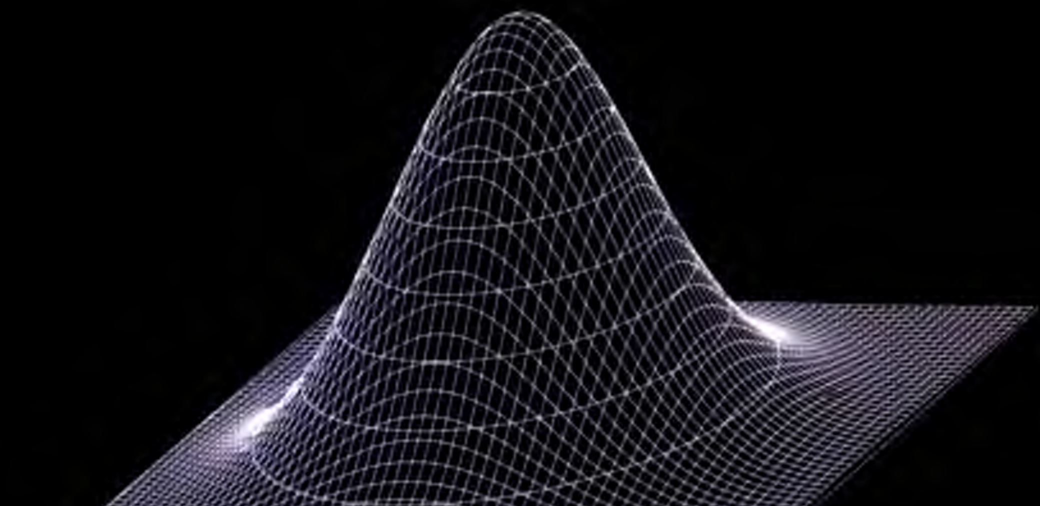
Michał Major

Elementy statystyki

Rachunek prawdopodobieństwa i wnioskowanie statystyczne



Krakowska Szkoła Wyższa im. Andrzeja Frycza Modrzewskiego
Kraków 2007

Krakowska Szkoła Wyższa
im. Andrzeja Frycza Modrzewskiego

Michał Major

Elementy statystyki

Rachunek prawdopodobieństwa i wnioskowanie statystyczne

Kraków 2007

Recenzja:
prof. AE dr hab. Józef Biolik
prof. KSW dr hab. Barbara Podolec

Projekt okładki i stron tytułowych
Wojciech Prażuch

Redaktor prowadzący:
Halina Baszak Jaroń

Korekta redakcyjna:
Zespół

ISBN 978-83-89823-27-4

Na zlecenie:
Krakowskiej Szkoły Wyższej im. Andrzeja Frycza Modrzewskiego
www.ksw.edu.pl

Wydawca:
Krakowskie Towarzystwo Edukacyjne Spółka z o.o.
Oficyna Wydawnicza AFM, Kraków 2007

Sprzedaż prowadzi:
Księgarnia Krakowskiego Towarzystwa Edukacyjnego Spółka z o.o.
Kampus Krakowskiej Szkoły Wyższej im. Andrzeja Frycza Modrzewskiego
ul. Gustawa Herlinga-Grudzińskiego 1
30-705 Krakow
tel./fax: (012) 252 45 93
e-mail: ksiegarnia@kte.pl

Skład i łamanie:
Wojciech Prażuch

Druk i oprawa: Drukarnia „Cenzus”

Spis treści

Od wydawcy

Drodzy Czytelnicy!

Przekazujemy Państwu podręcznik, który mamy nadzieję, ułatwi proces studiowania przedmiotu statystyka oraz dyscyplin pokrewnych, takich jak ekonometria, marketing, finanse czy rachunkowość.

Skrypt ten jest kontynuacją książki Michała Majora i Janusza Niezgody pt. *Elementy statystyki. Cz. I. Statystyka opisowa* wydanej w 2003 r. na zlecenie Krakowskiej Szkoły Wyższej im. Andrzeja Frycza Modrzewskiego, przez Krakowskie Towarzystwo Edukacyjne sp. z. o.o., w Krakowie. W podręczniku opisane zostały zagadnienia związane z teorią rachunku prawdopodobieństwa i wnioskowaniem statystycznym. Materiał zebrany w tej pracy został podzielony na sześć rozdziałów.

Rozdział pierwszy zawiera opis podstawowych pojęć i zagadnień związanych z teorią rachunku prawdopodobieństwa. Omówiono tutaj – między innymi – pojęcie prawdopodobieństwa, własności prawdopodobieństwa oraz prawdopodobieństwo zupełne i wzór Bayesa.

W rozdziale drugim opisano pojęcie zmiennej losowej, rozkładu zmiennej losowej skokowej i ciągłej, dystrybuanty zmiennej losowej, wartości oczekiwanej i wariancji zmiennej losowej, jak również dwuwymiarowe zmienne losowe.

Rozdział trzeci poświęcony został opisowi wybranych rozkładów zmiennych losowych oraz najważniejszych twierdzeń granicznych.

W rozdziale czwartym zebrano podstawowe pojęcia związane z wnioskowaniem statystycznym. Rozdział ten stanowi wstęp do zagadnień poruszanych w kolejnych dwóch rozdziałach. Omówiono tutaj również różne techniki pobierania próby losowej, którą wykorzystuje się w trakcie estymacji i weryfikacji hipotez statystycznych.

W kolejnym rozdziale – piątym – dokonano przeglądu zagadnień należących do tzw. teorii estymacji. W rozdziale tym przedstawiono techniki punktowej i przedziałowej estymacji podstawowych parametrów opisujących zbiorowość generalną, takich jak m.in.: wartość oczekiwana, wariancja i odchylenie standardowe, frakcja, współczynnik korelacji itd.

Ostatni z rozdziałów – szósty – zawiera przegląd metod służących weryfikacji hipotez statystycznych. Omówione zostały tutaj podstawowe parametryczne i nieparametryczne testy istotności, ponadto w zarysie przedstawiono podstawowe zasady tzw. sekwencyjnych testów.

Podkreślić należy, że w trakcie pisania podręcznika Autor celowo wprowadził jak najwięcej odpowiednio dobranych przykładów liczbowych, aby omawiane zagadnienia stały się łatwiejsze do zrozumienia i przyswojenia. I w tym przypadku, podobnie jak we wspomnianym podręczniku z 2003 roku, Autor zastosował strukturę przykładów wiodących, do których odwołuje się wielokrotnie w kolejnych rozdziałach. Pewne przykłady zostały zaczerpnięte z książki M. Majora i J. Niezgody pt. *Elementy statystyki, cz. I. Statystyka opisowa*.

Wszystkie te działania mają na celu podkreślenie spójności tych trzech zagadnień, jakimi są statystyka opisowa, rachunek prawdopodobieństwa i wnioskowanie statystyczne. Dla lepszej czytelności miejsca w tekście, gdzie kończą się przykłady, a zaczyna tekst o charakterze ogólnym, zaznaczono znakiem {kp}. Integralną częścią podręcznika są przypisy i odwołania do literatury, której studiowanie zalecamy osobom pragnącym lepiej poznać zagadnienia statystyki i dyscyplin z nią powiązanych.

Część I

ELEMENTY RACHUNKU PRAWDOPODOBIEŃSTWA

Wybrane zagadnienia z teorii rachunku prawdopodobieństwa

Rozdział

1

1.1. Wprowadzenie

Pierwszym z ważnych pojęć wykorzystywanych w teorii rachunku prawdopodobieństwa jest pojęcie **doświadczenia losowego**, przez które rozumieć będziemy powtarzalny eksperyment fizyczny lub myślowy. W literaturze przedmiotu[1] spotkać można definicję, według której doświadczeniem losowym jest realizacja (rzeczywista bądź myślowa) określonego zespołu warunków, wraz z góry określonym zbiorem wyników.

W innym opracowaniu[2] doświadczenie losowe to każde przedsięwzięcie empiryczne, rzeczywiste lub symulacyjne, którego wyniku nie można przewidzieć mimo sprecyzowanych warunków, w jakich jest ono realizowane. Doświadczenie losowe może polegać na przykład na rzucie kostką do gry, obserwacji, czy światło jest zapalone czy zgaszone, ustaleniu liczby jednostek wadliwych w partii produktu, przeprowadzeniu ankiety na określony temat itp.

W wyniku przeprowadzonego doświadczenia otrzymujemy **wynik doświadczenia**. Wynikiem może być odpowiedź „tak" lub „nie", odczyt pomiarów przyrządowych, wartość z określonego przedziału itp.

Wyniki doświadczenia nazywane są też **zdarzeniami elementarnymi** i oznaczane najczęściej symbolem „ω". Zbiór wszystkich zdarzeń elementarnych (wyników doświadczeń) określany jest mianem **przestrzeni zdarzeń elementarnych** (przestrzenią próby doświadczenia) i oznaczany symbolem „Ω". Przestrzeń zdarzeń elementarnych może być przestrzenią skończoną lub nieskończoną (np. w przypadku pomiarów przyjmujących wartości ciągłe).

Wyniki doświadczeń można prezentować algebraicznie lub graficznie (za pomocą punktów i symboli graficznych). Wariantom wyników doświadczenia możemy

---

[1] W. Krysicki i inni: *Rachunek prawdopodobieństwa i statystyka matematyczna w zadaniach*, cz. I, PWN, Warszawa 1986, s. 7.

[2] Zob. A. Iwasiewicz, Z. Paszek: *Statystyka z elementami statystycznych metod monitorowania procesów*, Akademia Ekonomiczna, Kraków 2004, s. 9.

przypisywać oznaczenia w postaci liczb typu 1, 2, 3, ..., n, liter a, b, c, ..., lub innych symboli[3].

Przykład 1.1

Doświadczenie polega na przeprowadzeniu ankiety poprzedzającej budowę supermarketu. Ankietowani mogą wybrać jedną z trzech możliwych odpowiedzi:

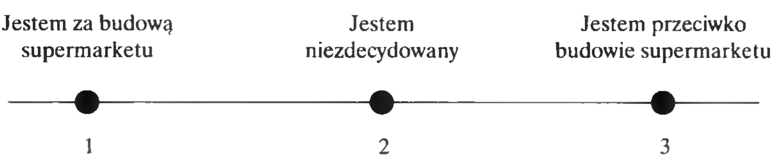
1) jestem za budową supermarketu,
2) jestem niezdecydowany,
3) jestem przeciwko budowie supermarketu.

Sposób zdefiniowania zdarzeń elementarnych i przestrzeni zdarzeń elementarnych zależy w tym przypadku od liczby respondentów, wśród których została przeprowadzona ankieta. Jeżeli rozpatrujemy ankietę dla jednej osoby, to wówczas przestrzeń zdarzeń elementarnych składa się tylko z trzech zdarzeń elementarnych ω_1, ω_2 i ω_3, odpowiadających odpowiedziom numer 1, 2 i 3.

Mamy zatem:

$$\Omega = \{\omega_1, \omega_2, \omega_3\}.$$

Zbiór wszystkich możliwych wyników ankiety dla jednej osoby można przedstawić również graficznie (zob. rys. 1).



Rys. 1. Zbiór możliwych wyników ankiety dla jednej osoby

Źródło: opracowanie własne.

Jeżeli analizujemy wyniki ankiety dla dwóch osob oraz gdy oznaczymy kolejne typy odpowiedzi cyframi 1, 2 i 3, to wówczas zbiór Ω składa się z dziewięciu zdarzeń elementarnych.

Mamy wówczas:

$$\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5, \omega_6, \omega_7, \omega_8, \omega_9\},$$

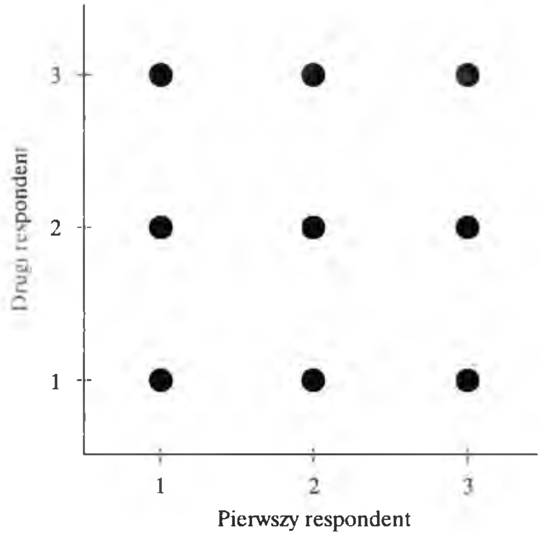
gdzie:

$$\omega_1 = (1,1), \omega_2 = (1,2), \omega_3 = (1,3), \omega_4 = (2,1), \omega_5 = (2,2),$$
$$\omega_6 = (2,3), \omega_7 = (3,1), \omega_8 = (3,2), \omega_9 = (3,3).$$

Przejrzystość analizowanego przypadku można zwiększyć, przedstawiając wyniki ankiety graficznie (zob. rys. 2).

---

[3] Zob. np. J. E. Freund: *Podstawy nowoczesnej statystyki*, PWE, Warszawa 1968, s. 88.



Rys. 2. Zbiór możliwych wyników ankiety dla dwóch osób

Źródło: opracowanie własne.

W podobny sposób zdefiniować można przestrzeń zdarzeń elementarnych dla trzech i więcej respondentów, przy czym graficzna prezentacja wyników w konwencji rysunków 1 i 2 – dla czterech lub więcej osób – staje się niemożliwa. Na przykład zapis (1,3,1,2) oznacza zdarzenie elementarne, w którym pierwszy i trzeci z czterech badanych respondentów jest za budową supermarketu, drugi jest przeciwnikiem budowy, natomiast czwarty jest niezdecydowany. W przypadku trzech osób liczba zdarzeń elementarnych wyniesie 27, czterech 81, pięciu 729 itd. Ogólnie liczba możliwych zdarzeń elementarnych wynosi 3^n, gdzie n jest liczbą osób uczestniczących w ankiecie.

Często w zagadnieniach praktycznych wykorzystuje się nie pojedyncze zdarzenia elementarne (wyniki doświadczeń), lecz ich zbiory, będące podzbiorami przestrzeni Ω. Każdy taki podzbiór, dla skończonej lub przeliczalnej przestrzeni zdarzeń losowych, nazywa się **zdarzeniem losowym**. Podzbiór przestrzeni zdarzeń elementarnych Ω, może stanowić także podzbiór pusty (**zdarzenie niemożliwe**) oraz całą przestrzeń Ω (**zdarzenie pewne**). Do oznaczenia zdarzeń wykorzystuje się zwykle duże litery początku alfabetu A, B, C, D... itd.

Jeżeli przestrzeń Ω jest nieprzeliczalna, to wtedy nie wszystkie jej podzbiory są zdarzeniami losowymi. Wówczas spośród wszystkich podzbiorów wyodrębnia się określoną klasę B (zbiór zbiorów, rodzinę), której elementy nazywamy zdarzeniami losowymi. Klasę taką zwykło nazywać się **przeliczalnym addytywnym ciałem zda-**

rzeń (lub **borelowskim ciałem zdarzeń** lub **sigma ciałem zdarzeń**). Borelowskie ciało zdarzeń musi spełniać następujące warunki[4]:

I. Cała przestrzeń Ω należy do tej klasy:

$$\Omega \in B. \tag{1.1}$$

II. Dopełnienie A' dowolnego zbioru A należącego do klasy B jest elementem tej klasy:

$$A \in B \Rightarrow A' \in B. \tag{1.2}$$

III. Suma co najwyżej przeliczalnej (skończonej lub przeliczalnej) liczby zbiorów należących do klasy B również należy do tej klasy:

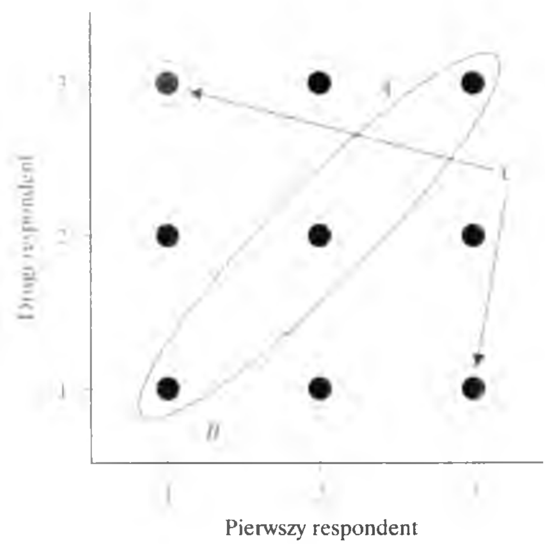
$$A_1 \in B, A_2 \in B, A_3 \in B, \ldots \Rightarrow (A_1 \cup A_2 \cup A_3 \cup \ldots) \in B. \tag{1.3}$$

Przykład 1.1 cd.

Na podstawie dwuwymiarowej przestrzeni zdarzeń elementarnych można zdefiniować – na przykład – zdarzenia losowe:

A – co najmniej jeden z uczestników ankiety jest niezdecydowany,
B – obaj respondenci dają taką samą odpowiedź,
C – jedna osoba popiera budowę supermarketu, a druga jest przeciw.

Graficzną prezentację tych zdarzeń przedstawiono na rys. 3.



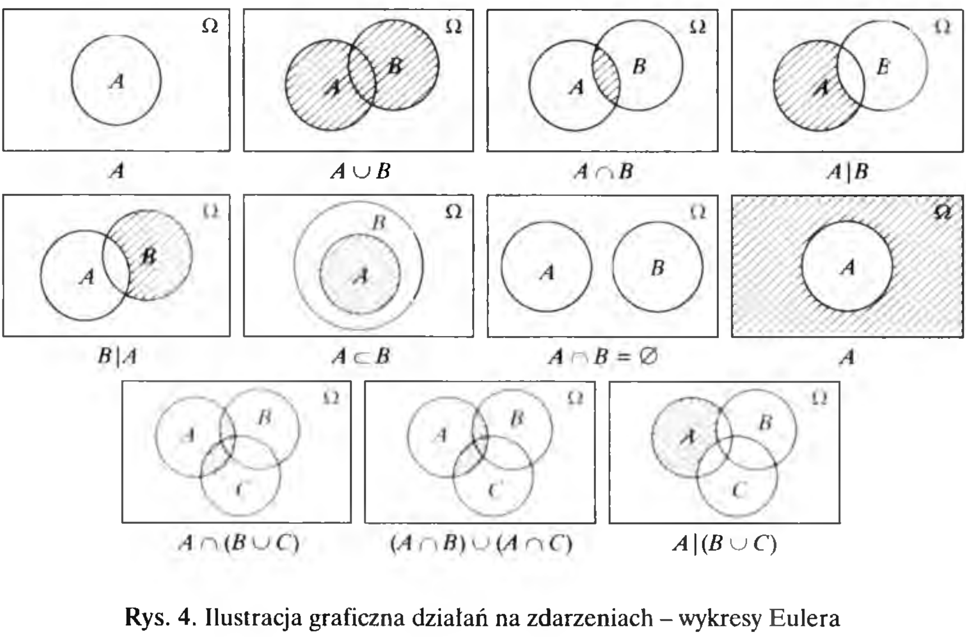
Rys. 3. Zdarzenia losowe A, B i C

Źródło: opracowanie własne.

[4] Zob. Krysicki i inni: *op. cit.*, s. 7.

Zdarzenia są zbiorami, a więc można dla nich stosować działania określone w teorii zbiorów. Do oznaczenia koniunkcji (iloczynu) używamy znaku $\cap$ a alternatywy (sumy) znaku $\cup$ (np. $A \cap B, A \cap B$). Znaku | używamy natomiast do oznaczenia różnicy pomiędzy zdarzeniami (np. $A \mid B$ oznacza, wszystkie zdarzenia elementarne należące do A, lecz nie należące do B). Znak $\subset$ stosujemy do oznaczenia przypadku, kiedy jedno zdarzenie pociąga drugie zdarzenie (np. $A \subset B$ oznacza, że A pociąga B, lub B jest następstwem zdarzenia A). Jeżeli dwa zdarzenia pociągają się nawzajem, to wówczas mówimy o nich, że są równe, i stawiamy pomiędzy nimi znak $=$ (np. jeżeli $A \subset B$ i $B \subset A \Rightarrow A = B$). O zdarzeniach możemy również powiedzieć, że się wykluczają (lub wyłączają się), jeżeli ich koniunkcja jest zdarzeniem niemożliwym. (np. A i B wykluczają się, jeżeli $A \cap B = \varnothing$).

Przestrzeń zdarzeń elementarnych i zdarzenia oraz związki pomiędzy nimi przedstawiane są często w postaci wykresów Eulera[5], w których do oznaczenia przestrzeni Ω używa się prostokąta, a do zdarzeń koła lub jego fragmentu[6]. Zacieniowany obszar ilustruje rozważane zdarzenie. Przykłady takich wykresów ilustruje rys. 4.



Rys. 4. Ilustracja graficzna działań na zdarzeniach – wykresy Eulera

Źródło: opracowanie własne.

[5] Można się również spotkać z nazwą: diagram Venna (zob. np. J.E. Freund: *op. cit.*, s. 92).
[6] Dopuszczalne jest również do oznaczenia zdarzeń stosowanie innych nieregularnych kształtów.

Przykład 1.1 cd.

Określić dla zdarzeń losowych A, B i C: $B \cup C, A \cap B, A \cap C, B \cap C$.

$B \cup C$ oznacza zdarzenie, w którym obydwie osoby odpowiadają tak samo lub jedna jest za, a druga przeciw budowie sklepu. Zdarzeniu takiemu odpowiada następujący podzbiór:

$$B \cup C = \{(1,1);(2,2);(3,3);(1,3);(3,1)\}.$$

$A \cap B$ oznacza zdarzenie, w którym obie osoby odpowiedziały tak samo i co najmniej jedna z nich była niezdecydowana. Zdarzeniu takiemu odpowiada podzbiór jednoelementowy:

$$X \cap Y = \{(2,2)\}.$$

Zdarzenia A i C oraz B i C to zdarzenia rozłączne (wzajemnie się wykluczające), zatem ich część wspólna jest równa zbiorowi pustemu, co można zapisać:

$$A \cap C = B \cap C = \varnothing.$$

Kolejnym ważnym pojęciem występującym podczas badania zbiorów lub zdarzeń jest **funkcja zbioru**, poprzez którą należy rozumieć funkcję przypisującą liczby różnym podzbiorom zbioru[7]. Liczby te mogą odpowiadać np. liczbie elementów zbioru (liczebności zbioru).

Przykład 1.1 cd.

Określić liczebność następujących zdarzeń losowych:

$$\Omega, A, B, C, A \cap B, A \cap C.$$

Oznaczając liczebność zbioru symbolem n, otrzymamy:

$$n_\Omega = 9, n_A = 5, n_B = 3, n_C = 2, n_{(A \cap B)} = 1, n_{(A \cap C)} = 0.$$

Funkcja zbioru charakteryzuje się następującymi własnościami:

1) liczby przyporządkowane poszczególnym zbiorom są zawsze dodatnie lub równe zeru,
2) wszystkie liczby są równe lub mniejsze od liczebności całej przestrzeni zdarzeń elementarnych n_Ω,
3) jeżeli dwa podzbiory nie posiadają wspólnych elementów, to liczba przyporządkowana sumie tych podzbiorów jest równa sumie liczb przyporządkowanych poszczególnym podzbiorom.

---

[7] Zob. J.E. Freund: *op. cit.*, s. 93.

1.2. Pojęcie prawdopodobieństwa

Prawdopodobieństwo można określić jako umowną miarę szansy realizacji w rzeczywistości pewnego zdarzenia. Jest ono liczbą przypisaną różnym zdarzeniom, takim jak na przykład wyrzucenie szóstki kostką sześcienną do gry, trafienie „trójki" w grze hazardowej, zdanie egzaminu z danego przedmiotu, osiągnięcie korzyści na giełdzie itp. Na liczbę tę nakłada się jednak pewne ograniczenia nazwane pewnikami lub aksjomatami[8]. Pewniki te przedstawiają się następująco[9]:

1. Każdemu zdarzeniu losowemu $A \in B$ można przypisać jednoznacznie nieujemną liczbę $P(A)$ określaną jako prawdopodobieństwo zdarzenia A.

$$_{A \in B}(A) \geq 0 \tag{1.4}$$

2. Jeśli Ω stanowi przestrzeń próby doświadczenia, to jego prawdopodobieństwo wynosi 1.

$$P(\Omega) = 1 \tag{1.5}$$

3. Jeżeli $A_1, A_2, A_3,\ldots$ są zdarzeniami wzajemnie wyłączającymi się (są parami rozłącznych zdarzeń), to:

$$P(A_1 \cup A_2 \cup A_3 \cup \ldots) = P(A_1) + P(A_2) + P(A_3) + \ldots. \tag{1.6}$$

Na podstawie powyższych pewników oraz z własności sigma ciała B wnioskujemy, że:

1. Jeżeli A i B są zdarzeniami wzajemnie się wykluczającymi, to wówczas:

$$P(A \cup B) = P(A) + P(B) \tag{1.7}$$

2. Prawdopodobieństwo zdarzenia niemożliwego wynosi zero:

$$P(\varnothing) = 0 \tag{1.8}$$

Podsumowując, można powiedzieć, że prawdopodobieństwem nazwiemy taką funkcję P, która jest nieujemną i przeliczalnie addytywną miarą zdarzeń. Jest to tak zwana **aksjomatyczna definicja prawdopodobieństwa**. Dzięki funkcji prawdopodobieństwa (P) zdefiniowaną we wcześniejszym paragrafie przestrzeń zdarzeń elementarnych możemy zastąpić pełniejszą charakterystyką doświadczeń losowych, jaką jest **przestrzeń probabilistyczna** postaci:

$$(\Omega, B, P) \tag{1.9}$$

[8] Zob. J.E. Freund: *op. cit.*, s. 98, zob. także M. Sobczyk: *Statystyka*, PWN, Warszawa 1998, s. 67.

[9] Aksjomaty te zostawały sformułowane w 1933 r. przez A. Kołmogorowa.

Przykład 1.2

Określić przestrzeń probabilistyczną dla doświadczenia losowego, polegającego na jednokrotnym rzucie monetą.

Aby określić przestrzeń probabilistyczną typu (1.9), należy zdefiniować jej trzy składowe: Ω, B oraz P.

W wyniku rzutu monetą realizuje się jedno z dwóch możliwych zdarzeń elementarnych: wypadnie orzeł (o), wypadnie reszka (r).

Przestrzeń zdarzeń elementarnych (Ω) tworzy tu zbiór dwuelementowy: $\Omega = \{o, r\}$. Ciało zdarzeń B tworzą cztery zbiory $B = \{\varnothing, \Omega, \{o\}, \{r\}\}$, gdzie $\varnothing$ oznacza zdarzenie niemożliwe, Ω – zdarzenie pewne (wyrzucono orła (o) lub reszkę (r)).

Prawdopodobieństwa zdarzeń losowych są równe odpowiednio:

$$P(o) = P(r) = \tfrac{1}{2} \text{ (oczywiście } P(\Omega) = 1 \text{ i } P(\varnothing) = 0). \{\text{kp}\}$$

Definicja aksjomatyczna nie jest jednak jedyną definicją prawdopodobieństwa, jaką można znaleźć, studiując literaturę przedmiotu. Funkcjonują trzy zasadnicze definicje prawdopodobieństwa. Najstarszą z nich jest definicja sformułowana w 1912 roku przez Laplace'a (tzw. **definicja klasyczna**), która głosi, że jeżeli przestrzeń zdarzeń elementarnych Ω jest skończonym zbiorem, jednakowo możliwych i wzajemnie wykluczających się n zdarzeń, oraz jeśli m z nich tworzy zdarzenie A, to

$$P(A) = \frac{m}{n}. \tag{1.10}$$

Definicja ta może być stosowana tylko wtedy, gdy każdy wynik doświadczenia jest jednakowo prawdopodobny.

Przykład 1.3

Niech zdarzenie A oznacza zdarzenie losowe polegające na wylosowaniu asa z talii 52 kart do gry w brydża. Przestrzeń zdarzeń Ω składa się z 52 elementów (kart), natomiast zdarzeniu A sprzyjają cztery wyniki (as karo, as kier, as pik, as trefl). Korzystając z klasycznej definicji, prawdopodobieństwo zajścia zdarzenia A wyniesie:

$$P(A) = \frac{4}{52} = 0{,}0769\,. \{\text{kp}\}$$

Definicja klasyczna zawiera jednak pewien błąd logiczny, gdyż do opisu pojęcia prawdopodobieństwa wykorzystuje pojęcie prawdopodobieństwa. W niektórych pozycjach literatury, aby wybrnąć z tej niedogodności, zastępuje się słowo prawdopodobnych słowem możliwych, co nie usuwa jednak tej niedogodności, gdyż prawdopodobny i możliwy to synonimy. Wadą tej definicji jest również to, że przestrzeń Ω oraz zdarzenie A muszą być zbiorami znanymi i skończonymi.

Rozszerzając klasyczną definicję prawdopodobieństwa na nieprzeliczalne zbiory zdarzeń elementarnych utworzono tzw. **geometryczną definicję prawdopodobieństwa**. Głosi ona, że jeżeli przestrzeń Ω jest zbiorem geometrycznym G o znanej mierze (tj. długość, pole, objętość), to prawdopodobieństwo zdarzenia A polegającego na losowym wyborze punktu (zbioru punktów) $g \in G$ wynosi:

$$P(A) = \frac{mes(g)}{mes(G)}, \tag{1.11}$$

gdzie *mes* jest skrótem angielskiego słowa *measure* (miara). {kp}

Przykład 1.4

Załóżmy, że przestrzeń zdarzeń elementarnych tworzy przedział na osi liczbowej $G = [10; 20]$, a zdarzenie A polega na tym, że losowo wybrany punkt będzie należał do przedziału $g = [12;15]$. Jako miarę zbioru przyjmiemy długości rozważanych przedziałów, które oznaczymy odpowiednio ΔG i Δg.

Wykorzystując wzór (1.10), obliczamy iloraz długości:

$$P(A) = \frac{\Delta g}{G} = \frac{15 - 12}{20 - 10} = 0{,}3\,. \text{ \{kp\}}$$

Prawdopodobieństwo można również zdefiniować, wykorzystując pojęcie częstości względnej (frakcji) – tzw. **definicja częstościowa**[10]. U podstaw tej definicji leży **prawo wielkich liczb**. Głosi ono, że jeśli liczba powtórzeń doświadczenia rośnie nieograniczenie, to frakcja liczby przypadków, w którym otrzymano dany wynik (częstość względna danego wyniku) dąży do prawdopodobieństwa pojawienia się tego wyniku w jednym doświadczeniu. Można zatem zapisać, że:

$$P(A) = \lim_{n \to \infty} \frac{n(A)}{n}, \tag{1.12}$$

gdzie

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

n – liczba wykonanych doświadczeń.

Zaletą definicji częstościowej jest to, że usuwa ona błąd *idem per idem* charakterystyczny dla definicji klasycznej. Wadą natomiast jest to, że aby ustalić dokładną

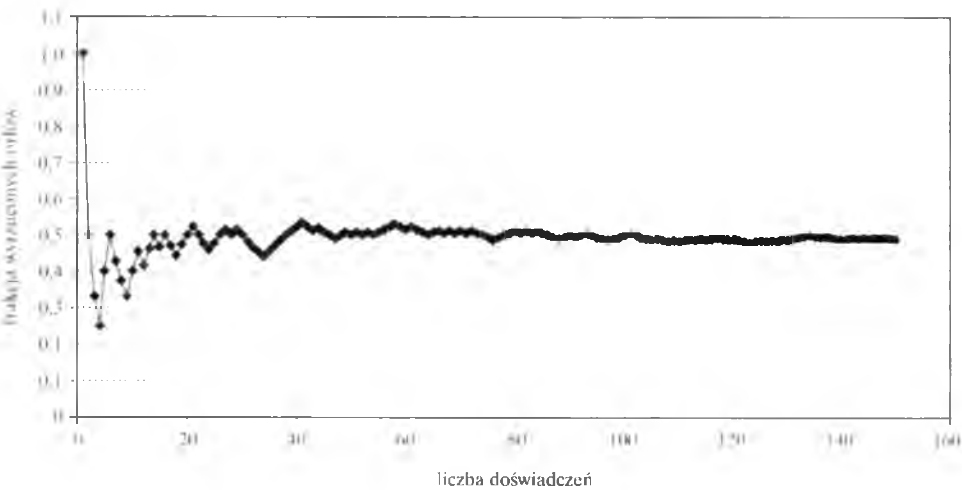
---

[10] Definicję w ujęciu częstościowym przypisuje się R.V. Misesowi, austriackiemu statystykowi żyjącemu w latach 1883–1953.

wartość prawdopodobieństwa – wykorzystując definicję czystościową – należy przeprowadzić nieskończoną liczbę doświadczeń, co w praktyce jest niemożliwe.

Przykład 1.5

Postanowiono oszacować prawdopodobieństwo zdarzenia A polegającego na wyrzuceniu orła w pojedynczym rzucie monetą. W tym celu przeprowadzono doswiadczenie polegające na 150-krotnym rzucie monetą jednozłotową i obserwowano, co wypadnie – „orzeł” czy „reszka”. W każdym kroku doświadczenia wyznaczano i nanoszono na wykres wartość frakcji wyrzuconych orłów, dzieląc liczbę wyrzuconych orłów przez liczbę doświadczeń. Wyniki doświadczenia przedstawia rys. 5.



Rys. 5. Frakcje „orłów” w kolejnych 150 rzutach monetą

Źródło: opracowanie na podstawie badań własnych.

Patrząc na rysunek 5, można wyraźnie zauważyć, że w miarę zwiększania liczby rzutów frakcja wyrzuconych orłów dąży do wartości 0,5, która jest oszacowaną wartością prawdopodobieństwa wyrzucenia orła w pojedynczym rzucie.

Zauważmy również, że prawdopodobieństwo zdarzenia A można również wyznaczyć, stosując definicję klasyczną. Spełnione są bowiem warunki jej stosowalności. Przestrzeń zdarzeń elementarnych jest skończona. Składa się ona z dwóch rozłącznych zdarzeń elementarnych (o – wyrzucono orła i r – wyrzucono reszkę), Liczebność zdarzeń sprzyjających zajściu zdarzenia A wynosi $m = 1$.

Zatem: $P(A) = m/n = ½$. {kp}

Coraz częściej w literaturze można spotkać się z jeszcze jednym ujęciem prawdopodobieństwa, tzw. **ujęciem subiektywnym**[11], przez które rozumie się pewną postawę albo wątpliwość dotyczącą jakiegoś zdarzenia przyszłego. Na przykład formułując twierdzenia w rodzaju: „prawdopodobnie jutro będę w pracy", „małe jest prawdopodobieństwo, że zdam za pierwszym razem ten egzamin" czy „uważam, że mam 90 procent szans na 100, że skończę studia" lub „prawdopodobieństwo, że pacjent wyzdrowieje, wynosi ½", dajemy wyraz pewnym odczuciom lub wątpliwościom w odniesieniu do zdarzeń przyszłych.

W wielu przypadkach subiektywne odczucie szansy zajścia określonego zjawiska w przyszłości staje się początkiem (zaczynem) przeprowadzanych eksperymentów i badań empirycznych, które mają potwierdzić lub zaprzeczyć subiektywnie oszacowaną wartość prawdopodobieństwa.

1.3. Podstawowe własności prawdopodobieństwa

Z trzech aksjomatów prawdopodobieństwa wynikają dalsze własności[12]:

1. Prawdopodobieństwo każdego zdarzenia A jest liczbą z przedziału $[0;1]$:

$$_{A \in B}\; 0 \leq P(A) \leq 1. \tag{1.13}$$

3. Prawdopodobieństwo zdarzenia niemożliwego wynosi zero:

$$P(\varnothing) = 0. \tag{1.14}$$

4. Prawdopodobieństwo zdarzenia A', które jest zdarzeniem przeciwnym do zdarzenia A, jest równe:

$$P(A') = 1 - P(A). \tag{1.15}$$

5. Jeżeli zdarzenie losowe A zawiera się w zdarzeniu losowym B ($A \subset B$), to wówczas:

$$P(B) \geq P(A). \tag{1.16}$$

6. Prawdopodobieństwo sumy dowolnych dwóch zdarzeń losowych A i B jest równe sumie prawdopodobieństw tych zdarzeń pomniejszonych o prawdopodobieństwo ich iloczynu:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B). \tag{1.17}$$

[11] Zob. np. G. A. Ferguson, Y. Takane: *Analiza statystyczna w psychologii i pedagogice*, PWN, Warszawa 2002.

[12] Zob. także A. Iwasiewicz, Z. Paszek: *Statystyka...*, *op. cit.*, s. 15.

7. Jeżeli A i B są zdarzeniami losowymi należącymi do pewnej przestrzeni Ω, oraz jeżeli $P(B) \neq 0$, to prawdopodobieństwo warunkowe (względne) zdarzenia A pod warunkiem B jest określane wzorem:

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

stąd

$$P(A \cap B) = P(A) \cdot P(A|B). \tag{1.19}$$

Podobnie

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

stąd

$$P(A \cap B) = P(A) \cdot P(B|A). \tag{1.21}$$

8. Jeżeli A i B są zdarzeniami niezależnymi, to wówczas:

$$P(A|B) = P(A), \tag{1.22}$$

oraz

$$P(B|A) = P(B), \tag{1.23}$$

oraz

$$P(A \cap B) = P(A)\ P(B). \tag{1.24}$$

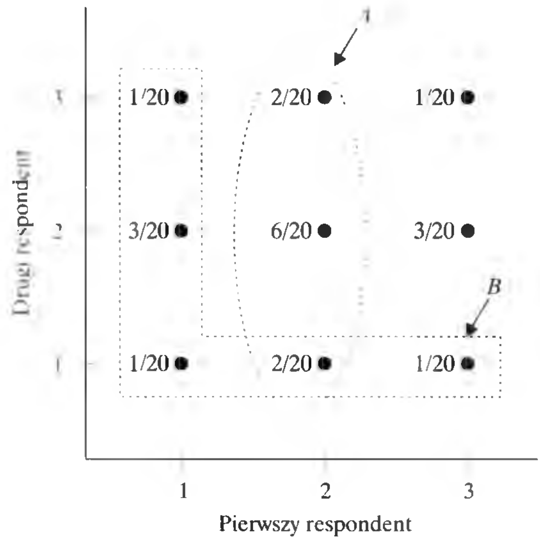
Przykład 1.6

Załóżmy, że zdarzeniom elementarnym opisanym w przykładzie 1.1 przyporządkujemy następujące wartości prawdopodobieństw:

$$P(\omega_1) = P(1,1) = 1/20, P(\omega_2) = P(1,2) = 2/20, P(\omega_3) = P(1,3) = 1/20,$$
$$P(\omega_4) = P(2,1) = 3/20, P(\omega_5) = P(2,2) = 6/20, P(\omega_6) = P(2,3) = 3/20,$$
$$P(\omega_7) = P(3,1) = 1/20, P(\omega_8) = P(3,2) = 2/20, P(\omega_9) = P(3,3) = 1/20.$$

Rozważmy następujące zdarzenia losowe:
A – pierwsza osoba jest niezdecydowana,
B – co najmniej jedna z badanych osób poprze budowę supermarketu.

Wartości prawdopodobieństw oraz zdarzenia losowe A i B prezentuje rys. 6.



Rys. 6. Graficzna prezentacja zdarzeń losowych

Źródło: opracowanie własne.

Wyznaczyć prawdopodobieństwo następujących zdarzeń losowych:

$$P(A), P(B), P(A \cap B), P(A \cup B).$$

Prawdopodobieństwo każdego zdarzenia losowego równa się sumie prawdopodobieństw wyników do niego należących. Mamy zatem:

$$P(A) = 2/20 + 6/20 + 2/20 = 10/20 = 1/2,$$

oraz

$$P(B) = 1/20 + 3/20 + 1/20 + 2/20 + 1/20 = 8/20 = 2/5.$$

Prawdopodobieństwo iloczynu zdarzeń losowych A i B jest równe prawdopodobieństwu przypisanemu punktowi o współrzędnych (2,1), co daje:

$$P(A \cap B) = 2/20 = 1/10.$$

Korzystając z własności (1.17), można wyznaczyć, że:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 1/2 + 2/5 - 2/20 = 16/20 = 4/5. \ \{\text{kp}\}$$

Przykład 1.7

Wśród dziesięciu osób przeprowadzono ankietę na temat budowy supermarketu. Dwie osoby opowiadały się za budową supermarketu, trzy były niezdecydowane, natomiast pięć kolejnych było przeciw budowie supermarketu. Ankiety zostały umieszczone w urnie.

Obliczyć prawdopodobieństwo następujących zdarzeń losowych:

a) losując trzykrotnie bez zwracania ankietę z urny, trzy razy otrzymano ankietę z zaznaczoną odpowiedzią: jestem niezdecydowany,

b) losując ankietę dwukrotnie bez zwracania do urny, dokładnie za drugim razem wylosowano ankietę z zaznaczoną odpowiedzią: jestem przeciwny budowie supermarketu.

ad a)

Przyjmijmy następujące oznaczenia poszczególnych zdarzeń losowych:

C_1 – w pierwszym losowaniu wyciągnięto ankietę z zaznaczoną odpowiedzią: „jestem niezdecydowany";

$C_2|C_1$ – w drugim losowaniu wyciągnięto ankietę z zaznaczoną odpowiedzią: „jestem niezdecydowany", przy założeniu (pod warunkiem), że w pierwszym losowaniu wyciągnięto ankietę z zaznaczoną identyczną odpowiedzią;

$C_3|C_2C_1$ – w trzecim losowaniu wyciągnięto ankietę z zaznaczoną odpowiedzią: „jestem niezdecydowany", przy założeniu (pod warunkiem), że w pierwszym i drugim losowaniu wyciągnięto ankiety z zaznaczoną identyczną odpowiedzią;

C – wylosowano po kolei trzy ankiety z zaznaczonymi odpowiedziami: „jestem niezdecydowany".

