,3114	1,6991	2,0452	2,4620	2,7564
30	1,3104	1,6973	2,0423	2,4573	2,7500
31	1,3095	1,6955	2,0395	2,4528	2,7440
32	1,3086	1,6939	2,0369	2,4487	2,7385
33	1,3077	1,6924	2,0345	2,4448	2,7333
34	1,3070	1,6909	2,0322	2,4411	2,7284
35	1,3062	1,6896	2,0301	2,4377	2,7238
36	1,3055	1,6883	2,0281	2,4345	2,7195
37	1,3049	1,6871	2,0262	2,4314	2,7154
38	1,3042	1,6860	2,0244	2,4286	2,7116
39	1,3036	1,6849	2,0227	2,4258	2,7079
40	1,3031	1,6839	2,0211	2,4233	2,7045

6.5 Kwantyle rozkładu F-Snedecora

Tablica zawiera kwantyle $F_{0,95}^{(n_1, n_2)}$ rozkładu F-Snedecora o (n_1, n_2) stopniach swobody; rzędu 0.95 oraz 0.99.

$n_1 \backslash n_2$	1	2	3	4	5	6	7	8	9
1	161,4 4052	199,5 4999	215,7 5404	224,6 5624	230,2 5764	234,0 5859	236,8 5928	238,9 5981	240,5 6022
2	18,513 98,502	19,000 99,000	19,164 99,164	19,247 99,251	19,296 99,302	19,329 99,331	19,353 99,357	19,371 99,375	19,385 99,390
3	10,128 34,116	9,552 30,816	9,277 29,457	9,117 28,710	9,013 28,237	8,941 27,911	8,887 27,671	8,845 27,489	8,812 27,345
4	7,709 21,198	6,944 18,000	6,591 16,694	6,388 15,977	6,256 15,522	6,163 15,207	6,094 14,976	6,041 14,799	5,999 14,659
5	6,608 16,258	5,786 13,274	5,409 12,060	5,192 11,392	5,050 10,967	4,950 10,672	4,876 10,456	4,818 10,289	4,772 10,158
6	5,987 13,745	5,143 10,925	4,757 9,780	4,534 9,148	4,387 8,746	4,284 8,466	4,207 8,260	4,147 8,102	4,099 7,976
7	5,591 12,246	4,737 9,547	4,347 8,451	4,120 7,847	3,972 7,460	3,866 7,191	3,787 6,993	3,726 6,840	3,677 6,719
8	5,318 11,259	4,459 8,649	4,066 7,591	3,838 7,006	3,688 6,632	3,581 6,371	3,500 6,178	3,438 6,029	3,388 5,911
9	5,117 10,562	4,256 8,022	3,863 6,992	3,633 6,422	3,482 6,057	3,374 5,802	3,293 5,613	3,230 5,467	3,179 5,351
10	4,965 10,044	4,103 7,559	3,708 6,552	3,478 5,994	3,326 5,636	3,217 5,386	3,135 5,200	3,072 5,057	3,020 4,942
12	4,747 9,330	3,885 6,927	3,490 5,953	3,259 5,412	3,106 5,064	2,996 4,821	2,913 4,640	2,849 4,499	2,796 4,388
15	4,543 8,683	3,682 6,359	3,287 5,417	3,056 4,893	2,901 4,556	2,790 4,318	2,707 4,142	2,641 4,004	2,588 3,895
20	4,351 8,096	3,493 5,849	3,098 4,938	2,866 4,431	2,711 4,103	2,599 3,871	2,514 3,699	2,447 3,564	2,393 3,457
24	4,260 7,823	3,403 5,614	3,009 4,718	2,776 4,218	2,621 3,895	2,508 3,667	2,423 3,496	2,355 3,363	2,300 3,256
30	4,171 7,562	3,316 5,390	2,922 4,510	2,690 4,018	2,534 3,699	2,421 3,473	2,334 3,305	2,266 3,173	2,211 3,067
40	4,085 7,314	3,232 5,178	2,839 4,313	2,606 3,828	2,449 3,514	2,336 3,291	2,249 3,124	2,180 2,993	2,124 2,888
60	4,001 7,077	3,150 4,977	2,758 4,126	2,525 3,649	2,368 3,339	2,254 3,119	2,167 2,953	2,097 2,823	2,040 2,718