Na realizację zdarzenia A składa się koniunkcja trzech pierwszych zdarzeń. Można zatem zapisać:

$$C = \{C_1\} \cap \{C_2|C_1\} \cap \{C_3|C_2C_1\}.$$

Prawdopodobieństwa tych zdarzeń będą równe:

$$P(C_1) = \frac{3}{10},\ P(C_2|C_1) = \frac{2}{9},\ P(C_3|C_2C_1) = \frac{1}{8},$$

i w konsekwencji

$$P(C) = \frac{3}{10} \cdot \frac{2}{9} \cdot \frac{1}{8} = \frac{6}{720} \approx 0{,}008.$$

ad b)

Przyjmijmy następujące oznaczenia:

B_2 – w drugim losowaniu wyciągnięto ankietę z zaznaczoną odpowiedzią: „jestem przeciwny budowie supermarketu";

B_1 – w pierwszym losowaniu wyciągnięto ankietę z zaznaczoną odpowiedzią: „jestem przeciwny budowie supermarketu";

A_1 – w pierwszym losowaniu wyciągnięto ankietę z zaznaczoną odpowiedzią: „jestem za budową supermarketu";

C_1 – w pierwszym losowaniu wyciągnięto ankietę z zaznaczoną odpowiedzią: „jestem niezdecydowany".

Zdarzenie B_2 tj. dokładnie za drugim razem wylosowano ankietę z zaznaczoną odpowiedzią: „jestem przeciw budowie supermarketu", jest konsekwencją realizacji następujących zdarzeń:

$$B_2 = \{C_1 \cap B_2\} \cup \{A_1 \cap B_2\} \cup \{B_1 \cap B_2\}.$$

Prawdopodobieństwo zdarzenia B_2, będzie sumą iloczynów poszczególnych prawdopodobieństw:

$$P(B_2) = P(C_1 \cap B_2) + P(A_1 \cap B_2) + P(B_1 \cap B_2).$$

Korzystając z zależności (1.21), można zapisać:

$$P(B_2) = P(C_1) \cdot P(B_2|C_1) + P(A_1) \cdot P(B_2|A_1) + P(B_1) \cdot P(B_2|B_1) =$$

$$= \frac{3}{10} \cdot \frac{5}{9} + \frac{2}{10} \cdot \frac{5}{9} + \frac{5}{10} \cdot \frac{44}{99} = \frac{5}{0} = 0{,}5. \ \{kp\}$$

1.4. Prawdopodobieństwo zupełne i wzór Bayesa

Jeżeli $A_1, A_2, ..., A_k$ są zdarzeniami wykluczającymi się parami, oraz zdarzenia A_1, $A_2, ..., A_k$ tworzą układ zupełny zdarzeń (tzn. $P(A_1) + P(A_2) + ... + P(A_k) = 1$), to wówczas prawdopodobieństwo zdarzenia B, mogącego się zrealizować pod warunkiem zajścia poszczególnych zdarzeń A_i, można obliczyć w sposób następujący:

$$P(B) = \sum_{i=1}^{k} P(A_i) \cdot P(B|A_i). \tag{1.25}$$

Jeżeli A_1, A_2, ..., A_k są zdarzeniami wykluczającymi się parami oraz $P(B) > 0$, to wówczas prawdopodobieństwo zajścia dowolnego zdarzenia A_i pod warunkiem, że zaszło B, można obliczyć ze wzoru:

$$P(A_i|B) = \frac{P(A_i)\cdot P(B|A_i)}{\sum\limits_{i=1}^{n} P(A_i)\cdot P(B|A_i)} = \frac{P(A_i)\cdot P(B|A_i)}{P(B)} \tag{1.26}$$

Zależność (1.25) określa się mianem **prawdopodobieństwa zupełnego (całkowitego)**, natomiast wzór (1.26) nazywany jest **wzorem *Bayesa***.

Przykład 1.8

W urnie znajduje się 1000 ankiet sondażu poprzedzającego budowę supermarketu. Wśród badanych było 550 kobiet i 450 mężczyzn. Struktura odpowiedzi wśród kobiet kształtowała się następująco:

odpowiedź: jestem za budową supermarketu – 65% głosów,
odpowiedź: jestem przeciwny budowie supermarketu – 25% głosów,
odpowiedź: jestem niezdecydowany – 10% głosów.

Struktura odpowiedzi wśród mężczyzn przedstawiała się następująco:
odpowiedź: jestem za budową supermarketu – 55% głosów,
odpowiedź: jestem przeciwny budowie supermarketu – 25% głosów,
odpowiedź: jestem niezdecydowany – 20% głosów.

1) obliczyć prawdopodobieństwo zdarzenia losowego, że wybierając w sposób losowy jedną ankietę, trafimy na ankietę z zaznaczoną odpowiedzią: „jestem za budową supermarketu”,
2) wylosowano ankietę z zaznaczoną odpowiedzią: „jestem za budową supermarketu”, jakie jest prawdopodobieństwo, że tę ankietę wypełniała kobieta.

ad 1)
Używając symboli K, M, A, B, C, oznaczymy następujące zdarzenia losowe:
K – wylosowano ankietę wypełnioną przez kobietę,
M – wylosowano ankietę wypełnioną przez mężczyznę,
A – wyciągnięto ankietę z zaznaczoną odpowiedzią: „jestem za budową supermarketu”,
B – wyciągnięto ankietę z zaznaczoną odpowiedzią: „jestem przeciw budowie supermarketu”,
C – wyciągnięto ankietę z zaznaczoną odpowiedzią: „jestem niezdecydowany”.

Zauważmy, że zdarzenie A może zrealizować się w następujących układach:

$$A = \{K \cap A|K\} \cup \{M \cap A|M\}.$$

Ponieważ zdarzenia K i M są zdarzeniami wykluczającymi się, prawdopodobieństwo zdarzenia A możemy obliczyć zgodnie z wzorem (1.24):

$$P(A) = P(K) \cdot P(A \mid K) + P(M) \cdot P(A \mid M).$$

W analizowanym przykładzie poszczególne prawdopodobieństwa wynoszą:

$$P(K) = \frac{550}{1000}, P(A \mid K) = 0{,}65,$$

$$P(M) = \frac{450}{1000}, P(A \mid M) = 0{,}55.$$

W konsekwencji, prawdopodobieństwo zdarzenia losowego polegającego na wyciągnięciu ankiety wypełnionej przez zwolennika budowy supermarketu wynosi:

$$P(A) = 0{,}55 \cdot 0{,}65 + 0{,}45 \cdot 0{,}55 = 0{,}3575 + 0{,}2475 = 0{,}605\ (60{,}5\%).$$

ad 2)

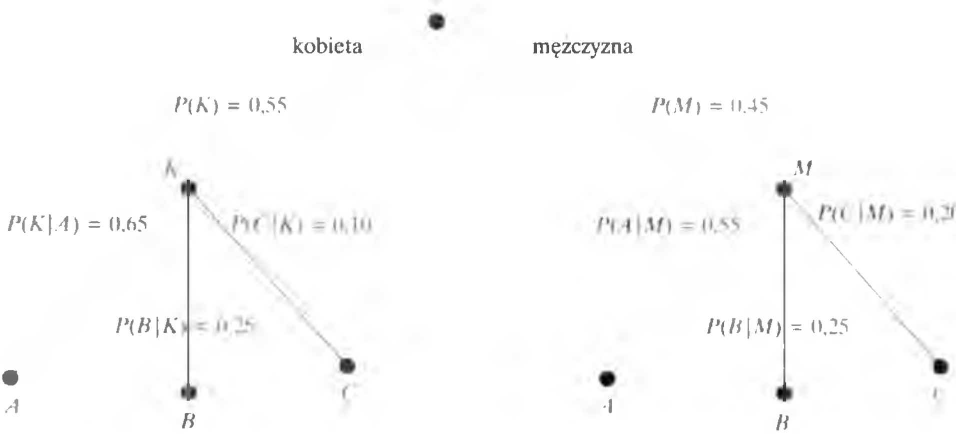
Prawdopodobieństwo zdarzenia losowego, że wylosowaną ankietę wypełniła kobieta, zakładając, że jest to ankieta respondenta popierającego budowę supermarketu, można wyznaczyć, stosując wzór *Bayesa* (1.26). Przyjmując powyższe oznaczenia z punktu 1 niniejszego przykładu, możemy zapisać:

$$P(K \mid A) = \frac{P(K) \cdot P(A \mid K)}{P(K) \cdot P(A \mid K) + P(M) \cdot P(A \mid M)} = \frac{P(K) \cdot P(A \mid K)}{P(A)}$$

W analizowanym przykładzie szukane prawdopodobieństwo wyniesie

$$P(K \mid A) = \frac{0{,}55 \cdot 0{,}65}{0{,}55 \cdot 0{,}65 + 0{,}45 \cdot 0{,}55} = \frac{0{,}3575}{0{,}605} \approx 0{,}59\ \ (59\%).$$

W procesie wyznaczania prawdopodobieństwa określonych zdarzeń losowych dosyć często wykorzystuje się technikę **dendrytu** (drzewa zdarzeń). Drzewo zdarzeń analizowanych w przykładzie 1.8 przedstawiono na rys. 7.



Rys. 7. Dendryt

Źródło: opracowanie własne.

Zauważmy, że do zdarzenia A prowadzić mogą dwie alternatywne ścieżki (zaznaczone linią przerywaną) biegnące przez punkt (zdarzenie) K lub przez punkt (zdarzenie) M. Po przemnożeniu prawdopodobieństw przypisanych gałęziom ścieżki pierwszej, a następnie drugiej i zsumowaniu tych iloczynów otrzymamy prawdopodobieństwo zupełne zdarzenia A.

Możemy zatem zapisać:

$$P(A) = 0{,}55 \cdot 0{,}65 + 0{,}45 \cdot 0{,}55 = 0{,}3575 + 0{,}2475 = 0{,}605. \{\text{kp}\}$$

Zmienne losowe

Rozdział

2

2.1. Jednowymiarowe zmienne losowe

2.1.1. Pojęcie zmiennej losowej

Intuicyjnie zmienną losową możemy nazwać taką wielkość, która w wyniku doświadczenia przyjmuje określoną wartość, znaną po zrealizowaniu doświadczenia, a nie dającą się przewidzieć przed realizacją doświadczenia[1].

Jest to więc taka funkcja, która w wyniku doświadczenia przybierze jedną i tylko jedną wartość, ze zbioru tych wartości, które ta zmienna może przyjąć.

Do oznaczenia zmiennych losowych używa się najczęściej wielkich końcowych liter alfabetu: $X, Y, Z, \ldots$, a wartosci tych zmiennych (realizacje zmiennych losowych) oznacza się odpowiednimi małymi literami: x, y, z, które w razie potrzeby rozszerza się dodatkowo o indeksy dolne np.: $x_1, x_2, x_3, \ldots; y_1, y_2, y_3, \ldots$ itd.

W literaturze przedmiotu znajdujemy następującą formalną definicję zmiennej losowej[2].

Niech (Ω, B, P) będzie dowolną przestrzenią probabilistyczną. **Zmienną losową** nazywać będziemy dowolną funkcję X, określoną na przestrzeni zdarzeń elementarnych Ω, o wartościach ze zbioru liczb rzeczywistych R, posiadającą następujące własności: dla dowolnej, ustalonej liczby rzeczywistej x zbiór zdarzeń elementarnych ω, dla których spełniona jest nierówność $X(\omega) < x$, jest zdarzeniem, czyli

$$\{\omega: X(\omega) < x\} \in B \text{ dla każdego } x \in R. \tag{2.1}$$

Warunek (2.1) staje się nieistotny, jeżeli przestrzeń zdarzeń elementarnych jest skończona. Wówczas **każda** funkcja X, która odwzorowuje zbiór zdarzeń elementarnych Ω w zbiór liczb rzeczywistych R będzie zmienną losową.

---

[1] Z. Helwig: *Elementy rachunku prawdopodobieństwa i statystyki matematycznej*, Warszawa 1967, s. 59.

[2] Definicja ta pochodzi z pracy: W. Krysicki i inni: *Rachunek...*, *op. cit.*, s. 48.

Rozróżnić można dwa typy zmiennych losowych:
- **zmienne losowe skokowe** (dyskretne, ziarniste),
- **zmienne losowe ciągłe**.

Pierwszy rodzaj zmiennych (zmienne losowe skokowe) to takie, dla których daje się określić skończony lub przeliczalny zbiór wartości. Przykładem takich zmiennych może być: liczba dzieci w rodzinie, liczba ankiet z zaznaczoną odpowiedzią: „popieram budowę supermarketu", liczba urodzeń, liczba zawartych małżeństw itp.

Do zmiennych losowych ciągłych zaliczane są te zmienne, które mogą przybierać dowolne wartości liczbowe, z pewnego przedziału liczbowego (w szczególności może to być przedział nieskończony). Do zmiennych tych zalicza się na przykład: wzrost w cm, ciężar w gramach, temperaturę w stopniach Celsjusza, ciśnienie w milimetrach słupa rtęci, grubość w mm, szerokość w cm itp. W praktyce, zbiór wartości zmiennej losowej jest zawsze skończony lub przeliczalny. Wartości zmiennej losowej i różnice między nimi zależą od czułości metody badawczej. Im metoda badawcza jest bardziej czuła, tym zbiór wartości zmiennej jest liczniejszy, a różnice między dwiema kolejnymi dowolnymi wartościami są na tyle małe, że zbiór wartości może być traktowany jak przedział na osi liczb rzeczywistych. Wynika z tego, że właściwie zmienne losowe powinno dzielić się na zmienne ciągłe i tzw. zmienne *quasi*-ciągłe. W dalszych rozważaniach pozostaniemy jednak, przy nazwie zmienne losowe ciągle, pamiętając jednocześnie, że taki rodzaj zmiennej może istnieć jedynie w czysto teoretycznej formie.

Przykład 2.1

W jednej z grup dziekańskich studiow magisterskich przeprowadzono egzamin z „Zarządzania jakością" i otrzymano następujące wyniki zestawione w tablicy 2.1.

Tablica 2.1.

nr (ω_i)	1	2	3	4	5	6	7	8	9	10
ocena	4,0	2,0	4,0	3,0	4,0	2,0	4,5	2,0	3,0	3,5
nr (ω_i)	11	12	13	14	15	16	17	18	19	20
ocena	3,5	3,0	4,5	3,0	2,0	3,0	3,0	4,0	3,0	4,0
nr (ω_i)	21	22	23	24	25	26	27	28	29	30
ocena	2,0	3,5	2,0	3,0	5,0	3,0	4,5	5,0	5,0	4,0

Zrodlo: opracowanie na podstawie badań własnych.

Przestrzenią zdarzeń elementarnych Ω jest tutaj zbiór 30 studentów wchodzących w skład badanej grupy studentów. Doświadczenie polega na wylosowaniu jednego studenta z badanej grupy dziekańskiej. Dyskretną zmienną losową X jest ocena egzaminowanego studenta. Zmienna X może przyjmować wartości z sześcioelementowego skończonego zbioru {2,0; 3,0; 3,5; 4,0; 4,5; 5,0}. Jeżeli przez ω_i oznaczymy numer studenta na liście obecności, to wówczas zmienną X możemy określić następująco:

$X(\omega_2) = X(\omega_6) = X(\omega_8) = X(\omega_{15}) = X(\omega_{21}) = X(\omega_{23}) = 2{,}0;$

$X(\omega_4) = X(\omega_9) = X(\omega_{12}) = X(\omega_{14}) = X(\omega_{16}) = X(\omega_{17}) = X(\omega_{19}) = X(\omega_{24}) = X(\omega_{26}) = 3{,}0;$

$X(\omega_{10}) = X(\omega_{11}) = X(\omega_{22}) = 3{,}5;$

$X(\omega_1) = X(\omega_3) = X(\omega_5) = X(\omega_{18}) = X(\omega_{20}) = X(\omega_{30}) = 4{,}0;$

$X(\omega_7) = X(\omega_{13}) = X(\omega_{27}) = 4{,}5;$

$X(\omega_{25}) = X(\omega_{28}) = X(\omega_{29}) = 5{,}0.$ {kp}

Zdefiniujmy obecnie pojęcie **niezależności zmiennych losowych**. Dwie zmienne losowe X i Y nazywamy niezależnymi, jeżeli określone są na tej samej przestrzeni probabilistycznej (Ω, B, P) i dla dowolnych zbiorów borelowskich prostej B_1, B_2, zachodzi równość[3]:

$$P(X \in B_1 \wedge X \in B_2) = P(X \in B_1)\ P(X \in B_2). \tag{2.1}$$

2.1.2. Funkcja rozkładu prawdopodobieństwa i dystrybuanta zmiennej losowej skokowej

Zmienna losowa X opisana w przykładzie 2.1 to zmienna skokowa, ponieważ może przyjmować tylko wybrane sześć wartości z przedziału od 2,0 do 5,0. Zauważmy, że zdarzeniu losowemu polegającemu na tym, że losowo wybrany student otrzyma jedną ze zbioru sześciu ocen, możemy przypisać określone prawdopodobieństwo (częstość) realizacji. Prawdopodobieństwo to można określić, wykorzystując zdefiniowaną w punkcie 1.3 częstościową definicję prawdopodobieństwa. Regułę (metodę), na podstawie której odbywa się rozdział masy prawdopodobieństwa, na poszczególne wartości zmiennej losowej, nazywa się **funkcją rozkładu prawdopodobieństwa** (rozkładem prawdopodobieństwa) lub krócej **rozkładem zmiennej**.

W przypadku zmiennej dyskretnej funkcję rozkładu prawdopodobieństwa można zdefiniować następująco:

$$P(X = x_i) = p(x_i) = p_i, \; i = 1, \ldots, k \tag{2.3}$$

lub w formie tabelarycznej:

---

[3] Zob. np. M. Krzyśko: *Wykłady z teorii prawdopodobieństwa*, Warszawa 2000, s. 122–128.

x_1	x_2	...	x_i	...	x_k
p_1	p_2	...	p_i	...	p_k

(2.4)

W obydwu przypadkach musi być spełniony warunek:

$$\sum_{i=1}^{k} p_i = p_1 + p_2 + \ldots + p_i + \ldots + p_k = 1 \qquad (2.4)$$

Przykład 2.2

Wykorzystując dane z przykładu 2.1, określić rozkład prawdopodobieństwa zmiennej losowej X w formie analitycznej i tabelarycznej.

Rozwiązanie:

$$p_1 = P(X = 2{,}0) = \frac{6}{30}, \qquad p_2 = P(X = 3{,}0) = \frac{9}{30}, \qquad p_3 = P(X = 3{,}5) = \frac{3}{30},$$

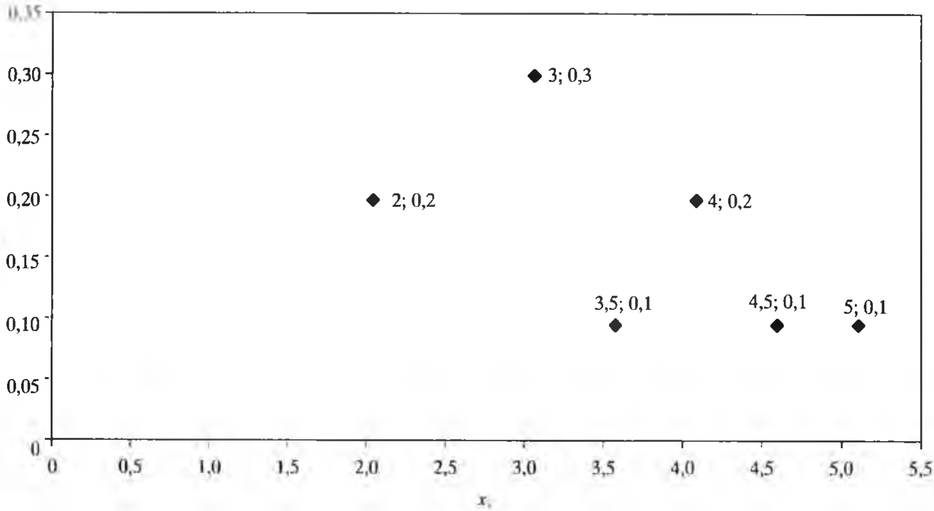
$$p_4 = P(X = 4{,}0) = \frac{6}{30}, \qquad p_5 = P(X = 4{,}5) = \frac{3}{30}, \qquad p_6 = P(X = 5{,}0) = \frac{3}{30}.$$

Funkcję rozkładu prawdopodobieństwa zmiennej X można przedstawić następująco:

x_i	2,0	3,0	3,5	4,0	4,5	5,0
p_i	0,2	0,3	0,1	0,2	0,1	0,1

Oczywiście $\sum_{i=1}^{6} p_i = 1.$

Dla określonego rozkładu zmiennej losowej X, można również sporządzić wykres. W rozważanym przykładzie wykres taki zaprezentowano na rys. 8.



Rys. 8. Wykres funkcji rozkładu prawdopodobieństwa zmiennej X z przykładu 2.2

Źródło: opracowanie własne. {kp}

Zdefiniujmy obecnie pojęcie **dystrybuanty** zmiennej losowej skokowej.

Dystrybuanta zmiennej losowej X w punkcie x_0 ($x_0 \in R$) jest prawdopodobieństwem zdarzenia losowego polegającego na tym, że zmienna losowa X przybierze wartość mniejszą niż x_0. Do oznaczenia dystrybuanty używać będziemy dużej litery F_x bądź F, jeżeli nie ma wątpliwości, jakiej zmiennej losowej ona dotyczy. Możemy więc zapisać, że:

$$F_X : R \to R \text{ i } F_X(x_0) = P(X < x_0). \tag{2.5}$$

Jeżeli X jest zmienną losową dyskretną, to wówczas dystrybuanta wynosi:

$$F_x(x_x) = \sum_{x_i < x_0} p_i. \tag{2.6}$$

Dystrybuanta zmiennej losowej przyjmuje następujące własności:

1) $\bigwedge_{x \in R} F(x) \in [0;1]$, (2.7)

2) jeżeli $x_1 < x_2 \Rightarrow F(x_1) \leq F(x_2)$, (2.8)

3) $\lim_{n \to -\infty} F(x) = 0$, (2.9)

4) $\lim_{n \to +\infty} F(x) = 1$, (2.10)

5) $P(a \leq X < b) = F(b) - F(a)$, (2.11)

6) jest funkcją lewostronnie ciągłą. (2.12)

Przykład 2.3

Wykorzystując rozkład prawdopodobieństwa zmiennej losowej X określony w przykładzie 2.2, wyznaczyć analitycznie i graficznie dystrybuantę F_x.

Dystrybuantę znajdujemy, wykorzystując wzór (2.6). Zmienna losowa może przybierać sześć wartości, zatem dystrybuanta określona będzie sześcioma następującymi wzorami:

Jeżeli $x_0 \in (-\infty; 2{,}0]$ to $F(x_0) = \sum_{x_i < x_0} p_i = 0,$

Jeżeli $x_0 \in (2{,}0; 3{,}0]$ to $F(x_0) = \sum_{x_i < x_0} p_i = p_1 = 0{,}2,$

Jeżeli $x_0 \in (3{,}0; 3{,}5]$ to $F(x_0) = \sum_{x_i < x_0} p_i = p_1 + p_2 = 0;2 + 0{,}3 = 0{,}5,$

Jeżeli $x_0 \in (3{,}5; 4{,}0]$ to $F(x_0) = \sum_{x_i < x_0} p_i = p_1 + p_2 + p_3 = 0;2 + 0{,}3 + 0{,}1 = 0{,}6,$

Jeżeli $x_0 \in (4{,}0; 4{,}5]$ to $F(x_0) = \sum_{x_i < x_0} p_i = p_1 + p_2 + p_3 + p_4 =$

$= 0;2 + 0{,}3 + 0{,}1 \cdot + 0{,}2 = 0{,}8,$

Jeżeli $x_0 \in (4{,}5; 5{,}0]$ to $F(x_0) = \sum_{x_i < x_0} p_i = p_1 + p_2 + p_3 + p_4 + p_5 =$

$= 0;2 + 0{,}3 + 0{,}1 +\backslash\, 0{,}2 + 0{,}1 = 0{,}9,$

Jeżeli $x_0 \in (5{,}0; +\infty)$ to $F(x_0) = \sum_{x_i < x_0} p_i = p_1 + p_2 + p_3 + p_4 + p_5 + p_6 =$

$= 0;2 + 0{,}3 + 0{,}1 + 0{,}2 + 0{,}1 + 0{,}1 = 1.$

Dystrybuantę można również ująć w formie następującej tabelki:

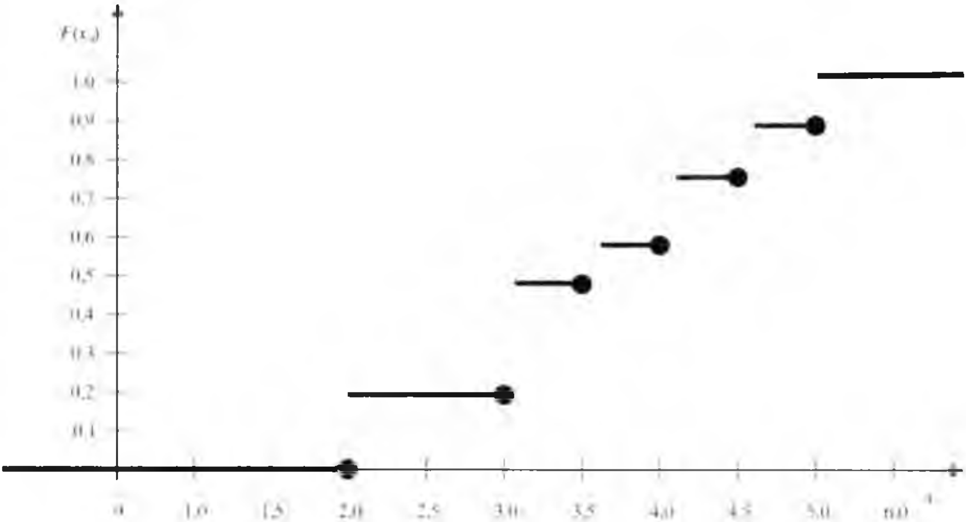
x_0	$(-\infty; 2{,}0]$	$(2{,}0; 3{,}0]$	$(3{,}0; 3{,}5]$	$(3{,}5; 4{,}0]$	$(4{,}0; 4{,}5]$	$(4{,}5; 5{,}0]$	$(5{,}0; +\infty)$
$F(x_0)$	0	0,2	0,5	0,6	0,8	0,9	1

Wykres dystrybuanty przedstawia rys. 9.

Zauważmy, że wykorzystując własność (2.11), można wyznaczyć, ile wynosi prawdopodobieństwo $P(3{,}5 < X < 4)$. Zgodnie z tą własnością:

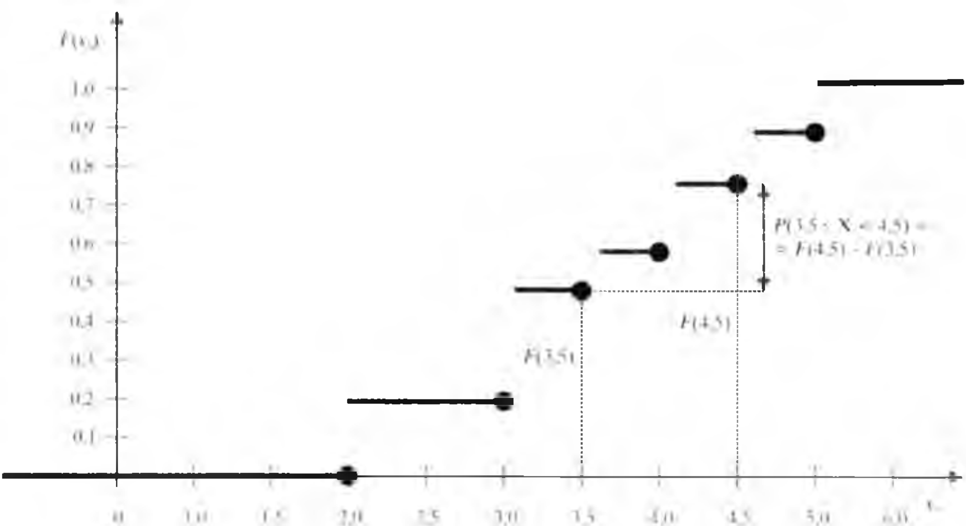
$$P(3{,}5 < X < 4{,}5) = F(4{,}5) - F(3{,}5) = 0{,}8 - 0{,}5 = 0{,}3.$$

Prawdopodobieństwo jest więc przyrostem dystrybuanty. Ilustruje to rys. 10.



Rys. 9. Wykres dystrybuanty zmiennej losowej X z przykładu 2.3

Źródło: opracowanie własne.



Rys. 10. Prawdopodobieństwo $P(3{,}5 < X < 4{,}5)$ jako przyrost dystrybuanty

Źródło: opracowanie własne.

Taki sam wynik otrzymamy, wykorzystując podczas obliczeń funkcję rozkładu prawdopodobieństwa, wyznaczoną w przykładzie 2.2. Mamy wówczas:

$$P(3{,}5 \leq X < 4{,}5) = P(X = 3{,}5) + P(X = 4{,}0) = 0{,}1 + 0{,}2 = 0{,}3.$$

Własność (2.11) można wykorzystać tylko wówczas, jeżeli przedział, dla którego chcemy wyznaczyć wartość prawdopodobieństwa, jest lewostronnie domknięty. Jeżeli np. będziemy chcieli wyznaczyć $P(3{,}5 \leq X \leq 4{,}5)$, to wówczas korzystając z funkcji rozkładu prawdopodobieństwa zmiennej, otrzymamy:

$$P(3{,}5 \leq X \leq 4{,}5) = P(X = 3{,}5) + P(X = 4) + P(X = 4{,}5) = 0{,}1 + 0{,}2 + 0{,}1 = 0{,}4.$$

To samo możemy otrzymać stosując następującą zależność:

$$P(3{,}5 \leq X \leq 4{,}5) = [F(4{,}5) - F(3{,}5)] + \mathrm{P}(X = 4{,}5) = (0{,}8 - 0{,}5) + 0{,}1 = 0{,}4. \ \{\mathrm{kp}\}$$

2.1.3. Funkcja gęstości i dystrybuanta zmiennej losowej ciągłej

W przypadku zmiennych losowych ciągłych nie jest możliwe wyznaczenie funkcji rozkładu prawdopodobieństwa, ponieważ, zgodnie z definicją zmienna ciągła to taka, która ma nieprzeliczalny zbiór wartości. Nie jest możliwe – tak jak w przypadku zmiennej dyskretnej – wymienienie wszystkich par $(x_i; p_i)$, gdyż zbiór wartości zmiennej losowej ciągłej jest nieskończenie liczny. Jeżeli zmienna losowa może przyjąć nieskończenie wiele wartości, to prawdopodobieństwo zdarzenia losowego, że przybierze ona jedną z nich, jest równe zero. Zatem możemy zapisać:

$$P(X = x_0) = 0, \tag{2.13}$$

gdzie:

X – zmienna losowa typu ciągłego,

x_0 – dowolna wartość z przedziału liczb rzeczywistych R.

W przypadku zmiennej losowej ciągłej wyznacza się tzw. **funkcję gęstości prawdopodobieństwa**.

Funkcją gęstości prawdopodobieństwa zmiennej losowej ciągłej X nazwiemy funkcję $f(x)$, określoną na zbiorze liczb rzeczywistych R zdefiniowaną następująco:

$$f(x) = \lim_{\Delta x \to 0} \frac{P(x < X < x + \Delta x)}{\Delta x}. \tag{2.14}$$

Funkcja $f(x)$ ma następujące własności:

1. $f(x) \geq 0,$ (2.15)
2. dla dowolnych $a < b$, $\int_a^b f(x)dx = P(a \leq X \leq b).$ (2.16)

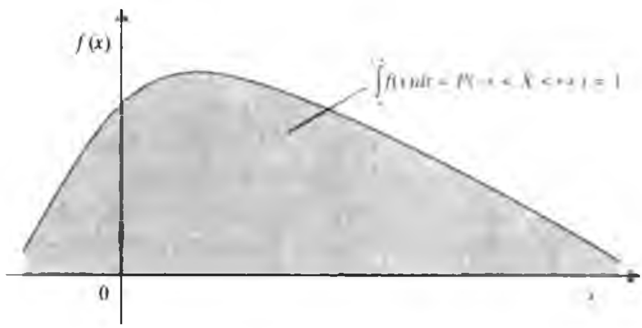
Z własności (2.16) wynika własność:

$$3. \int_{-\infty}^{+\infty} f(x)dx = P(-\infty < X < +\infty) = 1. \qquad (2.16a)$$

Ujmując, rzecz opisowo, powiemy, że funkcja gęstości jest funkcją nieujemną, a pole obszaru ograniczone jej wykresem i osią odciętych jest równe 1.

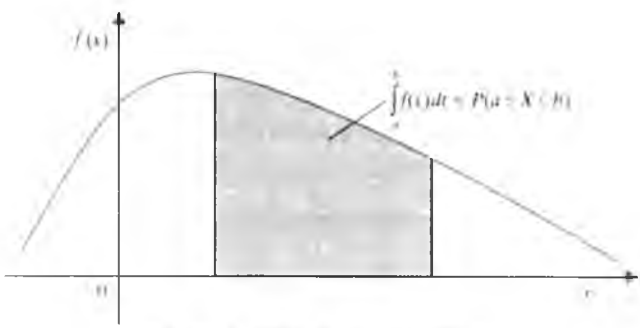
Ilustruje to rysunek 11, na którym został przedstawiony przykładowy wykres funkcji gęstości.

Natomiast na rysunku 12 przedstawiono graficzną interpretację własności (2.16).



Rys. 11. Przykładowy wykres funkcji gęstości

Źródło: opracowanie własne.



Rys. 12. Graficzna prezentacja

Źródło: opracowanie własne.

Zdefiniujmy obecnie pojęcie dystrybuanty zmiennej losowej ciągłej.

Dystrybuantą zmiennej losowej ciągłej X jest funkcja F: $\boldsymbol{R} \to \boldsymbol{R}$, taka że:

$$F(x_0) = P(X < x_0) = \int_{-\infty}^{x_0} f(t)dt, \text{ dla } x_0 \in R \qquad (2.17)$$

Pomiędzy funkcją dystrybuanty a funkcją gęstości zachodzi związek:

$$F'(x_0) = f(x_0), \tag{2.18}$$

co oznacza, że jeżeli funkcja dystrybuanty ma pochodną w punkcie x_0, to pochodna ta jest funkcją gęstości zmiennej losowej w punkcie x_0.

Własności funkcji dystrybuanty opisaliśmy już w punkcie 2.2. Modyfikacji ulegnie jedynie własność (2.11), która z uwagi na własność (2.13) będzie teraz następująca:

$$P(a \le X < b) = P(a \le X \le b) = P(a < X \le b) = P(a < X < b) = F(b) - F(a). \tag{2.19}$$

Przykład 2.4[4]

Tramwaj linii 15 w Krakowie w niedziele i święta pomiędzy godziną 9:00 a 18:00 kursuje regularnie co 20 minut (źródło: www.mpk.krakow.pl). Załóżmy, że pasażer przychodzi na przystanek w losowo wybranym momencie badanego okresu, nie kierując się rozkładem jazdy. Niech zmienna X oznacza czas oczekiwania (w minutach) na przystanku na przyjazd tramwaju. Określić, dla tak zdefiniowanej zmiennej losowej, funkcję gęstości oraz funkcję dystrybuanty a następnie obliczyć prawdopodobieństwo zdarzenia losowego, że czas oczekiwania pasażera na tramwaj będzie liczbą z przedziału [5, 10] minut.

Jeżeli przyjmiemy założenia, że tramwaj jeździ regularnie (co 20 minut), oraz że pasażer nie kieruje się rozkładem jazdy i przychodzi na przystanek w losowo wybranym momencie, to możemy stwierdzić że:

- prawdopodobieństwo zdarzenia losowego, że będzie czekał krócej niż 0 minut lub dłużej niż 20 jest równe zero, oraz
- gęstość prawdopodobieństwa w przedziale [0, 20] min. jest funkcją stałą postaci $f(x) = c$ (tzn. że wszystkie wartości zmiennej losowej w tym przedziale są jednakowo prawdopodobne).

Funkcja gęstości $f(x)$ ma zatem następującą postać:

$$f(x) = \begin{cases} 0 & dla \quad x < 0 \\ c & dla \quad 0 \le x \le 20. \\ 0 & dla \quad x > 20 \end{cases}$$

Aby wyznaczyć wartość c, można skorzystać z zależności (2.16a). Otrzymamy wowczas:

$$\int_{-\infty}^{+\infty} f(x)\,dx = \int_{-\infty}^{0} 0dx + \int_{0}^{20} cdx + \int_{20}^{+\infty} 0dx = 0 + cx\Big|_0^{20} + 0 = 20c,$$

---

[4] Inspiracją dla przykładu (2.4) był Przykład 1.9 z książki J. Jóźwiaka i J. Podgórskiego *Statystyka od podstaw*, PWE, Warszawa 1994.

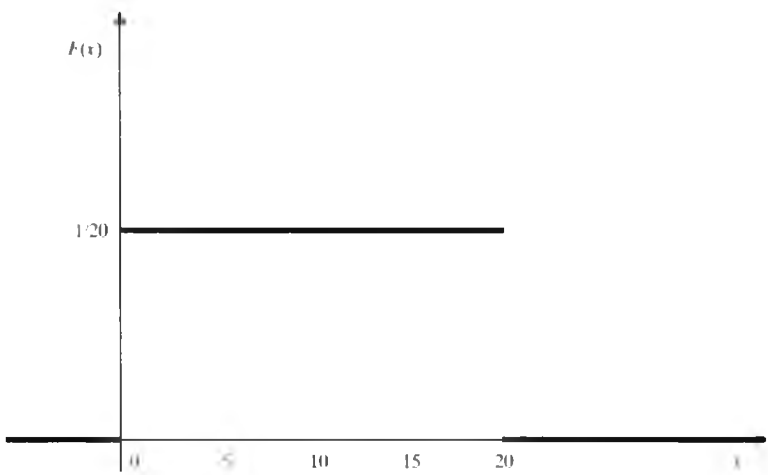
oraz

$$20c = 1 => c = \frac{1}{20}.$$

Po podstawieniu w miejsce stałej c wartości $\frac{1}{20}$ otrzymamy następującą funkcję gęstości:

$$f(x) = \begin{cases} 0 & dla \quad x < 0 \\ \frac{1}{20} & dla \quad 0 < x < 20. \\ 0 & dla \quad x > 20 \end{cases}$$

Graficznie wykres tej funkcji przedstawia rys. 13.



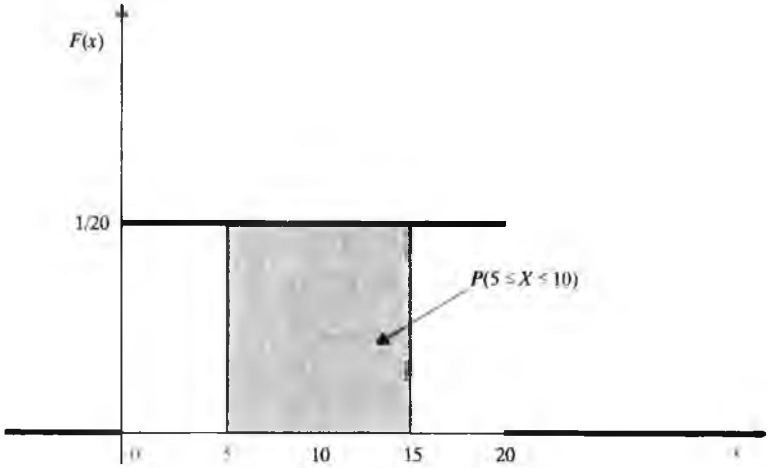
Rys. 13. Wykres funkcji gęstości do przykładu 2.4

Źródło: opracowanie własne.

Wartość prawdopodobieństwa $P(5 < X < 10)$ można obliczyć, wykorzystując własność (2.16); otrzymamy wówczas:

$$P(5 < X < 10) = \int_5^{10} f(x)dx = \int_5^{10} \frac{1}{20}\,dx = \frac{1}{20}x\Big|_5^{10} = \frac{10}{20} - \frac{5}{20} = \frac{5}{20} = \frac{1}{4}.$$

Na rysunku 14 przedstawiono graficzną interpretację szukanego prawdopodobieństwa.



Rys. 14. Graficzna interpretacja $P(5 < X < 10)$

Źródło: opracowanie własne.

Jak łatwo zauważyć, szukanemu prawdopodobieństwu odpowiada zacieniowany obszar będący w tym przypadku polem prostokąta o bokach równych 5 i 1/20 jednostek.

Aby wyznaczyć funkcję dystrybuanty, należy skorzystać ze wzoru (2.17).

Jeżeli $x_0 \le 0$, to dystrybuanta $F(x_0) = \int_{\infty}^{x_0} 0 dt = 0.$

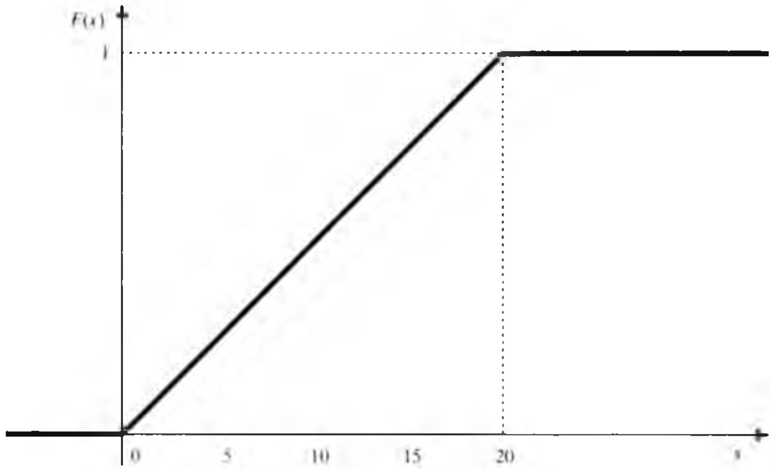
Jeżeli $0 < x_0 \le 20$, wówczas $F(x_0) = \int_{\infty}^{0} 0 dt + \int_{0}^{x_0} \frac{1}{20} dt = \frac{1}{20} t \Big|_{0}^{x_0} = \frac{1}{20} x_0.$

Jeżeli $x_0 > 20$, to $F(x_0) = \int_{\infty}^{0} 0 dt + \int_{0}^{20} \frac{1}{20} dt + \int_{20}^{x_0} 0 dt = 0 + \frac{1}{20} t \Big|_{0}^{20} + 0 = 1.$

Zastępując symbol x_0 symbolem x, możemy zapisać:

$$F(x) = \begin{cases} 0 & dla \quad x < 0 \\ \frac{1}{20} x & dla \quad 0 < x < 20. \\ 0 & dla \quad x > 20 \end{cases}$$

Wykres tej funkcji przedstawia rys. 15.



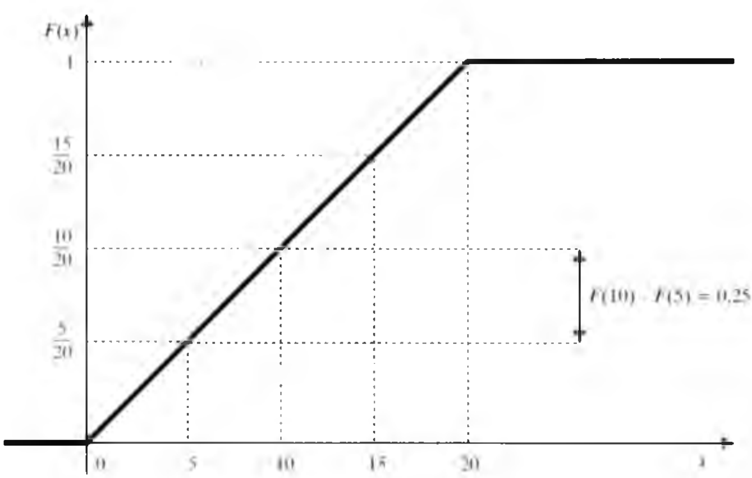
Rys. 15. Dystrybuanta czasu oczekiwania na tramwaj

Źródło: opracowanie własne.

Obliczane wcześniej prawdopodobieństwo $P(5 \le X \le 10)$ można również wyznaczyć, korzystając z dystrybuanty zmiennej X oraz z własności (2.19):

$$P(5 < X \le 10) = F(10) - F(5) = \frac{1}{20} \cdot 10 - \frac{1}{20} \cdot 5 = \frac{10}{20} - \frac{5}{20} = \frac{5}{20} = \frac{1}{4}.$$

Graficzną interpretację tego prawdopodobieństwa pokazano na rys. 16.



Rys. 16. Prawdopodobieństwo $P(5 < X \le 10)$ jako przyrost dystrybuanty

Źródło: opracowanie własne.

Szukane prawdopodobieństwo jest reprezentowane przez odcinek na osi rzędnych pomiędzy wartościami dystrybuanty w punktach 10 i 5. {kp}

2.1.4. Podstawowe charakterystyki liczbowe jednowymiarowej zmiennej losowej

2.1.4.1. Wartość oczekiwana i wariancja zmiennej losowej dyskretnej

Wartość oczekiwana[5] stanowi pewien odpowiednik – omawianej w ramach statystyki opisowej – średniej arytmetycznej. Do oznaczenia wartości oczekiwanej zmiennej X używać będziemy symbolu $E(X)$ i w przypadku zmiennej losowej dyskretnej definiować jako sumę iloczynów wartości zmiennej x_i oraz przypisanych im prawdopodobieństw realizacji p_i.

Można zatem zapisać:

$$E(X) = \sum_i x_i p_i, \tag{2.20}$$

pod warunkiem, że szereg $\sum_i x_i p_i$ jest bezwzględnie zbieżny. Jeżeli tak nie jest, to mówimy, że zmienna losowa nie posiada wartości oczekiwanej.

Podobnie – jak to ma miejsce w statystyce opisowej – wartość oczekiwaną można wykorzystać do wyznaczenia wariancji i następnie odchylenia standardowego. Wariancję w rachunku prawdopodobieństwa oznaczać będziemy symbolem $D^2(X)$, a odchylenie standardowe $D(X)$. Ogólnie wariancją zmiennej losowej nazywamy wyrażenie:

$$D^2(X) = E[X - E(X)]^2, \tag{2.21}$$

natomiast odchylenie standardowe to:

$$D(X) = \sqrt{D^2(X)}. \tag{2.22}$$

Jeżeli zmienna losowa jest typu skokowego i posiada wartość oczekiwaną, to wówczas wariancją nazywamy wyrażenie:

$$D^2(X) = \sum_i [x_i - E(X)]^2 p_i. \tag{2.23}$$

Dość często w praktyce do wyznaczenia wariancji wykorzystuje się równoważny wzór o postaci:

[5] W literaturze można spotkać też inne określenia tej miary, takie jak: wartość przeciętna, średnia, nadzieja matematyczna.

$$D^2(X) = E(X^2) - E^2(X) = \sum x_i^2 p_i - \left(\sum x_i p_i\right)^2 \qquad (2.23a)$$

Przykład 2.5

Wykorzystując funkcję rozkładu prawdopodobieństwa określoną w przykładzie 2.2, wyznaczyć wartość oczekiwaną, wariancję i odchylenie standardowe określonej tam zmiennej losowej X.

Przypomnijmy, że zmienna losowa zdefiniowana w przykładzie 2.2 oznacza ocenę uzyskaną przez losowo wybranego studenta, a jej rozkład przedstawia się następująco:

x_i	2,0	3,0	3,5	4,0	4,5	5,0
p_i	0,2	0,3	0,1	0,2	0,1	0,1

Aby wyznaczyć wartość oczekiwaną, a następnie wariancję, rozbudujemy powyższą tabelkę o kolejne trzy wiersze.

1	x_i	2,0	3,0	3,5	4,0	4,5	5,0
2	p_i	0,2	0,3	0,1	0,2	0,1	0,1
3	$x_i p_i$	0,4	0,9	0,35	0,8	0,45	0,5
4	x_i^2	4,0	9,0	12,25	16,0	20,25	25,0
5	$x_i^2 p_i$	0,8	2,7	1,225	3,2	2,025	2,5

Sumując wartości w wierszu 3, obliczamy wartość oczekiwaną zmiennej losowej X:

$$E(X) = 0,4 + 0,9 + 0,35 + 0,8 + 0,45 + 0,5 = 3,4.$$

Do obliczenia wariancji wykorzystamy natomiast równanie (2.23a).

Pierwsza składowa tego równania $\sum_i x_i^2 p_i$, to suma wartości w wierszu 5 powyższej tabelki, natomiast druga:

$$\left(\sum_i x_i p_i\right)^2 = (3,4)^2 = 11,56.$$

Wariancja zmiennej losowej X jest równa:

$$D^2(X) = (0,8 + 2,7 + 1,225 + 3,2 + 2,025 + 2,5) - 11,56 = 0,89,$$

natomiast odchylenie standardowe:

$$D(X) = \sqrt{0,89} \approx 0,94.$$

Wariancję można również obliczyć, stosując wzór (2.23). Otrzymamy wówczas:

$$D^2(X) = 0,2 \cdot (2-3,4)^2 + 0,3 \cdot (3-3,4)^2 + 0,1 \cdot (3,5-3,4)^2 +$$
$$+\ 0,2 \cdot (4-3,4)^2 + 0,1 \cdot (4,5-3,4)^2 + 0,1 \cdot (5-3,4)^2 = 0,89.$$

Wartość oczekiwana oznacza w tym przypadku, że przeciętna ocena w badanej grupie wynosi 3,4, natomiast wartość odchylenia standardowego informuje nas, że oceny badanych studentów odchylały się od oceny przeciętnej średnio o ±0,94 stopnia. {kp}

2.1.4.2. Wartość oczekiwana i wariancja zmiennej losowej ciągłej

W przypadku zmiennych losowych ciągłych wartość oczekiwaną obliczamy stosując następujący wzór:

$$E(X) = \int_{-\infty}^{+\infty} x f(x)\, dx, \tag{2.24}$$

natomiast podczas wyznaczania wariancji należy wykorzystać wzór:

$$D^2(X) = \int_{-\infty}^{+\infty} [x-E(X)]^2 f(x) dx, \tag{2.25}$$

lub alternatywnie wzór

$$D^2(X) = E(X^2) - E^2(X) = \int_{-\infty}^{+\infty} x^2 f(x) dx - \left(\int_{-\infty}^{+\infty} x f(x) dx \right)^2 \tag{2.26}$$

Przykład 2.6

Wykorzystując funkcję gęstości określoną w przykładzie 2.4, wyznaczyć wartość oczekiwaną czasu oczekiwania pasażera na tramwaj, a następnie wariancję i odchylenie standardowe.

Korzystając z równania (2.24), otrzymamy:

$$E(X) = \int_{-\infty}^{+\infty} x f(x) dx = \int_{-\infty}^{0} x \quad 0 dx + \int_{0}^{20} x \quad \frac{1}{20} dx + \int_{20}^{+\infty} x \quad 0 dx = \frac{x^2}{40}\bigg|_{0}^{20} = \frac{400}{40} = 10.$$

Wynika z tego, że jeżeli pasażer w badanym okresie będzie przychodził na ten przystanek wielokrotnie, to wówczas będzie on czekał na przyjazd tramwaju przeciętnie 10 minut.

Do obliczenia wariancji użyjemy równania (2.25):

$$D^2(X)=\int_{-\infty}^{+\infty}[x-E(X)]^2 f(x)dx=\int_0^{20}(x-10)^2\cdot\frac{1}{20}dx=\int_0^{20}\left(\frac{x^2}{20}-x+5\right)dx=$$

$$=\frac{x^3}{60}-\frac{x^2}{2}+5x\Big|_0^{20}=\frac{8000}{60}-200+100=33\frac{1}{3}$$

Pierwiastkując wartość wariancji, otrzymujemy wartość odchylenia standardowego:

$$D(X)=\sqrt{D^2(X)}=\sqrt{33\frac{1}{3}}\approx 5{,}77.$$

Z powyższych obliczeń wynika, że jeżeli pasażer będzie wielokrotnie korzystał w badanych godzinach w niedziele i święta z tego tramwaju, to wówczas jego czas oczekiwania na przystanku będzie się odchylał od średniego czasu oczekiwania (10 min) przeciętnie o 5,77 minuty. {kp}

2.1.4.3. Podstawowe własności wartości oczekiwanej i wariancji zmiennej losowej

Istnieje kilka podstawowych własności ujmowanych niekiedy jako twierdzenia dotyczące wartości oczekiwanej i wariancji zmiennej losowej[6].

1. Wartość oczekiwana dowolnej stałej c jest równa tej stałej, natomiast wariancja stałej c jest równa zero:

$$E(c)=c, \tag{2.27}$$

$$D^2(c)=0. \tag{2.28}$$

2. Wartość oczekiwana sumy dwóch zmiennych losowych X i Y jest równa sumie wartości przeciętnych tych zmiennych losowych, oraz wariancja sumy dwóch niezależnych zmiennych losowych X i Y jest równa sumie wariancji tych zmiennych:

$$E(X+Y)=E(X)+E(Y), \tag{2.29}$$

a gdy X i Y są niezależne, to:

$$D^2(X+Y)=D^2(X)+D^2(Y). \tag{2.30}$$

Powyższe twierdzenie można również uogólnić dla n zmiennych losowych.

[6] Dowody zamieszczonych poniżej twierdzeń można znaleźć np. w Z. Helwig: *Elementy rachunku...*, *op. cit.*, rozdział 4.2.2. i 4.3.2., s. 97–111.

Mamy wówczas:

$$E\left(\sum_{i=1}^{n} X_i\right)=\sum_{i=1}^{n} E(X_i), \tag{2.31}$$

$$D^2\left(\sum_{i=1}^{n} X_i\right)=\sum_{i=1}^{n} D^2(X_i), \text{ gdy } X_1, X_2, \dots, X_n \text{ są niezależne.} \tag{2.32}$$

W przypadku wariancji rozważamy oczywiście n wzajemnie niezależnych zmiennych losowych $X_1, \dots, X_n$.

3. Wariancja różnicy dwóch niezależnych zmiennych losowych X i Y jest równa sumie wariancji tych zmiennych:

$$D^2(X-Y) = D^2(X) + D^2(Y). \tag{2.33}$$

4. Wartość oczekiwana iloczynu dwóch niezależnych zmiennych losowych X i Y jest równa iloczynowi wartości oczekiwanych tych zmiennych:

$$E(XY) = E(X) \cdot E(Y). \tag{2.34}$$

Własność (4) można uogólnić dla przypadku n wzajemnie niezależnych zmiennych losowych. Mamy wówczas:

$$E\left(\prod_{i=1}^{n} X_i\right)=\prod_{i=1}^{n} E(X_i). \tag{2.35}$$

5. Wartość oczekiwana iloczynu stałej c przez zmienną losową X jest równa iloczynowi tej stałej przez wartość oczekiwaną zmiennej losowej X, natomiast wariancja iloczynu stałej c przez zmienną losową X jest równa iloczynowi kwadratu tej stałej przez wariancję zmiennej losowej X:

$$E(cX) = cE(X), \tag{2.36}$$

$$D^2(cX) = c^2D^2(X). \tag{2.37}$$

6. Wartość oczekiwana sumy zmiennej losowej X i stałej c, jest równa sumie wartości oczekiwanej zmiennej i stałej c, natomiast wariancja sumy zmiennej losowej X i stałej c, jest równa wariancji zmiennej X:

$$E(X + c) = E(X) + c, \tag{2.38}$$

$$D^2(X + c) = D^2(X). \tag{2.39}$$

7. Wartość oczekiwana różnicy zmiennej losowej X i wartości oczekiwanej tej zmiennej jest równa zero.

$$E[X - E(X)] = 0. \tag{2.40}$$

2.2. Dwuwymiarowe zmienne losowe

Obecnie przejdziemy do krótkiego omówienia zasad wyznaczania funkcji rozkładu prawdopodobieństwa lub funkcji gęstości oraz dystrybuanty zmiennej losowej wielowymiarowej. Następnie w punktach 2.2.1 oraz 2.2.2 omówimy również zasady wyznaczania podstawowych miar położenia i zmienności dla wielowymiarowych zmiennych losowych. Ze względów praktycznych rozpatrzymy jedynie najprostszy przypadek zmiennej wielowymiarowej, jakim jest zmienna dwuwymiarowa. Zmienną losową dwuwymiarową definiuje się podobnie jak zmienną losową jednowymiarową, z tym że poszczególnym zdarzeniom elementarnym danego doświadczenia przyporządkowuje się nie jedną, lecz dwie (w ogólnym przypadku tyle, ile jest zmiennych losowych) wartości rzeczywiste.

Zmienną dwuwymiarową oznaczać będziemy symbolem (X, Y), a jej wartości odpowiednio $(x_i; y_i)$.

2.2.1. Dwuwymiarowa zmienna losowa skokowa

Mówimy, że zmienna dwuwymiarowa (X, Y) jest typu skokowego (dyskretnego), jeżeli przyjmuje ona skończoną lub co najwyżej przeliczalną liczbę wartości $(x_i; y_i)$ $(i, j = 1,2,3,\ldots)$ odpowiednio z prawdopodobieństwem p_{ij}, przy czym:

$$\sum_i \sum_j p_{ij} = 1. \qquad (2.41)$$

Określony w ten sposób zbiór wszystkich prawdopodobieństw:

$$p_{ij} = P(X = x_i, Y = y_j) \text{ dla } i, j = 1,2,3,\ldots, \qquad (2.42)$$

spełniających warunek (2.41), określa się mianem funkcji rozkładu prawdopodobieństwa dwuwymiarowej zmiennej losowej.

Dystrybuantą dwuwymiarowej zmiennej losowej skokowej (X,Y) nazywamy funkcję określoną wzorem:

$$F(x,y) = \sum_{x_i < x} \sum_{y_i < y} p_{ij}, \qquad (2.43)$$

Bardzo często rozkład prawdopodobieństwa dwuwymiarowej zmiennej skokowej przedstawia się w postaci tabelarycznej. Przykład takiego zapisu prezentuje tablica 2.2.

Tablica 2.2. Rozkład prawdopodobieństw dwuwymiarowej zmiennej skokowej

x_i \ y_j	y_1	y_2	………………	y_l	$\sum_j p_{ij}$
x_1	p_{11}	p_{12}	………………	p_{1l}	$p_{1.}$
x_2	p_{21}	p_{22}	………………	p_{2l}	$p_{2.}$
			………………		
			………………		
			………………		
x_k	p_{k1}	p_{k2}	………………	p_{kl}	$p_{k.}$
$\sum_i p_{ij}$	$p_{.1}$	$p_{.2}$	………………	$p_{.l}$	1

Źródło: opracowanie własne.

Podczas konstrukcji tej tablicy założono, że zmienna losowa X może przyjąć k różnych wartości $x_1, ..., x_k$, natomiast zmienna losowa Y, l różnych wartości $y_1, ..., y_l$. W ostatniej kolumnie i w ostatnim wierszu tablicy 2.2 umieszczono prawdopodobieństwa brzegowe $p_{i.}$ i $p_{.j}$ obliczone zgodnie z regułami:

$$p_{i.} = \sum_j p_{ij}, \tag{2.44}$$

$$p_{.j} = \sum_i p_{ij}. \tag{2.45}$$

Prawdopodobieństwa te obliczono poprzez zsumowanie prawdopodobieństw p_{ij} odpowiednio w wierszach i w kolumnach tablicy 2.2. Prawdopodobieństwo $p_{i.}$ interpretuje się jako prawdopodobieństwo, że zmienna X przyjmie wartość x_i, niezależnie od tego jaką wartość przyjmie zmienna Y. Podobnie interpretujemy $p_{.j}$, z tym że określa ono wartość prawdopodobieństwa zdarzenia losowego, że zmienna Y przyjmie wartość y_j, podczas gdy zmienna X przyjmie dowolną ze swych wartości.

Oczywiście suma prawdopodobieństw brzegowych $p_{i.}$ oraz $p_{.j}$ jest równa jednosci, co zapiszemy:

$$\sum_i p_{i.} = \sum_j p_{.j} = \sum_i \sum_j p_{ij} = 1. \tag{2.46}$$

Korzystając z wartości prawdopodobieństw p_{ij} oraz prawdopodobieństw brzegowych $p_{i.}$ i $p_{.j}$, można wyznaczyć następujące prawdopodobieństwa warunkowe:

$$P(X = x_i \mid Y = y_j) = \frac{P(X = x_i, Y = y_j)}{P(y = y_j)} = \frac{p_{ij}}{p_{.j}}, \tag{2.47}$$

$$P(Y = y_j \mid X = x_i) = \frac{P(X = x_i, Y = y_j)}{P(X = x_i)} = \frac{p_{ij}}{p_{i.}}. \tag{2.48}$$

Wyrażenie (2.47) i (2.48) czytamy jako prawdopodobieństwo zrealizowania się wartości jednej zmiennej przy założeniu stałej wartości drugiej zmiennej.

Podobnie jak w przypadku prawdopodobieństw warunkowych zachodzi własność:

$$\sum_i P(X=x_i\,|Y=y_j)=\sum_j P(Y=y_j\,|X=x_i)=1. \tag{2.49}$$

Przykład 2.7

Z grupy 30 studentów – opisanych w przykładzie 2.1 – którzy zdawali egzamin z „Zarządzania jakością", dwukrotnie wylosowano niezależnie jednego studenta i sprawdzono jego ocenę. Na zbiorze wyników doświadczenia określono następującą dwuwymiarową zmienną losową (X, Y):

X = 0, jeżeli wybrany w pierwszym losowaniu student nie zdał egzaminu,
X = 1, jeśli wylosowany w pierwszym losowaniu student zdał egzamin,
Y = ocenie studenta (przy sześciostopniowej skali ocen) wylosowanego w drugim losowaniu.

Przypomnijmy, że liczba ocen: bdb, +db, db, +dst, dst, ndst, wynosiła odpowiednio: 3, 3, 6, 3, 9, 6.

Każdemu zdarzeniu elementarnemu przyporządkowana jest tylko jedna para wartosci zmiennej (X,Y) z następującego zbioru 12 elementów:

$$\{(0; 2{,}0), (0; 3{,}0), (0; 3{,}5), (0; 4{,}0), (0; 4{,}5), (0; 5{,}0), (1; 2{,}0), (1; 3{,}0), (1; 3{,}5), (1; 4{,}0), (1; 4{,}5), (1; 5{,}0)\}$$

Każda para typu (0, ocena) oznacza, że pierwszy z wylosowanych studentów otrzymał oceną niedostateczną. Natomiast para (1, ocena) oznacza, że pierwszy z wylosowanych studentów otrzymał jedną z pięciu pozytywnych ocen.

Na przykład pod parą (1; 3,0), kryje się 5 następujących zdarzeń elementarnych:

$$(3{,}0; 3{,}0), (3{,}5; 3{,}0), (4{,}0; 3{,}0), (4{,}5; 3{,}0), (5{,}0; 3{,}0).$$

Prawdopodobieństwo tego zdarzenia losowego wynosi:

$$P(X = 1, Y = 3{,}0) = 0{,}3 \cdot 0{,}3 + 0{,}1 \cdot 0{,}3 + 0{,}2 \cdot 0{,}3 + 0{,}1 \cdot 0{,}3 + 0{,}1 \cdot 0{,}3 = 0{,}24.$$

To prawdopodobieństwo i obliczone w podobny sposób pozostałe 11 prawdopodobieństw umieszczono w tablicy 2.3.

Tablica 2.3. Rozkład dwuwymiarowej zmiennej losowej

$x_i \backslash y_j$	2,0	3,0	3,5	4,0	4,5	5,0	$\sum_j p_{ij}$
0	0,04	0,06	0,02	0,04	0,02	0,02	0,2
1	0,16	0,24	0,08	0,16	0,08	0,08	0,8
$\sum_i p_{ij}$	0,2	0,3	0,1	0,2	0,1	0,1	1

Źródło: obliczenia własne.

W ostatnim wierszu tej tablicy i w ostatniej kolumnie umieszczono rozkłady brzegowe zmiennej Y i zmiennej X.

Korzystając z powyższych danych oraz z własności (2.47) i (2.48), obliczymy poszczególne rozkłady warunkowe X przy warunku Y i Y przy warunku X.

$$P(X = 0 | Y = 2{,}0) = \frac{P(X = 0, Y = 2{,}0)}{P(Y = 2{,}0)} = \frac{p_{11}}{p_{.1}} = \frac{0{,}04}{0{,}2} = 0{,}2,$$

$$P(X = 1 | Y = 2{,}0) = \frac{P(X = 1, Y = 2{,}0)}{P(Y = 2{,}0)} = \frac{p_{21}}{p_{.1}} = \frac{0{,}16}{0{,}2} = 0{,}8,$$

$$P(X = 0 | Y = 3{,}0) = \frac{P(X = 0, Y = 3{,}0)}{P(Y = 3{,}0)} = \frac{p_{12}}{p_{.2}} = \frac{0{,}06}{0{,}3} = 0{,}2,$$

$$P(X = 1 | Y = 3{,}0) = \frac{P(X = 1, Y = 3{,}0)}{P(Y = 3{,}0)} = \frac{p_{22}}{p_{.2}} = \frac{0{,}24}{0{,}3} = 0{,}8,$$

$$P(X = 0 | Y = 3{,}5) = \frac{P(X = 0, Y = 3{,}5)}{P(Y = 3{,}5)} = \frac{p_{13}}{p_{.3}} = \frac{0{,}01}{0{,}1} = 0{,}2,$$

$$P(X = 1 | Y = 3{,}5) = \frac{P(X = 1, Y = 3{,}5)}{P(Y = 3{,}5)} = \frac{p_{23}}{p_{.3}} = \frac{0{,}08}{0{,}1} = 0{,}8,$$

$$P(X = 0 | Y = 4{,}0) = \frac{P(X = 0, Y = 4{,}0)}{P(Y = 4{,}0)} = \frac{p_{14}}{p_{.4}} = \frac{0{,}04}{0{,}2} = 0{,}2,$$

$$P(X = 1 | Y = 4{,}0) = \frac{P(X = 1, Y = 4{,}0)}{P(Y = 4{,}0)} = \frac{p_{24}}{p_{.4}} = \frac{0{,}16}{0{,}2} = 0{,}8,$$

$$P(X = 0 | Y = 4{,}5) = \frac{P(X = 0, Y = 4{,}5)}{P(Y = 4{,}5)} = \frac{p_{15}}{p_{.5}} = \frac{0{,}02}{0{,}1} = 0{,}2,$$

$$P(X = 1 | Y = 4{,}5) = \frac{P\,(X = 1, Y = 4{,}5)}{P\,(Y = 4{,}5)} = \frac{p_{25}}{p_{.5}} = \frac{0{,}08}{0{,}1} = 0{,}8,$$

$$P(X = 0 | Y = 5{,}0) = \frac{P\,(X = 0, Y = 5{,}0)}{P\,(Y = 5{,}0)} = \frac{p_{16}}{p_{.6}} = \frac{0{,}02}{0{,}1} = 0{,}2,$$

$$P(X = 1 | Y = 5{,}0) = \frac{P\,(X = 1, Y = 5{,}0)}{P\,(Y = 5{,}0)} = \frac{p_{26}}{p_{.6}} = \frac{0{,}08}{0{,}1} = 0{,}8,$$

$$P(Y = 2{,}0 | X = 0) = \frac{P\,(Y = 2{,}0, X = 0)}{P\,(X = 0)} = \frac{p_{11}}{p_{1}} = \frac{0{,}04}{0{,}2} = 0{,}2,$$

$$P(Y = 3{,}0 | X = 0) = \frac{P\,(Y = 3{,}0, X = 0)}{P\,(X = 0)} = \frac{p_{12}}{p_{1.}} = \frac{0{,}06}{0{,}2} = 0{,}3,$$

$$P(Y = 3{,}5 | X = 0) = \frac{P\,(Y = 3{,}5, X = 0)}{P\,(X = 3{,}0)} = \frac{p_{13}}{p_{1}} = \frac{0{,}02}{0{,}2} = 0{,}1,$$

$$P(Y = 4{,}0 | X = 0) = \frac{P\,(Y = 4{,}0, X = 0)}{P\,(X = 0)} = \frac{p_{14}}{p_{1.}} = \frac{0{,}04}{0{,}2} = 0{,}2,$$

$$P(Y = 4{,}5 | X = 0) = \frac{P\,(Y = 4{,}5, X = 0)}{P\,(X = 0)} = \frac{p_{15}}{p_{1.}} = \frac{0{,}02}{0{,}2} = 0{,}1,$$

$$P(Y = 5{,}0 | X = 0) = \frac{P\,(Y = 5{,}0, X = 0)}{P\,(X = 0)} = \frac{p_{16}}{p_{1.}} = \frac{0{,}02}{0{,}2} = 0{,}1,$$

$$P(Y = 2{,}0 | X = 1) = \frac{P\,(Y = 2{,}0, X = 1)}{P\,(X = 1)} = \frac{p_{21}}{p_{2}} = \frac{0{,}16}{0{,}8} = 0{,}2,$$

$$P(Y = 3{,}0 | X = 1) = \frac{P\,(Y = 3{,}0, X = 1)}{P\,(X = 1)} = \frac{p_{22}}{p_{2}} = \frac{0{,}24}{0{,}8} = 0{,}3,$$

$$P(Y = 3{,}5 | X = 1) = \frac{P\,(Y = 3{,}5, X = 1)}{P\,(X = 1)} = \frac{p_{23}}{p_{2.}} = \frac{0{,}08}{0{,}8} = 0{,}1,$$

$$P(Y = 4{,}0 | X = 1) = \frac{P\,(Y = 4{,}0, X = 1)}{P\,(X = 1)} = \frac{p_{24}}{p_{2}} = \frac{0{,}16}{0{,}8} = 0{,}2$$

$$P(Y = 4{,}5 | X = 1) = \frac{P\,(Y = 4{,}5, X = 1)}{P\,(X = 1)} = \frac{p_{25}}{p_{2.}} = \frac{0{,}08}{0{,}8} = 0{,}1,$$

$$P(Y = 5{,}0 | X = 1) = \frac{P\,(Y = 5{,}0, X = 1)}{P\,(X = 1)} = \frac{p_{26}}{p_{2.}} = \frac{0{,}08}{0{,}8} = 0{,}1. \quad \{\text{kp}\}$$

Analizując obliczone rozkłady brzegowe oraz rozkłady warunkowe, przykładu 2.7, można dojść do wniosku, że spełnione są następujące zależności:

$$P(X=x_i\,|Y=y_j)=\frac{P(X=x_i,Y=y_j)}{P(y=y_i)}=\frac{p_{ij}}{p_{.j}}=p_{i.}=P(X=x_i),\qquad(2.50)$$

$$P(Y=y_j\,|X=x_i)=\frac{P(X=x_i,Y=y_j)}{P(X=x_i)}=\frac{p_{ij}}{p_{i.}}=p_{.j}=P(Y=y_j),\qquad(2.51)$$

oraz

$$p_{ij}=p_{i.}\,p_{.j}.\qquad(2.52)$$

Spełnienie powyższych warunkow gwarantuje, że badane zmienne X i Y są zmiennymi niezależnymi.

Dla zmiennej dwuwymiarowej (X, Y) można wyznaczać tak jak dla zmiennych jednowymiarowych wartość oczekiwaną. W przypadku, gdy zmienne losowe są dyskretne, wartość oczekiwaną wyznaczamy, stosując wzór:

$$E(XY)=\sum_i\sum_j x_i y_j p_{ij}.\qquad(2.53)$$

Wśród miar opisujących rozkłady zmiennych dwuwymiarowych ważną rolę odgrywają dwie kolejne: **kowariancja i współczynnik korelacji liniowej**. Dzięki nim dowiadujemy się, czy istnieje i jak silny jest związek pomiędzy badanymi zmiennymi.

Kowariancja zmiennych losowych $\mathrm{cov}(X,Y)$, lub inaczej wspólna wariancja tych zmiennych określana jest wzorem:

$$\mathrm{cov}(X,Y)=E\big[(X-E(X))(Y-E(Y))\big],\qquad(2.54)$$

lub równoważnym wzorem:

$$\mathrm{cov}(X,Y)=E(XY)-E(X)\,E(Y).\qquad(2.55)$$

Jeżeli zmienne X i Y są zmiennymi dyskretnymi, to wowczas wzory (2.54) i (2.55) przybierają postać:

$$\mathrm{cov}(X,Y)=\sum_i\sum_j\big[x_i-E(X)\big]\big[y_j-E(Y)\big]\,p_{ij},\qquad(2.56)$$

$$\mathrm{cov}(X,Y)=\sum_i\sum_j x_i y_j p_{ij}-\sum_i x_i p_{i.}\sum_j y_j p_{.j}.\qquad(2.57)$$

Jeżeli zmienne losowe są niezależne, to wówczas prawdziwa jest równość:

$$E(X,Y)=E(X)E(Y),\qquad(2.58)$$

i kowariancja jest równa zeru. Oznacza to, że zmienne losowe X i Y są nieskorelowane.

Współczynnik korelacji to stosunek kowariancji cov(X,Y) do iloczynu odchyleń standardowych rozkładów brzegowych zmiennych losowych X i Y. Możemy zatem zapisać:

$$\rho = \frac{\mathrm{cov}(X,Y)}{D(X)D(Y)} \tag{2.59}$$

Współczynnik korelacji przybiera wartości z przedziału [–1; 1]. Jeżeli $\rho = 0$, wówczas zmienne są nieskorelowane. Jeżeli $\rho = 1$ lub $\rho = -1$, to o zmiennych losowych X i Y można powiedzieć, że są powiązane zależnością funkcyjną, przy czym, gdy $\rho = 1$, to zależność ta jest rosnąca, natomiast, gdy $\rho = -1$, wówczas zależność funkcyjna ma charakter malejący.

Wartości $\rho \in (-1, 0)$ lub $\rho \in (0, 1)$, oznacza, że pomiędzy badanymi zmiennymi istnieje współzależność ujemna lub dodatnia, przy czym nie ma ona charakteru funkcyjnego. Generalnie im wartość ρ jest bliższa (–1) lub (+1), tym korelacja pomiędzy zmiennymi jest silniejsza.

Przykład 2.8

Wykorzystując dane z przykładu 2.7, obliczyć: wartość oczekiwaną $E(XY)$, kowariancję cov(XY) i współczynnik korelacji ρ.

Wartość oczekiwana iloczynu zmiennych losowych X i Y wynosi:

$$E(XY) = 0 \cdot 2{,}0 \cdot 0{,}04 + 0 \cdot 3{,}0 \cdot 0{,}06 + 0 \cdot 3{,}5 \cdot 0{,}02 + 0 \cdot 4{,}0 \cdot 0{,}04 + 0 \cdot 4{,}5 \cdot 0{,}02 + \\ + 0 \cdot 5{,}0 \cdot 0{,}02 + 1 \cdot 2{,}0 \cdot 0{,}16 + 1 \cdot 3{,}0 \cdot 0{,}24 + 1 \cdot 3{,}5 \cdot 0{,}08 + 1 \cdot 4{,}0 \cdot 0{,}04 + \\ + 1 \cdot 4{,}5 \cdot 0{,}08 + 1 \cdot 5{,}0 \cdot 0{,}08 = 0{,}32 + 0{,}72 + 0{,}28 + 0{,}64 + 0{,}36 + 0{,}4 = 2{,}72.$$

Aby wyznaczyć kowariancję zmiennych losowych X i Y, wyznaczymy wartości oczekiwane rozkładów brzegowych $E(X)$ i $E(Y)$.

$$E(X) = 0 \cdot 0{,}2 + 1 \cdot 0{,}8 = 0{,}8,$$

$$E(Y) = 2{,}0 \cdot 0{,}2 + 3{,}0 \cdot 0{,}3 + 3{,}5 \cdot 0{,}1 + 4{,}0 \cdot 0{,}2 + 4{,}5 \cdot 0{,}1 + 5{,}0 \cdot 0{,}1 = \\ = 0{,}4 + 0{,}9 + 0{,}35 + 0{,}8 + 0{,}45 + 0{,}5 = 3{,}4.$$

Zatem wartość kowariancji wyniesie:

$$\mathrm{cov}(X,Y) = E(XY) - E(X) \cdot E(Y) = 2{,}72 - 0{,}8 \cdot 3{,}4 = 0.$$

Jeżeli kowariancja jest równa zero, to również współczynnik korelacji $\rho = 0$.

Potwierdza to już wcześniej pojawiającą się prawidłowość, że zmienne losowe X i Y są niezależne (nieskorelowane). {kp}

W analogiczny sposób można również wyznaczyć warunkową wartość oczekiwaną zmiennej losowej X przy założeniu $Y = y_j$, lub też zmiennej losowej Y przy założeniu $X = x_i$. W sytuacji gdy zmienne losowe X i Y są skokowe, wartości oczekiwane wynoszą odpowiednio:

$$E(X=x_i \mid Y=y_j) = \sum_i x_i \frac{p_{ij}}{p_j}, \tag{2.60}$$

$$E(Y=y_j \mid X=x_i) = \sum_j x_j \frac{p_{ij}}{p_i}. \tag{2.61}$$

Jeżeli, tak jak w przykładach 2.7 i 2.8, zmienne są niezależne, to wówczas:

$$E(X=x_i \mid Y=y_j) = \sum_i x_i \frac{p_{ij}}{p_j} = \sum_i x_i p_i = E(X=x_i), \tag{2.62}$$

$$E(Y=y_j \mid X=x_i) = \sum_j x_j \frac{p_{ij}}{p_i} = \sum_j x_j p_j = E(Y=y_j), \tag{2.63}$$

co oznacza, że warunkowa wartość oczekiwana jest identyczna jak wartość oczekiwana rozkładu brzegowego. Czytelnikowi pozostawiamy numeryczne sprawdzenie powyższych zależności dla danych z przykładów 2.7 i 2.8.

2.2.2. Dwuwymiarowa zmienna losowa ciągła

Obecnie zostaną omówione w skrócie pojęcia, związane z dwuwymiarową zmienną losową ciągłą.

Rozkład dwuwymiarowej zmiennej losowej ciągłej opisujemy za pomocą łącznej funkcji gęstości prawdopodobieństwa postaci:

$$f(x,y) = \lim_{\substack{\Delta x \to 0 \\ \Delta y \to 0}} \frac{P(x \le X < x + \Delta x, y \le Y < y + \Delta y)}{\Delta x \Delta y} \tag{2.64}$$

przy czym:

$$\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x,y)\,dxdy = 1, \tag{2.65}$$

gdzie: $f(x, y) > 0$.

Dystrybuanta dwuwymiarowej zmiennej losowej ciągłej (X, Y) ma postać:

$$F(x,y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(x,y)\,dxdy. \tag{2.66}$$

Rozkłady brzegowe prawdopodobieństwa zmiennych losowych ciągłych X i Y mają następujące funkcje gęstości prawdopodobieństwa:

$$f_x(x)=\int\limits_{-\infty}^{+\infty} f(x,y)dx, \tag{2.67}$$

$$f_y(y)=\int\limits_{-\infty}^{+\infty} f(x,y)dy. \tag{2.68}$$

Warunkowe funkcje gęstości prawdopodobieństwa opisywane są przez następujące wzory:

$$f(x\,|\,y)=\frac{f(x,y)}{f_y(y)}, \tag{2.69}$$

$$f(y\,|\,x)=\frac{f(x,y)}{f_x(x)}. \tag{2.70}$$

Podobnie jak w przypadku dwuwymiarowej zmiennej losowej skokowej, zachodzą następujące zależności:

$$f_x(x)=\int\limits_{-\infty}^{+\infty} f(x,y)dx=\int\limits_{-\infty}^{+\infty}\left[\int\limits_{-\infty}^{+\infty} f(x,y)dy\right]dx=1, \tag{2.71}$$

$$f_y(y)=\int\limits_{-\infty}^{+\infty} f(x,y)dy=\int\limits_{-\infty}^{+\infty}\left[\int\limits_{-\infty}^{+\infty} f(x,y)dx\right]dy=1, \tag{2.72}$$

$$\int\limits_{-\infty}^{+\infty} f(x\,|\,y)dx=\int\limits_{-\infty}^{+\infty}\frac{f(x,y)}{f_y(y)}dx=1, \tag{2.73}$$

$$\int\limits_{-\infty}^{+\infty} f(y\,|\,x)dx=\int\limits_{-\infty}^{+\infty}\frac{f(x,y)}{f_x(x)}dy=1. \tag{2.74}$$

Jeżeli zmienne ciągłe X i Y są niezależne, to wówczas zachodzi zależność:

$$f(x,y)=f_x(x)\,f_y(y). \tag{2.75}$$

Wartość oczekiwaną i kowariancję dwuwymiarowej zmiennej losowej ciągłej obliczamy następująco:

$$E(X\,Y)=\int\limits_{-\infty}^{+\infty}\int\limits_{-\infty}^{+\infty} xyf(x,y)\,dxdy, \tag{2.76}$$

$$\operatorname{cov}(X\,Y)=\int\limits_{-\infty}^{+\infty}\int\limits_{-\infty}^{+\infty}(x-E(X))(y-E(Y))f(x,y)dxdy. \tag{2.77}$$

Wartości oczekiwane rozkładów warunkowych są równe:

$$E(X=x_i|Y=y_j)=\int\limits_{-\infty}^{+\infty} x\,\frac{f(x,y)}{f_y(y)}dx, \tag{2.78}$$

$$E(Y=y_i|X=x_j)=\int_{-\infty}^{+\infty} y\,\frac{f(x,y)}{f_x(x)}\,dy. \tag{2.79}$$

Do dwuwymiarowych zmiennych losowych powrócimy w rozdziale trzecim, przy omawianiu wybranych rozkładów zmiennych losowych[7].

[7] Szersza analiza tego zagadnienia np. w: T. Gerstenkorn, T. Środka: *Kombinatoryka i rachunek prawdopodobieństwa*, rozdział 7, wyd. 5, Warszawa 1979.

Wybrane rozkłady zmiennych losowych oraz podstawowe twierdzenia graniczne

Rozdział

3

3.1. Wybrane rozkłady jednowymiarowych zmiennych losowych dyskretnych

Opiszemy obecnie kilka najważniejszych rozkładów zmiennej losowej dyskretnej. Będą to w kolejności ich omawiania: rozkład dwupunktowy, rozkład zero-jedynkowy, rozkład dwumianowy nazywany od nazwiska jego twórcy rozkładem Bernoulliego. W następnej kolejności opiszemy rozkład Poissona, rozkład hipergeometryczny i na koniec rozkład Pascala oraz rozkład geometryczny. Każdy typ rozkładu postaramy się zobrazować krótkim przykładem liczbowym. Podamy również wzory, dzięki którym można szybko wyznaczyć podstawowe parametry badanych zmiennych losowych, takie jak: wartość oczekiwana, wariancja i odchylenie standardowe.

3.1.1. Rozkład dwupunktowy i rozkład zero-jedynkowy

Zmienna losowa X ma **rozkład dwupunktowy**, jeżeli może ona przyjąć tylko dwie wartości x_1 i x_2 odpowiednio z prawdopodobieństwami[1]:

$$P(X = x_1) = p, P(X = x_2) = q, \text{ przy czym } p + q = 1. \tag{3.1}$$

Jeżeli $x_1 = 1$, natomiast $x_2 = 0$, to wówczas mamy do czynienia ze szczególnym przypadkiem rozkładu dwupunktowego, który nazywany jest **rozkładem zero-jedynkowym**. Funkcja prawdopodobieństwa ma wówczas następującą postać:

$$P(X = 1) = p, P(X = 0) = 1 - p = q. \tag{3.2}$$

Wartość oczekiwana takiej zmiennej losowej wynosi:

$$E(X) = p, \tag{3.3}$$

[1] Zob. np. A. Iwasiewicz, Z. Paszek: *Statystyka...*, *op. cit.*, s. 27.

natomiast wariancja:

$$D^2(X) = pq. \tag{3.4}$$

Zmienna losowa o rozkładzie (3.2) jest związana z doświadczeniem losowym, którego wyniki mogą być dwojakiego rodzaju: posiadające interesującą nas cechę (wówczas zmiennej X nadajemy wartość 1) i nie mające tej cechy (przyjmujemy, że $X = 0$). Przypadek pierwszy zwykło nazywać się „sukcesem", natomiast drugi „porażką" lub „niepowodzeniem"[2].