$n_1 \backslash n_2$	10	12	15	20	24	30	40	60	80
1	241,9 6056	243,9 6107	245,9 6157	248,0 6209	249,1 6234	250,1 6260	251,1 6286	252,2 6313	252,7 6326
2	19,40 99,40	19,41 99,42	19,43 99,43	19,45 99,45	19,45 99,46	19,46 99,47	19,47 99,48	19,48 99,48	19,48 99,48
3	8,785 27,23	8,745 27,05	8,703 26,87	8,660 26,69	8,638 26,60	8,617 26,50	8,594 26,41	8,572 26,32	8,561 26,27
4	5,964 14,55	5,912 14,37	5,858 14,20	5,803 14,02	5,774 13,93	5,746 13,84	5,717 13,75	5,688 13,65	5,673 13,61
5	4,735 10,051	4,678 9,888	4,619 9,722	4,558 9,553	4,527 9,466	4,496 9,379	4,464 9,291	4,431 9,202	4,415 9,157
6	4,060 7,874	4,000 7,718	3,938 7,559	3,874 7,396	3,841 7,313	3,808 7,229	3,774 7,143	3,740 7,057	3,722 7,013
7	3,637 6,620	3,575 6,469	3,511 6,314	3,445 6,155	3,410 6,074	3,376 5,992	3,340 5,908	3,304 5,824	3,286 5,781
8	3,347 5,814	3,284 5,667	3,218 5,515	3,150 5,359	3,115 5,279	3,079 5,198	3,043 5,116	3,005 5,032	2,986 4,989
9	3,137 5,257	3,073 5,111	3,006 4,962	2,936 4,808	2,900 4,729	2,864 4,649	2,826 4,567	2,787 4,483	2,768 4,441
10	2,978 4,849	2,913 4,706	2,845 4,558	2,774 4,405	2,737 4,327	2,700 4,247	2,661 4,165	2,621 4,082	2,601 4,039
12	2,753 4,296	2,687 4,155	2,617 4,010	2,544 3,858	2,505 3,780	2,466 3,701	2,426 3,619	2,384 3,535	2,363 3,493
15	2,544 3,805	2,475 3,666	2,403 3,522	2,328 3,372	2,288 3,294	2,247 3,214	2,204 3,132	2,160 3,047	2,137 3,004
20	2,348 3,368	2,278 3,231	2,203 3,088	2,124 2,938	2,082 2,859	2,039 2,778	1,994 2,695	1,946 2,608	1,922 2,563
24	2,255 3,168	2,183 3,032	2,108 2,889	2,027 2,738	1,984 2,659	1,939 2,577	1,892 2,492	1,842 2,403	1,816 2,357
30	2,165 2,979	2,092 2,843	2,015 2,700	1,932 2,549	1,887 2,469	1,841 2,386	1,792 2,299	1,740 2,208	1,712 2,160
40	2,077 2,801	2,003 2,665	1,924 2,522	1,839 2,369	1,793 2,288	1,744 2,203	1,693 2,114	1,637 2,019	1,608 1,969
60	1,993 2,632	1,917 2,496	1,836 2,352	1,748 2,198	1,700 2,115	1,649 2,028	1,594 1,936	1,534 1,836	1,502 1,783