Przykład 3.1

Doświadczenie polega na losowym wyborze jednej osoby z grupy 30 studentów i zbadaniu oceny, jaką otrzymała ona z egzaminu z „Zarządzania jakością". Załóżmy, że badanie dotyczy grupy studentów opisanej w przykładzie 2.1, a rozkład ocen przedstawia tablica 2.1. Załóżmy, że „sukcesem" nazwiemy takie zdarzenie losowe, podczas którego wylosowana osoba otrzymała ocenę bardzo dobrą (5,0), natomiast „porażką" pozostałe przypadki. Ponieważ w przykładzie 2.1 symbolem X oznaczono zmienną losową będącą oceną losowo wybranego studenta, w celu uniknięcia konfliktu oznaczeń przyjmijmy obecnie, że zmienną o rozkładzie zero-jedynkowym oznaczymy symbolem Y. Zmienna Y przyjmie wartość 1, jeżeli $X = 5{,}0$, natomiast $Y = 0$, jeżeli $X \neq 5{,}0$. Dokładniej:

$$Y = \begin{cases} 1 \text{ gdy } & X = 5{,}0 \\ 0 \text{ gdy } & X = 2{,}0 \vee X = 3{,}0 \vee X = 3{,}5 \vee X = 4{,}0 \vee X = 4{,}5. \end{cases}$$

Korzystając z tablicy 2.1, możemy wyznaczyć częstość rozważanych tu zdarzeń losowych. Ponieważ w badanej grupie 30 studentów były tylko 3 osoby z oceną bardzo dobrą, a pozostałych 27 studentów otrzymało oceny niższe, dlatego:

$$P(Y = 1) = p = \frac{3}{30} = 0{,}1,$$

oraz

$$D(Y) = \sqrt{pq} = \sqrt{0{,}09} = 0{,}3.$$

Wartość oczekiwana, wariancja i odchylenie standardowe badanej zmiennej wyniosą odpowiednio:

$$E(Y) = p = 0{,}1,$$

$$D^2(Y) = pq = 0{,}1 \cdot 0{,}9 = 0{,}09,$$

$$D(Y) = \sqrt{pq} = \sqrt{0{,}09} = 0{,}3. \ \{\text{kp}\}$$

---

[2] Zob. np. Z. Helwig: *op. cit.*, s. 63–64.

3.1.2. Rozkład dwumianowy (binomialny)

Załóżmy, że rozpatrzymy obecnie n zero-jedynkowych zmiennych losowych $X_1, X_2, \ldots, X_n$.

Załóżmy ponadto, że:

$$P(X_1 = 1) = p_1 = P(X_2 = 1) = p_2 = \ldots = P(X_n = 1) = p_n = p. \qquad (3.5)$$

Niech zmienna losowa Z będzie sumą rozważanych n zmiennych zero-jedynkowych:

$$Z = X_1 + X_2 + \ldots + X_n \qquad (3.6)$$

Zmienna Z może przyjąć dowolną wartość $z = 0, 1, 2, \ldots, n$.

Rozkład prawdopodobieństwa tej zmiennej opisuje wzór:

$$P(Z=z;n,p) = \binom{n}{z} p^z (1-p)^{n-z} = \binom{n}{z} p^z q^{n-z}, \qquad (3.7)$$

gdzie:

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

Powyższy rozkład prawdopodobienstwa zmiennej losowej Z nazywa się **rozkładem dwumianowym** (binominalnym)[3], z parametrami n i p. W skrócie fakt, że zmienna losowa Z ma rozkład dwumianowy, można zapisać, stosując następują symbolikę: $Z \sim B(n, p)$.

Zmienną losową Z można interpretować również jako możliwą liczbę sukcesów (realizacji określonego zdarzenia losowego) uzyskanych w dowolnej kolejności, w n niezależnych doświadczeniach, przy założeniu, że prawdopodobieństwo sukcesu w każdym doświadczeniu jest jednakowe i wynosi p.

Powyższy schemat postępowania zwykło się nazywać schematem Bernoulliego[4].

Niezależność doświadczeń w praktyce oznacza najczęściej losowanie ze zwracaniem z populacji ograniczonej (skończonej) lub bezzwrotne, ale z populacji nieograniczonej, przy czym wynik pojedynczego losowania jest zmienną losową o rozkładzie zero-jedynkowym.

Wartość oczekiwana, wariancja i odchylenie standardowe zmiennej Z są równe odpowiednio:

---

[3] Nazwa rozkładu wiąże się z faktem, że prawa strona (1.71) jest rozwinięciem dwumianu Newtona postaci $(p + q)^n$. Szerzej zob. np. Z. Hellwig: *op. cit.* s. 19.

[4] Jakub Bernoulli (1654–1705), szwajcarski matematyk i fizyk, prof. uniwersytetu w Bazylei; twórca podstaw rachunku prawdopodobieństwa, za: *Encyklopedia PWN*, Warszawa 2005, t. 2.

$$E(Z) = np, \tag{3.8}$$

$$D^2(Z) = npq, \tag{3.9}$$

$$D(Z) = \sqrt{npq}. \tag{3.10}$$

Przykład 3.2

Załóżmy, że rozważamy zbiorowość studentów badaną ze względu na ocenę z „Zarządzania jakością" opisaną w przykładach 2.1 i 3.1. Załóżmy, że doświadczenie polega na trzykrotnym niezależnym losowaniu studenta i sprawdzaniu oceny, jaką otrzymał z egzaminu z „Zarządzania jakością". Wyznaczyć prawdopodobieństwo zdarzenia losowego polegającego na tym, że mniej niż dwa razy natrafiono w badaniu na studenta z oceną bardzo dobrą. Ponieważ założyliśmy, że losowanie będzie niezależne (co oznacza, że ta sama osoba może być poddawana badaniu wielokrotnie), oraz badana zmienna Z jest sumą trzech zmiennych zero-jedynkowych (Y_1, Y_2, Y_3) o prawdopodobieństwach $P(Y_1 = 1) = P(Y_2 = 1) = P(Y_3 = 1) = p = 0{,}1$, możemy stwierdzić, opisane powyżej doświadczenie spełnia warunki schematu Bernoulliego.

Można zatem w skrócie zapisać $Z \sim B(n = 3, p = 1)$. Poszukujemy wartości prawdopodobieństwa następującego zdarzenia losowego:

$$P(Z < 2; n = 3, p = 0{,}1) = P(Z = 0) + P(Z = 1) = \binom{3}{0} \cdot 0{,}1^0 \; 0{,}9^3 + \binom{3}{1} \; 0{,}1 \cdot 0{,}9^2 =$$

$$= 0{,}9^3 + 3 \cdot 0{,}1 \cdot 0{,}9^2 = 0{,}729 + 0{,}243 = 0{,}972.$$

Jak łatwo zauważyć, szukane prawdopodobieństwo, zgodnie z definicją (2.5), jest dystrybuantą zmiennej Z w punkcie $x_0 = 2$, co zapiszemy:

$$F_Z(x_0 = 2) = P(Z < 2) = P(Z = 0) + P(Z = 1).$$

Wartość oczekiwana, wariancja i odchylenie standardowe zmiennej Z wyniosą odpowiednio:

$$E(Z) = np = 3 \cdot 0{,}1 = 0{,}3,$$

$$D^2(Z) = npq = 3 \cdot 0{,}1 \cdot 0{,}9 = 0{,}27,$$

$$D(Z) = \sqrt{npq} = \sqrt{0{,}27} \approx 0{,}5196. \text{ \{kp\}}$$

3.1.3. Rozkład Poissona

Zmienna losowa Z posiada rozkład Poissona[5], jeżeli przyjmuje wartości $z = 1, 2, 3, \ldots$ z prawdopodobieństwami[6]:

$$P(Z = z; \lambda) = e^{-\lambda}\frac{\lambda^z}{z!}, \tag{3.11}$$

gdzie $\lambda = np$ jest parametrem tego rozkładu, natomiast $e \approx 2{,}718$.

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

W praktyce rozkładu Poissona używa się w sytuacji, gdy liczba serii niezależnych doświadczeń wzrasta do nieskończoności ($n \to \infty$), natomiast prawdopodobieństwo „sukcesu" p maleje do zera ($p \to 0$), przy czym $np = \lambda$ jest wielkością stałą ($\lambda > 0$).

Z powyższego stwierdzenia wynika, że rozkład Poissona można traktować, jako rozkład graniczny rozkładu dwumianowego. Można zatem zapisać:

$$\lim_{n\to\infty}\binom{n}{z}p^z q^{n-z} = e^{-\lambda}\frac{\lambda^z}{z!}. \tag{3.12}$$

Rozkład Poissona stanowi tym lepsze przybliżenie rozkładu dwumianowego, im n jest większe, a p bliższe zera.

Problematyczne pozostaje ustalenie minimalnych wartości n i λ, przy których można zastąpić wzór (3.7) wzorem (3.11). W literaturze znajdujemy kilka propozycji. Na przykład T. Gerstenkorn i T. Środka[7] proponują, aby $n \geq 100$ i $\lambda \leq 20$; Krysicki i współautorzy książki *Rachunek prawdopodobieństwa i statystyka matematyczna w zadaniach*[8] podają, że wystarczy aby $n > 50$ i $\lambda < 10$. Na jeszcze większe ustępstwa pozwalają A. Iwasiewicz i Z. Paszek, autorzy książki *Statystyka z elementami statystycznych metod sterowania jakością*[9], proponując, aby: $n > 20$, a $\lambda < 4$.

Na użytek niniejszej publikacji w dalszej części podręcznika przyjmiemy, że to ostatnie – najbardziej liberalne założenie – jest wystarczające, aby dość dobrze przybliżyć rozkład dwumianowy.

Wartość oczekiwana oraz wariancja zmiennej losowej Z o rozkładzie Poissona wynosi:

$$E(Z) = D^2(Z) = \lambda, \tag{3.13}$$

[5] Denis Simon Poisson (1781–1840), francuski mechanik, fizyk i matematyk, za: *Encyklopedia PWN*, Warszawa 2005, t. 13.

[6] Zob. A. Iwasiewicz, Z. Paszek: *Statystyka...*, *op. cit.*, s. 33.

[7] T. Gerstenkorn, T. Środka: *Kombinatoryka i rachunek prawdopodobieństwa*, wyd. 5, *op. cit.*

[8] W. Krysicki i inni: *Rachunek prawdopodobieństwa...*, *op. cit*, s. 85

[9] A. Iwasiewicz, Z. Paszek: *Statystyka...*, *op. cit.* s. 34.

natomiast odchylenie standardowe:

$$D(Z) = \sqrt{\lambda}. \tag{3.14}$$

Przykład 3.3

Podczas kontroli jakości sprzedawanych paliw zakłada się, że 1% stacji paliwowych w kraju sprzedaje paliwo złej jakości. Obliczyć prawdopodobieństwo, że w losowo wybranej niezależnej próbie 100 stacji znajdą się:

a) dwie stacje sprzedające paliwo złej jakości,

b) co najwyżej jedna stacja handlująca paliwami złej jakości.

ad a)

Niech Z oznacza zmienną losową będącą liczbą stacji, które w próbie 100 pobranych do kontroli sprzedawały paliwa złej jakości.

Ponieważ $n = 100$, a $\lambda = 1$, do wyznaczenia wartości szukanego prawdopodobieństwa możemy wykorzystać wzór (3.11) i otrzymamy:

$$P(Z = 2; \lambda = 1) = e^{-1}\frac{1}{2!} \approx 0{,}184.$$

ad b)

Określenie, że co najwyżej jedna stacja sprzedaje paliwa złej jakości, oznacza, że w pobranej próbie znajdzie się jedna stacja sprzedająca paliwa złej jakości lub wszystkie ze zbadanych stu stacji handlowały paliwami spełniającymi normy jakościowe.

Szukamy prawdopodobieństwa następującego zdarzenia losowego:

$$P(Z < 1; \lambda = 1) = F_Z(2) = P(Z = 0) + P(Z = 1) =$$
$$= e^{-1}\frac{1^0}{0!} + e^{-1}\frac{1}{1!} = 2\ e^{-1} = 2 \cdot 0{,}368 = 0{,}736.$$

Do obliczenia szukanych prawdopodobieństw, można wykorzystac rownież rozkład dwumianowy. Wówczas prawdopodobieństwo realizacji zdarzenia opisanego w podpunkcie *a* wyniesie:

$$P(Z = 2; n = 100, p = 0{,}01) = \binom{100}{2} \cdot 0{,}01^2 \cdot 0{,}99^{98} =$$

$$= 4950 \cdot 0{,}0001 \cdot 0{,}0001 \cdot 0{,}37346428 = 0{,}184864819 \approx 0{,}185.$$

Korzystając z rozkładu dwumianowego, można również wyznaczyć dokładną wartość prawdopodobieństwa zdarzenia losowego, wyznaczonego w punkcie *b* tego przykładu.

Otrzymamy wówczas:

$$P(Z < 2; n = 100, p = 0{,}01) = P(Z = 0) + P(Z = 1) =$$

$$= \binom{100}{0} \cdot 0{,}01^0 \; 0{,}99^{100} + \binom{100}{1} \, 0{,}01 \cdot 0{,}99^{99} = 0{,}99^{100} + 100 \; 0{,}1 \; 0{,}99^{99} =$$

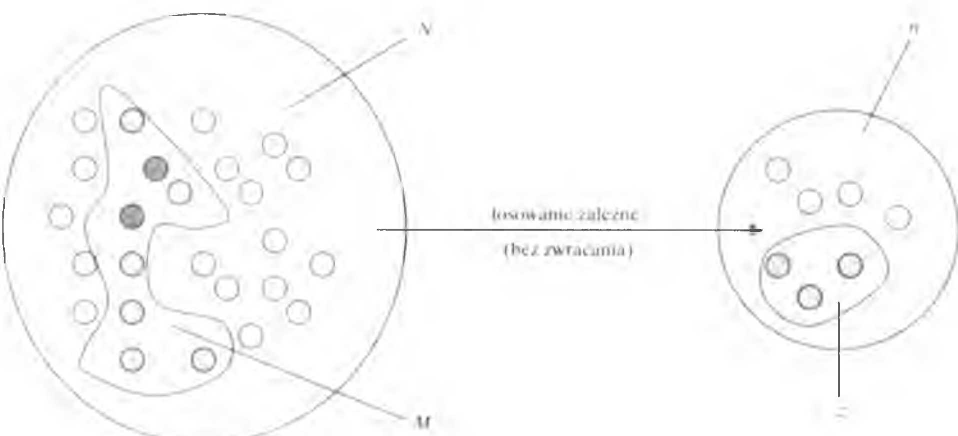
$$= 0{,}366032341 + 0{,}369729638 = 0{,}735761979 \approx 0{,}736.$$

Porównując powyższe wyniki, wyznaczone z zastosowaniem schematów rozkładu Poissona i rozkładu dwumianowego, można dojść do wniosku, że różnice pomiędzy wynikami przybliżonymi za pomocą schematu Poissona i wynikami dokładnymi obliczonymi z wykorzystaniem schematu Bernoulliego są na tyle małe, że można je zignorowac na rzecz zmniejszenia pracochłonności obliczeń. {kp}

3.1.4. Rozkład hipergeometryczny

Załóżmy obecnie, że z populacji N-elementowej, w której znajduje się M elementów posiadających określoną cechę, losujemy w sposób **zależny** (bezzwrotny) n-elementów. Interesujące nas zdarzenie losowe to wylosowanie z elementów o określonej cesze.

Poniższe doświadczenie zilustrowano na rys. 17.



Rys. 17. Schemat losowania zależnego

Źródło: opracowanie własne.

Niech zmienna Z oznacza liczbę elementów o określonej cesze (na rysunku elementów zacieniowanych) znajdujących się w wylosowanej zależnie próbie o liczności n. Zmienna ta może przyjmować wartości całkowite z przedziału $[0, n]$, a jej rozkład prawdopodobieństwa nazywany jest **rozkładem hipergeometrycznym** ($Z \sim H(N, M, n)$). Prawdopodobieństwo zdarzenia losowego $P(Z = z; N, M, n)$ wyznaczamy, stosując następujący wzór kombinatoryczny:

$$P(Z = z; N, M, n) = \frac{\binom{M}{z}\binom{N-M}{n-z}}{\binom{N}{n}} \tag{3.15}$$

gdzie:

$N - M$ – liczba elementów w populacji pozbawionych określonej cechy (na rysunku elementy białe),

$n - z$ – liczba elementów w próbie nieposiadających określonej cechy,

przy czym:

$$z = 0, 1, ..., n; n \le N; z \le M; z \le n; n - z \le N - M.$$

Jeżeli założymy, że $p = \frac{M}{n}$, to wówczas wzór (3.15) można zapisać następująco[10]:

$$P(Z = z; N, M, n) = \frac{\binom{Np}{z}\binom{Nq}{n-z}}{\binom{N}{n}} \tag{3.16}$$

Wartość oczekiwaną oraz wariancję zmiennej losowej Z wyznacza się, stosując wzory (3.17) i (3.18):

$$E(Z) = \frac{nM}{N} = np, \tag{3.17}$$

$$D^2(Z) = \frac{N-n}{N-1} npq. \tag{3.18}$$

W pewnych przypadkach, zwłaszcza, gdy $\frac{n}{N} \to 0$, rozkład hipergeometryczny można przybliżyć za pomocą rozkładu dwumianowego.

W przypadku, gdy $N \to \infty$, $M \to \infty$, $n \to \infty$, a także $E(Z) \to \lambda$, $\lambda > 0$, wówczas

$$P(Z = z; N, M, n) \to P(Z = z; \lambda), \tag{3.19}$$

co oznacza, że dokładną wartość uzyskaną z rozkładu hipergeometrycznego można przybliżyć, używając rozkładu Poissona[11].

[10] Zob. A. Iwasiewicz, Z. Paszek: *Statystyka..., op. cit.*, s. 38.

[11] W. Krysicki i inni: *Rachunek prawdopodobieństwa..., op. cit.*, s. 86.

Przykład 3.4

Wykorzystując informacje zawarte w przykładzie 3.2 oraz powiązanych z nim przykładach 2.1 i 3.1, wyznaczyć prawdopodobieństwo zdarzenia losowego polegającego na tym, że w trzykrotnym zależnym (bez zwracania) losowaniu studentów, mniej niż dwa razy natrafiono w badaniu na studenta z oceną bardzo dobrą.

Ponieważ założyliśmy obecnie, że losowanie odbywa się bez zwracania, do wyznaczenia szukanego prawdopodobieństwa, nie można zastosować schematu Bernoulliego, lecz należy skorzystać ze wzoru na rozkład hipergeometryczny.

Wartość szukanego prawdopodobieństwa wyniesie:

$$P(Z < 1;\ N = 30, M = 3, n = 3) = P(Z = 0) + P(Z = 1) =$$

$$= \frac{\binom{3}{0}\binom{27}{3}}{\binom{30}{3}} + \frac{\binom{3}{1}\binom{27}{2}}{\binom{30}{3}} = \frac{2925}{4060} + \frac{1053}{4060} = 0{,}72 + 0{,}259 = 0{,}979.$$

Jak widać, różnica pomiędzy wartością otrzymaną przy wykorzystaniu rozkładu hipergeometrycznego i wartością obliczoną na podstawie rozkładu dwumianowego (zob. Przykład 3.2) jest bardzo mała i wynosi zaledwie 0,007 (0,979 – 0,972 = 0,007). Jest to efekt tego, że stosunek $n/N = 0{,}1$ jest wartością bliską zera, co pozwala na aproksymację rozkładu hipergeometrycznego za pomocą rozkładu dwumianowego. {kp}

3.1.5. Rozkład Pascala (ujemny rozkład dwumianowy) i rozkład geometryczny

Zagadnienie Pascala jest nieco podobne do zagadnienia Bernoulliego. W schemacie Bernoulliego liczba niezależnych doświadczeń n jest z góry ustalona, a szukamy prawdopodobieństwa, że podczas tych doświadczeń z razy osiągniemy sukces. W schemacie Pascala szukamy natomiast prawdopodobieństwa zdarzenia losowego, polegającego na tym, że liczba przeprowadzonych zgodnie ze schematem Bernoulliego doświadczeń będzie równa $N = n$, przy czym zakłada się, że losowanie przeprowadza się do momentu otrzymania ustalonej liczby sukcesów z ($z > 1$).

T. Gerstenkorn i T. Środka rozwiązanie tego problemu ujmują w formie następującego twierdzenia[12]:

[12] T. Gerstenkorn, T. Środka: *Kombinatoryka..., op. cit.*, s. 172.

Jeżeli przeprowadzamy doświadczenia według schematu Bernoulliego o stałym prawdopodobieństwie sukcesu w poszczególnym doświadczeniu równym p aż do momentu uzyskania z góry ustalonej liczby z sukcesów ($z > 1$), to prawdopodobieństwo, że liczba doświadczeń będzie równa n ($n > z$), wyraża się wzorem:

$$P(N=n;z,p)=\binom{n-1}{z-1}p^z q^{n-z}, \tag{3.20}$$

przy czym:

$$n > z > 1, q = 1-p.$$

Fakt, że zmienna losowa N ma rozkład Pascala z parametrami p i z (ujemny dwumianowy, ang. *negative binomial*), będziemy zapisywać w skrócie $N \sim NB(p, z)$.

Wartość oczekiwana i wariancja zmiennej losowej N jest równa odpowiednio:

$$E(N)=\frac{z}{p}, \tag{3.21}$$

$$D^2(N)=\frac{z\cdot q}{p^2}. \tag{3.22}$$

W przypadku, gdy liczba sukcesów $z = 1$, wzór (3.20) redukuje się do postaci:

$$P(N = n; z = 1, p) = pq^{n-1}. \tag{3.23}$$

Rozkład prawdopodobieństwa zmiennej losowej N wyrażony wzorem (3.23) zwykło nazywać się **rozkładem geometrycznym**, a fakt ten zapisywać w skrócie $N \sim G(p, z)$.

Wartość oczekiwaną oraz wariancję zmiennej losowej o rozkładzie geometrycznym opisują wzory:

$$E(N)=\frac{1}{p}, \tag{3.24}$$

$$D^2(N)=\frac{q}{p^2}. \tag{3.25}$$

Przykład 3.5

Doświadczenie polega na losowaniu niezależnym studenta z grupy 30-osobowej i sprawdzeniu oceny otrzymanej z egzaminu z „Zarządzania jakością". Losowanie prowadzi się do momentu dwukrotnego wylosowania studenta, który otrzymał ocenę bardzo dobrą (5,0). Zakładając, że rozkład ocen studentów jest identyczny jak w przykładzie 2.1, obliczyć prawdopodobieństwo, że losowanie będzie sześciokrotne.

Szukane prawdopodobieństwo wyznaczamy, stosując wzór (3.20) i otrzymujemy:

$$P(N = 6; z = 2, p = 0{,}1) = \binom{5}{1} 0{,}1^2\, 0{,}9^4 = 5 \cdot 0{,}01 \cdot 0{,}6561 = 0{,}032805 \approx 0{,}033.$$

Załóżmy obecnie, że prowadzimy losowanie niezależne do momentu natrafienia na pierwszego studenta, który otrzymał oceną bardzo dobrą. Wyznaczmy prawdopodobieństwo zdarzenia losowego, że potrzebne będzie również sześć losowań. Intuicyjnie prawdopodobieństwo takiego zdarzenia powinno być dwa razy większe od prawdopodobieństwa wyznaczonego w pierwszej części tego przykładu.

Ponieważ losowanie przeprowadzamy do momentu osiągnięcia pierwszego „sukcesu" ($z = 1$), wzór (3.20) można zastąpić wzorem (3.23). Otrzymamy wówczas:

$$P(N = 6; z = 1, p = 0{,}1) = 0{,}1 \cdot 0{,}9^5 \approx 0{,}059.$$

Potwierdza się intuicyjne przypuszczenie, że prawdopodobieństwo uzyskania jednego sukcesu w sześciu niezależnych próbach (ok. 6%) jest dwukrotnie większe niż uzyskanie dwóch sukcesów (ok. 3%). {kp}

3.2. Wybrane rozkłady jednowymiarowych ciągłych zmiennych losowych

Spośród licznych rozkładów ciągłej zmiennej losowej omówimy trzy ważniejsze. Będą to: rozkład prostokątny (jednostajny), rozkład normalny i rozkład wykładniczy. Podobnie jak zrobiliśmy to w punkcie 3.1, obok wzorów opisujących rozkłady zmiennych losowych podamy sposoby wyznaczania wartości oczekiwanej i wariancji tych zmiennych. Każdy z opisanych rozkładów zilustrujemy krótkim przykładem liczbowym.

3.2.1. Rozkład prostokątny (jednostajny, równomierny)

Zmienna losowa X ma **rozkład prostokątny** (jednostajny, równomierny) w sytuacji, gdy funkcję gęstości tej zmiennej opisuje wzór:

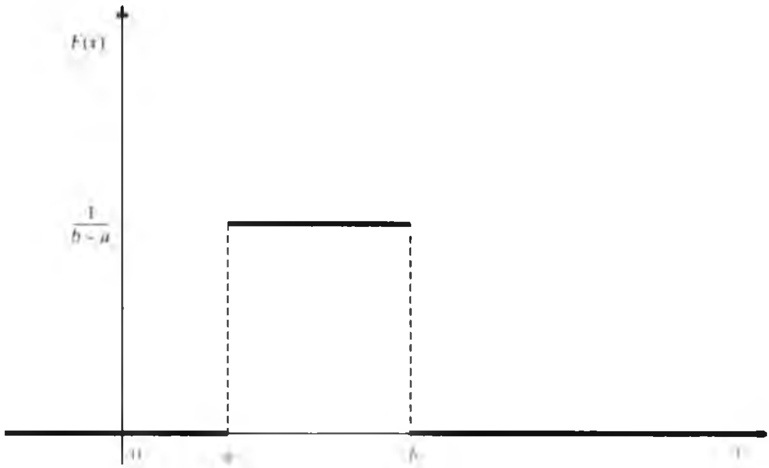
$$f(x) = \begin{cases} 0 & dla \quad x < a \\ \dfrac{1}{b-a} & dla \quad a \le x \le b. \\ 0 & dla \quad x > b \end{cases} \tag{3.26}$$

Dystrybuanta zmiennej losowej X opisana jest wzorem:

$$F(x) = \begin{cases} 0 & dla \quad x < a \\ \dfrac{x-a}{b-a} & dla \quad a < x < b. \\ 1 & dla \quad x > b \end{cases} \tag{3.27}$$

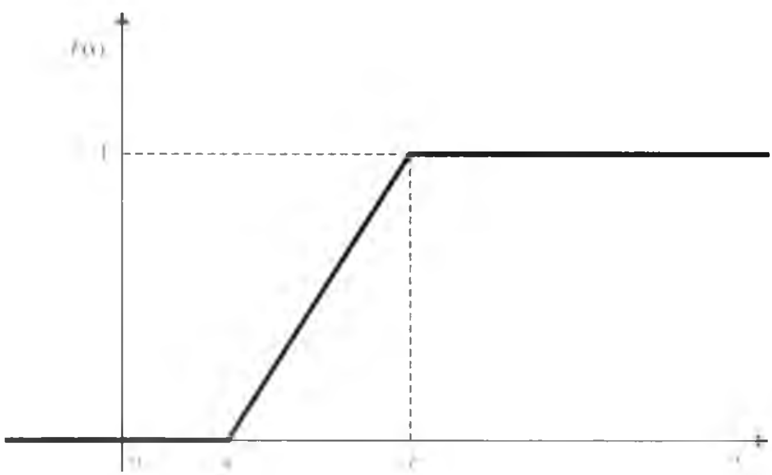
Fakt, że zmienna losowa X ma rozkład prostokątny, zapiszemy w skrócie $X \sim R(a, b)$.

Wykresy funkcji gęstości oraz funkcji dystrybuanty przedstawiają rysunki 18 oraz 19.



Rys. 18. Funkcja gęstości prawdopodobieństwa rozkładu zmiennej losowej o rozkładzie prostokątnym

Źródło: opracowanie własne.



Rys. 19. Funkcja dystrybuanty rozkładu zmiennej losowej o rozkładzie prostokątnym

Źródło: opracowanie własne.

Wystarczy popatrzeć na rysunek 3.2, aby zrozumieć, skąd wzięła się nazwa rozkładu. Pole powierzchni zawartej pomiędzy funkcją gęstości, w przedziale $[a, b]$, a osią odciętych, jest polem prostokąta o bokach *b-a* na 1/(*b-a*). Oczywiście, iloczyn tych boków wynosi 1.

Wartość oczekiwana i wariancja zmiennej losowej o rozkładzie prostokątnym wynoszą odpowiednio:

$$E(X)=\frac{a+b}{2}, \tag{3.28}$$

$$D^2(X)=\frac{(b-a)^2}{12}. \tag{3.29}$$

Przykład 3.6

Powróćmy do przykładu 2.4, w którym zmienną losową X był czas oczekiwania pasażera na przyjazd tramwaju linii 15. Przypomnijmy, że wyznaczona w przykładzie 2.4, funkcja gęstości prawdopodobieństwa zmiennej losowej X miała postać:

$$f(x)=\begin{cases} 0 & dla \quad x<0 \\ \frac{1}{20} & dla \quad 0\le x\le 20, \\ 0 & dla \quad x>20 \end{cases}$$

Zmienna losowa X ma więc rozkład jednostajny o parametrach $a = 0$ i $b = 20$. Korzystając ze wzoru (3.27), możemy wyznaczyć – bez konieczności stosowania rachunku całkowego – postać dystrybuanty:

$$F(x)=\begin{cases} 0 & dla \quad x\le 0 \\ \frac{x-0}{20-0} & dla \quad 0<x<20, \\ 1 & dla \quad x\ge 20 \end{cases}$$

z czego wynika:

$$F(x)=\begin{cases} 0 & dla \quad x\le 0 \\ \frac{1}{20}x & dla \quad 0<x<20. \\ 1 & dla \quad x\ge 20 \end{cases}$$

Wzory (3.28) i (3.29) pozwalają natomiast szybko określić wartość oczekiwaną i wariancję badanej zmiennej losowej X. Otrzymamy wówczas:

$$E(X)=\frac{0+20}{2}=\frac{20}{2}=10,$$

$$D^2(X)=\frac{(20-0)^2}{12}=\frac{400}{12}=33\frac{1}{3}.$$

Otrzymaliśmy – szybciej i łatwiej – identyczne wyniki jak podczas stosowania rachunku całkowego w przykładzie 2.4. {kp}

3.2.2. Rozkład normalny

Zmienna losowa X ma rozkład normalny, nazywany też rozkładem Moivre'a--Gaussa[13], o parametrach μ i σ (co zapisujemy w skrócie $X \sim N(\mu; \sigma)$), jeżeli jej funkcja gęstości prawdopodobieństwa ma postać:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2 \right] \tag{3.30}$$

przy czym:

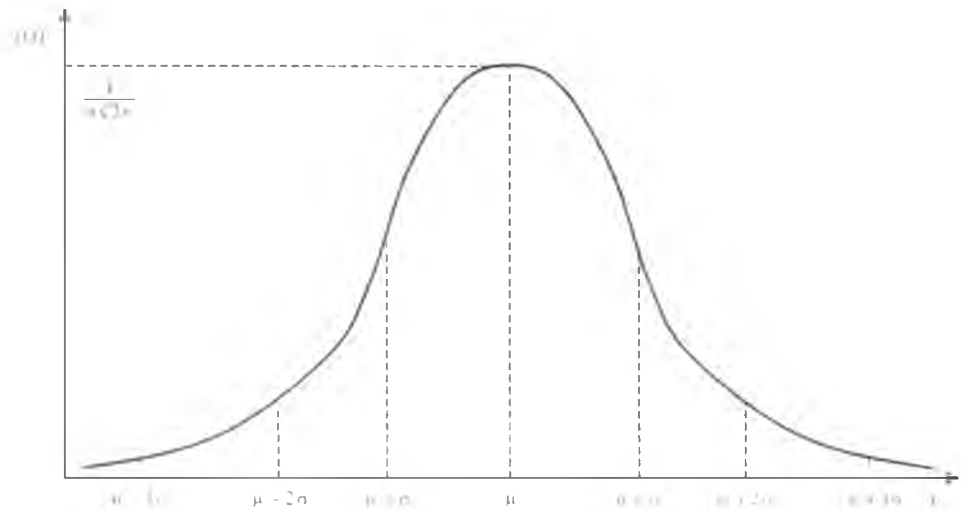
$$x \in R,\ \mu \in R, \text{ natomiast } \sigma \in R^+$$

Wartość oczekiwana oraz wariancja zmiennej losowej X wynoszą odpowiednio:

$$E(X) = \mu, \tag{3.31}$$

$$D^2(X) = \sigma^2. \tag{3.32}$$

Charakterystyczny wygląd krzywej rozkładu normalnego prezentuje rys. 20:



Rys. 20. Przykładowa krzywa gęstości rozkładu normalnego

Źródło: opracowanie własne.

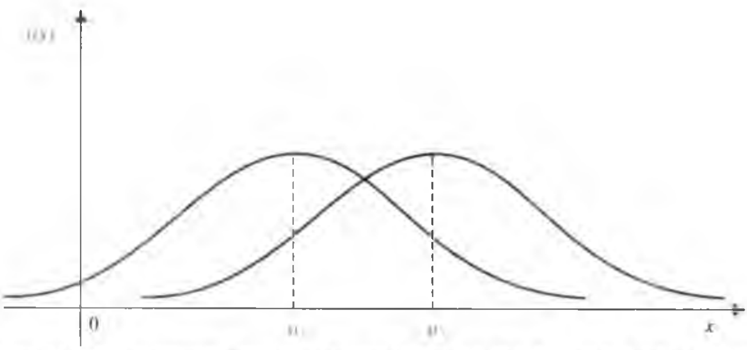
Jak widać, rozkład normalny jest rozkładem symetrycznym względem prostej przebiegającej $x = \mu$. Funkcja ta w punkcie μ osiąga swoje maksimum wynoszące[14]:

[13] Carl Friedrich Gauss (1777–1855), jeden z najwybitniejszych niemieckich matematyków; Abraham de Moivre (1667–1754), matematyk angielski pochodzenia francuskiego.

[14] Wartość maksimum można obliczyć, wyznaczając pochodną funkcji (3.30) i przyrównując ją do zera.

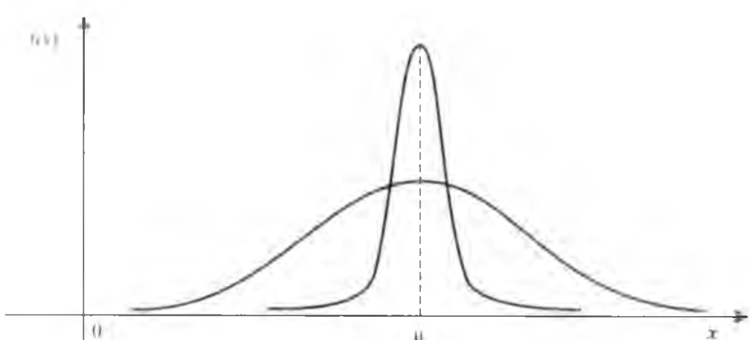
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}. \tag{3.33}$$

Im dalej na lewo i prawo od wartości μ – w której funkcja gęstości osiąga maksimum – tym krzywa gęstości bardziej opada i zbliża się asymptotycznie do osi odciętych. Punkty przegięcia krzywej gęstosci rozkładu normalnego mają odcięte $\mu \pm \sigma$. Parametr μ decyduje o położeniu krzywej na wykresie, natomiast odchylenie standardowe σ decyduje o kształcie tej krzywej (wpływa na wartość maksymalną funkcji gęstości). Im odchylenie standardowe jest mniejsze, tym krzywa gęstości jest bardziej „wysmukła”, im większe, tym bardziej „spłaszczona”. Na rysunku 21 przedstawiono dwa wykresy rozkładu normalnego o tym samym odchyleniu standardowym, lecz różniące się wartością oczekiwaną. Z kolei na rysunku 22 wartość oczekiwana jest taka sama, natomiast rozkłady różnią się wartością odchylenia standardowego.



Rys. 21. Dwie krzywe rozkładu normalnego o jednakowych odchyleniach standardowych i różnych wartościach oczekiwanych $\mu_1 < \mu_2$

Źródło: opracowanie własne.



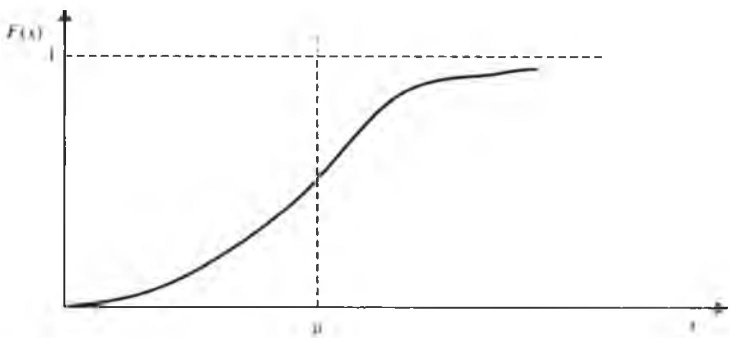
Rys. 22. Dwie krzywe rozkładu normalnego o identycznych wartościach oczekiwanych i różnych odchyleniach standardowych

Źródło: opracowanie własne.

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

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] dx. \tag{3.34}$$

Przykładowy wykres dystrybuanty przedstawiono na rysunku 23.



Rys. 23. Przykładowy wykres dystrybuanty rozkładu normalnego

Źródło: opracowanie własne.

Funkcję gęstości rozkładu normalnego charakteryzują następujące właściwości[15]:

$$P(\mu - \sigma < X < \mu + \sigma) = 0{,}6826, \tag{3.35}$$

$$P(\mu - 2\sigma < X < \mu + 2\sigma) = 0{,}9545, \tag{3.36}$$

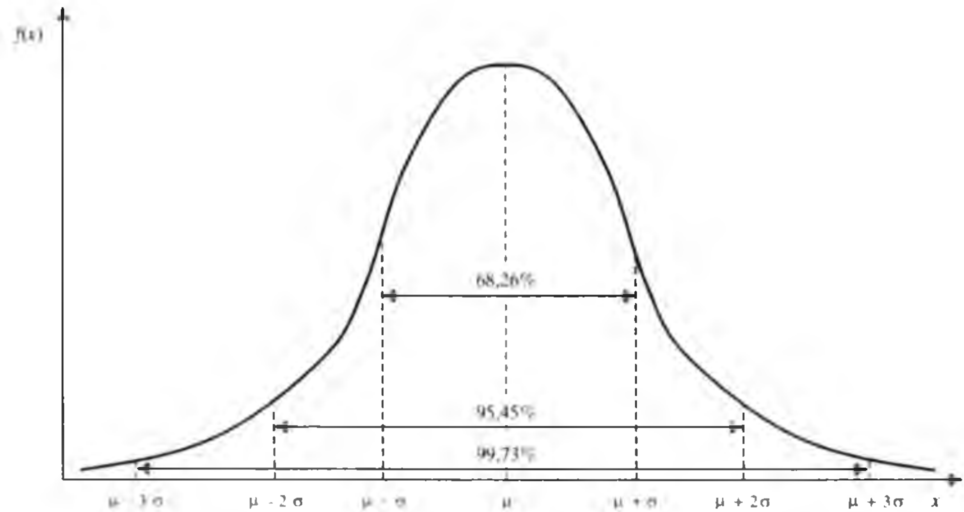
$$P(\mu - 3\sigma < X < \mu + 3\sigma) = 0{,}9973. \tag{3.37}$$

Powyższe zależności ilustruje rysunek 24.

Własność (3.37) określa się mianem reguły „trzech sigm". Oznacza ona, że prawie sto procent wartości, jakie przyjmuje zmienna losowa X, znajduje się w przedziale

$$(\mu - 3\sigma, \mu + 3\sigma).$$

[15] A. Iwasiewicz, Z. Paszek: *Statystyka...*, *op. cit.*, s. 42.



Rys. 24. Własności funkcji gęstości rozkładu normalnego

Źródło: A. Iwasiewicz, Z. Paszek: *Statystyka*..., *op. cit.*, s. 43.

W praktyce korzysta się najczęściej ze zmiennej standaryzowanej $U \sim N(\mu = 0; \sigma = 1)$, która powstaje w wyniku następującej transpozycji:

$$U = \frac{X - E(X)}{D(X)}. \tag{3.38}$$

Wartości zmiennej losowej U wynoszą:

$$u = \frac{x - \mu}{\sigma}. \tag{3.39}$$

Wartość oczekiwana oraz wariancja zmiennej U wynoszą odpowiednio:

$$E(U) = 0, \tag{3.40}$$

$$D^2(U) = 1. \tag{3.41}$$

Modyfikacji ulegnie również funkcja gęstości i dystrybuanta zmiennej standaryzowanej U.

Funkcja gęstości prawdopodobieństwa ma postać:

$$\varphi(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right). \tag{3.42}$$

przy czym $u \in R$, natomiast dystrybuanta (nazywana często funkcją lub całką Laplace'a[16]) zmiennej losowej U przedstawia się następująco:

[16] Pierre S. de Laplace'a (1749–1855).

$$\Phi(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{u} \exp\left(-\frac{u^2}{\sigma}\right) du \tag{3.43}$$

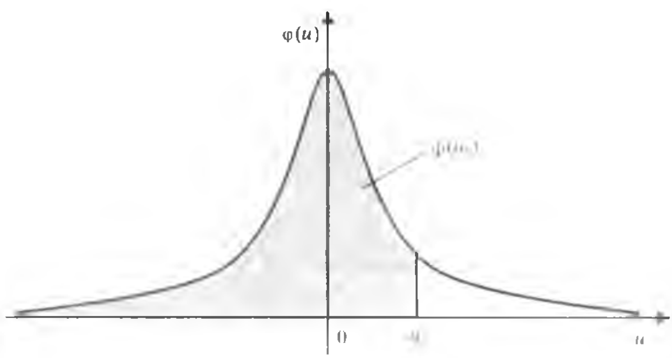
W większości podręczników do statystyki publikuje się tablice prezentujące wartości funkcji (3.43) w przedziale od $u \approx -3$ do $u \approx 3$, lub w przedziale od $u = 0$ do $u \approx 3$.

W dodatku do tego podręcznika (zob. tablica I) zamieszczono tablice dystrybuanty zmiennej standaryzowanej U w przedziale od $u = -3$ do $u = 3$.

Dzięki tablicom nie musimy wielokrotnie wyznaczać całki zgodnie ze wzorem (3.43). W przypadku, gdy mamy do wyznaczenia wartość dystrybuanty zmiennej losowej U dla dowolnej wartości $-u_0$, i korzystamy z tablic podających wartości dystrybuanty tylko dla przedziału od $u = 0$ do $u \approx 3$, to wówczas można skorzystać z następującej zależności:

$$\Phi(-u_0) = 1 - \Phi(u_0). \tag{3.44}$$

Powyższa własność wynika z faktu, że funkcja gęstosci jest funkcją symetryczną względem osi rzędnych. Na rysunku 25 przedstawiono graficzną prezentację dystrybuanty $\Phi(u_0)$.



Rys. 25. Graficzna prezentacja dystrybuanty $\Phi(u_0)$

Źródło: opracowanie własne.

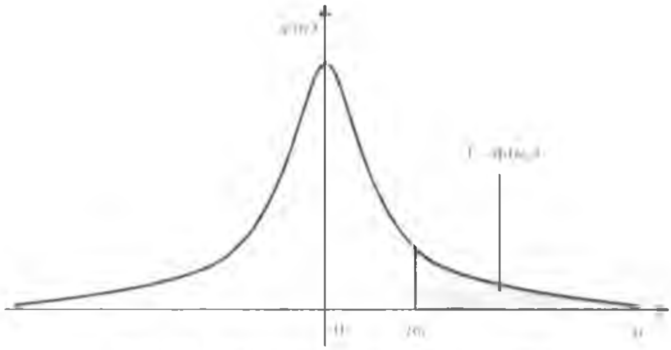
Obok własności (3.44) możemy również sformułować trzy kolejne własności:

$$P(X < x_0) = F(x_0) = \Phi\left(\frac{x_0 - \mu}{\sigma}\right) = \Phi(u_0), \tag{3.45}$$

$$P(X > x_0) = 1 - F(x_0) = 1 - \Phi\left(\frac{x_0 - \mu}{\sigma}\right) = 1 - \Phi(u_0), \tag{3.46}$$

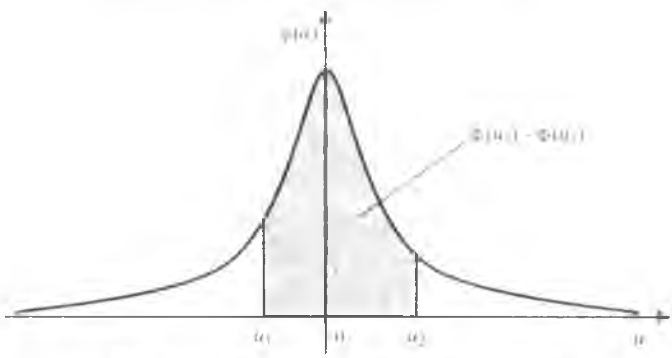
$$\begin{aligned} P(x_1 < X < x_2) &= P\left(\frac{x_1-\mu}{\sigma} < \frac{X-\mu}{\sigma} < \frac{x_2-\mu}{\sigma}\right) = \\ &= P\left(\frac{x_1-\mu}{\sigma} < U < \frac{x_2-\mu}{\sigma}\right) = \Phi\left(\frac{x_2-\mu}{\sigma}\right) - \Phi\left(\frac{x_1-\mu}{\sigma}\right) = \Phi(u_2) - \Phi(u_1). \end{aligned} \tag{3.47}$$

Własności (3.46) i (3.47) ilustrują rysunki 26 i 27.



Rys. 26. Graficzna interpretacja $1 - \Phi(u_0)$

Źródło: opracowanie własne.



Rys. 27. Graficzna ilustracja $\Phi(u_2) - \Phi(u_1)$

Źródło: opracowanie własne.

Przykład 3.7

W bardzo dużej grupie studentów przeprowadzono egzamin z „Zarządzania jakością" – mierząc wyniki na ciągłej skali od 0 do 40 punktów. Ustalono, że rozkład wyników jest zbliżony do normalnego ze średnią 29,5 i odchyleniem standardowym wynoszącym 6,4.

1) Obliczyć prawdopodobieństwo, zdarzenia losowego, że wybrany przypadkowo student otrzymał liczbę punktów:
   a) mniejszą niż 25,
   b) z przedziału (25; 35),
   c) większą niż 35;
2) wiedząc, że 10% studentów otrzymało oceny bardzo dobre (5,0), proszę obliczyć, ile minimalnie należało otrzymać punktów, aby otrzymać ocenę 5,0?

ad 1)

Niech zmienna losowa X oznacza liczbę punktów uzyskanych przez losowo wybranego studenta, przy czym $X \sim N(29{,}5;\, 6{,}4)$.

ad 1a)

$$P(X<25)=P\left(\frac{X-29{,}5}{6{,}4}<\frac{25-29{,}5}{6{,}4}\right)=P(U<-0{,}70)=$$
$$=\Phi(-0{,}70)=1-\Phi(0{,}70)=1-0{,}7580=0{,}242.$$

ad 1b)

$$P(25<X<35)=P\left(\frac{25-29{,}5}{6{,}4}<\frac{X-29{,}5}{6{,}4}<\frac{35-29{,}5}{6{,}4}\right)=P(-0{,}70<U<0{,}86)=$$
$$=\Phi(0{,}86)-\Phi(-0{,}70)=0{,}8051-[1-\Phi(0{,}70)]=$$
$$=0{,}8051-1+0{,}7580=0{,}5631.$$

ad 1c)

$$P(X>35)=P\left(\frac{X-29{,}5}{6{,}4}>\frac{35-29{,}5}{6{,}4}\right)=P(U>0{,}86)=$$
$$=1-\Phi(0{,}86)=1-0{,}8051=0{,}1949.$$

Uzyskane wyniki oznaczają, że ok. 24% studentów otrzymało mniej niż 25 punktów, ponad 56% studentów otrzymało liczbę punktów z przedziału (25; 35) i ok. 19,5% badanych miało liczbę punktów większą niż 35.

ad 2)

Wiedząc, że 10% studentów otrzymało ocenę bardzo dobrą, możemy wnioskować, że oceny 90% badanych są niższe niż 5,0. Można zatem zapisać:

$$P(X<x_d)=0{,}9,$$

gdzie x_d oznacza dolną granicę przedziału punktowego, któremu przyporządkowano ocenę 5,0.

Utrzymując wcześniejsze założenie: $X \sim N(29{,}5;\, 6{,}4)$, możemy zapisać:

$$P(X < x_d) = 0{,}9 \Rightarrow P\left(\frac{X-29{,}5}{6{,}4} < \frac{x_d-29{,}5}{6{,}4}\right) = 0{,}9 \Rightarrow$$

$$\Rightarrow P\left(U < \frac{x_d-29{,}5}{6{,}4}\right) = 0{,}9 \Rightarrow \Phi\left(\frac{x_d-29{,}5}{6{,}4}\right) = 0{,}9.$$

Obecnie naszym zadaniem jest znalezienie takiej wartości u_0, dla której dystrybuanta $\Phi(u_0) = 0{,}9$. Korzystając z tablicy I, szukamy wśród wartości dystrybuanty najbliższej 0,9 i odczytujemy wartość u_0 jako sumę wartości występujących w nagłówku odpowiedniej kolumny i odpowiedniego wiersza dającego na przecięciu wartość najbardziej zbliżoną do 0,9. W badanym przypadku jest to wiersz 1,2 i kolumna 0,08, co po zsumowaniu daje wartość 1,28. Można więc zapisać:

$$\Phi(1{,}28) \approx 0{,}9 \Rightarrow \frac{x_d-29{,}5}{6{,}4} = 1{,}28.$$

Po rozwiązaniu powyższej równości względem x_d otrzymamy:

$$x_d = 1{,}28 \cdot 6{,}4 + 29{,}5 = 37{,}692 \approx 37{,}7.$$

Powyższy wynik oznacza, że w przybliżeniu od 38 punktów na 40 możliwych stawiano studentom ocenę bardzo dobrą. {kp}

W pewnych przypadkach rozkład normalny można stosować jako dosyć dobre przybliżenie rozkładu dwumianowego. Dzieje się tak w przypadkach, gdy liczność próby (n) jest duża, a prawdopodobieństwo „sukcesu" bliskie 0,5. J.E. Freund[17] podaje, że rozkład normalny o średniej $\mu = np$ i odchyleniu standardowym $\sigma = \sqrt{np\,(1-p)}$, można stosować do aproksymowania rozkładów dwumianowych, nawet gdy n jest „niezbyt duże", a p różne od 0,5, lecz nie bliskie 0 ani 1. Cytowany autor podaje, że aby możliwe było użycie rozkładu normalnego do aproksymacji rozkładu dwumianowego, to w praktyce wystarczy sprawdzić, czy np oraz $n(1\text{-}p)$ jest większe niż 5.

Przykład 3.8

Załóżmy, że badamy opisaną w przykładzie 2.1 grupę 30 studentów, ze względu na otrzymaną ocenę z egzaminu z „Zarządzania jakością". Przeprowadzono doświadczenie polegające na 12-krotnym niezależnym wylosowaniu studenta i sprawdzeniu jego oceny z egzaminu. Obliczyć prawdopodobieństwo zdarzenia losowego, że podczas losowania czterokrotnie natrafiono na studenta z oceną dostateczną (3.0) lub studenta z oceną dobrą (4.0). „Sukcesem" w tym doświadczeniu jest wylosowanie studenta

[17] Zob. np. J. E. Freund: *Podstawy nowoczesnej statystyki...*, *op. cit.*, s. 185.

z oceną dostateczną lub studenta z oceną dobrą. Ponieważ (zob. przykład 2.1, tablica 2.1) w badanej grupie było 9 studentów z ocenami dostatecznymi i 6 z ocenami dobrymi, prawdopodobieństwo sukcesu w pojedynczym losowaniu wyniesie:

$$p = (9 + 6)/30 = 0{,}5.$$

Obliczmy najpierw szukane prawdopodobieństwo, posługując się schematem Bernoulliego. Niech zmienna $Z \sim B(n = 12,\ p = 0{,}5)$ oznacza liczbę studentów z oceną 3,0 lub 4,0 w niezależnej próbie $n = 12$ osób. Poszukiwane prawdopodobieństwo wyniesie:

$$P(Z = 4; n = 12, p = 0{,}5) = \binom{12}{4} 0{,}5^4 \cdot 0{,}5^8 = 495 \cdot 0{,}0625 \quad 0{,}00390625 \approx 0{,}1208.$$

Ponieważ $np = 6 > 5$ oraz $n(1-p) = 6 > 5$, do przybliżenia wartości szukanego prawdopodobieństwa możemy użyć rozkładu normalnego o parametrach $\mu = np = 6$ i $\sigma = \sqrt{np(1-p)} = \sqrt{12 \cdot 0{,}5 \quad 0{,}5} = \sqrt{3} \approx 1{,}732$. Ponieważ zmienna Z jest zmienną dyskretną (skokową), mogącą przyjmować wartości $z = 0,1,2, \ldots, n$, musimy zastosować tu *poprawkę na ciągłość*, która będzie polegała na zamianie kolejnych wartości z na przedziały od $z - 0{,}5$ do $z + 0{,}5$. I tak na przykład $z = 1$ będzie odpowiadał przedział od 0,5 do 1,5, $z = 2$ to przedział od 1,5 do 2,5 itd.

W miejsce szukanego prawdopodobieństwa zdarzenia losowego $P(Z = 4; n = 12, p = 0{,}5)$, będziemy mieli obecnie $P(3{,}5 < Z < 4{,}5; \mu = 6{,}0, \sigma = 1{,}732)$, przy czym $Z \sim N(6{,}0; 1{,}732)$.

Postępując analogicznie jak w punkcie 1b przykładu 3.7, otrzymamy:

$$P(3{,}5 < Z < 4{,}5) = P\left(\frac{3{,}5 - 6{,}0}{1{,}732} < \frac{Z - 6{,}0}{1{,}732} < \frac{4{,}5 - 6{,}0}{1{,}732}\right) = P(-1{,}44 < U < -0{,}87) =$$

$$= \Phi(-0{,}87) - \Phi(-1{,}44) = 1 - \Phi(0{,}87) - [1 - \Phi(1{,}44)] =$$

$$= 1 - 0{,}8078 - 0{,}9251 = 0{,}1173.$$

Jak łatwo zauważyć, różnica pomiędzy wartością dokładną – obliczoną na podstawie rozkładu dwumianowego – a wartością przybliżoną, wyznaczoną za pomocą rozkładu normalnego, jest bardzo mała i wynosi zaledwie 0,0035 (0,1208 – 0,1173).

Wykorzystując rozkład normalny, możemy również szybko obliczyć np. $P(Z > 4)$, co po uwzględnieniu poprawki na ciągłość zapiszemy $P(Z > 3{,}5)$. Jeżeli tak jak poprzednio założymy, że $Z \sim N(6{,}0; 1{,}732)$, to wówczas szukane prawdopodobieństwo wyniesie:

$$P(Z > 3{,}5) = P\left(Z > \frac{3{,}5 - 6{,}0}{1{,}732}\right) = P(U > -1{,}44) = 1 - P(U < -1{,}44) =$$
$$= 1 - \Phi(-1{,}44) = 1 - [1 - \Phi(1{,}44)] = \Phi(1{,}44) = 0{,}9251.$$

Zauważmy, że gdybyśmy próbowali rozwiązać to zagadnienie, stosując wzór Bernoulliego, to szukane prawdopodobieństwo $P(Z > 4)$ należałoby rozbić na sumę prawdopodobieństw:

$$P(Z > 4) = P(Z = 4) + P(Z = 5) + \ldots + P(Z = 12)$$

lub

$$P(Z \geq 4) = 1 - P(Z < 4) = 1 - F_z(4) = 1 - [P(Z = 0) + (Z = 1) + (Z = 2) + P(Z = 3)].$$

Liczba koniecznych obliczeń byłaby oczywiście niewspółmiernie większa do tych, jakie musimy wykonać, stosując jako przybliżenie rozkład normalny.

3.2.3. Rozkład wykładniczy

Mówimy, że zmienna losowa T ma rozkład wykładniczy, jeżeli funkcja gęstości prawdopodobieństwa ma postać:

$$f(t) = \begin{cases} 0 & dla \quad t < 0 \\ \lambda e^{-\lambda t} & dla \quad t \geq 0 \end{cases}, \tag{3.48}$$

natomiast dystrybuantę opisuje wzór:

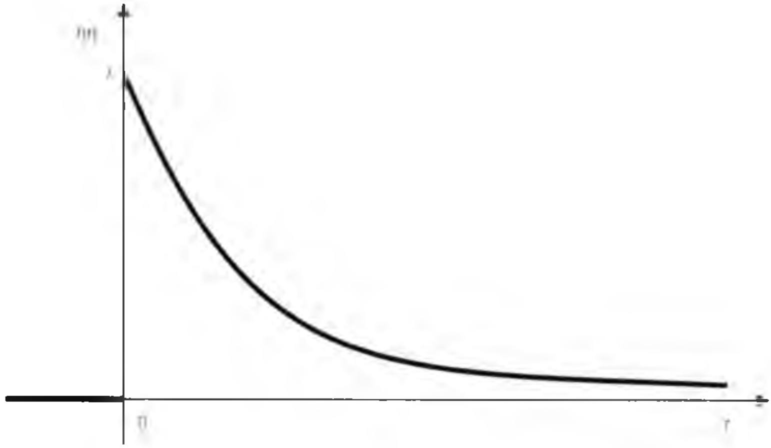
$$F(t) = \begin{cases} 0 & dla \quad t \leq 0 \\ 1 - e^{-\lambda t} & dla \quad t > 0 \end{cases}. \tag{3.49}$$

Wartość oczekiwana i wariancja zmiennej T wynoszą odpowiednio:

$$E(T) = \lambda^{-1}, \tag{3.50}$$

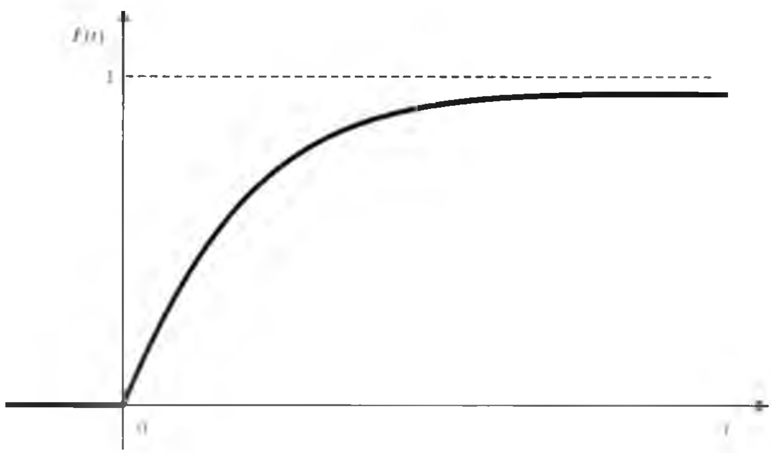
$$D^2(T) = \lambda^{-2}. \tag{3.51}$$

Graficzną ilustrację funkcji gęstości prawdopodobieństwa i dystrybuanty rozkładu wykładniczego zaprezentowano na rysunkach 28 i 29.



Rys. 28. Funkcja gęstości prawdopodobieństwa rozkładu wykładniczego

Źródło: opracowanie własne.



Rys. 29. Funkcja dystrybuanty rozkładu wykładniczego

Źródło: opracowanie własne.

Przykład 3.9

Czas bezawaryjnej pracy telewizorów marki NONAME jest zmienną losową o rozkładzie wykładniczym. Producent deklaruje, że przeciętny czas bezawaryjnej pracy wynosi 150 miesięcy. Obliczyć prawdopodobieństwo następujących zdarzeń losowych:

a) losowo wybrany telewizor będzie działał poprawnie przez co najmniej 5 lat,

b) losowo wybrany telewizor ulegnie awarii przed końcem okresu gwarancyjnego trwającego 2 lata,
c) losowo wybrany telewizor ulegnie awarii pomiędzy czwartym i piątym rokiem użytkowania.

Niech T oznacza zmienną losową oznaczającą czas niezawodnej pracy telewizora.

Wartość oczekiwana zmiennej losowej T: $E(T) = 150$ miesięcy.

Korzystając z własności (3.61), możemy zapisać:

$$E(T) = \lambda^{-1} = \frac{1}{\lambda} = 150 \Rightarrow \lambda = \frac{1}{150}.$$

ad a)

Pięć lat to inaczej 60 miesięcy. Zatem poszukiwane prawdopodobieństwo to:

$$P(T \geq 60) = 1 - P(T < 60) = 1 - F_T(60) = 1 - (1 - e^{-\lambda t}) = e^{-\lambda t} = e^{-\frac{60}{150}} = e^{-0,4} \approx 0,67.$$

Prawdopodobieństwo, że telewizor będzie działał niezawodnie przez co najmniej 5 lat, wynosi ok. 0,67.

ad b)

Prawdopodobieństwo, że telewizor ulegnie awarii w okresie dwóch pierwszych lat (24 miesięcy) użytkowania, wyniesie:

$$P(T < 24) = F_T(24) = 1 - e^{-\lambda t} = 1 - e^{-\frac{24}{150}} = 1 - e^{-0,16} \approx 0,15.$$

Prawdopodobieństwo, że telewizor zepsuje się w okresie gwarancji, wynosi ok. 0,15.

ad c)

Obecnie poszukiwane prawdopodobieństwo to:

$$P(48 < T < 60) = F_T(60) - F_T(40) = 1 - e^{-\frac{60}{150}} - (1 - e^{-\frac{48}{150}}) = e^{-0,32} - e^{-0,4} \approx 0,056.$$

Prawdopodobieństwo awarii telewizora pomiędzy czwartym a piątym rokiem użytkowania wyniesie ok. 0,056. {kp}

3.3. Dwuwymiarowy rozkład normalny

Jednym z ważniejszych rozkładów dwuwymiarowej zmiennej losowej ciągłej jest **rozkład normalny**, którego funkcja gęstości przedstawia się następująco:

$$f(x,y)=\frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}e^{-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_x)^2}{\sigma_x^2}-\frac{2\rho(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y}+\frac{(y-\mu_y)^2}{\sigma_y^2}\right]} \tag{3.52}$$

gdzie:

μ_x, μ_y – wartości oczekiwane zmiennych losowych X i Y,
σ_x, σ_y – odchylenia standardowe badanych zmiennych,
ρ – współczynnik korelacji.

Odpowiednie funkcje brzegowe $f_x(x)$ oraz $f_y(y)$ definiuje się następująco:

$$f_X(x)=\frac{1}{\sigma_x\sqrt{2\pi}}e^{-\frac{(x-\mu_x)^2}{2\sigma_x^2}}, \tag{3.53}$$

$$f_Y(y)=\frac{1}{\sigma_y\sqrt{2\pi}}e^{-\frac{(y-\mu_y)^2}{2\sigma_y^2}}, \tag{3.54}$$

Jeżeli współczynnik korelacji ρ badanych zmiennych losowych X i Y jest równy zero (zmienne są nieskorelowane), to wówczas funkcję gęstości prawdopodobieństwa (3.52) można przedstawić jako:

$$\begin{aligned}f(x,y)=f_X(x)\cdot f_Y(y)&=\frac{1}{\sigma_x\sqrt{2\pi}}e^{-\frac{(x-\mu_x)^2}{2\sigma_x^2}}\cdot\frac{1}{\sigma_y\sqrt{2\pi}}e^{-\frac{(y-\mu_y)^2}{2\sigma_y^2}}=\\&=\frac{1}{2\pi\sigma_x\sigma_y}e^{-\frac{1}{2}\left[\frac{(x-\mu_x)^2}{\sigma_x^2}+\frac{(y-\mu_y)^2}{\sigma_y^2}\right]}\end{aligned}, \tag{3.55}$$

Brak korelacji pomiędzy zmiennymi oznacza w przypadku tego rozkładu, że zmienne losowe X i Y są niezależne.

Jeżeli natomiast współczynnik korelacji $\rho = 1$ lub $\rho = -1$, to wówczas funkcja gęstości prawdopodobieństwa nie jest określona, co oznacza, że pomiędzy zmiennymi losowymi X i Y zachodzi zależność funkcyjna. Otrzymany w ten sposób rozkład nazywa się **osobliwym dwuwymiarowym rozkładem normalnym**[18].

Podobnie jak w przypadku jednowymiarowej zmiennej losowej ciągłej, funkcję (3.52) można poddawać standaryzacji poprzez podstawienie: $\mu_x = \mu_y = 0$ oraz $\sigma_x = \sigma_y = 1$. Dzięki standaryzacji zmniejsza się liczba parametrów decydujących o postaci funkcji gęstości tylko do jednego, którym jest współczynnik korelacji ρ.

Przeanalizujmy na koniec rozkłady warunkowe otrzymane z dwuwymiarowego rozkładu normalnego. Opierając się na wzorach (2.69) i (2.70), otrzymamy:

$$f(x|y)=\frac{f(x,y)}{f_y(y)}=\frac{1}{\sigma_x\sqrt{2\pi(1-\rho^2)}}e^{-\frac{1}{2(1-\rho^2)}\left(\frac{x-\mu_x}{\sigma_x}-\rho\frac{y-\mu_y}{\sigma_y}\right)^2}, \tag{3.56}$$

[18] Zob. np. A. Iwasiewicz, Z. Paszek: *Statystyka... op. cit.*, s. 58.

$$f(y|x)=\frac{f(x,y)}{f_x(x)}=\frac{1}{\sigma_y\sqrt{2\pi(1-\rho^2)}}e^{-\frac{1}{2(1-\rho^2)}\left(\frac{y-\mu_y}{\sigma_y}-\rho\frac{x-\mu_x}{\sigma_x}\right)^2}. \tag{3.57}$$

Otrzymany w ten sposób warunkowy rozkład normalny (3.56), określany jest przez parametry:

$$\mu_{x|y}=\mu_x+\rho\frac{\sigma_x}{\sigma_y}(y-\mu_y), \tag{3.58}$$

oraz

$$\sigma_{y|x}=\sigma_x\sqrt{1-\rho^2}, \tag{3.59}$$

Analogicznie parametrami rozkładu (3.57) są:

$$\mu_{y|x}=\mu_y+\rho\frac{\sigma_y}{\sigma_x}(y-\mu_x), \tag{3.60}$$

oraz

$$\sigma_{x|y}=\sigma_y\sqrt{1-\rho^2}, \tag{3.61}$$

Parametry te to oczywiście warunkowe wartości oczekiwane i warunkowe odchylenia standardowe[19].

3.4. Wybrane twierdzenia graniczne

Twierdzenia graniczne mają ważne znaczenie w szeroko pojętym rachunku prawdopodobieństwa. Dotyczą one problemów omawiających własności graniczne ciągów zmiennych losowych. Twierdzenia graniczne podają warunki, po spełnieniu których dla ciągu zmiennych losowych istnieje asymptotyczny rozkład o określonej postaci. Zwykle wśród twierdzeń granicznych wyróżnia się dwie grupy: twierdzenia lokalne i twierdzenia integralne. Pierwsze z nich (lokalne) dotyczą zbieżności funkcji prawdopodobieństwa lub funkcji gęstości, natomiast drugie (twierdzenia integralne) poruszają problem zbieżności ciągu dystrybuant zmiennych losowych do pewnej dystrybuanty granicznej. Odrębną klasę twierdzeń granicznych stanowią tzw. prawa wielkich liczb, które dotyczą zbieżności ciągu zmiennych losowych do rozkładu jednopunktowego.

Niektóre z twierdzeń granicznych wprowadziliśmy już podczas omawiania wybranych rozkładów zmiennych losowych dyskretnych i funkcji gęstości zmiennych

[19] Czytelnika zainteresowanego praktycznym zastosowaniem ciągłych rozkładów dwuwymiarowych odsyłamy np. do podręcznika W. Krysicki i inni: *Rachunek prawdopodobieństwa i statystyka matematyczna w zadaniach*, cz. I., PWN, Warszawa 1986, rozdział 5 lub do T. Gerstenkorn, T. Środka: *Kombinatoryka i rachunek prawdopodobieństwa*, *op. cit.*, rozdział 7.

losowych ciągłych. Na przykład w rozdziale 3.1.3 pisaliśmy o zbieżności rozkładu dwumianowego do rozkładu Poissona, w rozdziale 3.1.4 o aproksymacji rozkładu hipergeometrycznego za pomocą rozkładów dwumianowego i Poissona, natomiast w rozdziale 3.2.2 o możliwości przybliżenia rozkładu dwumianowego za pomocą rozkładu normalnego. Podczas omawiania częstościowej definicji prawdopodobieństwa wspomnieliśmy także o prawie wielkich liczb Bernoulliego. Obecnie usystematyzujemy te pojęcia i w kolejnych trzech podrozdziałach omówimy trzy ważne twierdzenia graniczne: prawo wielkich liczb Bernoulliego, nierówność Czybyszewa i twierdzenie Moivre'a-Láplace'a.

3.4.1. Prawo wielkich liczb Bernoulliego

Załóżmy, że rozważamy zmienną losową W, zdefiniowaną jako:

$$W_n = \frac{X_1}{n} + \frac{X_2}{n} + \ldots + \frac{X_n}{n} = \frac{Z}{n}, \tag{3.62}$$

gdzie:

$X_1, \ldots, X_n$, jest ciągiem niezależnych zmiennych zero-jedynkowych, o identycznych wartościach oczekiwanych wynoszących p.

Zmienna W_n przyjmuje następujący ciąg wartości: $\frac{1}{n}$, $\frac{2}{n}$, ..., 1 i ma dwumianowy rozkład prawdopodobieństwa postaci (3.7), o parametrach n i p ($W_n \sim B(n, p)$).

Wartość oczekiwana oraz wariancja zmiennej losowej W_n są równe:

$$E(W_n) = \frac{E(Z)}{n} = \frac{np}{n} = p, \tag{3.63}$$

$$D^2(W_n) = \frac{D^2(Z)}{n^2} = \frac{npq}{n^2} = \frac{pq}{n} \tag{3.64}$$

Prawo wielkich liczb Bernoulliego głosi, że ciąg zmiennych losowych W_n jest zbieżny stochastycznie (według prawdopodobieństwa) do wartości p. Oznacza to, że:

$$\lim_{n\to\infty} P\left(|W_n - p| < \varepsilon\right) = 1, \tag{3.65}$$

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

Własność (3.65) oznacza, że wraz ze wzrostem liczebności próby n do nieskończoności, prawdopodobieństwo zdarzenia losowego, będzie dążyć do jedności.

Przykład 3.10

Załóżmy, że rozpatrujemy grupę 30 studentów opisaną w przykładzie 2.1, badaną ze względu na ocenę otrzymaną z egzaminu z „Zarządzania jakością". Załóżmy, że „sukcesem" jest z-krotne wylosowanie (w n-krotnym losowaniu ze zwracaniem) studenta z oceną dostateczną lub dobrą. Niech zmienna W_n oznacza częstość (frakcję) „sukcesów" uzyskanych w n-krotnym losowaniu ze zwracaniem (losowaniu niezależnym).

Zbadajmy, czy dla n równego kolejno: 5, 10 i 20 prawdziwa jest własność (3.65). Podczas analizy przyjmiemy, że $\varepsilon = 0{,}1$.

Ponieważ w badanej grupie było 9 studentów z oceną dostateczną i 6 z oceną dobrą, zatem prawdopodobieństwo „sukcesu" w pojedynczej próbie wyniesie $p = 15/30 = 0{,}5$.

Rozważymy następujące zmienne losowe:

$$W_{n=5} \sim B(n = 5, p = 0{,}5),\ W_{n=10} \sim B(n = 10, p = 0{,}5),\ W_{n=20} \sim B(n = 20, p = 0{,}5).$$

Ponieważ $\varepsilon = 0{,}1$, otoczeniem punktu $p = 0{,}5$ będzie przedział $[0{,}4; 0{,}6]$.

Jeżeli $n = 5$, to wówczas zmienna W_5 ma następujący zbiór wartości: $\{0; 0{,}2; 0{,}4; 0{,}6; 0{,}8; 1\}$.

Do otoczenia p należą jedynie wartości 0,4 i 0,6, w przypadku, których liczba „sukcesów" $z = 2$ lub $z = 3$. Mamy zatem:

$$\begin{aligned} P(|W_5 - 0{,}5| < 0{,}1) &= P(W_5 = 0{,}4) + P(W_5 = 0{,}6) = \\ &= \binom{5}{2} 0{,}5^2 \cdot 0{,}5^3 + \binom{5}{3} 0{,}5^3 \cdot 0{,}5^2 = \\ &= 2 \cdot (10 \cdot 0{,}25 \cdot 0{,}125) = 0{,}625. \end{aligned}$$

Dla $n = 10$, W_{10}: $\{0; 0{,}1; 0{,}2; 0{,}3; 0{,}4; 0{,}5; 0{,}6; 0{,}7; 0{,}8; 1\}$.

Do otoczenia p należą obecnie trzy punkty: 0,4; 0,5 i 0,6, które powstały w wyniku podzielenia trzech wartości zmiennej Z równych odpowiednio: 4, 5 i 6. Szukane prawdopodobieństwo będzie teraz równe:

$$\begin{aligned} P(|W_{10} - 0{,}5| < 0{,}1) &= P(W_{10} = 0{,}4) + P(W_{10} = 0{,}5) + P(W_{10} = 0{,}6) = \\ &= \binom{10}{4} \cdot 0{,}5^4 \cdot 0{,}5^6 + \binom{10}{5} \cdot 0{,}5^5 \cdot 0{,}5^5 \binom{10}{6} \cdot 0{,}5^6 \cdot 0{,}5^4 = \\ &= 210 \cdot 0{,}0625 \quad 0{,}015625 + 252 \cdot 0{,}03125 \cdot 0{,}03125 + \\ &+ 210 \cdot 0{,}015625 \cdot 0{,}0625 = 0{,}205 + 0{,}246 + 0{,}205 = 0{,}656. \end{aligned}$$

Jeżeli $n = 20$, to W_{20}: $\{0; 0{,}05; 0{,}1; 0{,}15; 0{,}2; 0{,}25; 0{,}3; 0{,}35; 0{,}4; 0{,}45; 0{,}5; 0{,}55; 0{,}6; 0{,}65; 0{,}7; 0{,}75; 0{,}8; 0{,}85; 0{,}9; 0{,}95; 1\}$.

Do otocznia punktu p należą obecnie wartości: 0,4; 0,45; 0,5; 0,55; 0,6, dla których liczba „sukcesów" z jest równa odpowiednio: 8, 9, 10, 11, 12. Otrzymamy następujące prawdopodobieństwo:

$$P(|W_{20}-0{,}5|<0{,}1) = P(W_{20}=0{,}4) + P(W_{20}=0{,}45) + P(W_{20}=0{,}5) +$$

$$+ P(W_{20}=0{,}55) + P(W_{20}=0{,}6) = \binom{20}{8}\cdot 0{,}5^{8}\cdot 0{,}5^{12} + \binom{20}{9}\, 0{,}5^{9}\ 0{,}5^{11} +$$

$$+ \binom{20}{10}\cdot 0{,}5^{10}\cdot 0{,}5^{10} + \binom{20}{11}\cdot 0{,}5^{11}\cdot 0{,}5^{9} + \binom{20}{12}\cdot 0{,}5^{12}\cdot 0{,}5^{8} =$$

$$= 0{,}1201 + 0{,}1602 + 0{,}1762 + 0{,}1602 + 0{,}1201 + 0{,}1201 = 0{,}7368.$$

Jak widać, im liczebność próby n jest większa, tym prawdopodobieństwo $P(|W_n - p| < \varepsilon)$ jest bliższe jedności. Można się również spodziewać, że dla n równego 30, 40, 50, 100 itd. otrzymane różnice $1 - P(|W_n - p| < \varepsilon)$ będą zmierzać do zera. {kp}

3.4.2. Nierówność Czybyszewa

Nierówność Czybyszewa[20] pozwala oszacować prawdopodobieństwa, przy założeniu, że znana jest wartość oczekiwana badanej zmiennej losowej oraz że zmienna ta przyjmuje wartości dodatnie. Nie jest wymagana natomiast znajomość rozkładu tej zmiennej.

W literaturze[21] znanych jest kilka zapisów postaci nierówności Czybyszewa. Najprostsza z nich to taka, że:

$$P(X < k) \geq 1-\frac{\mu}{k}, \tag{3.66}$$

gdzie:

X – jest zmienną losową przyjmującą wartości nieujemne,
μ – skończona wartość oczekiwana zmiennej losowej X,
k – dowolna stała dodatnia.

Jeżeli obok wartości oczekiwanej μ znana jest również wartość odchylenia standardowego σ, to wówczas wygodna w wielu zagadnieniach jest następująca postać nierówności Czybyszewa:

$$P(|X-\mu| > k\sigma) < \frac{1}{k^2}, \tag{3.67}$$

lub równoważna:

$$P(|X-\mu| > k\sigma) \geq 1-\frac{1}{k^2}. \tag{3.68}$$

---

[20] Pafnutij Czybyszew (1821–1894) – rosyjski matematyk, twórca petersburskiej szkoły matematycznej, profesor uniwersytetu w Petersburgu.

[21] Zob. np. T. Gerstenkorn, T. Śródka: *Kombinatoryka...*, *op. cit.*, s. 388.

Nierówność (3.68), można zapisać również:

$$P(\mu - k\sigma < X < \mu + k\sigma) > 1 - \frac{1}{k^2}, \tag{3.69}$$

Interpretacja powyższego zapisu świadczy, że jeżeli dany jest dowolny rozkład ze średnią μ i odchyleniem standardowym σ, to prawdopodobieństwo otrzymania wartości w granicach k odchyleń standardowych od średniej jest większe bądź równe $1 - \frac{1}{k^2}$.

Przykład 3.11

Niech zmienna losowa X oznacza liczbę punktów uzyskanych podczas egzaminu z „Zarządzania jakością" przez losowo wybranego studenta, przy czym (tak jak w przykładzie 3.7) $X \sim N(29{,}5;\ 6{,}4)$.

Z jakim prawdopodobieństwem można twierdzić, że zmienna losowa X odchyli się bezwzględnie od swej wartości oczekiwanej o mniej niż jedno, dwa i trzy odchylenia standardowe.

Przyjmując w nierówności (3.68), kolejno za k wartości 1, 2 i 3, otrzymamy:

$$P(|X - 29{,}5| < \sigma) \geq 1 - \frac{1}{1^2} = 0,$$

$$P(|X - 29{,}5| < 2\sigma) > 1 - \frac{1}{2^2} = 0{,}75,$$

$$P(|X - 29{,}5| < 3\sigma) \geq 1 - \frac{1}{3^2} \approx 0{,}89.$$

Uzyskane oszacowania wartości prawdopodobieństwa, na podstawie nierówności Czybyszewa są znacznie mniej dokładne, niż te same prawdopodobieństwa uzyskane opisaną w podrozdziale 3.2.2 metodą trzech sigm (porównaj z (3.35), (3.36) i (3.37)).

Pamiętać jednak należy, że nierówność Czybyszewa jest uniwersalna dla wszystkich typów rozkładów, podczas gdy reguła trzech sigm odnosi się jedynie do rozkładu normalnego. Dlatego też można ją stosować we wszystkich przypadkach, gdy nie jesteśmy pewni, jakim typem rozkładu charakteryzuje się badana zmienna losowa. {kp}

3.4.3. Twierdzenie Moivre'a-Laplace'a

Twierdzenie to już częściowo zostało omówione w podrozdziale 3.2.2. Obecnie usystematyzujemy jedynie pewne związane z nim pojęcia. Załóżmy, że rozpatrujemy ciąg zmiennych Z_n postaci (3.7), (Z_n jest sumą rozważanych n zmiennych zero-jedynkowych: $Z_n = X_1 + ... + X_n$) o rozkładzie dwumianowym o parametrach n i p.

Zmienne te poddajemy standaryzacji, transformując je w ciąg zmiennych U_n postaci:

$$U_n = \frac{Z_n - np}{\sqrt{npq}}. \tag{3.70}$$

Twierdzenie Moivre'a-Laplace'a głosi, że ciąg dystrybuant $F_n(u)$ tak określonych zmiennych losowych U_n, w miarę wzrostu n do nieskończoności dąży do dystrybuanty rozkładu normalnego z parametrowymi $\mu = 0$ i $\sigma = 1$. Można więc zapisać:

$$\lim_{n \to \infty} F_n(u) = \Phi(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{u} \exp\left(\frac{-u^2}{2}\right) du. \tag{3.71}$$

Powyższe twierdzenie nazywa się **integralnym twierdzeniem granicznym Moivre'a-Laplace'a**.

Z powyższego twierdzenia można wyciągnąć następujące wnioski:

– ciąg zmiennych losowych X_n o rozkładzie dwumianowym o parametrach n i p dąży do rozkładu normalnego o parametrach $\mu = np$ i $\sigma = \sqrt{npq}$,

– ciąg zmiennych losowych $W_n = \frac{X_1}{n} + \frac{X_2}{n} + \ldots + \frac{X_n}{n} = \frac{Z_n}{n}$ ma asymptotyczny rozkład normalny z parametrami: $\mu = p$ i $\sigma = \sqrt{\frac{pq}{n}}$.

Zalecamy czytelnikom ponowne przestudiowanie przykładu 3.8, w którym pokazaliśmy jedną z możliwości zastosowania twierdzenia granicznego Moivre'a-Laplace'a.

Przedstawione w podrozdziale 3.4 twierdzenia graniczne nie wyczerpują całkowicie tego tematu[22].

[22] Czytelnikowi zainteresowanemu tą ważną tematyką zalecamy przestudiowanie np. rozdziału 9 wspominanej wielokrotnie pracy: T. Gerstenkorn, T. Środka: *Kombinatoryka i rachunek prawdopodobieństwa*, *op. cit*.

Część II

WNIOSKOWANIE STATYSTYCZNE

Wnioskowanie statystyczne – podstawowe pojęcia

Rozdział 4

Wnioskowanie statystyczne nazywane też popularnie statystyką matematyczną uznaje się za dział statystyki ściśle związany z rachunkiem prawdopodobieństwa. Rachunek prawdopodobieństwa od statystyki matematycznej odróżnia przede wszystkim to, że w rachunku prawdopodobieństwa podczas wyznaczania prawdopodobieństwa zajścia zdarzeń losowych zakłada się zwykle pełną znajomość rozkładu badanej zmiennej losowej. W obrębie statystyki matematycznej wyróżnia się dwie grupy zagadnień: **teorię estymacji** (szacowania) oraz **teorię weryfikacji hipotez statystycznych**.