6.6 Wartości krytyczne testu znaków

n	α	0,005	0,010	0,025	0,050
4		-	-	-	-
5		-	-	-	0
6		-	-	0	0
7		-	0	0	0
8		0	0	0	1
9		0	0	1	1
10		0	0	1	1
11		0	1	1	2
12		1	1	2	2
13		1	1	2	3
14		1	2	2	3
15		2	2	3	3
16		2	2	3	4
17		2	3	4	4
18		3	3	4	5
19		3	4	4	5
20		3	4	5	5
21		4	4	5	6
22		4	5	5	6
23		4	5	6	7
24		5	5	6	7
25		5	6	7	7
26		6	6	7	8
27		6	7	7	8
28		6	7	8	9
29		7	7	8	9
30		7	8	9	10
31		7	8	9	10
32		8	8	9	10
33		8	9	10	11
34		9	9	10	11
35		9	10	11	12
36		9	10	11	12
37		10	10	12	12
38		10	11	12	13
39		11	11	12	13
40		11	12	13	14

6.7 Wartości krytyczne testu rangowanych znaków

n	α	0,05	0,01
8		4	0
9		6	2
10		8	3
11		11	5
12		14	7
13		17	10
14		21	13
15		25	16
16		30	20
17		35	23
18		40	28
19		46	32
20		52	38
21		59	43
22		66	49
23		73	55
24		81	61
25		89	68

6.8 Wartości krytyczne testu Wilcoxona

$\alpha = 0.025$		m								
		3	4	5	6	7	8	9	10	
n	3	-	-	6	7	7	8	8	9	
	4	-	10	11	12	13	14	14	15	
	5	-	-	17	18	20	21	22	23	
	6	-	-	-	26	27	29	31	32	
	7	-	-	-	-	36	38	40	42	
	8	-	-	-	-	-	49	51	53	
	9	-	-	-	-	-	-	62	65	
	10	-	-	-	-	-	-	-	78	

$\alpha = 0.050$		m								
		3	4	5	6	7	8	9	10	
n	3	6	6	7	8	8	9	10	10	
	4	-	11	12	13	14	15	16	17	
	5	-	-	19	20	21	23	24	26	
	6	-	-	-	28	29	31	33	35	
	7	-	-	-	-	39	41	43	45	
	8	-	-	-	-	-	51	54	56	
	9	-	-	-	-	-	-	66	69	
	10	-	-	-	-	-	-	-	82	

$\alpha = 0.100$		m								
		3	4	5	6	7	8	9	10	
n	3	7	7	8	9	10	11	11	12	
	4	-	13	14	15	16	17	19	20	
	5	-	-	20	22	23	25	27	28	
	6	-	-	-	30	32	34	36	38	
	7	-	-	-	-	41	44	46	49	
	8	-	-	-	-	-	55	58	60	
	9	-	-	-	-	-	-	70	73	
	10	-	-	-	-	-	-	-	87	