Punktem wyjścia wnioskowania statystycznego jest założenie, że populacja (nazywana częściej w tym dziale statystyki zbiorowością generalną) poddawana jest badaniom niewyczerpującym (częściowym). Jest to istotna cecha odróżniająca ten dział statystyki od statystyki opisowej[1], w której zakłada się najczęściej, że badania populacji mają charakter wyczerpujący (stuprocentowy). Badania wyczerpujące populacji w wielu przypadkach są niecelowe czy niewykonalne. Przykładem mogą być tutaj badania o charakterze niszczącym lub badania populacji składających się z nieskończonej liczby elementów. W takich sytuacjach, źródłem informacji opisujących zbiorowość generalną jest pochodząca z niej **próba** (próbka). Próbę definiuje się jako skończony podzbiór populacji, podlegający bezpośredniemu badaniu ze względu na interesujące nas właściwości populacji. Liczba jednostek, elementów zbiorowości statystycznej wybranej do próby nazywana jest **liczebnością próby** i oznaczana symbolem n. Zaobserwowane wartości badanej cechy u tych elementów zbiorowości generalnej, które weszły w skład próby, nazywane są **wynikami próby**. Zbiór wszystkich możliwych wyników próby o liczności n nazywa się **przestrzenią próby**. W praktyce operuje się najczęściej tzw. **próbą losową**, której dobór elementów odbywa się w drodze losowania (np. za pomocą tzw. tablic liczb losowych), a nie na drodze celowego doboru. Wyniki próby losowej o liczności n stanowią wartość n – wymiarowej zmiennej losowej lub n – wymiarowego wektora losowego.

[1] Zob. M. Major, J. Niezgoda: *Elementy statystyki, Cz. I. Statystyka opisowa*, Krakowskie Towarzystwo Edukacyjne sp. z o.o., Kraków 2003.

Podczas doboru próby dba się również o to, aby jej struktura pod względem badanej cechy przypominała strukturę populacji, z której próba pochodzi (próba ma być „miniaturą" populacji). Próba o takich własnościach nazywana jest **próbą reprezentatywną**.

O tym, czy próba jest reprezentatywna czy też nie, decydują zwykle techniki i schematy losowania próby.

4.1. Techniki doboru próby

Próby mogą być dobierane w sposób celowy, losowy lub też celowo-losowy. Próbki celowe dobierane są w sytuacji, gdy badane cechy są skorelowane z określonymi cechami kwalifikującymi. W takiej sytuacji wnioskowanie odbywa się na podstawie wartości ekstremalnych i nazywane jest wnioskowaniem ekstremalnym[2].

Przykład 4.1

Klasycznym przykładem losowania celowego może być np. losowanie mające ustalić, czy w pewnej zbiorowości zwierząt danego gatunku nie występuje choroba zakaźna. Do próby celowej zostaną zaliczone te zwierzęta, które swoim zachowaniem (np. są osowiałe, nie przyjmują pokarmu, mają zwiększone łaknienie itp.) wyraźnie odbiegają od zachowania pozostałych zwierząt badanego stada. Jeżeli po dokładnym przebadaniu tych osobników (np. za pomocą odpowiednich badań morfologicznych), okaże się, że nie są one chore, to można przypuszczać z dużym prawdopodobieństwem, że całe stado jest również zdrowe. {kp}

Przykład 4.2

Załóżmy, że konstruktor samochodów jest zainteresowany opinią użytkowników na temat rozmieszczenia określonych przycisków sterujących w samochodzie. Jest rzeczą oczywistą, że najwięcej ciekawych opinii może on uzyskać od kierowców, którzy często prowadzą samochód. W powyższej sytuacji dobór próby będzie dwustopniowy. W pierwszym kroku z całej zbiorowości należy **celowo** wybrać tylko osoby spełniające założone kryterium (tzn. np. mają prawo jazdy i przynajmniej 1 raz w tygodniu prowadzą samochód). W kroku drugim z wyodrębnionej grupy osób należy **wylosować** n – osób i wśród nich przeprowadzić badania (np. badania ankietowe). Opisany schemat doboru próby ma charakter celowo-losowy. {kp}

---

[2] Zob. A. Iwasiewicz, Z. Paszek: *Statystyka...*, *op. cit.*, s. 76.

Przykład 4.3

Nauczyciel chce poznać strukturę ocen swoich studentów, lecz brakuje mu czasu, aby przebadać wyniki wszystkich studentów. Postanowił więc, że pobierze próbę o charakterze losowym. Aby zagwarantować losowość próby, wprowadził następujące zasady:

- każdy student należący do badanej populacji ma w sposób jednoznaczny przyporządkowany numer (np. numer indeksu na liście obecności),
- numery zostały naniesione na oddzielne karki (losy), które trafiły do specjalnej urny.

Podczas losowania kartek z numerami studentów nauczyciel stosuje jedną z dwóch technik:

- technika pierwsza: po wyciągnięciu kartki identyfikuje studenta i jego ocenę (oceny), a następnie wrzuca kartkę na powrót do urny,
- technika druga: po wyciągnięciu kartki identyfikuje studenta i jego ocenę (oceny), a następnie odkłada kartkę na bok i nie wrzuca jej do urny.

Proces losowania jest wykonywany n razy. {kp}

Opisany w przykładzie 4.3 sposób doboru próby ma charakter losowy. O tym, czy dany student stanie się obiektem badań, decyduje przypadek, a nie subiektywny wybór badacza. Jeżeli pobrana w opisany sposób próba jest dostatecznie duża, to można przypuszczać, że będzie ona reprezentatywna. Reprezentatywność pobranej próby oznacza, że powinna być ona miniaturą całej zbiorowości. Jednostki, które się w niej znajdą, muszą być reprezentatywne dla całej populacji. Struktura próby i struktura populacji powinny być niemal identyczne.

Opisane dwie techniki losowania poznaliśmy już w trakcie omawiania rozkładów zmiennych losowych. Są to oczywiście:

- **losowanie ze zwracaniem** (losowanie niezależne),
- **losowanie bez zwracania** (losowanie zależne).

Podczas pierwszej techniki losowania, w kolejnych aktach losowania, każda jednostka zbiorowosci generalnej posiada jednakową szansę trafienia do próby. Wynik kolejnego losowania nie zależy od historii wcześniejszych losowań. Odmiennie jest w przypadku losowania bez zwracania. Tutaj wynik kolejnych losowań jest ściśle powiązany z wynikami wcześniejszych losowań. Różnice pomiędzy wynikami otrzymanymi tymi technikami zacierają się wraz ze wzrostem liczebności zbiorowości generalnej i stosunku liczebności próby (n) do liczebności populacji (N). Do oceny tego zjawiska można wykorzystać następujący wskaźnik:

$$\varepsilon = \frac{N-n}{N-1} \tag{4.1}$$

Jeżeli wartość ε jest bliska jedności, to wówczas różnica pomiędzy wynikami otrzymanymi za pomocą opisanych technik losowania przestaje być istotna.

Opisany w przykładzie 4.3 nauczyciel, wybierając jednostki do próby, miał dostęp do wszystkich studentów należących do badanej populacji. Jego losowanie można nazwać **losowaniem indywidualnym nieograniczonym**. Jeżeli zdecyduje się on na technikę losowania ze zwracaniem (niezależnego), to uzyskaną próbę można określić mianem **próby losowej prostej**. W pewnych przypadkach losowanie indywidualne nieograniczone nie jest możliwe, gdyż jednostki zbiorowości są ściśle pogrupowane w pewne zbiorowości, a ich podział jest nieuzasadniony lub niemożliwy. W takiej sytuacji przeprowadza się tzw. **losowanie zespołowe**, podczas którego wybiera się nie indywidualne jednostki, lecz ich zbiorowości. Próbę tworzą wszystkie jednostki badania wchodzące w skład wylosowanych zespołów. Jednostka badania nie pokrywa się z jednostką losowania. Na przykład podczas badania zjawiska, jakim jest przyszła praca zawodowa studentów, zespołem (jednostką badania) może być grupa dziekańska, a jednostką badania student. Innym przykładem, w którym można zastosować losowanie zespołowe, może być kwestia badania poprawności naliczania podatku dochodowego w urzędzie skarbowym, gdzie zespołem może być grupa nazwisk zaczynająca się od liter A, B, C itd., natomiast jednostką badania są zeznania podatkowe zaliczone do wylosowanego zespołu. Sposób tworzenia zespołów musi pozostawać w zgodzie z merytoryczną koncepcją badań. Jeżeli w trakcie kolejnych losowań zespołowych maleje liczebność pobieranych zbiorów, to takie losowanie zwykło określać się mianem losowania wielostopniowego. W przypadku zbiorowości, z wyraźnie zaznaczonym podziałem na warstwy (podzbiory) zalecane jest tzw. **losowanie warstwowe**. Podział populacji na warstwy (nazywany też stratyfikacją) ma za zadanie wyodrębnienie spośród bardzo zróżnicowanej zbiorowości generalnej, względnie jednorodnych grup, z których w kolejnych krokach wybiera się bezzwrotnie elementy i tworzy próbę wspólną dla wszystkich warstw[3].

Dość powszechnym sposobem doboru próby jest także **dobór losowy z interwałem**, nazywany też losowaniem sekwencyjnym. Jest to metoda oparta na zasadzie, że do próby trafia co k – ta jednostka zbiorowości, przy czym pierwsza jednostka – od której rozpoczyna się pobieranie – jest wybierana losowo. Symbol k oznacza tutaj odstęp (interwał losowania) i jest on wyznaczany z relacji: $k = N/n$, gdzie: N oznacza liczebność populacji, natomiast n – jest liczebnością próby. Wartość k musi być liczbą całkowitą i dlatego jest ona zaokrąglana w dół do najbliższej liczby całkowitej. Wybór pierwszej jednostki determinuje wybór pozostałych. Z tego względu ten typ losowania jest losowaniem ograniczonym.

Spotyka się tu dwa algorytmy wyboru punktu startu losowania. W pierwszym przypadku punkt ten przypada na k pierwszych jednostek zbiorowości generalnej (tzn. leży w pierwszej strefie pomiędzy 0 a k). Kolejne elementy pobierane do próby

[3] Technika losowania warstwowego oraz inne techniki losowań zostały opisane szczegółowo np. w: J. Steczkowski: *Metoda reprezentacyjna w badaniach zjawisk ekonomiczno-społecznych*, Warszawa–Kraków 1995.

są oddalone na prawo od punktu startu o kolejne wielokrotności k (k, $2k$, $3k$, itd.). W drugim przypadku punkt startu może znajdować się w dowolnej strefie zbiorowości generalnej, a jednostki dobierane do próby są oddalone na prawo i na lewo od punktu startu o kolejne wielokrotności k. Niekiedy dla poprawy reprezentatywności próbki wybiera się kilka różnych punktów startu.

W praktyce pobieranie próby losowej nie zawsze odbywa się – tak jak opisano w przykładzie 4.3 – przy użyciu urny. Dość często w procesie losowania wykorzystuje się tzw. tablice liczb losowych czy też elektroniczne generatory liczb losowych, w które wyposażona jest większość arkuszy kalkulacyjnych typu Excel czy Works. W następnym z podrozdziałów zaprezentujemy podstawowe techniki wykorzystania tablic liczb losowych w trakcie wyboru próby losowej.

4.2. Tablice liczb losowych

Dość często tablice liczb losowych pogrupowane są w ponumerowane bloki składające się np. z 25 wierszy i 10 kolumn czterocyfrowych liczb[4]. Tablice liczb losowych zostały zamieszczone w osobnym dodatku na końcu podręcznika (zob. tablice Va–Vj). Wcześniej każdej jednostce zaliczanej do badanej zbiorowości przyporządkowuje się w sposób jednoznaczny symbol literowy lub częściej liczbę całkowitą z przedziału $[1; N]$.

Technikę takiego pobierania zilustrujemy kilkoma przykładami opisującymi różne przypadki.

Przypadek I

Najwyższy numer w populacji jest nie większy od 99.

Przykład 4.4

W skrzyni znajduje się zapakowanych pojedynczo 85 nieponumerowanych opakowań szklanych. Celem sprawdzenia poprawności zapakowania należy pobrać w sposób losowy próbę o liczności 7 elementów.

Etapy doboru próby:

1. W pierwszym kroku należy ponumerować poszczególne sztuki. W tym celu nanosimy kolejne numery na poszczególne opakowania (opisujemy, przyklejamy etykiety z numerami, lub umawiamy się, że znajdują się one w odpowiedniej kolejności). Numerowanie wykonujemy liczbami o stałej liczbie cyfr. W tym przykładzie 01; 02; 03; ...; 83; 84; 85.
2. Ustalamy początek odczytywania. W tym celu wybieramy w sposób dowolny blok, wiersz i kolumnę tablic liczb losowych.

---

[4] Zob. np. PN N-03010; Statystyczna kontrola jakości. Losowy wybór sztuk do próbek; PKN 1951.

Np: blok 5
wiersz 6
kolumna 3

3. Odczytujemy kolejne liczby losowe zapisując tylko dwie pierwsze cyfry.

08 32 77 34 31 75 (89) (88) 46

Liczby w nawiasach są większe od 85 i nie mogą być wykorzystane. Po ich pominięciu pozostają:

08 32 77 34 31 75 46

Jeżeli któryś z numerów zostałby wylosowany dwukrotnie, jednostka produktu badana jest jednokrotnie, natomiast wynik badania uwzględniany jest dwa razy. Ważne jest to, że jednostka taka nie zostaje wyeliminowana z badania ani też nie zostaje dolosowana jednostka dodatkowa. {kp}

Przypadek II
Najwyższy numer w populacji jest nie większy od 999.

Przykład 4.5

Wylosować 10 elementową próbę z transportu zegarków o końcówkach kolejnych numerów seryjnych od 000 do 299 oraz od 400 do 650.

1. Jako kolejne numery wykorzystujemy końcówki numerów seryjnych od 000 do 299, oraz od 400 do 650
2. Ustalamy początek odczytywania.
   Np: blok 1
   wiersz 3
   kolumna 2
3. Odczytujemy kolejne liczby losowe, uwzględniając pierwsze trzy cyfry:

   511 (683) (896) (355) (753) (462) 110 (797) 512 (479) (728) 449 150
   424 035 (859) 096 555 281

Liczby w nawiasach nie należą do przedziałów określonych przez końcówki numerów seryjnych i nie mogą być wykorzystane. Po ich pominięciu pozostają:

511 110 512 449 150 424 035 096 555 281. {kp}

Przypadek III
Najwyższy numer w populacji jest nie większy od 9 999.

Przykład 4.6

Należy wylosować 12 sztuk z partii produktów o liczności 8533 sztuk ponumerowanych kolejno.

1. Ustalamy początek odczytywania pierwszego.
   Np: blok 3
   wiersz 3
   kolumna 6
2. Odczytujemy kolejne liczby losowe, uwzględniając pierwsze dwie cyfry.

   30 02 48 33 07 19 28 92 03 61 51 93 14
3. Ustalamy początek odczytywania drugiego.
   Np: blok 4
   wiersz 10
   kolumna 4
4. Odczytujemy kolejne liczby losowe, uwzględniając pierwsze dwie cyfry.

   81 51 56 14 44 19 28 84 81 20 92 54 42
5. Składamy każdy numer czterocyfrowy z dwu grup po dwie cyfry.

   3081 0251 4856 3314 0744 1919 2828 9284 0381 6120 5192 (9354) 1442

Liczba w nawiasach jest większa od liczebności badanej partii i nie może być wykorzystana. Po jej pominięciu pozostają:

3081 0251 4856 3314 0744 1919 2828 9284 0381 6120 5192 1442

Przypadek IV
Najwyższy numer w populacji jest nie większy od 99 999.

Przykład 4.7

Należy wylosować 11 sztuk z partii produktów o liczności 84545 sztuk ponumerowanych kolejno.
1. Ustalamy początek odczytywania pierwszego.
   Np: blok 1
   wiersz 16
   kolumna 1
2. Odczytujemy kolejne liczby losowe, uwzględniając pierwsze dwie cyfry.

   62 41 57 25 28 34 30 35 23 53

3. Ustalamy początek odczytywania drugiego.
   Np: blok 10
   wiersz 6
   kolumna 5
4. Odczytujemy kolejne liczby losowe, uwzględniając pierwsze trzy cyfry.

   706 821 365 719 427 628 255 098 212 529 096

5. Składamy każdy numer pięciocyfrowy z dwu grup po dwie i trzy cyfry i ewentualnie eliminujemy liczby większe niż liczebność populacji.

   62706 41821 57365 14719 25427 28628 34255 30098 35212 23529 53096

Przypadek V
Najwyższy numer w populacji jest nie większy od 999 999.

Przykład 4.8

Należy wylosować 10 sztuk z partii produktów o liczności 855000 sztuk ponumerowanych kolejno.

1. Ustalamy początek odczytywania pierwszego.
   Np: blok 10
   wiersz 5
   kolumna 8
2. Odczytujemy kolejne liczby losowe, uwzględniając pierwsze trzy cyfry.

   710 304 524 628 269 840 701 706 821 365

3. Ustalamy początek odczytywania drugiego.
   Np: blok 1
   wiersz 5
   kolumna 5
4. Odczytujemy kolejne liczby losowe, uwzględniając pierwsze trzy cyfry.

   605 793 573 246 225 629 187 518 259 550

5. Składamy każdy numer sześciocyfrowy z dwu kolejnych odczytów i ewentualnie eliminujemy liczby większe niż liczebność populacji:

   710605 304793 524573 628246 269225 840629 701187 706518 821259 365550

Przypadek VI

Losowanie bez użycia tablic liczb losowych „losowanie bezpośrednie”

W pewnych przypadkach, np. uniemożliwiających ponumerowanie sztuk należących do badanej populacji, dopuszcza się także tzw. losowanie bezpośrednie. Ten typ losowania może mieć zastosowanie np. w odbiorczej kontroli jakości produktów, podczas którego poszczególne jednostki należy pobrać do badania niezależnie od tego, czy pobierający przypuszcza lub przewiduje, że dane jednostki okażą się zgodne lub niezgodne z wymaganiami jakościowymi. Jeżeli istnieje możliwość, że różne warstwy produktu wewnątrz opakowania, stosu lub sterty cechują się różnym poziomem jakości, to należy do próbki pobierać jednostki ze środka z wierzchu i ze spodu każdego wylosowanego opakowania, stosu lub sterty. Metody tej nie należy natomiast stosować w sytuacji, gdy wady poszczególnych jednostek należących do badanej partii są na tyle wyraźne, że mogą być rozpoznane bez zaawansowanych metod badawczych.

Losowanie próby może przebiegać również wielostopniowo. Z tego typu losowaniem mamy do czynienia na przykład, gdy jednostki produktu pakowane są w opakowania jednostkowe, te w większe itd. Wówczas w pierwszym losowaniu należy pobrać największe opakowanie, następnie coraz mniejsze, a na końcu sztuki z najmniejszych opakowań zbiorowych. Niektóre z tych losowań (zwłaszcza ostatnie) mogą być przeprowadzane „bezpośrednio”.

Przykład 4.9

Partia spinaczy biurowych jest dostarczana w ten sposób, że pakuje się je najpierw w małe pudełka (zawierające średnio 100 sztuk), te w większe pudła (zawierające 100 pudelek), te następnie w skrzynie (zawierające 5000 pudeł)

W pierwszym kroku losuje się skrzynie, po ich otwarciu losuje się pudła, następnie pudełka i na końcu pobiera „bezpośrednio” kilka czy kilkanaście spinaczy. {kp}

Rozdział

5

Wybrane zagadnienia z teorii estymacji

5.1. Uwagi wstępne

Załóżmy, że rozpatrujemy n elementową próbę prostą, którą badamy ze względu na zmienną losową X. Uzyskane w trakcie badania wyniki utworzą skończony zbiór $x_1, x_2, ..., x_n$.

Pobierając dwie lub więcej n – elementowe próby losowe, otrzymuje się z reguły tyleż samo różnych ciągów wynikow opisujących badaną zmienną losową X. Tak jak zaznaczono to we wstępie do rozdziału 4, ciągi otrzymanych wyników $x_1, ..., x_n$ traktować będziemy jako realizację ciągu n zmiennych losowych $X_1, ..., X_n$ i nazywać próbą losową lub krótko próbką.

Uzyskane tą drogą wyniki mogą posłużyć do wyciągania wniosków o badanej zbiorowości generalnej. W przypadku estymacji wnioskowanie ma służyć szacowaniu nieznanych wartości parametrów lub ich funkcji, które charakteryzują rozkład badanej cechy populacji. Jeżeli szacunek dotyczy nieznanego parametru populacji, to wówczas mówimy o **estymacji parametrycznej**, jeżeli, natomiast efektem szacunku ma być postać funkcji opisującej rozkład badanej zmiennej opisującej zbiorowość generalną, to wówczas takie postępowanie określimy **estymacją nieparametryczną**. W dalszej części podręcznika zajmiemy się estymacją parametryczną. Jeżeli efektem szacunku nieznanego parametru populacji jest jedna liczba, czyli punkt na osi liczb rzeczywistych, to estymację nazywamy **punktową**. Jeżeli natomiast efektem szacunku jest przedział na osi liczb rzeczywistych (tzw. **przedział ufności**), który z określonym prawdopodobieństwem pokrywa nieznaną wartość parametru, to wówczas estymację nazywamy **przedziałową**.

5.2. Estymatory i ich własności

Podstawowymi pojęciami estymacji są: estymator, szacowany parametr i ocena parametru (oszacowanie).

Estymatorem, nazywanym też statystyką czy charakterystyką z próby, jest funkcja próby losowej $X_1, ..., X_n$, której rozkład prawdopodobieństwa zależy od szacowanego parametru. Ogólnie estymator szukanego parametru Q oznaczymy symbolem Q_n. Wykorzystując założenia przyjęte w podrozdziale w 5.1, można zapisać:

$$Q_n = f(X_1, X_2, ..., X_n). \tag{5.1}$$

Szacowanym parametrem Q, może być np. średnia μ lub wariancja σ^2 w zbiorowości o rozkładzie normalnym, p w rozkładzie dwumianowym czy λ w rozkładzie Poissona.

Jak wynika z zapisu (5.1), estymator Q_n jest funkcją ciągu zmiennych losowych $X_1, ..., X_n$, co pociąga wniosek, że należy go uważać również za zmienną losową.

Oceną parametru Q nazwiemy konkretną wartość liczbową

$$q_n = f(x_1, x_2, ..., x_n), \tag{5.2}$$

jaką przyjmuje estymator Q_n parametru Q dla realizacji próby $(x_1, x_2, ..., x_n)$.

Ponieważ szacunek nieznanej wartości parametru Q opiera się na próbie losowej, wartość liczbowa (5.2) może być obarczona błędem szacunku. Błędem szacunku parametru Q nazwiemy różnicę $Q - q_n$. Spośród dostępnych estymatorów parametru Q należy wybrać ten, dla którego błąd szacunku jest najmniejszy.

Dobroć określonego estymatora opisywana jest za pomocą jego własności. Od dobrego estymatora wymaga się głównie, aby był: **nieobciążony**, **zgodny** i **efektywny**.

Estymator $\hat{Q}_n$ jest nieobciążony, jeżeli jego wartość oczekiwana jest równa szacowanemu parametrowi, tzn.:

$$E\left(Q_n\right) = Q. \tag{5.3}$$

Jeżeli wartość estymatora jest różna od szacowanego parametru, to wówczas estymator nazywamy obciążonym, a wyrażenie:

$$\delta_n = E\left(Q_n\right) - Q, \tag{5.4}$$

określane jest mianem obciążenia estymatora.

Jeżeli obciążenie δ_n, wraz ze wzrostem liczebności próby maleje do zera tzn.:

$$\lim_{n \to \infty} \delta_n = 0, \tag{5.5}$$

to wówczas estymator nazywany jest estymatorem asymptotycznie nieobciążonym.

Jeżeli estymator jest asymptotycznie nieobciążony, oznacza to, że otrzymywane za pomocą niego szacunki wolne są od błędu systematycznego.

Z powyższego wynika, że jeżeli chcemy otrzymywać prawidłowe szacunki, to w przypadku małych prób należy stosować estymatory nieobciążone, natomiast gdy próba jest duża, estymator powinien być co najmniej asymptotycznie nieobciążony.

Estymator nazywany jest zgodnym, jeżeli:

$$\lim_{n \to \infty} P\left(\left|Q_n - Q\right| < \varepsilon\right) = 1, \tag{5.6}$$

gdzie ε jest dowolnie małą, dodatnią liczbą rzeczywistą.

Wyrażenie (5.6) oznacza, że zwiększając liczebność próby, otrzymuje się coraz większe prawdopodobieństwo tego, że estymator przyjmie wartości niewiele różniące się od wartości szacowanego parametru Q.

Pomiędzy własnościami nieobciążoności i zgodności można wyróżnić następujące zależności[1]:

1. Jeżeli estymator Q_n parametru Q jest zgodny, to jednocześnie jest on asymptotycznie nieobciążony, przy czym zależność odwrotna nie jest prawdziwa.
2. Jeżeli estymator $\hat{Q}_n$ parametru Q jest nieobciążony lub asymptotycznie nieobciążony, oraz gdy w miarę wzrostu liczebności próby do nieskończoności wariancja estymatora dąży do zera, to wówczas estymator Q_n jest również zgodny.

Jeżeli parametr Q może być oszacowany za pomocą k różnych nieobciążonych estymatorów, to wówczas najefektywniejszym estymatorem jest ten, który posiada najmniejszą wariancję.

Chcąc ocenić efektywność badanego estymatora $\hat{Q}^*_n$, należy skorzystać z ilorazu:

$$e\left(\hat{Q}_n\right) = \frac{D^2\left(\hat{Q}^*_n\right)}{D^2\left(\hat{Q}_n\right)} \tag{5.7}$$

gdzie:

$D^2\left(\hat{Q}^*_n\right)$ – wariancja estymatora najefektywniejszego,

$D^2\left(\hat{Q}_n\right)$ – wariancja estymatora badanego.

Im wartość wyrażenia (5.7) jest bliższa jedności, tym estymator jest efektywniejszy. Jeżeli natomiast w miarę wzrostu liczebności próby wartość $e\left(\hat{Q}_n\right)$ zmierza do jedności, to estymator nazwiemy asymptotycznie najefektywniejszym. Możemy zatem zapisać, że $\hat{Q}_n$ jest asymptotycznie najefektywniejszy, jeżeli:

$$\lim_{n \to \infty} e\left(\hat{Q}_n\right) = \frac{D^2\left(\hat{Q}^*_n\right)}{D^2\left(\hat{Q}_n\right)} = 1. \tag{5.8}$$

[1] Zob. np. M. Sobczyk: *Statystyka*, PWN 1998, s. 120.

Obok tych trzech własności wymienia się niekiedy czwartą, jaką jest **wystarczalność (dostateczność)**. Estymator Q_n, parametru Q jest dostateczny, jeśli zawiera wszystkie wiadomości, jakie na temat szacowanego parametru zostały zawarte w próbie i żaden inny estymator nie może dać o nim (parametrze) dodatkowych informacji[2].

Na przykład za estymator dostateczny wartości oczekiwanej populacji może uznać średnią arytmetyczną z próby, w odróżnieniu od statystyki $(x_{min} + x_{max})/2$, która opiera się tylko na dwóch wartościach (minimalnej i maksymalnej) z próby.

5.3. Metody uzyskiwania estymatorów

Istnieją trzy metody uzyskiwania estymatorów:
1) metoda momentów (MM),
2) metoda największej wiarygodności (MNW),
3) metoda najmniejszych kwadratów (MNK).

Pierwszą z nich – metodę momentów – datuje się na przełom XIX i XX wieku, a za jej twórcę uznaje się K. Pearsona. Istota MM polega na tym, że wartość momentu z próby przyjmuje się za oszacowanie odpowiedniego momentu zbiorowości generalnej. Na przykład, aby oszacować wartość oczekiwaną populacji wykorzystuje się jako estymator moment zwykły z próby, natomiast do oszacowania wariancji, zgodnie z metodą momentów, należy użyć momentu centralnego rzędu drugiego.

Uzyskane metodą momentów estymatory są najczęściej zgodne, lecz obciążone i mało efektywne. Dużo lepsze estymatory otrzymuje się, stosując metodę największej wiarygodnosci (MNW). Powstała ona w latach dwudziestych XX wieku, a za twórcę uważa się R.A. Fiszera. Podstawowym pojęciem metody jest tzw. **wiarygodność** (*ang. likelihood*) **próby**, którą określa się – w zależności od rodzaju zmiennej losowej – jako łączne prawdopodobieństwo lub łączną gęstość prawdopodobieństwa wyników $x_1, ..., x_n$ otrzymanych na podstawie próby. Wiarygodność próby jest oczywiście uzależniona od prawdziwej wartości szacowanego parametru Q. Systematyzując powyższe stwierdzenia, możemy zapisać:

$$L(x_1,.. x_n;Q)=\prod_{i=1}^{n} f(x_i;Q), \tag{5.9}$$

dla ciągłych zmiennych losowych, lub

$$L(x_1,.. x_n;Q)=\prod_{i=1}^{n} p(x_i;Q), \tag{5.9a}$$

dla dyskretnych zmiennych losowych.

2 *Ibidem*, s. 121.

Ideą MNW jest znalezienie takiej wartości parametru Q, przy której funkcja (5.9) lub (5.9a) osiąga maksimum. Wadą omawianej metody jest to, że warunkiem koniecznym, pozwalającym zastosować MNW jest znajomość postaci funkcyjnej rozkładu populacji, co w praktyce często jest utrudnione. W kolejnych podrozdziałach przedstawimy przebieg uzyskiwania estymatorów metodą MNW dla zmiennych losowych o rozkładzie dwupunktowym, Poissona, normalnym i wykładniczym. Najpierw jednak scharakteryzujmy krótko ostatnią trzecią metodę uzyskiwania estymatorów.

Z metodą najmniejszych kwadratów (MNK) spotkaliśmy się już w pierwszej części podręcznika, podczas omawiania metod szacowania parametrów funkcji regresji i funkcji trendu. Metodę datuje się na początek XIX w, a za jej twórcę uznaje się K. Gaussa. Istotny wpływ na rozwój tej metody miał również A. Markow. Istotą MNK, jest taki dobór ocen szacowanych parametrów, aby suma kwadratów różnic wartości empirycznych danej funkcji i wartości teoretycznych była minimalna. Uzyskane tym sposobem estymatory parametrów liniowych funkcji regresji lub liniowych funkcji trendu są zgodne, nieobciążone i najefektywniejsze w klasie estymatorów liniowych.

5.3.1. Estymacja MNW parametru *p* w rozkładzie dwupunktowym

Załóżmy, że zbiorowość generalna charakteryzuje się rozkładem zero-jedynkowym z parametrem $p = P(X = 1)$. Z populacji pobrano n-elementową próbę losową $X_1, ..., X_n$. Funkcję wiarygodności opisywać będzie rownanie (5.10) postaci:

$$\begin{aligned} L = L(x_1,... x_n; p) &= p^{x_1}(1-p)^{1-x_1} + p^{x_2}(1-p)^{1-x_2} + ... + p^{x_n}(1-p)^{1-x_n} = \\ &= \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\sum_{i=1}^{n} x_i}(1-p)^{n-\sum_{i=1}^{n} x_i} \end{aligned} \tag{5.10}$$

Po obustronnym zlogarytmowaniu funkcję (5.10) można zapisać w postaci:

$$\ln L = \sum_{i=1}^{n} x_i \ln p + (n - \sum_{i=1}^{n} x_i)\ln(1-p).$$

Po obliczeniu pierwszej pochodnej względem p otrzymamy:

$$\frac{\partial \ln L}{\partial p} = \sum_{i=1}^{n} x_i \frac{1}{p} - (n - \sum_{i=1}^{n} x_i)\frac{1}{1-p}$$

Aby wyznaczyć wartość p, przy którym funkcja (5.10) osiąga maksimum, otrzymaną pochodną przyrównujemy do zera i mamy wówczas:

$$\sum_{i=1}^{n} x_i \frac{1}{p} - (n - \sum_{i=1}^{n} x_i)\frac{1}{1-p} = 0.$$

Po rozwiązaniu względem p równania (5.10) dostajemy:

$$p = \hat{p}_n = \frac{\sum_{i=1}^{n} x_i}{n}. \tag{5.11}$$

Z powyższego zapisu wynika, że estymatorem parametru p w rozkładzie dwupunktowym jest klasyczna średnia arytmetyczna z próby. Ponieważ każda ze zmiennych losowych X_i ($i = 1, \ldots, n$), może przyjmować 0 lub 1, wartość licznika jest sumą jedynek. Jeżeli sumę tę oznaczymy symbolem z (tzn. $z = \sum_{i=1}^{n} x_i$), to wówczas równanie (5.11), możemy zapisać następująco:

$$p = \hat{p}_n = \frac{z}{n}, \tag{5.12}$$

i interpretować jako frakcję jedynek w prostej próbie n-elementowej.

Otrzymany estymator (5.12) jest estymatorem nieobciążonym, zgodnym i najefektywniejszym. Warto zaznaczyć, że estymator (5.12) jest również estymatorem parametru p populacji o rozkładzie dwumianowym. Dzieje się tak, ponieważ zmienna losowa o rozkładzie dwumianowym jest sumą n niezależnych zmiennych losowych o rozkładzie zero-jedynkowym.

5.3.2. Estymacja MNW parametru λ w rozkładzie Poissona

Załóżmy obecnie, że populacja ma rozkład Poissona o parametrze λ. Funkcja wiarygodności L będzie miała postać:

$$L = L(x_1, \ldots x_n; \lambda) = \frac{\lambda^{x_1} e^{-\lambda}}{x_1!} \cdot \frac{\lambda^{x_2} e^{-\lambda}}{x_2!} \cdot \ldots \cdot \frac{\lambda^{x_n} e^{-\lambda}}{x_n!} =$$
$$= \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} = \frac{\lambda^{\sum_{i=1}^{n} x_i} e^{-n\lambda}}{\prod_{i=1}^{n} x_i!}. \tag{5.13}$$

Dokonując podobnych jak w punkcie 5.3.1 przekształceń i obliczeń, otrzymamy:

$$\ln L = \sum_{i=1}^{n} x_i \ln \lambda - n\lambda - \sum_{i=1}^{n} \ln x_i!,$$

$$\frac{\partial \ln L}{\partial \lambda} = \frac{\sum_{i=1}^{n} x_i}{\lambda} - n = 0,$$

$$\lambda = \hat{\lambda}_n = \frac{\sum_{i=1}^{n} x_i}{n}. \tag{5.14}$$

Otrzymany estymator parametru λ jest nieobciążony, zgodny i najefektywniejszy.

5.3.3. Estymacja MNW parametrów μ i σ^2 w rozkładzie normalnym

Stosując metodę największej wiarygodności, na podstawie n-elementowej próby prostej pobranej z populacji, której badana zmienna X ma rozkład normalny $N(\mu, \sigma)$, ustalimy estymatory parametrów μ i σ^2. Funkcja wiarygodności będzie miała postać:

$$L = L(x_1,..,x_n;\mu,\sigma) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2\right] = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right] \tag{5.15}$$

Po obustronnym zlogarytmowaniu otrzymamy:

$$\ln L = n\ln\frac{1}{\sigma} + n\ln\frac{1}{\sqrt{2\pi}} - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2.$$

Obliczając pierwsze pochodne względem μ i σ^2 i przyrównując je do zera, otrzymamy następujący układ równań:

$$\begin{cases} \dfrac{\partial \ln L}{\partial \mu} = \dfrac{1}{\sigma^2}\left(\sum_{i=1}^{n} x_i - n\mu\right) = 0 \\ \dfrac{\partial \ln L}{\partial \sigma^2} = -\dfrac{n}{2\sigma^2} + \dfrac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2 = 0, \end{cases}$$

z którego wyznaczamy szukane estymatory:

$$\mu = \mu_n = \frac{\sum_{i=1}^{n} x_i}{n}, \tag{5.16}$$

$$\sigma^2 = \sigma_n^2 = \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n} \tag{5.17}$$

Estymator (5.16) jest estymatorem nieobciążonym, zgodnym i najefektywniejszym, natomiast estymator (5.17) jest zgodny i tylko asymptotycznie nieobciążony.

5.3.4. Estymacja MNW parametru λ w rozkładzie wykładniczym

Na koniec rozważań dotyczących metody MNW przeanalizujemy proces wyznaczania estymatora parametru λ dla populacji o rozkładzie wykładniczym. Wyjściowa funkcja wiarygodności będzie miała postać:

$$L = L(x_1, .., x_n; \lambda) = \prod^{n} \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda \sum x_i} \tag{5.18}$$

Postępując analogicznie jak we wcześniejszych przypadkach, otrzymamy:

$$\frac{\partial \ln L}{\partial \lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i = 0,$$

a stąd:

$$\lambda = \lambda_n = \frac{n}{\sum\limits_{i=1}^{n} x_i}. \tag{5.19}$$

Otrzymany estymator parametru λ jest tym razem odwrotnością średniej arytmetycznej z próby.

5.4. Estymacja punktowa parametrów jednowymiarowej zmiennej losowej

5.4.1. Uwagi wstępne

Przedstawimy obecnie sposoby punktowej estymacji podstawowych parametrów opisujących rozkłady jednowymiarowej zmiennej losowej. Szacowanymi parametrami będą: wartość oczekiwana, wariancja i wynikające z niej odchylenie standardowe. W związku z tym, że rozważamy obecnie punktową estymację, za każdym razem efektem estymacji będzie jedna wartość (punkt na osi liczb rzeczywistych).

5.4.2. Estymacja wartości oczekiwanej

Istnieje kilka estymatorów pozwalających na punktowy szacunek nieznanej wartości oczekiwanej zbiorowości generalnej, jednak najlepszym jest średnia arytmetyczna z próby. Średnią arytmetyczną z próby, zmiennej losowej X, oznaczamy $(\bar{X}_n)$ i obliczamy analogicznie jak średnią arytmetyczną z całej populacji[3] z tą tylko różnicą, że obliczenia wykonywane są nie na całej populacji, lecz na jej części (próbie).

Sposób wyznaczania średniej z próby zależy od postaci szeregu statystycznego opisującego zbiór realizacji ciągu n zmiennych losowych X_n $\{X_1, ..., X_n\}$. Jeżeli jest to szereg szczegółowy, to wówczas wartość średniej arytmetycznej z próby obliczamy według wzoru:

$$\bar{x}_n = \frac{1}{n} \sum x_i. \tag{5.20}$$

[3] Zob. M. Major, J. Niezgoda: *Elementy statystyki...*, *op. cit.*, rozdz. 3.2.1.

gdzie:

x_i – jest realizacją w zbiorze X_n,

n – jest liczebnością badanej losowej próby.

W przypadku szeregów rozdzielczych punktowych i przedziałowych podczas obliczania wartości średniej arytmetycznej z próby należy skorzystać odpowiednio:

$$\bar{x}_n = \frac{1}{\sum\limits_{j=1}^{k} f_j} \sum_{j=1}^{k} x_j f_j = \sum_{j=1}^{k} x_j v_j, \tag{5.21}$$

w przypadku szeregu rozdzielczego punktowego

lub
$$\bar{x}_n = \frac{1}{\sum\limits_{j=1}^{k} f_j} \sum_{j=1}^{k} x_j^0 f_j = \sum_{j=1}^{k} x_j^0 v_j, \tag{5.22}$$

dla szeregu rozdzielczego przedziałowego.

Interpretacja oznaczeń użytych we wzorach (5.21) i (5.22) odpowiada zapisowi podręcznika M. Major, J. Niezgoda: *Elementy...*, *op. cit.*, w rozdz. 3.2.1. Przypomnijmy:

f_j – jest liczebością, z jaką występowała j-ta wartość zmiennej X (wzór 5.21) lub liczebnością j-tego przedziału klasowego (wzór 5.22),

v_j – jest częstością względną występowania j – tej wartości zmiennej X lub częstością przypisaną j-temu przedziałowi klasowemu,

x^o – jest środkiem j-tego przedziału klasowego.

Średnia arytmetyczna z próby jest estymatorem nieobciążonym, zgodnym i najefektywniejszym i może być stosowana dla dowolnego typu rozkładu.

W przypadku braku możliwości – w trakcie estymacji wartości oczekiwanej populacji – posłużenia się średnią arytmetyczną z próby, można zastosować określoną statystykę pozycyjną z próby. Może to być np. mediana z próby ($X_{me,n}$), której wartość ($x_{me,n}$) wyznaczamy analogicznie jak w badaniach stuprocentowych (zob. M. Major, J. Niezgoda: *Elementy...*, *op. cit.*, wzór (3.20) i (3.21)). Jeżeli wyniki otrzymane po przebadaniu n-elementowej próby losowej przedstawione są w postaci szeregu szczegółowego, to wówczas korzystamy ze wzoru:

$$x_{me,n} = \begin{cases} x_{\frac{n+1}{2}} & \textit{gdy} \quad n - \textit{nieparzyste} \\ \frac{1}{2}\left[x_{\frac{n}{2}} + x_{\frac{n}{2}+1} \right] & \textit{gdy} \quad n - \textit{parzyste} \end{cases} \tag{5.23}$$

Natomiast, jeżeli operujemy szeregiem rozdzielczym przedziałowym, to wówczas podczas wyznaczania $x_{me,n}$ skorzystamy ze wzoru[4]:

$$x_{me,n} = x_{d,r} + \frac{\frac{n}{2} - \sum_{i=1}^{r-1} f_i}{f_r} l. \tag{5.24}$$

gdzie:

$x_{d,r}$ – dolna granica przedziału, w którym znajduje się mediana,
f_r – liczebność przedziału, w którym znajduje się mediana,
l – długość przedziału, w którym znajduje się mediana,
$\sum_{i=1}^{r-1} f_i$ – suma liczebności przedziałów poprzedzających przedział, w którym znajduje się mediana.

Mediana z próby jest najczęściej gorszym estymatorem wartości oczekiwanej populacji niż średnia arytmetyczna z próby. Mediana jest estymatorem zgodnym, a w przypadku ciągłych zmiennych losowych o umiarkowanej asymetrii, asymptotycznie nieobciążonym. Jeżeli chodzi o jego efektywność, to wynosi ona $2/\pi \approx 0{,}64$ efektywności estymatora średniej z próby.