6.9 Współczynniki dla testu Shapiro–Wilka

$i \setminus n$	2	3	4	5	6	7	8	9	10
1	0,7071	0,7071	0,6872	0,6646	0,6431	0,6233	0,6052	0,5888	0,5739
2	0,0000	0,1677	0,2413	0,2806	0,3031	0,3164	0,3244	0,3291	
3		0,0000	0,0875	0,1401	0,1743	0,1976	0,2141		
4			0,0000	0,0561	0,0947	0,1224			
5				0,0000	0,0399				
$i \setminus n$	11	12	13	14	15	16	17	18	19
1	0,5601	0,5475	0,5359	0,5251	0,5150	0,5056	0,4968	0,4486	0,4808
2	0,3315	0,3325	0,3325	0,3318	0,3306	0,3290	0,3273	0,3253	0,3232
3	0,2260	0,2347	0,2412	0,2460	0,2495	0,2521	0,2540	0,2533	0,2561
4	0,1429	0,1586	0,1707	0,1802	0,1878	0,1939	0,1988	0,2027	0,2059
5	0,0695	0,0922	0,1099	0,1240	0,1353	0,1447	0,1524	0,1587	0,1641
6	0,0000	0,0303	0,0539	0,0727	0,0880	0,1005	0,1109	0,1197	0,1271
7		0,0000	0,0240	0,0433	0,0593	0,0725	0,0837	0,0932	0,1013
8			0,0000	0,0196	0,0359	0,0496	0,0612	0,0711	
9				0,0000	0,0013	0,0303	0,0422		
10					0,0000	0,0140			
$i \setminus n$	21	22	23	24	25	26	27	28	29
1	0,4643	0,4390	0,4542	0,4493	0,4450	0,4407	0,4366	0,4328	0,4291
2	0,3815	0,3156	0,3126	0,3098	0,3069	0,3043	0,3018	0,2992	0,2944
3	0,2578	0,2571	0,2563	0,2554	0,2543	0,2533	0,2522	0,2510	0,2499
4	0,2199	0,2131	0,2139	0,2145	0,2148	0,2151	0,2152	0,2151	0,2148
5	0,1736	0,1764	0,1787	0,1807	0,1822	0,1836	0,1848	0,1857	0,1864
6	0,1399	0,1443	0,1480	0,1512	0,1539	0,1563	0,1584	0,1601	0,1616
7	0,1092	0,1150	0,1201	0,1245	0,1283	0,1316	0,1346	0,1372	0,1395
8	0,0804	0,0878	0,0941	0,0997	0,1046	0,1089	0,1128	0,1162	0,1192
9	0,0530	0,0618	0,0696	0,0764	0,0823	0,0876	0,0923	0,0965	0,1002
10	0,0263	0,0368	0,0459	0,0539	0,0610	0,0672	0,0728	0,0778	0,0822
11	0,0000	0,0122	0,0228	0,0321	0,0403	0,0476	0,0540	0,0598	0,0650
12			0,0000	0,0107	0,0200	0,0284	0,0358	0,0424	0,0483
13				0,0000	0,0094	0,0178	0,0253	0,0320	0,0381
14					0,0000	0,0084	0,0159	0,0227	
15						0,0000	0,0076		