5.4.3. Estymacja wariancji i odchylenia standardowego

Sposób szacowania wariancji populacji zależy od kilku czynników. Pierwszym z nich jest znajomość rzeczywistej wartości oczekiwanej całej populacji. Jeżeli znamy wartość oczekiwaną populacji μ, to wówczas należy skorzystać ze statystyki $\underline{S}^2$. Postać tej statystyki jest uzależniona od postaci zbioru realizacji X_n. Jeżeli jest to szczegółowy szereg statystyczny, to realizację zmiennej losowej otrzymujemy ze wzoru:

$$\underline{s}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2. \tag{5.25}$$

W przypadku szeregów rozdzielczych punktowych i przedziałowych wzór (5.25) należy zastąpić odpowiednio wzorami:

$$\underline{s}^2 = \frac{1}{\sum_{j=1}^{k} f_j}\sum_{j=1}^{k} f_j(x_j - \mu)^2 = \sum_{j=1}^{k} v_j(x_j - \mu)^2, \tag{5.26}$$

$$\underline{s}^2 = \frac{1}{\sum_{j=1}^{k} f_j}\sum_{j=1}^{k} f_j(x_j^o - \mu)^2 = \sum_{j=1}^{k} v_j(x_j^o - \mu)^2. \tag{5.27}$$

[4] Zob. M. Major, J. Niezgoda: *Elementy...*, *op. cit.*, s. 52.

Statystyka S^2 jest estymatorem nieobciążonym, zgodnym i najefektywniejszym. Warunkiem jego stosowania jest jednak znajomość wartości oczekiwanej badanej zmiennej losowej. W praktyce nie zawsze jest to możliwe, gdyż wymaga wcześniejszych badań stuprocentowych.

Jeżeli wartość oczekiwana populacji μ jest nieznana, to należy ją wcześniej oszacować na podstawie próby, korzystając z opisanej we wcześniejszym punkcie średniej z próby. Po oszacowaniu średniej z próby wartość wariancji populacji estymujemy, stosując statystykę S^2, lub statystykę S^{*2}. Realizacje pierwszego z estymatorów S^2 w przypadku szczegółowego szeregu statystycznego wyznaczamy ze wzoru:

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2, \tag{5.28}$$

natomiast w przypadku szeregów rozdzielczych odpowiednio:

$$s^2 = \frac{1}{\sum_{j=1}^{k} f_j}\sum_{j=1}^{k} f_j (x_j - \bar{x}_n)^2 = \sum_{j=1}^{k} v_j (x_j - \bar{x}_n)^2, \tag{5.29}$$

$$s^2 = \frac{1}{\sum_{j=1}^{k} f_j}\sum_{j=1}^{k} f_j (x_j^o - \bar{x}_n)^2 = \sum_{j=1}^{k} v_j (x_j^o - \bar{x}_n)^2. \tag{5.30}$$

Statystyka S^2 jest estymatorem obciążonym, wielkość tego obciążenia wynika ze wzoru[5]:

$$\delta_n = E(S^2) - \sigma^2 = \frac{n-1}{n}\sigma^2 - \sigma^2 = -\frac{\sigma^2}{n}. \tag{5.31}$$

Ponieważ wielkość obciążenia (5.31) maleje do zera wraz ze wzrostem liczebności próby, tzn. spełniona jest zależność (5.5), zatem estymator S^2 możemy zaliczyć do grupy asymptotycznie nieobciążonych. Oznacza to, że można go efektywnie stosować, jeżeli losowo dobrana próba prosta jest odpowiednio duża, co zwykle oznacza liczność powyżej 30. W przypadku małych próbek zaleca się stosować drugi z estymatorow: S^{*2}. Realizacje statystyki S^{*2} – dla analizowanych trzech typów szeregow statystycznych – wyznaczamy ze wzorów:

$$s^{*2} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2, \tag{5.32}$$

$$s^{*2} = \frac{1}{\sum_{j=1}^{k} f_j - 1}\sum_{j=1}^{k} f_j (x_j - \bar{x}_n)^2, \tag{5.33}$$

$$s^{*2} = \frac{1}{\sum_{j=1}^{k} f_j - 1}\sum_{j=1}^{k} f_j (x_j^o - \bar{x}_n)^2. \tag{5.34}$$

[5] Zob. A. Iwasiewicz, Z. Paszek: *Statystyka..., op. cit.*. s. 180.

Estymator S^{*2} jest estymatorem zgodnym, nieobciążonym i najefektywniejszym. Mianownik we wzorach (5.32), (5.33) i (5.34) wynoszący $n - 1$, pozwala na usunięcie obciążenia badanego estymatora. Wyrażenie to, określa się **liczbą stopni swobody** i definiuje jako liczbę niezależnych obserwacji. Ponieważ w trakcie obliczeń wariancji z próby w oparciu o statystykę S^{*2} musieliśmy wcześniej oszacować wartość średniej arytmetycznej z próby, przez co liczba niezależnych obserwacji zmniejszyła się od początkowej n o 1 i wynosi obecnie $n - 1$. Wyrażenie $n - 1$ można również interpretować jako liczbę odchyleń od zmiennej, które mogą się swobodnie zmieniać[6]. Można to zilustrować prostym przykładem:

Przykład 5.1

Rozważmy na przykład, że w wyniku trzech pomiarów w próbie uzyskano wartości: 2, 4 i 12. Średnia z tych wartości wynosi $(2 + 4 + 12)/3 = 6$, a odchylenia od średniej odpowiednio: –4, –2 i 6. Suma tych odchyleń oczywiście jest równa zero $(-4 + (-2) + 6 = 0)$. Zatem, jeśli znamy dwa dowolne odchylenia, to wartość trzeciego jest ustalona. Suma kwadratów odchyleń od średniej wynosi: $16 + 4 + 36 = 56$. Zauważmy, że pomimo, iż w sumie tej mamy trzy składniki, tylko dwa z nich mają swobodę zmieniania się, a trzeci jest od nich uzależniony. Ową liczbę składników, które mogą swobodnie zmieniać się nazywa się liczbą stopni swobody. Zatem obliczając wariancję z tej próby powinniśmy w mianowniku wstawić wartość 2, a nie 3. Otrzymamy wówczas: $56/2 = 28$. {kp}

Mając oszacowaną wariancję populacji, można wyciągnąć z niej pierwiastek kwadratowy i tym sposobem otrzymać charakterystykę określaną jako odchylenie standardowe próby. Przyjmując podobne oznaczenia jak w przypadku wariancji otrzymamy kolejno trzy estymatory: $\underline{S}=\sqrt{\underline{S}^2}$; $S=\sqrt{S^2}$; $S^*=\sqrt{S^{*2}}$.

Ze względu, że otrzymane estymatory $\underline{S}$ i S^* są jedynie estymatorami asymptotycznie nieobciążonymi, wprowadza się odpowiednie współczynniki c_r, które likwidują obciążoność tych estymatorów. Warunkiem koniecznym do stosowania takiego zabiegu jest, aby rozkład populacji był przynajmniej zbliżony do normalnego. Po przemnożeniu otrzymujemy dwa nieobciążone estymatory odchylenia standardowego, których wartości wynoszą odpowiednio:

$$s = c_r \cdot s, \tag{5.35}$$

$$s^{**} = c_r \cdot s^{*}. \tag{5.36}$$

---

[6] Zob. G.A. Ferguson, Y. Takane: *Analiza statystyczna w psychologii i pedagogice*, PWN, Warszawa 2002, s. 89.

A. Iwasiewicz i Z. Paszek w cytowanej już książce, *Statystyka...*, s. 181, podają następującą funkcję umożliwiającą wyznaczenie wartość parametru c_r:

$$c_r = \frac{\Gamma\left(\frac{r}{2}\right)\sqrt{r}}{\Gamma\left(\frac{r+1}{2}\right)\sqrt{2}}, \tag{5.37}$$

gdzie: r oznacza liczbę stopni. We wzorze (5.35) $r = n$, natomiast w przypadku wzoru (5.36) $r = n - 1$. Wartość funkcji gamma wyznaczona jest w następujący sposób:

$$\Gamma\left(\frac{r}{2}\right) = \begin{cases} \left(\frac{r}{2}-1\right)! & \textit{gdy r jest liczbą parzystą, } n \geq 2, \\ \left(\frac{r-2}{2}\right)\left(\frac{r-4}{2}\right)\ldots\left(\frac{1}{2}\right)\sqrt{\pi} & \textit{gdy r jest liczbą nieparzystą, } n \geq 3, \end{cases}$$

$$\Gamma\left(\frac{r+1}{2}\right) = \begin{cases} \left(\frac{r-1}{2}\right)\left(\frac{r-3}{2}\right)\ldots\left(\frac{1}{2}\right)\sqrt{\pi} & \textit{gdy r jest liczbą parzystą, } n > 2, \\ \left(\frac{r+1}{2}-1\right)! & \textit{gdy r jest liczbą nieparzystą, } n > 1. \end{cases}$$

W niektórych podręcznikach[7] podczas obliczania wartość współczynnika c_r zaleca się korzystanie ze specjalnych tablic umieszczonych na końcu podręcznika.

Ponieważ proces wyznaczania współczynnika c_r na podstawie wzoru (5.37), wizualnie wydaje się dość skomplikowany, prześledźmy go na podstawie krótkiego przykładu liczbowego.

Przykład 5.2

Załóżmy, że liczba stopni swobody $r = 7$.

$$c_r = \frac{\Gamma\left(\frac{7}{2}\right)\sqrt{7}}{\Gamma\left(\frac{7+1}{2}\right)\sqrt{2}} = \frac{\frac{5}{2}\cdot\frac{3}{2}\cdot\frac{1}{2}\sqrt{\pi}\sqrt{7}}{\left(\frac{8}{2}-1\right)!\sqrt{2}} =$$

$$\frac{1{,}875\cdot\sqrt{\pi}\sqrt{7}}{6\cdot\sqrt{2}} = \frac{8{,}792760186}{8{,}485281374} = 1{,}036236726 \approx 1{,}036.$$

---

[7] Zob. np. W. Krysicki i inni: *Rachunek prawdopodobieństwa i statystyka matematyczna w zadaniach*, t. II, s. 48.

Otrzymaną wartość c_r należy następnie podstawić do jednego ze wzorów (5.35) lub (5.36) i otrzymamy wówczas nieobciążoną ocenę odchylenia standardowego populacji. {kp}

Odchylenie standardowe z próby nie jest jedynym możliwym estymatorem odchylenia standardowego populacji. Do tego celu można również użyć inną miarę rozproszenia opartą na statystykach pozycyjnych. Taką miarą jest np. rozstęp z próby (R_n), którego wartości wyznaczamy następująco:

$$r_n = x_{max} - x_{min}, \tag{5.38}$$

gdzie: x_{max} i x_{min} odpowiadają największej i najmniejszej wartości w zbiorze wartości X_n.

W przypadku gdy obserwowana zmienna losowa ma rozkład normalny z odchyleniem standardowym σ, to wówczas estymatorem parametru σ jest statystyka:

$$\sigma_n = \frac{\bar{r}_n}{d_n} = \frac{\bar{r}_n}{\frac{E(R_n)}{\sigma}}, \tag{5.39}$$

gdzie: $\bar{r}_n$ jest średnią arytmetyczną rozstępu w n-elementowej próbce.

Wartość d_n dla $n = 1, ..., 12$ zostały podane w tablicy 5.1:

Tablica 5.1 Parametry rozkładu rozstępu

n	2	3	4	5	6	7
d_n	1,12838	1,69257	2,05875	2,32593	2,53441	2,70436
n	8	9	10	11	12	
d_n	2,84720	2,97003	3,07751	3,17287	3,25846	

Źródło: opracowanie własne na podstawie: A. Iwasiewicz: *Zarządzanie jakością*, Warszawa–Kraków 1999, s. 294.

Rozstęp z próby jest efektywnym estymatorem dla prób o małej liczebności. W praktyce oznacza to próbę nie większą niż 10. W przypadku, gdy $n < 10$, w miejsce we wzorze (5.39) wstawia się charakterystykę r_n. Natomiast, gdy $n > 10$, to wówczas dzieli się ją na „l" mniejszych podzbiorów (o jednakowej liczności nie większej niż 10), dla których wyznacza się wartości rozstępów $r_1, r_2, \ldots, r_l$. W następnym kroku z rozstępów tych oblicza się wartość średniego rozstępu, stosując wzór:

$$\bar{r}_n = \frac{r_1 + r_2 + \ldots + r_l}{l}. \tag{5.40}$$

Uzyskaną w ten sposób wartość charakterystyki wstawia się do wzoru (5.39).

Przykład 5.3

Załóżmy, że producent świeczek postanowił ustalić, jaka jest rzeczywista wartość odchylenia standardowego zmiennej losowej X będącej czasem spalania świeczki w minutach. W tym celu z bardzo dużej partii świeczek pobrał próbę losową liczącą 20 sztuk. Próbę tę podzielił następnie na 4 równe podzbiory o liczebności pięcioelementowej. Po przeprowadzeniu eksperymentu otrzymano następujące wyniki:

Tablica 5.2. Dane liczbowe oraz technika obliczania średniego rozstępu z próby

Podzbiór 1			Podzbiór 2			Podzbiór 3			Podzbiór 4		
i	j	x_{ij}	i	j	x_{ij}	i	j	x_{ij}	i	j	x_{ij}
1	1	721,0	6	1	723,0	11	1	721,5	16	1	722,0
2	2	721,5	7	2	720,0	12	2	721,0	17	2	719,0
3	3	719,0	8	3	719,5	13	3	719,5	18	3	720,5
4	4	720,6	9	4	721,0	14	4	721,0	19	4	721,5
5	5	718,5	10	5	719,5	15	5	719,0	20	5	720,0
$x_{j\,max}$		721,5	$x_{j\,max}$		723,0	$x_{j\,max}$		721,5	$x_{j\,max}$		722,0
$x_{i\,min}$		718,5	$x_{j\,min}$		719,5	$x_{j\,min}$		719,0	$x_{j\,min}$		719,0
r_1		3	r_2		3,5	r_3		2,5	r_4		3

Źródło: opracowanie własne na podstawie badań własnych.

Podstawiając wartości czterech kolejnych rozstępów z próby do wzoru (5.40), otrzymujemy średni rozstęp z próby:

$$\bar{r}_5 = \frac{3 + 35 + 2,5 + 3}{4} = \frac{12}{4} = 3.$$

Otrzymaną wartość charakterystyki podstawiamy następnie do wzoru (5.39) i otrzymujemy:

$$\sigma_n = \frac{3}{2,32593} = 1,289807.$$

Otrzymana wartość oznacza, że czasy palenia świeczek powinny odbiegać średnio od przeciętnego czasu palenia świeczki o ok. 1,3 minuty. Należy pamiętać jednak, że szacunek ten będzie prawdziwy, jeżeli rozkład czasu spalania będzie normalny lub zbliżony do rozkładu normalnego. {kp}

5.5. Rozkłady podstawowych statystyk z próby

5.5.1. Rozkład średniej z próby

Rozkład statystyki z próby należy do najbardziej podstawowych pojęć wnioskowania statystycznego. Rozkład statystyki z próby można ustalić w drodze wielokrotnego przeprowadzania eksperymentu (tzw. **eksperymentalny (empiryczny) rozkład z próby**), bądź też przez teoretyczne przeanalizowanie rozkładu prawdopodobieństwa z wszystkich możliwych do zrealizowania kombinacji wyników próby (tzw. **teoretyczny rozkład z próby**). Za każdym razem można rozważać dwa przypadki, w których próba losowa pobierana jest ze skończonej lub nieskończenie wielkiej populacji. Pobieranie próby z populacji nieskończenie wielkiej jest w zasadzie tym samym, co pobieranie prób z populacji skończonej, ale ze zwracaniem, podczas którego każdy element na powrot wędruje do populacji, zanim zostanie wylosowany następny. Zauważmy, że gdy operujemy populacją nieskończoną, to wówczas podczas losowania bez zwracania prawdopodobieństwo (podobnie jak przy losowaniu ze zwracaniem ze skonczonych populacji) wybrania konkretnego elementu pozostanie niezmienione, niezależnie od liczności pobieranej próby. Dzieje się tak dlatego, że w obydwu analizowanych przypadkach niezmienną pozostaje liczność badanej populacji i prawdopodobieństwo nie ulega zmianie w zależności od liczby pobranych już elementów.

Przykład 5.4

Założmy, że w urnie znajduje się pięć kartek ponumerowanych kolejno liczbami 1 do 5. Średnia dla populacji skonczonej – składającej się z 5 kartek – wyniesie:

$$x_N = (1 + 2 + 3 + 4 + 5)/5 = 3,$$

natomiast wariancja:

$$s_N^2 = [(1-3)^2 + (2-3)^2 + (3-3)^2 + (4-3)^2 + (5-3)^2]/5 =$$
$$= (4 + 1 + 0 + 1 + 4)/5 = 10/5 = 2.$$

Doświadczenie polega na wylosowaniu bez zwracania, trzech kartek. Liczba rożnych prob trzyelementowych, pobieranych bez zwracania do populacji jest liczbą kombinacji 5 kartek po 3 naraz, czyli:

$$\binom{5}{3} = \frac{5!}{3!\,2!} = 10.$$

Możemy wyciągnąć kartki z liczbami: {(1,2,3); (1,2,4); (1,2; 5); (2; 3; 4); (2,3,5); (2,4,5); (3,4,5); (1,3,4); (1,3,5); (1,4,5)}.

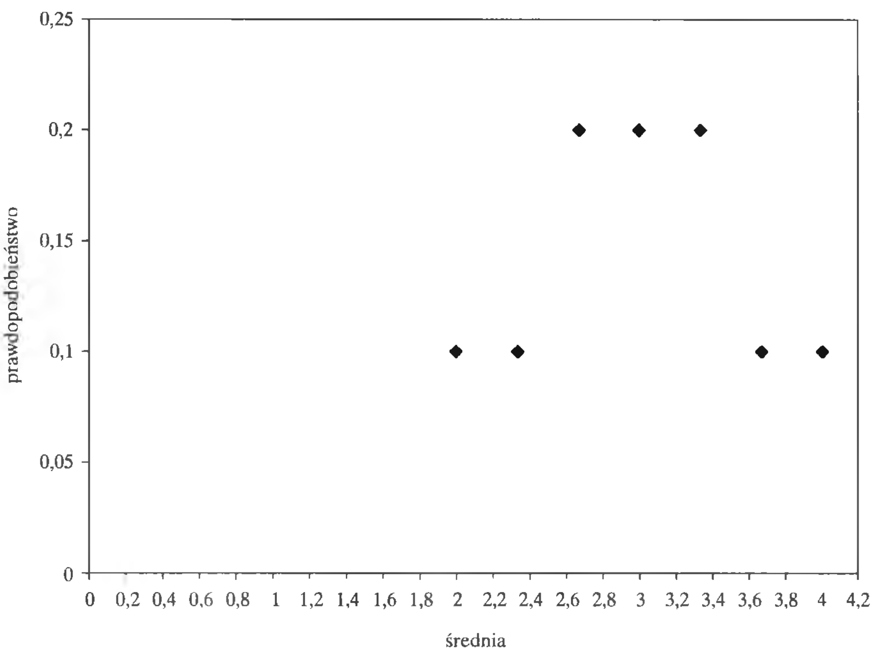
Po obliczeniu z tych realizacji średnich otrzymamy następujący ciąg wartości:

$$\left\{2; \frac{7}{3}; \frac{8}{3}; 3; \frac{10}{3}; \frac{11}{3}; 4; \frac{8}{3}; 3; \frac{10}{3}\right\}.$$

Korzystając z założenia, że pobrana próba jest losowa, można stwierdzić, że prawdopodobieństwo otrzymania każdej z 10 prób trzyelementowych jest jednakowe i wynosi 0,1. Takie samo prawdopodobieństwo realizacji ma również każda z otrzymanych średnich. Wykorzystując te wnioski, budujemy teoretyczny rozkład średniej z próby:

$\bar{x}_{in}$	2	$\frac{7}{3}$	$\frac{8}{3}$	3	$\frac{10}{3}$	$\frac{11}{3}$	4
$p_{\bar{x}_i}$	0,1	0,1	0,2	0,2	0,2	0,1	0,1

Wykres tego rozkładu został przedstawiony na rysunku 30.



Rys. 30. Rozkład prawdopodobieństwa średniej z próby

Źródło: opracowanie własne.

Na podstawie powyższych danych możemy wyznaczyć wartość oczekiwaną oraz wariancję otrzymanego rozkładu prawdopodobieństwa.

$$E(\overline{X}_n) = 2 \cdot 0{,}1 + \frac{7}{3} \cdot 0{,}1 + \frac{8}{3} \cdot 0{,}2 + 3 \cdot 0{,}2 + \frac{10}{3} \cdot 0{,}2 + \frac{11}{3} \cdot 0{,}1 + 4 \cdot 0{,}2 = 3;$$

$$D^2(\overline{X}_n) = (2-3)^2 \cdot 0{,}1 + \left(\frac{7}{3}-3\right)^2 \cdot 0{,}1 + \left(\frac{8}{3}-3\right)^2 \cdot 0{,}2 +$$

$$+ (3-3)^2 \cdot 0{,}2 + \left(\frac{10}{3}-3\right)^2 \cdot 0{,}2 + \left(\frac{11}{3}-3\right)^2 \cdot 0{,}1 + (4-3)^2 \cdot 0{,}2 = \frac{1}{3}.$$

Odchylenie standardowe:

$$D(\overline{X}_n) = \sqrt{D^2(\overline{X}_n)} = \sqrt{\frac{1}{3}} \approx 0{,}58.$$

Zauważmy, że średnia badanego rozkładu z próby jest identyczna, tak jak średnia całej populacji; natomiast odchylenie standardowe jest inne niż odchylenie rozkładu średniej z próby. {kp}

Obserwując wyniki doświadczenia przeprowadzonego w przykładzie 5.4, można sformułować kilka zasadniczych wniosków. Po pierwsze, jeżeli doświadczenie polega na pobieraniu n-elementowej bezzwrotnej losowej próby, ze skończonej populacji o średniej $E(X)$ i odchyleniu standardowym $D(X)$, to wówczas teoretyczny rozkład średniej z próby ma wartość oczekiwaną $E(\overline{X}_n)$ oraz odchylenie standardowe $D(\overline{X}_n)$ wynoszące:

$$D(\overline{X}_n) = \frac{D(X)}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}, \tag{5.41}$$

gdzie:
N – jest licznością skończonej populacji,
n – licznością pobieranej próby.

Po drugie, jeżeli populacja jest nieskończona, to wówczas wzór (5.37) należy zastąpić wzorem:

$$D(\overline{X}_n) = \frac{D(X)}{\sqrt{n}}. \tag{5.42}$$

Dzieje się tak dlatego, że w miarę wzrostu liczebności populacji do nieskończoności człon $\sqrt{\frac{N-n}{N-1}}$, we wzorze (5.41) dążyć będzie do jedności, co pozwala na redukcję wzoru (5.41) do postaci (5.42). W praktyce czynnik ten pomija się, jeśli próba stanowi nie więcej niż 5% populacji. Na przykład, jeżeli pobierzemy próbę losową $n = 10$ z populacji $N = 1000$, to próba stanowi jedynie 1% populacji, a wyrażenie

$\sqrt{\frac{N-n}{N-1}} = \sqrt{\frac{1000-10}{1000-1}} = \sqrt{\frac{990}{999}} \approx 0{,}995 \simeq 1$. Odchylenie standardowe rozkładu średniej z próby nazywane jest często **błędem standardowym średniej**. Nazwa ta powstała, ponieważ zadaniem statystyki (5.41) lub (5.42) jest pomiar stopnia zmienności średnich wywołanych czynnikami przypadkowymi (losowymi).

Przykład 5.4 cd.

Można obecnie sprawdzić numerycznie prawdziwość wzoru (5.41). Ponieważ wariancja badanej populacji wynosiła 2, liczebność próby $n = 3$, liczebność populacji $N = 5$, zatem:

$$D(\overline{X}_n) = \frac{\sqrt{2}}{\sqrt{3}}\sqrt{\frac{5-3}{5-1}} = \frac{\sqrt{2}}{\sqrt{3}}\sqrt{\frac{1}{2}} \approx 0{,}58.$$

Wartość obliczonego odchylenia standardowego statystyki $\overline{X}_n$ jest identyczna jak ta, którą obliczyliśmy, korzystając z rozkładu teoretycznego średniej z próby.

Średnią z próby i błąd standardowy średniej można również przybliżyć na podstawie rozkładu empirycznego. Jest to jednak dość pracochłonne, gdyż wymaga wielokrotnego przeprowadzenia doświadczenia polegającego na poborze próby losowej, na podstawie której wyznacza się średnie arytmetyczne, by później utworzyć z nich empiryczny rozkład średniej. Czytelnikom podręcznika polecamy przeprowadzenie doświadczenia podobnego do opisanego w przykładzie 5.4, które będzie polegało na np. 100-krotnym wylosowaniu bez zwracania 3 kartek z pięciu i 100-krotnym obliczeniu średniej arytmetycznej. Wielokrotne wykonanie tej operacji powinno gwarantować, że otrzymany empiryczny szereg rozdzielczy częstości średniej będzie zbliżony kształtem do rozkładu z rysunku 30.

Opisane powyżej zależności i wzory są oczywiście prawdziwe dla przypadku, gdy populacja ma rozkład normalny. Teoretyczny rozkład z prób pobranych z populacji normalnej jest rozkładem normalnym. Wynika z tego, że jeżeli wiemy, że rozkład zbiorowości generalnej jest normalny, to wówczas i rozkład z próby średnich będzie normalny o parametrach:

$$E(\overline{X}_n) = \mu_{\bar{x}} = \mu, \tag{5.43}$$

$$D(\overline{X}_n) = \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}. \tag{5.44}$$

Ponadto można stwierdzić, że niezależnie od kształtu rozkładu populacji, jeżeli liczebność próby pobieranej z populacji o średniej μ i odchyleniu standardowym σ wzrasta do nieskończoności, to wówczas rozkład z próby średnich zbliża się do rozkładu normalnego ze średnią μ i odchyleniem standardowym $\frac{\sigma}{\sqrt{n}}$.

Powyższe stwierdzenie określa się w literaturze jako **centralne twierdzenie graniczne**.

Ujmując rzecz nieco inaczej, można stwierdzić, że jeżeli n jest dostatecznie duże (w praktyce przynajmniej równe 30), to wówczas wartości u wyznaczone według równania:

$$u = \frac{\bar{x}_n - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\bar{x}_n - \mu}{\sigma}\sqrt{n}, \tag{5.45}$$

należy uznać za realizacje zmiennej losowej $U \sim N(0,1)$.

Stosowanie charakterystyki (5.45) wymaga jednak znajomości rzeczywistej wartości odchylenia standardowego populacji σ, co w praktyce jest często utrudnione. Jeżeli parametr σ jest nieznany, to wówczas należy go oszacować z próby wykorzystując estymator S lub estymator S^*. Jeżeli podczas szacunku stosujemy statystykę S, to oszacowanie parametru $s_{\bar{x}}$, uzyskuje się ze wzoru:

$$s_{\bar{x}} = \frac{s}{\sqrt{n-1}} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2}{n(n-1)}}, \tag{5.46}$$

natomiast gdy posługujemy się statystyką S^*, to do oszacowania $s_{\bar{x}}$ należy stosować wzór:

$$s_{\bar{x}} = \frac{s^*}{\sqrt{n}} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2}{n(n-1)}}. \tag{5.47}$$

Podstawiając w miejsce wartości odchylenia standardowego populacji jego oszacowanie, wzór (5.45) można przedstawić następująco:

$$u = t = \frac{\bar{x}_n - \mu}{s}\sqrt{n-1} = \frac{\bar{x}_n - \mu}{s^*}\sqrt{n}. \tag{5.48}$$

Zmienna losowa t[8] o realizacjach wyznaczonych według wzoru (5.48) nie będzie już miała rozkładu normalnego standaryzowanego, lecz tzw. **rozkład Studenta**, o $r = n - 1$ stopniach swobody. Nazwa tego rozkładu wywodzi się od pseudonimu naukowego jego twórcy, którym byl angielski statystyk William Goset żyjący na przełomie XIX i XX wieku. Funkcję prawdopodobieństwa rozkładu Studenta opisuje wzór:

$$f_r = \frac{\Gamma\left(\frac{r+1}{2}\right)}{\sqrt{\pi r}\,\Gamma\left(\frac{r}{2}\right)}\left(1 + \frac{t^2}{r}\right)^{-\frac{r+1}{2}} \tag{5.49}$$

[8] W tym przypadku wyjątkowo odstąpimy od ogolnej zasady i zarówno zmienną, jak i jej realizację oznaczać będziemy małą literą t.

gdzie:

$$r = n - 1, t \in (-\infty; +\infty),$$

zaś:

$$\Gamma(x) = \int_0^{+\infty} t^{x-1} e^{-t} dt \, (x > 0), \tag{5.50}$$

jest funkcją gamma.

Wartość oczekiwana i odchylenie standardowe statystyki t wynoszą odpowiednio:

$$E(t) = 0, \tag{5.51}$$

$$D(t) = \sqrt{\frac{r}{r-2}} = \sqrt{\frac{n-1}{n-3}}. \tag{5.52}$$

Krzywa rozkładu Studenta kształtem przypomina krzywą rozkładu normalnego $N(0,1)$. Podobnie jak krzywa rozkładu normalnego standaryzowanego jest ona symetryczna z osią symetrii o równaniu $t = 0$, natomiast jest nieco bardziej spłaszczona, gdyż jej odchylenie standardowe $D(t) > 1$. Rozkład Studenta jest stablicowany (zob. tablica II). Najczęściej w tablicach ujmuje się wartości prawdopodobieństwa:

$$P(|t| \geq t_\alpha) = \alpha, \tag{5.53}$$

gdzie: $\alpha \in [0; 1]$.

Rozkład Studenta przy wzroście liczebności próby dąży do rozkładu normalnego $N(0,1)$, przy czym zbieżność ta jest na tyle szybka, że przy $n > 30$, krzywe obu rozkładów są praktycznie nierozróżnialne. W praktyce oznacza to, że przy dużej liczebności próby, rozkład średniej arytmetycznej z próby, może być aproksymowany przez rozkład normalny, bez względu na to, czy znane jest odchylenie standardowe populacji σ, czy tylko jego szacunek.

5.5.2. Rozkład wariancji i odchylenia standardowego z próby

Załóżmy obecnie, że populacja ma rozkład normalny $N(\mu, \sigma)$. Można dowieść, że wariancja z próby ma również rozkład asymptotycznie normalny, przy czym średnie i odchylenia standardowe wynoszą odpowiednio[9]:

$$E(\underline{S}^2) = \sigma^2, D(\underline{S}^2) = \sigma^2 \sqrt{\frac{2}{n}}. \tag{5.54}$$

[9] Zob. np.: *Statystyka ogólna*, red. M. Woźniak, AE w Krakowie 1997, s. 162 i 163, lub A. Iwasiewicz, Z Paszek: *Statystka ...*, *op. cit.*, s. 188.

$$E(S^2) = \frac{n-1}{n}\sigma^2,\ D(S^2) = \frac{\sigma^2}{n}\sqrt{2(n-1)}, \tag{5.55}$$

$$E(S^{*2}) = \sigma^2,\ D(S^{*2}) = \sigma^2\sqrt{\frac{2}{n-1}}. \tag{5.56}$$

Zastosowanie powyższych wzorów wymaga jednak, aby liczebność próby była odpowiednio liczna (przynajmniej $n = 30$). W sytuacji gdy operujemy próbami małymi, wykorzystuje się statystykę postaci:

$$\frac{\frac{S^{*2}}{\sigma^2}}{n-1} = \frac{S^{*2}(n-1)}{\sigma^2}. \tag{5.57}$$

Statystyka (5.57) ma rozkład χ^2 (**chi-kwadrat**) o $n-1$ stopniach swobody i jest sumą kwadratów n niezależnych zmiennych losowych $U_1, ..., U_n$ o rozkładach normalnych $N(0,1)$. Zatem wartości statystyki (5.57) można wyznaczyć następująco:

$$\sum_{i=1}^{n} u_i^2 = \sum_{i=1}^{n}\left(\frac{x-\bar{x}_n}{\sigma}\right)^2 = \sum_{i=1}^{n}\frac{(x-\bar{x}_n)^2}{\sigma^2} = \frac{s^{*2}(n-1)}{\sigma^2}. \tag{5.58}$$

Funkcja gęstości $f(\chi^2)$ ma postać:

$$f_r(\chi^2) = \frac{(\chi^2)^{\frac{r}{2}-1}}{2^{\frac{r}{2}}\Gamma\left(\frac{r}{2}\right)}\exp\left(-\frac{r}{2}\chi^2\right), \tag{5.59}$$

gdzie: $r = n - 1$ stopni swobody.

Wartość oczekiwana i wariancja zmiennej losowej chi-kwadrat o r stopniach swobody wynoszą odpowiednio:

$$E(\chi_r^2) = r = n-1, \tag{5.60}$$

$$D^2(\chi_r^2) = 2r = 2(n-1). \tag{5.61}$$

Rozkład chi-kwadrat – podobnie jak rozkład Studenta – jest stablicowany. Najczęściej w tablicach podaje się prawdopodobieństwa relacji (zob. tablica III):

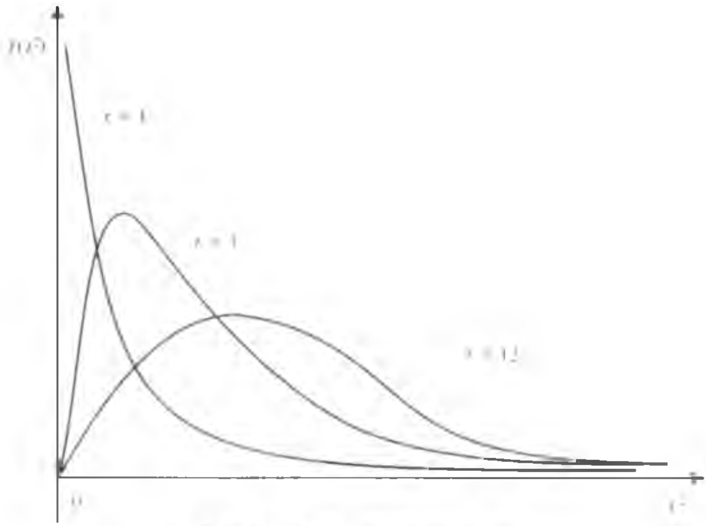
$$P(\chi^2 \geq \chi^2_{\alpha,r}) = \alpha, \tag{5.62}$$

lub

$$P(\chi^2 < \chi^2_{\alpha,r}) = 1-\alpha, \tag{5.63}$$

gdzie: $\alpha \in [0; 1]$.

Rozkład chi-kwadrat jest rozkładem prawostronnie asymetrycznym (zob. rys. 31), a jego asymetria maleje wraz ze wzrostem liczby stopni swobody.



Rys. 31. Rozkład chi-kwadrat

Źródło: opracowanie własne.

Dowodzi się, że wraz ze wzrostem liczby stopni swobody do nieskończoności rozkład chi-kwadrat zmierza do rozkładu normalnego o parametrach $\mu = 0$ i $\sigma = 1$. Zbieżność ta nie jest jednak tak szybka jak w przypadku rozkładu Studenta, dlatego też w praktyce wykorzystuje się zmienną losową $\sqrt{2\chi_r^2}$. Można dowieść, że jeżeli liczba stopni swobody $r \to \infty$, to wówczas zmienna $U = \sqrt{2\chi_r^2} - \sqrt{2r-1}$ ma asymptotyczny rozkład normalny $N(0,1)$.

5.6. Przedziałowa estymacja podstawowych parametrów zmiennej losowej

5.6.1. Uwagi wstępne

Jak już wspomniano w podrozdziale 5.1, drugim sposobem estymacji nieznanego parametru populacji Q jest estymacja przedziałowa. Estymacja przedziałowa, polega na konstrukcji przedziału liczbowego (tzw. przedziału ufności lub przedziału ufności Neymana[10]), który z zadanym prawdopodobieństwem $1 - \alpha$ pokrywa wartość nieznanego parametru Q. Można, więc zapisać:

[10] Nazwa wywodzi się od nazwiska polskiego matematyka i statystyka Jerzego Spławy-Neymana (1894–1980).

$$P(H_d(Q) < Q < H_g(Q)) = 1 - \alpha, \tag{5.64}$$

gdzie:

$H_d(Q)$, $H_g(Q)$ – zmienne losowe o realizacjach $h_d(Q)$ i $h_g(Q)$ decydujące o poziomie dolnej i górnej granicy przedziału ufności,

$1 - \alpha$ – współczynnik ufności.

Współczynnik ufności $1 - \alpha$ jest wartością dowolnie bliską jedności, ale różną od jedności (najczęściej 0,9; 0,95; 0,99; 0,995; 0,975 itp.), przy czym im współczynnik ufności jest bliższy jedności, tym większa jest wiarygodność otrzymanego szacunku. Granice przedziału ufności $h_d(Q)$ i $h_g(Q)$, w odróżnieniu szacowanego parametru Q należy traktowac jako realizacje zmiennych losowych $H_d(Q)$ i $H_g(Q)$, których rozkłady są zależne od parametru Q. Interpretując wyznaczony przedział ufności, należy stwierdzić, że przedział liczbowy o końcach $h_d(Q)$ i $h_g(Q)$, jest jednym z tych przedziałów otrzymanych z różnych prób, które to przedziały mają tę własność, że z dużym wynoszącym $1 - \alpha$ prawdopodobieństwem, pokrywają prawdziwą wartość parametru Q populacji generalnej. Oznacza to, że częstość przypadków, w których otrzymany przedział nie będzie pokrywał parametru Q wynosi α. W praktyce podczas formułowania interpretacji dopuszcza się pominięcie frazy: „jest jednym z tych przedziałow..." i wowczas brzmi ona: „przedział o końcach $h_d(Q)$ i $h_g(Q)$, pokrywa z ufnością $1 - \alpha$ prawdziwą wartość parametru Q populacji generalnej"[11].

5.6.2. Przedziałowa estymacja wartości oczekiwanej

Konstrukcja przedziału ufności dla wartości oczekiwanej (średniej) μ jest uzależniona od założeń dotyczących typu rozkładu zmiennej losowej X w populacji generalnej oraz od znajomości odchylenia standardowego (wariancji) populacji a także od liczebności pobranej losowej próby. W praktyce rozróżnia się najczęściej trzy zasadnicze sytuacje (modele sytuacyjne):

Model I. Zmienna losowa X opisująca populację ma rozkład normalny lub zbliżony do normalnego, z nieznaną wartością oczekiwaną μ i znanym odchyleniem standardowym σ.

Model II. Zmienna losowa X opisująca populację ma rozkład normalny lub zbliżony do normalnego, z nieznaną wartoscią oczekiwaną μ i nieznanym odchyleniem standardowym σ.

Model III. Zmienna losowa X opisująca populację ma nieznany rozkład, z nieznaną wartością oczekiwaną μ i nieznanym odchyleniem standardowym σ.

Model I. Załóżmy, że próba losowa $X_1, ..., X_n$, została pobrana z populacji generalnej o rozkładzie $N(\mu, \sigma)$, przy czym srednia μ jest nieznana, natomiast znane jest odchylenie standardowe σ. Przyjmijmy, że estymatorem parametru μ jest średnia

[11] Zob. np. J. Greń: *Statystyka matematyczna, modele i zadania*, PWN, Warszawa 1974, s. 24 i dalsze.

arytmetyczna z próby $\bar{X}_n \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ (zob. wzory (5.39) i (5.40)). Przypomnijmy, że w wyniku standaryzacji zmiennej $\bar{X}_n$, otrzymujemy statystykę U o rozkładzie $N(0,1)$ o wartościach (zob. wzór (5.41)):

$$u = \frac{\bar{x}_n - \mu}{\sigma}\sqrt{n}. \tag{5.65}$$

Przyjmując określony współczynnik ufności $1 - \alpha$ i korzystając z tablic dystrybuanty rozkładu normalnego standaryzowanego, odczytujemy taką liczbę $u_{\alpha/2}$, dla której spełniona jest relacja:

$$P(-u_{\alpha/2} < U < u_{\alpha/2}) = 1 - \alpha. \tag{5.66}$$

Po podstawieniu do wzoru (5.66) w miejsce u wyrażenia (5.65) otrzymamy:

$$P\left(-u_{\alpha/2} < \frac{\bar{X}_n - \mu}{\sigma}\sqrt{n} < u_{\alpha/2}\right) = 1 - \alpha. \tag{5.67}$$

Po przekształceniu relacji (5.67) w taki sposób, aby pośrodku nierówności znalazł się parametr μ, otrzymujemy:

$$P\left(\bar{X}_n - u_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X}_n + u_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha. \tag{5.68}$$

Tym sposobem otrzymaliśmy wzór pozwalający na przedziałowe szacowanie wartości oczekiwanej, przy założeniu, że populacja ma rozkład normalny ze znaną wartością odchylenia standardowego σ.