$i \backslash n$	31	32	33	34	35	36	37	38	39	40
1	0,4420	0,4188	0,4156	0,4127	0,4096	0,4068	0,4040	0,4015	0,3989	0,3964
2	0,2921	0,2898	0,2876	0,2854	0,2834	0,2813	0,2794	0,2774	0,2755	0,2737
3	0,2475	0,2463	0,2451	0,2439	0,2427	0,2415	0,2403	0,2391	0,2380	0,2368
4	0,2145	0,2141	0,2137	0,2132	0,2127	0,2121	0,2116	0,2110	0,2104	0,2098
5	0,1874	0,1878	0,1880	0,1882	0,1883	0,1883	0,1883	0,1881	0,1880	0,1878
6	0,1641	0,1651	0,1660	0,1667	0,1673	0,1678	0,1683	0,1686	0,1689	0,1691
7	0,1433	0,1449	0,1463	0,1475	0,1487	0,1496	0,1505	0,1513	0,1520	0,1526
8	0,1243	0,1265	0,1284	0,1301	0,1317	0,1331	0,1344	0,1356	0,1366	0,1376
9	0,1066	0,1093	0,1118	0,1140	0,1160	0,1170	0,1196	0,1211	0,1225	0,1237
10	0,0899	0,0931	0,0961	0,0988	0,1013	0,1036	0,1056	0,1075	0,1092	0,1108
11	0,0739	0,0777	0,0812	0,0844	0,0873	0,0900	0,0924	0,0947	0,0967	0,0986
12	0,0585	0,0629	0,0669	0,0706	0,0739	0,0770	0,0798	0,0824	0,0848	0,0870
13	0,0435	0,0485	0,0530	0,0572	0,0610	0,0645	0,0677	0,0706	0,0733	0,0759
14	0,0280	0,0344	0,0395	0,0441	0,0484	0,0523	0,0559	0,0592	0,0622	0,0651
15	0,0144	0,0206	0,0262	0,0314	0,0361	0,0404	0,0444	0,0481	0,0515	0,0546
16	0,0000	0,0068	0,0131	0,0187	0,0239	0,0287	0,0331	0,0372	0,0409	0,0444
17										0,0000 0,0062 0,0119 0,0172 0,0220 0,0264 0,0305 0,0343
18										0,0000 0,0057 0,0110 0,0158 0,0203 0,0244
19										0,0000 0,0053 0,0101 0,0146
20										0,0000 0,0049
$i \backslash n$	41	42	43	44	45	46	47	48	49	50
1	0,3940	0,3917	0,3894	0,3872	0,3850	0,3830	0,3808	0,3789	0,3770	0,3751
2	0,2719	0,2701	0,2684	0,2667	0,2651	0,2635	0,2620	0,2604	0,2589	0,2574
3	0,2357	0,2345	0,2334	0,2323	0,2313	0,2302	0,2291	0,2281	0,2271	0,2260
4	0,2091	0,2085	0,2078	0,2072	0,2065	0,2058	0,2052	0,2045	0,2038	0,2032
5	0,1876	0,1874	0,1871	0,1868	0,1865	0,1862	0,1859	0,1855	0,1851	0,1847
6	0,1693	0,1694	0,1695	0,1695	0,1695	0,1695	0,1693	0,1692	0,1691	
7	0,1531	0,1535	0,1539	0,1542	0,1545	0,1548	0,1550	0,1551	0,1553	0,1554
8	0,1384	0,1392	0,1398	0,1405	0,1410	0,1415	0,1420	0,1423	0,1427	0,1430
9	0,1249	0,1259	0,1269	0,1278	0,1286	0,1293	0,1300	0,1306	0,1312	0,1317
10	0,1123	0,1136	0,1149	0,1160	0,1170	0,1180	0,1189	0,1197	0,1205	0,1212
11	0,1004	0,1020	0,1035	0,1049	0,1062	0,1073	0,1085	0,1093	0,1105	0,1113
12	0,0891	0,0909	0,0927	0,0943	0,0959	0,0972	0,0986	0,0998	0,1010	0,1020
13	0,0782	0,0804	0,0824	0,0842	0,0860	0,0876	0,0892	0,0906	0,0919	0,0932
14	0,0677	0,0701	0,0724	0,0745	0,0765	0,0783	0,0801	0,0817	0,0832	0,0846
15	0,0575	0,0602	0,0628	0,0651	0,0673	0,0694	0,0713	0,0731	0,0748	0,0764
16	0,0476	0,0508	0,0534	0,0560	0,0584	0,0607	0,0628	0,0648	0,0667	0,0685
17	0,0379	0,0411	0,0442	0,0471	0,0497	0,0522	0,0546	0,0568	0,0588	0,0608
18	0,0283	0,0318	0,0352	0,0383	0,0412	0,0439	0,0465	0,0489	0,0511	0,0532
19	0,0188	0,0227	0,0263	0,0296	0,0328	0,0357	0,0385	0,0411	0,0436	0,0459
20	0,0094	0,0136	0,0175	0,0211	0,0245	0,0277	0,0307	0,0335	0,0361	0,0386
21	0,0000	0,0045	0,0087	0,0126	0,0163	0,0197	0,0229	0,0259	0,0288	0,0314
22										0,0000 0,0042 0,0081 0,0118 0,0153 0,0185 0,0215 0,0244
23										0,0000 0,0039 0,0076 0,0011 0,0143 0,0174
24										0,0000 0,0037 0,0071 0,0104
25										0,0000 0,0035

6.10 Wartości krytyczne testu Shapiro–Wilka

$n \backslash \alpha$	0,01	0,02	0,05	0,95	0,98	0,99
3	0,753	0,756	0,767	0,999	1,000	1,000
4	0,687	0,707	0,748	0,992	0,996	0,997
5	0,686	0,715	0,762	0,986	0,991	0,993
6	0,713	0,743	0,788	0,981	0,986	0,989
7	0,730	0,760	0,803	0,979	0,985	0,988
8	0,749	0,778	0,818	0,978	0,984	0,987
9	0,764	0,791	0,829	0,978	0,984	0,986
10	0,781	0,806	0,842	0,978	0,983	0,986
11	0,792	0,817	0,850	0,979	0,984	0,986
12	0,803	0,828	0,859	0,979	0,984	0,986
13	0,814	0,837	0,866	0,979	0,984	0,986
14	0,825	0,846	0,874	0,980	0,984	0,986
15	0,835	0,855	0,881	0,980	0,984	0,987
16	0,844	0,863	0,887	0,981	0,985	0,987
17	0,851	0,869	0,892	0,981	0,985	0,987
18	0,858	0,874	0,897	0,982	0,986	0,988
19	0,863	0,879	0,901	0,982	0,986	0,988
20	0,868	0,884	0,905	0,983	0,986	0,988
21	0,873	0,888	0,908	0,983	0,987	0,989
22	0,878	0,892	0,911	0,984	0,987	0,989
23	0,881	0,895	0,914	0,984	0,987	0,989
24	0,884	0,898	0,916	0,984	0,987	0,989
25	0,888	0,901	0,918	0,985	0,988	0,989
26	0,891	0,904	0,920	0,985	0,988	0,989
27	0,894	0,906	0,923	0,985	0,988	0,990
28	0,896	0,908	0,924	0,985	0,988	0,990
29	0,898	0,910	0,926	0,985	0,988	0,990
30	0,900	0,912	0,927	0,985	0,988	0,990

6.11 Wartości krytyczne testu Kołmogorowa

n	α	0,20	0,10	0,05	0,02	0,01
1		0,9000	0,9500	0,9750	0,9900	0,9950
2		0,6838	0,7764	0,8419	0,9000	0,9293
3		0,5648	0,6360	0,7076	0,7846	0,8290
4		0,4927	0,5652	0,6239	0,6889	0,7342
5		0,4470	0,5095	0,5633	0,6272	0,6685
6		0,4104	0,4680	0,5193	0,5774	0,6166
7		0,3815	0,4361	0,4834	0,5384	0,5758
8		0,3583	0,4096	0,4543	0,5065	0,5418
9		0,3391	0,3875	0,4300	0,4796	0,5133
10		0,3226	0,3687	0,4093	0,4566	0,4889
11		0,3083	0,3524	0,3912	0,4367	0,4677
12		0,2958	0,3382	0,3754	0,4192	0,4491
13		0,2847	0,3255	0,3614	0,4036	0,4325
14		0,2748	0,3142	0,3489	0,3897	0,4176
15		0,2659	0,3040	0,3376	0,3771	0,4042
16		0,2578	0,2947	0,3273	0,3657	0,3920
17		0,2504	0,2863	0,3180	0,3553	0,3809
18		0,2436	0,2785	0,3094	0,3457	0,3706
19		0,2374	0,2714	0,3014	0,3369	0,3612
20		0,2316	0,2647	0,2941	0,3287	0,3524
21		0,2262	0,2586	0,2872	0,3210	0,3443
22		0,2212	0,2528	0,2809	0,3139	0,3367
23		0,2165	0,2475	0,2749	0,3073	0,3295
24		0,2121	0,2424	0,2693	0,3010	0,3229
25		0,2079	0,2377	0,2640	0,2952	0,3166
26		0,2040	0,2332	0,2591	0,2896	0,3106
27		0,2003	0,2290	0,2544	0,2844	0,3050
28		0,1968	0,2250	0,2499	0,2794	0,2997
29		0,1935	0,2212	0,2457	0,2747	0,2947
30		0,1903	0,2176	0,2417	0,2702	0,2899

6.12 Wartości krytyczne testu Kołmogorowa-Smirnowa

n_2	n_1	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
3	3	-																	
4	4	-	-																
5	5	-	-	25															
6	6	-	24	30	36														
7	7	-	28	35	36	42													
8	8	-	32	35	40	48	56												
9	9	27	36	40	45	49	55	63											
10	10	30	36	45	48	56	60	63	80										
11	11	33	40	45	54	59	64	70	77	88									
12	12	36	44	50	60	60	68	75	80	86	96								
13	13	39	48	52	59	65	72	78	84	91	95	107							
14	14	42	48	56	64	77	76	84	90	96	104	104	126						
15	15	42	52	60	69	75	81	90	100	102	108	115	123	135					
16	16	45	56	64	72	77	88	94	100	106	116	121	126	133	160				
17	17	48	60	68	73	84	88	99	106	110	119	127	134	142	143	170			
18	18	51	60	70	84	87	94	108	108	118	126	131	140	147	154	164	180		
19	19	54	64	71	83	91	98	113	113	122	130	138	148	152	160	166	176	190	
20	20	57	68	80	88	93	104	120	120	127	140	143	152	160	168	175	182	187	
																		220	
n_2	n_1	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
3	3	-																	
4	4	-	16																
5	5	15	20	25															
6	6	18	20	24	30														
7	7	21	24	29	30	42													
8	8	21	28	29	34	40	48												
9	9	24	28	35	39	42	46	54											
10	10	27	30	40	40	46	48	53	70										
11	11	30	33	39	43	48	53	59	60	77									
12	12	30	36	43	48	53	60	63	66	72	84								
13	13	33	39	45	52	56	62	65	70	75	81	91							
14	14	36	42	46	54	63	65	70	74	82	86	89	112						
15	15	36	44	55	57	62	67	75	80	84	93	96	98	120					
16	16	39	48	54	60	64	80	78	82	89	96	101	106	114	128				
17	17	42	48	55	62	68	77	82	89	93	100	105	111	116	124	136			
18	18	45	50	60	72	72	80	90	92	97	108	110	116	123	128	133	162		
19	19	45	53	61	70	76	82	89	94	102	108	114	121	127	133	141	142	171	
20	20	48	60	65	72	79	88	99	100	107	116	120	126	135	140	146	152	160	

 $\alpha = 0,01$ $\alpha = 0,05$

Literatura

- [1] Bartoszewicz J., Wykłady ze statystyki matematycznej, PWN, Warszawa 1989.
- [2] Domański C., Testy statystyczne, PWE, Warszawa 1990.
- [3] Feller W., Wstęp do rachunku prawdopodobieństwa, PWN, Warszawa 1980.
- [4] Fisz M., Rachunek prawdopodobieństwa i statystyka matematyczna, PWN, Warszawa 1969.
- [5] Gajek L., Kałuszka M., Wnioskowanie statystyczne. Modele i metody, WNT, Warszawa 2000.
- [6] Greń J., Statystyka matematyczna – modele i zadania, PWN, Warszawa, 1984.
- [7] Hryniewicz O., Wykłady ze statystyki, WSISiZ, Warszawa 1999.
- [8] Jakubowski J., Sztencel R., Wstęp do teorii prawdopodobieństwa, Script, Warszawa 2000.
- [9] Jóźwiak J., Podgórski J., Statystyka od podstaw, PWE, Warszawa 1998.
- [10] Klonecki W., Statystyka dla inżynierów, PWN, Warszawa 1999.

- [11] Krysicki W., Bartos J., Dyczka W., Królikowska K., Wasilewski M., Rachunek prawdopodobieństwa i statystyka matematyczna w zadaniach, PWN, Warszawa 1998.
- [12] Krzyśko M., Wykłady z teorii prawdopodobieństwa, WNT, Warszawa, 2000.
- [13] Lehmann E.L., Testowanie hipotez statystycznych, PWN, Warszawa 1963.
- [14] Plucińska A., Pluciński E., Probabilistyka, WNT, Warszawa 2000.
- [15] Rao C.R., Statystyka i prawda, PWN, Warszawa 1994.
- [16] Silvey S.D., Wnioskowanie statystyczne, PWN, Warszawa 1978.
- [17] Sobczyk M., Statystyka, PWN, Warszawa 1996.
- [18] Zieliński R., Siedem wykładów wprowadzających do statystyki matematycznej, PWN, Warszawa 1990.
- [19] Zieliński R., Zieliński W., Tablice statystyczne, PWN, Warszawa 1990.