Model II. Pozostawiając założenie z modelu I o normalności rozkładu zbiorowości generalnej, uchylmy teraz założenie, że znane jest odchylenie standardowe tego rozkładu. Jak pamiętamy z podrozdziału 5.4.3, w przypadku, gdy nie znamy rzeczywistej wartości odchylenia standardowego i nieznana jest wartość oczekiwana populacji μ, to wówczas należy oszacować odchylenie standardowe, korzystając alternatywnie z estymatora S lub S^*. Podstawą budowy przedziału ufności będzie wówczas charakterystyka (zob. wzór 5.48):

$$t_r = \frac{\bar{x}_n - \mu}{s}\sqrt{n-1} = \frac{\bar{x}_n - \mu}{s^*}\sqrt{n}, \tag{5.69}$$

będąca realizacją zmiennej losowej o rozkładzie Studenta z $r = n - 1$ stopniami swobody. Podobnie jak w modelu I, dla ustalonego współczynnika ufności $1 - \alpha$ oraz liczby stopni swobody $r = n - 1$, z tablic rozkładu Studenta, można odczytać taką wartość $t_{r,\alpha/2}$, przy których spełniona będzie zależność:

$$P(-t_{r,\alpha/2} < t_r < t_{r,\alpha/2}) = 1 - \alpha. \tag{5.70}$$

Jeżeli podczas estymacji odchylenia standardowego został użyty estymator S^*, to wówczas relację (5.70) należy przekształcić do postaci:

$$P\left(-t_{r,\alpha/2} < \frac{\overline{X}_n - \mu}{S^*}\sqrt{n} < t_{r,\alpha/2}\right) = 1 - \alpha, \tag{5.71}$$

natomiast jeśli do szacowania zastosowano statystykę S, to wówczas relacja ta będzie miała postać:

$$P\left(-t_{r,\alpha/2} < \frac{\overline{X}_n - \mu}{S}\sqrt{n-1} < t_{r,\alpha/2}\right) = 1 - \alpha. \tag{5.72}$$

Dokonując prostych przekształceń, relacje (5.71) i (5.72) można zapisać następująco:

$$P\left(\overline{X}_n - t_{r,\alpha/2}\frac{S^*}{\sqrt{n}} < \mu < \overline{X}_n + t_{r,\alpha/2}\frac{S^*}{\sqrt{n}}\right) = 1 - \alpha. \tag{5.73}$$

$$P\left(\overline{X}_n - t_{r,\alpha/2}\frac{S}{\sqrt{n-1}} < \mu < \overline{X}_n + t_{r,\alpha/2}\frac{S}{\sqrt{n-1}}\right) = 1 - \alpha. \tag{5.74}$$

Jeśli opieramy się na tych samych wynikach z próby, to wzory (5.73) i (5.74) dają identyczne przedziały ufności.

Model III. Załóżmy obecnie, że nie znamy rozkładu populacji, z której została pobrana próba. Do oszacowania wartości oczekiwanej populacji została użyta średnia arytmetyczna z próby $\overline{X}_n$. Jeżeli liczebność próby, pobieranej z populacji o średniej μ i odchyleniu standardowym σ wzrasta do nieskończoności ($n > 30$) to, zgodnie z centralnym twierdzeniem granicznym, rozkład z próby średnich zbliża się do rozkładu normalnego ze średnią μ i odchyleniem standardowym $\frac{\sigma}{\sqrt{n}}$. Gdy znana jest rzeczywista wartość odchylenia standardowego σ, to wówczas podczas tworzenia przedziału ufności można skorzystać ze statystyki U (5.65) i otrzymać przedział o wzorze identycznym jak (5.68). Jeżeli natomiast nie znamy wartości parametru σ, to przy dużej próbie można przyjąć, że w przybliżeniu jest on równy szacunkowi otrzymanemu przy użyciu estymatora S ($s \approx \sigma$) i przedział ufności zbudować na podstawie wzoru:

$$P\left(\overline{X}_n - u_{\alpha/2}\frac{S}{\sqrt{n}} < \mu < \overline{X}_n + u_{\alpha/2}\frac{S}{\sqrt{n}}\right) = 1 - \alpha. \tag{5.75}$$

Jak łatwo się domyślić, nic nie stoi na przeszkodzie, aby wzór (5.75) można było stosować dla przypadków z modelu II, przy dodatkowym założeniu, że próba jest dostatecznie duża. Jak już zostało wcześniej wspomniane rozkład Studenta można dobrze aproksymować rozkładem normalnym, gdy próba jest dostatecznie liczna (przynajmniej 30 elementów).

Przykład 5.5

W bardzo dużej grupie studentow (zob. przykład 3.7) przeprowadzono egzamin z „Zarządzania jakością” – mierząc wyniki na ciągłej skali od 0 do 40 punktów. Ustalono, że rozkład punktów z egzaminu jest w przybliżeniu normalny z odchyleniem standardowym wynoszącym $\sigma = 6{,}4$. Z grupy tej wybrano prostą próbę losową o liczności $n = 25$ studentów. Średnia liczba punktów w badanej grupie wynosiła $\bar{x}_{25} = 30{,}5$. Zakładając współczynnik ufności $1 - \alpha = 0{,}95$, oszacować przedział ufności pokrywający nieznaną średnią liczbę punktów w całej populacji.

Ponieważ znamy rzeczywistą wartość odchylenia standardowego populacji σ oraz wiemy, że zmienna losowa X – oznaczająca liczbę punktów – ma w przybliżeniu rozkład normalny, dlatego do zbudowania przedziału ufności można wykorzystać wzór (5.68) z modelu I.

Ponieważ $1 - \alpha = 0{,}95$ to $\alpha = 0{,}05$ i $\alpha/2 = 0{,}025$ oraz $1 - \alpha/2 = 0{,}975$. Z tablic dystrybuanty rozkładu normalnego odczytujemy, że $\Phi(u_{\alpha/2} = 1{,}96) = 0{,}975$.

Podstawiając odpowiednie liczby do wzoru (5.68), otrzymamy:

$$30{,}5 - 1{,}711\frac{6{,}5}{\sqrt{25}} < \mu < 30{,}5 + 1{,}711\frac{6{,}5}{\sqrt{25}}$$

z czego

$$27{,}99 < \mu < 33{,}01.$$

Ten przedział o końcach 27,99; 33,01 jest jednym ze wszystkich możliwych do otrzymania przedziałów, który z prawdopodobieństwem 0,95 pokrywa nieznaną średnią liczbę punktów w całej populacji. Ujmując rzecz nieco inaczej, można powiedzieć, że w dużej serii niezależnych 25-elementowych prób, w 95 przypadkach na 100, oszacowany przedział będzie pokrywać nieznaną średnią liczbę punktów w populacji, natomiast w 5 przypadkach na 100 można spodziewać się, że oszacowanie będzie błędne.

Przykład 5.6

Przyjmując dane wejściowe z przykładu 5.5 oraz zakładając, że nie znamy odchylenia standardowego całej populacji, a jedynie odchylenie standardowe z badanej próby 25-elementowej $s^* = 6{,}5$, oszacować przedział ufności pokrywający z prawdopodobieństwem 0,9 nieznaną średnią liczbę punktów w całej populacji studentow.

Ponieważ nie jest znana rzeczywista wartość odchylenia standardowego w całej zbiorowości generalnej, a jedynie jej punktowy szacunek (został użyty estymator S^*), oraz próba losowa jest statystycznie mała, to do oszacowania przedziału ufności należy użyć wzór (5.73).

Przyjmując $1 - \alpha = 0,9$, możemy wnioskować, że $\alpha = 0,1$, oraz $\alpha/2 = 0,05$. Z tablic rozkładu Studenta, przy przyjęciu $r = n - 1 = 24$ stopni swobody odczytujemy $t_{r,\alpha/2} = 1,711$. Podstawiając wartości do wzoru (5.70), otrzymujemy:

$$30,5 - 1,711\frac{6,5}{\sqrt{25}} < \mu < 30,5 + 1,711\frac{6,5}{\sqrt{25}},$$

$$28,276 < \mu < 32,724.$$

W dużej serii niezależnych 25-elementowych prób, w 90 przypadkach na 100 otrzymany przedział będzie pokrywał nieznaną średnią liczbę punktów w całej populacji studentów, a tylko w 10 przypadkach na 100 możemy otrzymać oszacowanie błędne.

Przykład 5.7

W przykładzie tym wykorzystamy dane, z przykładu 2.2 z podręcznika M. Major, J. Niezgoda, *Elementy statystyki, Cz. I Statystyka opisowa*, przedstawiające liczbę osób korzystających z Biblioteki Miejskiej w jednym z miast województwa podkarpackiego. Dane te potraktujemy obecnie jako losowy podzbiór większej zbiorowości generalnej. Przypomnijmy, że obserwację liczby osób odwiedzających bibliotekę prowadzono w ciągu $n = 100$ dni roboczych i otrzymano średnią arytmetyczną $\bar{x}_{100} = 76,1$, oraz odchylenie standardowe z próby $s = 18,74$. Naszym zadaniem jest zbudowanie 95 procentowego przedziału ufności pokrywającego nieznaną średnią dzienną liczbę osób korzystających z biblioteki. Ponieważ próba losowa jest duża i nie jest znana rzeczywista wartość odchylenia standardowego populacji, to do wyznaczenia przedziału ufności należy wykorzystać procedury opisane w modelu III (wzór 5.75)

Z tablic dystrybuanty rozkładu normalnego odczytujemy, że $\Phi(u_{\alpha/2} = 1,96) = 0,975$. Szukany przedział ufności będzie miał postać:

$$76,1 - 1,96\frac{18,74}{\sqrt{100}} < \mu < 76,1 + 1,96\frac{18,74}{\sqrt{100}},$$

a więc

$$72,43 < \mu < 79,77.$$

Można zatem stwierdzić, że przedział o końcach 72,43; 79,77 z prawdopodobieństwem 0,95 będzie pokrywał nieznaną średnią dzienną liczbę osób odwiedzających bibliotekę. {kp}

5.6.3. Przedziałowa estymacja wariancji i odchylenia standardowego

Załóżmy, że rozpatrujemy populację generalną o rozkładzie normalnym $N(\mu, \sigma)$ o nieznanych parametrach μ i σ. Z populacji została pobrana próba losowa prosta $X_1, \ldots, X_n$, na podstawie której należy zbudować przedział ufności pokrywający z prawdopodobieństwem $1 - \alpha$ nieznaną wartość parametru σ^2. Sposób budowy przedziału ufności jest uzależniony od rodzaju estymatora użytego do oszacowania wariancji populacji oraz od liczebności próby. Można wyróżnić dwa modele sytuacyjne:

Model I. Zmienna losowa X opisująca populację ma rozkład normalny lub zbliżony do normalnego, z nieznaną wartością oczekiwaną μ i nieznanym odchyleniem standardowym σ. Próba losowa jest mała ($n \leq 30$).

Model II. Zmienna losowa X opisująca populację ma rozkład normalny lub zbliżony do normalnego, z nieznaną wartością oczekiwaną μ i nieznanym odchyleniem standardowym σ. Próba losowa jest duża ($n > 30$).

W przypadku **modelu I** budowa przedziału ufności opiera na się na założeniu, że statystyka (5.57) ma rozkład chi-kwadrat o $r = n - 1$ stopniach swobody. W konsekwencji wartość tej statystyki można wyznaczyć jako:

$$\chi_r^2 = \frac{s^{*2}(n-1)}{\sigma^2} = \frac{s^2 n}{\sigma^2}. \tag{5.76}$$

Przy przyjętym współczynniku ufności $1 - \alpha$, oraz $r = n - 1$ stopniach swobody, z tablic rozkładu chi-kwadrat należy odczytać takie dwie wartości $\chi^2_{r,\alpha/2}$, $\chi^2_{r,1-\alpha/2}$, dla których spełniona jest zależność:

$$P(\chi^2_{r,1-\alpha/2} < \chi^2_r < \chi^2_{r,\alpha/2}) = 1 - \alpha. \tag{5.77}$$

Jeżeli estymatorem parametru σ^2 jest wariancja z próby S^{*2}, to wówczas wzór (5.77) należy zastąpić wzorem:

$$P\left(\chi^2_{r,1-\alpha/2} < \frac{S^{*2}(n-1)}{\sigma^2} < \chi^2_{r,\alpha/2}\right) = 1 - \alpha, \tag{5.78}$$

z czego po przekształceniu otrzymamy:

$$P\left(\frac{S^{*2}(n-1)}{\chi^2_{r,\alpha/2}} < \sigma^2 < \frac{S^{*2}(n-1)}{\chi^2_{r,1-\alpha/2}}\right) = 1 - \alpha. \tag{5.79}$$

Natomiast, jeżeli estymatorem parametru σ^2 jest wariancja z próby S^2, to wówczas otrzymamy:

$$P\left(\chi^2_{r,1-\alpha/2} < \frac{S^2 n}{\sigma^2} < \chi^2_{r,\alpha/2}\right) = 1 - \alpha, \tag{5.80}$$

lub po przekształceniach:

$$P\left(\frac{S^2 n}{\chi^2_{r,\alpha/2}} < \sigma^2 < \frac{S^2 n}{\chi^2_{r,1-\alpha/2}}\right) = 1 - \alpha. \tag{5.81}$$

Wzory (5.79) i (5.81) dają jednakowe oszacowania, pod warunkiem że opierają się na tej samej losowej próbie. W celu otrzymania przedziału ufności dla odchylenia standardowego σ, wystarczy wyciągnąć kwadratowy pierwiastek ze wszystkich członów nierówności podwójnej znajdujących się pod znakiem prawdopodobieństwa we wzorach (5.79) i (5.81).

W modelu II zakładamy, że liczebność pobranej próby jest statystycznie duża (tzn. $n > 30$). Przedział ufności dla odchylenia standardowego populacji generalnej σ, budujemy na podstawie założenia, że odchylenie standardowe z próby S, pobrane z populacji o rozkładzie normalnym ma graniczny rozkład normalny o parametrach:

$$E(S) = \sigma, \quad D(S) = \frac{\sigma}{\sqrt{2n}}. \tag{5.82}$$

Jak pamiętamy z rozdziału 5.5.2, statystyka:

$$U = \sqrt{2\chi_r^2} - \sqrt{2r-1}, \tag{5.83}$$

przy liczbie stopni swobody $r \to \infty$ ma rozkład asymptotycznie zbieżny do rozkładu normalnego $N(0,1)$. Wartości $\chi^2_{r,1-\alpha/2}$ i $\chi^2_{r,\alpha/2}$ można wyznaczyć, wykorzystując zależności:

$$\sqrt{2\chi^2_{r,1-\alpha/2}} - \sqrt{2r-1} = -u_{\alpha/2}, \tag{5.84}$$

$$\sqrt{2\chi^2_{r,\alpha/2}} - \sqrt{2r-1} = u_{\alpha/2}, \tag{5.85}$$

gdzie: $r = n - 1$, $u_{\alpha/2}$ jest wartością odczytaną z tablic dystrybuanty rozkładu $N(0,1)$.

Standaryzując odchylenie standardowe z próby S, otrzymamy statystykę U o wartościach:

$$u = \frac{s-\sigma}{\frac{\sigma}{\sqrt{2n}}} = \frac{s-\sigma}{\sigma}\sqrt{2n}, \tag{5.86}$$

która ma rozkład asymptotycznie normalny o parametrach $\mu = 0$ i $\sigma = 1$.

Podstawiając wzór (5.86) do zależności (5.66), a następnie go odpowiednio przekształcając, otrzymamy:

$$P\left(u_{\alpha/2} < \frac{S-\sigma}{\sigma}\sqrt{2n} < u_{\alpha/2}\right) = 1-\alpha, \tag{5.87}$$

$$P\left(S - u_{\alpha/2}\frac{S}{\sqrt{2n}} < \sigma < S + u_{\alpha/2}\frac{S}{\sqrt{2n}}\right) = 1-\alpha. \tag{5.88}$$

Wzór (5.88) jest szukaną formułą budowy przedziału ufności dla parametru σ, przy założeniu że $n > 30$ oraz rozkład populacji jest normalny.

Przykład 5.8

Uwzględniając dane z przykładu 5.6, oszacować 95-procentowy przedział ufności pokrywający wariancję liczby punktów z egzaminu dla wszystkich studentów zdających egzamin z „Zarządzania jakością". Przypomnijmy, że próba losowa liczyła $n = 25$ osób, średnia liczba punktów w badanej grupie wynosiła $\bar{x}_{25} = 30,5$, natomiast odchylenie standardowe z próby $s^* = 6,5$. Ponieważ próba jest statystycznie mała ($n < 30$), dlatego w czasie konstrukcji przedziału ufności skorzystamy z własności (5.79). Przyjmując współczynnik ufności na poziomie $1 - \alpha = 0,95$, oraz liczbę stopni swobody $r = n - 1 = 24$, z tablic rozkładu chi-kwadrat (zob. tablica III) odczytujemy wartości $\chi^2_{r,\alpha/2=0,025} = 39,364$ oraz $\chi^2_{r,1-\alpha/2=0,975} = 12,401$. Po podstawieniu danych do wzoru (5.76), otrzymamy:

$$\frac{(6,5)^2 \cdot 24}{39,364} < \sigma^2 < \frac{(6,5)^2 \cdot 24}{12,401},$$

a stąd

$$25,76 < \sigma^2 < 81,77.$$

Przedział o końcach 25,76; 81,77 z prawdopodobieństwem 0,95 będzie zawierał nieznaną wartość wariancji liczby punktów dla wszystkich studentów zdających egzamin z „Zarządzania jakością". {kp}

Przykład 5.9

Korzystając z danych z przykładu 5.7, oszacować przedział ufności dla odchylenia standardowego dziennej liczby osób odwiedzających bibliotekę. Podczas konstrukcji przedziału założyć 90 procentowy współczynnik ufności.

Ponieważ tym razem próba $n = 100$ jest duża, dlatego przedział ufności dla odchylenia standardowego populacji możemy oszacować, wykorzystując własność (5.88).

Przypomnijmy, że średnia z próby i odchylenie standardowe z próby wynoszą Odpowiednio: $\bar{x}_{100} = 76,1$ oraz $s = 18,74$. Wartość $u_{\alpha/2}$ występującą we wzorze (5.88) wyznaczamy, korzystając z tablic dystrybuanty rozkładu $N(0,1)$. W tym przypadku jest ona równa $u_{\alpha/2} = 1,64$. Po podstawieniu wszystkich danych do wzoru (5.88) otrzymamy przedział:

$$18,74 - 1,64\,\frac{18,74}{\sqrt{200}} < \sigma < 18,74 + 1,64\,\frac{18,74}{\sqrt{200}},$$

$$16,57 < \sigma < 20,91,$$

który z prawdopodobieństwem $1 - \alpha = 0,9$ pokrywa odchylenie standardowe dziennej liczby osób odwiedzających bibliotekę. {kp}

5.6.4. Przedziałowa estymacja frakcji (wskaźnika struktury)

Załóżmy obecnie, że zbiorowość generalna będzie opisywana przez zmienną losową X o rozkładzie zero-jedynkowym, dla którego $P(X = 1) = p$. Naszym zadaniem jest oszacowanie parametru p, przy założeniu, że oznacza on wskaźnik struktury (frakcję) zbiorowości generalnej. Jak pamiętamy z podrozdziału 5.3.1, estymatorem parametru p w zbiorowości generalnej jest wskaźnik struktury:

$$p = W_n = \frac{Z}{n}, \tag{5.89}$$

gdzie:

$Z = \sum_{i=1}^{n} X_i$ jest sumą n zero-jedynkowych zmiennych losowych, których realizacje otrzymuje się poprzez badanie kolejnych elementów próby.

Wartość oczekiwana oraz wariancja statystyki p wynoszą odpowiednio (zob. wzory: (3.66) i (3.67)):

$$E\left(p\right) = \frac{E(Z)}{n} = \frac{np}{n} = p, \tag{5.90}$$

$$D^2\left(p\right) = \frac{D^2(Z)}{n^2} = \frac{npq}{n^2} = \frac{pq}{n}. \tag{5.91}$$

Zgodnie z twierdzeniem granicznym Moivre'a-Laplace'a statystyka $p = W_n$ ma asymptotyczny rozkład normalny z parametrami: $\mu = p$ i $\sigma = \sqrt{\frac{pq}{n}}$.

Standaryzując estymator $p = W_n$, otrzymamy:

$$u = \frac{w - p}{\sqrt{\frac{w(1-w)}{n}}}, \tag{5.92}$$

gdzie: $w = z/n$ jest wartością statystyki uzyskana w rezultacie badania próby.

Wyrażenie (5.92) jest realizacją standaryzowanej zmiennej losowej U o rozkładzie $N(0,1)$.

Przy ustalonym współczynniku ufnosci $1 - \alpha$, z tablic dystrybuanty rozkładu $N(0,1)$, należy odczytać takie wartości $u_{\alpha/2}$ i $-u_{\alpha/2}$, dla których spełniona jest zależność:

$$P\left(-u_{\alpha/2} < \frac{W_n - p}{\sqrt{\frac{W_n(1-W_n)}{n}}} < u_{\alpha/2}\right) = 1 - \alpha. \tag{5.93}$$

Dokonując odpowiednich przekształceń, w taki sposób, aby po środku nierówności (5.90) znalazł się parametr p, otrzymamy relację:

$$P\left(W_n - u_{\alpha/2}\sqrt{\frac{W_n(1-W_n)}{n}} < p < W_n + u_{\alpha/2}\sqrt{\frac{W_n(1-W_n)}{n}}\right) = 1-\alpha, \tag{5.94}$$

będącą przybliżonym wzorem na przedział ufności dla parametru p w zbiorowości generalnej, a przybliżenie to jest tym lepsze, im liczniejsza będzie pobrana próba losowa[12].

Przykład 5.10

Z populacji studentów zdających egzamin z „Zarządzania jakością" pobrano losową próbę liczącą $n = 100$ osób i stwierdzono, że $z = 12$ osób otrzymało ocenę niedostateczną. Przyjmując poziom ufności $1 - \alpha = 0{,}9$, oszacować przedziałowo frakcję studentów, którzy w badanej populacji nie zdali egzaminu.

W pierwszym kroku postępowania należy obliczyć frakcję studentów, którzy w badanej próbie nie zdali egzaminu. W analizowanym przypadku frakcja ta wyniesie: $w = z/n = 12/100 = 0{,}12$.

Z tablic dystrybuanty rozkładu normalnego odczytamy, że $u_{\alpha/2} = 1{,}64$. Po podstawieniu wszystkich danych do wzoru (5.94) otrzymamy przedział:

$$0{,}12 - 1{,}64\sqrt{\frac{0{,}12\cdot 0{,}88}{100}} < p < 0{,}12 + 1{,}64\sqrt{\frac{0{,}12\cdot 0{,}88}{100}},$$

$$0{,}07 < p < 0{,}17,$$

który z 90-procentową ufnością będzie zawierał frakcję studentów całej populacji, którzy nie zdali pozytywnie egzaminu z „Zarządzania jakością". {kp}

5.6.5. Przedziałowa estymacja różnicy między dwiema wartościami oczekiwanymi

Załóżmy, że rozważamy dwie populacje opisywane przez dwie zmienne losowe X_1 i X_2, przy czym $X_1{\sim}N(\mu_1, \sigma_1)$ oraz $X_2{\sim}N(\mu_2, \sigma_2)$. Różnica pomiędzy tymi zmiennymi $X_1 - X_2$ jest również zmienną losową o rozkładzie normalnym $N\left(\mu_1 - \mu_2;\ \sqrt{\sigma_1^2 + \sigma_2^2}\right)$. Z populacji pobiera się dwie próby losowe o liczności n_1 i n_2

[12] Zwykle zaleca się próbę składającą się co najmniej ze 100 elementów. Zob. np. J. Greń: *Statystyka..., op. cit.*, s. 35.

i oblicza się z nich średnie arytmetyczne $\overline{X}_1$ i $\overline{X}_2$, przy czym $\overline{X}_1 \sim N\left(\mu_1; \frac{\sigma_1}{\sqrt{n_1}}\right)$, oraz $\overline{X}_2 \sim N\left(\mu_2; \frac{\sigma_2}{\sqrt{n_2}}\right)$. W konsekwencji różnica średnich: $\overline{X}_1 - \overline{X}_2$ ma również rozkład normalny o parametrach: $\mu_1 - \mu_2$ i $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$. Dokonując standaryzacji zmiennej losowej $\overline{X}_1 - \overline{X}_2$ – otrzymujemy następujące wyrażenie:

$$u = \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}, \tag{5.95}$$

gdzie:

u – jest realizacją zmiennej losowej $U \sim N(0,1)$.

Postępując podobnie jak podczas konstrukcji przedziału ufności dla jednej wartości oczekiwanej i wykorzystując relację (5.66) oraz dokonując niezbędnych przekształceń możemy zapisać:

$$P\left(-u_{\alpha/2} < \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} < u_{\alpha/2}\right) = 1 - \alpha, \tag{5.96}$$

z czego:

$$P\left((\overline{X}_1 - \overline{X}_2) - u_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < (\overline{X}_1 - \overline{X}_2) + u_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right) = 1 - \alpha. \tag{5.97}$$

Wzór (5.97) można stosować do estymacji różnicy dwóch wartości oczekiwanych tylko wówczas, gdy znane są wariancje badanych populacji σ_1^2 i σ_2^2. Jeżeli ich nie znamy, to należy je oszacować za pomocą estymatora S^{*2} i wyrażenie (5.95) zastąpić wzorem[13]:

$$t_r = \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{s_p\sqrt{\frac{n_1 + n_2}{n_1 n_2}}}, \tag{5.98}$$

gdzie:

$r = n_1 + n_2 - 2$, natomiast s_p jest tzw. uogólnionym oszacowaniem odchylenia standardowego $\sqrt{\sigma_1^2 + \sigma_2^2}$ i wynosi:

[13] Zob. A. Iwasiewicz, Z. Paszek: *Statystyka...*, *op. cit.*, s. 195.

$$s_p = \sqrt{\frac{(n_1-1)s_1^{*2} + (n_2-1)s_2^{*2}}{(n_1-1)+(n_2-1)}} \tag{5.99}$$

Korzystając z zależności (5.70), można zapisać:

$$P\left(-t_{r,\alpha/2} < \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{s_p\sqrt{\frac{n_1+n_2}{n_1 n_2}}} < t_{r,\alpha/2} \right) = 1-\alpha, \tag{5.100}$$

z czego po odpowiednich przekształceniach otrzymamy:

$$P\left((\overline{X}_1 - \overline{X}_2) - t_{r,\alpha/2}\, s_p\sqrt{\frac{n_1+n_2}{n_1 n_2}} < \mu_1 - \mu_2 < (\overline{X}_1 - \overline{X}_2) + t_{r,\alpha/2}\, s_p\sqrt{\frac{n_1+n_2}{n_1 n_2}} \right) = 1-\alpha. \tag{5.101}$$

5.6.6. Przedziałowa estymacja stosunku dwóch wariancji

Załóżmy, podobnie jak we wcześniejszym podrozdziale, że rozpatrujemy dwie zmienne losowe X_1 i X_2, opisujące dwie populacje, przy czym $X_1 \sim N(\mu_1, \sigma_1)$ oraz $X_2 \sim N(\mu_2, \sigma_2)$.

Jeżeli podczas porównania dwóch populacji opieramy się na ilorazie wariancji tych populacji, to można wykazać, że statystyka:

$$\frac{\chi^2_{r_1}}{r_1} : \frac{\chi^2_{r_2}}{r_2} = \frac{(n_1-1)S_1^{*2}}{(n_1-1)\sigma_1^2} : \frac{(n_2-1)S_2^{*2}}{(n_2-1)\sigma_2^2} = \frac{S_1^{*2}}{\sigma_1^2} : \frac{S_2^{*2}}{\sigma_2^2} = \frac{S_1^{*2}}{S_2^{*2}} \cdot \frac{\sigma_2^2}{\sigma_1^2}, \tag{5.102}$$

ma rozkład F – **Snedecora** o $r_1 = n_1 - 1$ i $r_2 = n_2 - 1$ stopniach swobody. Rozkład F – Snedecora jest stablicowany (zob. tablice IVa i IVb). Przyjmując r_1 i r_2 stopnie swobody, oraz zakładając współczynnik ufności na poziomie $1 - \alpha$, z tablic rozkładu F – Snedecora, należy odczytać takie dwie wartości $F_{r_1, r_2, 1-\alpha/2}$, $F_{r_1, r_2, \alpha/2}$, dla których spełniona jest zależność:

$$P\left(F_{r_1, r_2, 1-\alpha/2} < \frac{S_1^{*2}}{\sigma_1^2} : \frac{S_2^{*2}}{\sigma_2^2} < F_{r_1, r_2, \alpha/2} \right) = 1-\alpha, \tag{5.103}$$

$$P\left(\frac{\frac{S_1^{*2}}{S_2^{*2}}}{F_{r_1, r_2, \alpha/2}} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{\frac{S_1^{*2}}{S_2^{*2}}}{F_{r_1, r_2, 1-\alpha/2}} \right) = 1-\alpha. \tag{5.104}$$

Otrzymany wzór (5.104) pozwala na przedziałowe oszacowanie ilorazu dwóch wariancji opisujących dwie różne populacje o rozkładach normalnych.

5.6.7. Przedziałowa estymacja różnicy między dwoma wskaźnikami struktury

Załóżmy, że chcemy porównać dwie populacje opisywane, przez dwie zmienne losowe X_1 i X_2 o rozkładach zero-jedynkowych z parametrami wynoszącymi kolejno: p_1 oraz p_2. Z populacji pobrano dostatecznie liczne próby n_1 i n_2. Korzystając w wzoru (5.86) wyznaczamy wartości wskaźników w pobranych próbach wynoszące odpowiednio $w_1 = z_1/n_1$ i $w_2 = z_2/n_2$. Aby oszacować przedział ufności dla różnicy między dwoma wskaźnikami $p_1 - p_2$, należy skorzystać z wzoru:

$$u = \frac{(w_1 - w_2) - (p_1 - p_2)}{\sqrt{\dfrac{w_1(1-w_1)}{n_1} + \dfrac{w_2(1-w_2)}{n_2}}}, \tag{5.105}$$

gdzie: u jest realizacją zmiennej losowej U o rozkładzie normalnym standaryzowanym.

Wykorzystując relację (5.66) oraz przyjmując określony poziom ufności $1 - \alpha$, możemy zapisać:

$$P\left(-u_{\alpha/2} < \frac{(W_1 - W_2) - (p_1 - p_2)}{\sqrt{\dfrac{W_1(1-W_1)}{n_1} + \dfrac{W_2(1-W_2)}{n_2}}} < u_{\alpha/2}\right) = 1 - \alpha. \tag{5.106}$$

Dokonując odpowiednich przekształceń, otrzymamy następujący wzór na przedział ufności dla różnicy dwoch frakcji $p_1 - p_2$:

$$P\left((W_1 - W_2) - u_{\alpha/2}\sqrt{\frac{W_1(1-W_1)}{n_1} + \frac{W_2(1-W_2)}{n_2}} < p_1 - p_2 < (W_1 - W_2) + \right. \tag{5.107}$$

$$\left. + u_{\alpha/2}\sqrt{\frac{W_1(1-W_1)}{n_1} + \frac{W_2(1-W_2)}{n_2}}\right) = 1 - \alpha.$$

5.6.8. Przedziałowa estymacja współczynnika korelacji liniowej

Ze współczynnikiem korelacji liniowej spotkaliśmy się już podczas omawiania rozkładów dwuwymiarowej zmiennej losowej. Przypomnijmy (zob. wzór (2.59)), że współczynnik korelacji liniowej ma postać:

$$r_{xy} = \frac{\operatorname{cov}(X,Y)}{D(X)D(Y)} \tag{5.108}$$

Estymatorem współczynnika korelacji liniowej w populacji jest współczynnik korelacji z próby (R_{xy}), którego realizacje obliczany w następujący sposób[14]:

$$r_{xy} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2}} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}. \qquad (5.109)$$

Jeżeli liczebność próby jest dostatecznie liczna (kilkaset elementów), to wówczas można udowodnić, że rozkład współczynnika korelacji z próby jest asymptotycznie zbieżny do rozkładu normalnego $N\left(\rho_{xy}; \frac{1-\rho_{xy}^2}{\sqrt{n}}\right)$. Dokonując standaryzacji współczynnika korelacji z próby otrzymamy wyrażenie:

$$u = \frac{r_{xy}-\rho_{xy}}{\frac{1-\rho_{xy}^2}{\sqrt{n}}} \qquad (5.110)$$

będące wartością zmiennej $U \sim N(0; 1)$. Korzystając z relacji (5.66), możemy zapisać:

$$P\left(-u_{\alpha/2} < \frac{R_{xy}-\rho_{xy}}{\frac{1-\rho_{xy}^2}{\sqrt{n}}} < u_{\alpha/2}\right) = 1-\alpha. \qquad (5.111)$$

Po dokonaniu niezbędnych przekształceń przedział ufności dla współczynnika korelacji liniowej, będzie przedstawiał się następująco:

$$P\left(r_{xy} - u_{\alpha/2}\frac{1-R_{xy}^2}{\sqrt{n}} < \sigma_{xy} < r_{xy} + u_{\alpha/2}\frac{1-R_{xy}^2}{\sqrt{n}}\right) = 1-\alpha. \qquad (5.112)$$

Przykład 5.11

Z grupy studentów kierunku Towaroznawstwa uczęszczających na zajęcia ze statystyki pobrano losową próbę liczącą 103 osoby. Na podstawie pobranej próby zbadano zależność pomiędzy oceną ze statystyki uzyskaną na koniec pierwszego i drugiego semestru. Obliczony współczynnik korelacji liniowej wyniósł r_{xy} = 0,46. Zakładając poziom ufności $1 - \alpha = 0{,}95$, oszacować przedziałowo współczynnik korelacji pomiędzy oceną ze statystyki w pierwszym i w drugim semestrze w całej populacji studentów studiujących ten przedmiot.

[14] Porównaj ze wzorami (4.3) i (4.4) w: M. Major, J. Niezgoda: *Elementy...*, *op. cit.*

Przedział ufności wyznaczymy, korzystając z relacji (5.112). Po podstawieniu danych do nierówności otrzymamy:

$$0{,}46 - 1{,}96\frac{1-0{,}46^2}{\sqrt{103}} < \sigma_{xy} < 0{,}46 + 1{,}96\frac{1-0{,}46^2}{\sqrt{103}},$$

czyli:

$$0{,}31 < \rho_{xy} < 0{,}61.$$

Można stwierdzić, że w ciągu n niezależnych prób przedział liczbowy o końcach 0,31; 0,61 będzie pokrywał nieznany współczynnik korelacji między oceną z pierwszego i drugiego semestru w 95 przypadkach na 100. W pięciu przypadkach na 100 możemy jednak otrzymać błędne oszacowanie nieznanego parametru.

5.6.9. Szacowanie minimalnej liczebności próby

Na dokładność estymacji można wpływać, odpowiednio dobierając współczynnik ufności $1-\alpha$, a także przez zmianę liczebności próby. Im mniejszy jest współczynnik ufności, tym krótszy jest przedział ufności. Zmniejszenie współczynnika ufności zwiększa jednak prawdopodobieństwo zdarzenia, że otrzymany przedział nie pokryje szacowanego parametru. Poprawę precyzji szacunku możemy również uzyskać, zwiększając liczebności pobranej próby. Pojawia się pytanie, jak liczną należy przyjąć próbę, aby otrzymać przedział ufności o zadanej z góry długości. Naszą analizę rozpoczniemy od sposobu ustalania minimalnej liczności próby potrzebnej do oszacowania wartości oczekiwanej μ w populacji normalnej ze znanym odchyleniem standardowym σ (model I). Jak ustaliliśmy (zob. wzór (5.68)) przedział ten ma postać: $\bar{x}_n - u_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x}_n + u_{\alpha/2}\frac{\sigma}{\sqrt{n}}$, a jego długość wynosi:

$$\bar{x}_n + u_{\alpha/2}\frac{\sigma}{\sqrt{n}} - \left(\bar{x}_n - u_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 2u_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \tag{5.113}$$

Dąży się do tego, aby połowa przedziału ufności $\left(u_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$ określana jako maksymalny błąd szacunku nie przekraczała z góry ustalonej wartości d. Zatem można zapisać:

$$u_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le d, \tag{5.114}$$

z czego po przekształceniu otrzymamy:

$$n \ge \frac{u_{\alpha/2}^2\sigma^2}{d^2}. \tag{5.115}$$

Z powyższego zapisu wynika, że jeżeli chcemy, aby maksymalny błąd szacunku nie przekroczył założonej wartości d, to wówczas należy wylosować próbę nie mniejszą niż n.

W podobny sposób można ustalić minimalną liczebność próby, która jest niezbędna do oszacowania przedziału ufności dla wartości oczekiwanej μ w populacji normalnej z nieznanym odchyleniem standardowym σ (model II).

Zamiast relacji (5.114) zapiszemy wówczas:

$$n > \frac{t_{\alpha,r}^2 s^{*2}}{d^2}, \tag{5.116}$$

gdzie:

$r = n_0 - 1$ stopni swobody,

s^{*2} – jest wartością wariancji obliczonej na podstawie próby pilotażowej o liczności n_o.

Jeżeli minimalna liczebność próby n wyznaczona na podstawie wzoru (5.116) będzie większa niż liczebność próby pilotażowej n_0, to wówczas zachodzi konieczność dolosowania do próby pilotażowej $n - n_0$ elementów.

Na koniec rozważmy problem minimalnej liczności próby potrzebnej do oszacowania frakcji p (wskaźnika struktury). Jak ustaliliśmy (zob. wzór (5.94)), przedział ufności w tym przypadku ma końce opisane wzorami: $w - u_{\alpha/2}\sqrt{\frac{w(1-w)}{n}}$ i $w + u_{\alpha/2}\sqrt{\frac{w(1-w)}{n}}$

Polowa długości przedziału wyniesie (maksymalny błąd szacunku): $u_{\alpha/2}\sqrt{\frac{w(1-w)}{n}}$.

Zatem, zachodzi relacja: $u_{\alpha/2}\sqrt{\frac{w(1-w)}{n}} \le d$, z której po przekształceniu otrzymamy:

$$n > \frac{u_{\alpha/2}^2 \cdot w(1-w)}{d^2}. \tag{5.117}$$

Parametr w we wzorze (5.117) traktuje się tutaj jako przypuszczalny rząd wartości parametru p. W sytuacji, gdy brak jest jakichkolwiek przesłanek do ustalenia rzędu wartości parametru p, to wówczas przyjmuje się że $w = \frac{1}{2}$, czyli wartość, przy której iloczyn $w(1 - w)$ jest największy. Wzór (5.117) należy zastąpić wzorem:

$$n > \frac{\frac{1}{4}u_{\alpha/2}^2}{d^2} = \frac{u_{\alpha/2}^2}{4d^2}. \tag{5.118}$$

Przyjęcie za w wartości $\frac{1}{2}$ gwarantuje, że w większości przypadków (gdy $p \ne \frac{1}{2}$) oszacowana liczebność próby zapewni, że dokładność estymacji będzie większa od założonej.

Przykład 5.12

Oszacować minimalną liczność próby, potrzebną do oszacowania średniej liczby punktów otrzymanych z egzaminu z „Zarządzania jakością", jeżeli założymy, że maksymalny błąd szacunku wynosi 2,5, a poziom ufności $1 - \alpha = 0{,}95$. Załóżmy, że na podstawie wcześniejszych badań generalnych wiadomo, że rozkład uzyskanych punktów jest zbliżony do normalnego, z odchyleniem standardowym $\sigma = 6{,}4$.

Aby ustalić minimalną liczebność próby potrzebną do oszacowania średniej liczby punktów, należy skorzystać ze wzoru (5.115) (znane jest odchylenie standardowe populacji σ). Po podstawieniu do wzoru otrzymamy:

$$n = \frac{u_{\alpha/2}^2\, \sigma^2}{d^2} = \frac{1{,}96^2 \cdot 6{,}4^2}{2{,}5^2} = 25{,}17631.$$

W celu obliczenia średniej liczby punktów z egzaminu z „Zarządzania jakością" należy wylosować niezależną próbę o liczności większej niż 25 osób, czyli minimalnie $n = 26$ osób. {kp}

Przykład 5.13

Z grupy studentów zdających egzamin z „Zarządzania jakością", wylosowano próbę pilotażową o liczności $n_0 = 10$ osób i obliczono wariancję liczby uzyskanych punktow. Wariancja ta wynosiła $s^{*2} = 42{,}25$. Zbadać, czy liczebność pobranej próby pilotażowej jest wystarczająca do oszacowania średniej liczby punktów, jeżeli założymy, że średni błąd szacunku wynosi 2,5, a poziom ufności 0,95.

Ponieważ, nieznana jest wartość parametru σ^2, a jedynie wartość szacunkowa s^{*2}, dlatego do wyznaczenia minimalnej liczności próby użyjemy wzoru (5.116).

Wartosć $t_{r,\alpha/2}$ odczytamy z tablic rozkładu Studenta, zakładając, że $\alpha/2 = 0{,}025$ oraz liczba stopni swobody $r = 10 - 1 = 9$.

Podstawiając do wzoru, otrzymamy:

$$n = \frac{(2{,}262)^2 \cdot 42{,}25}{2{,}5^2} \approx 35.$$

Z powyższych obliczeń wynika, że do próby pilotażowej należy dolosować jeszcze dodatkowo 25 osób. {kp}

Przykład 5.14

Ustalić minimalną liczebność próby, na podstawie której można będzie określić frakcję studentów, którzy otrzymali oceny niedostateczne z egzaminu z „Zarządzania jakością". Na podstawie analiz przeprowadzonych w latach wczesniejszych wynika, że frakcja tych studentów oscyluje na poziomie około 12%. Podczas szacunku przyjąć błąd szacunku 5% oraz współczynnik ufności 0,9.

Ponieważ znany jest szacunkowy rząd wielkości frakcji w całej populacji, do oszacowania minimalnej liczności próby możemy wykorzystać wzór (5.117). Po podstawieniu wartości otrzymamy:

$$n = \frac{1{,}64^2 \cdot 0{,}12 \cdot 0{,}88}{0{,}05^2} \approx 114.$$

Jest to minimalna liczebność próby, która przy założonym poziomie ufności 0,9 i średnim błędzie szacunku równym 5%, pozwoli na oszacowanie frakcji studentów z ocenami niedostatecznymi. {kp}

Weryfikacja hipotez statystycznych

Rozdział

6

6.1. Uwagi wstępne

Postępowanie weryfikacyjne rozpoczyna się od sformułowania **hipotez statystycznych**.

Mianem **hipotezy statystycznej** można określić sąd (pogląd) odnoszący się do zbiorowości generalnej sformułowany bez przeprowadzenia badania generalnego. Jeżeli ten sąd dotyczy parametrów opisujących populację, to wówczas hipotezy nazywać będziemy **parametrycznymi**. W przypadku, gdy sąd odnosi się do klasy rozkładu (postaci funkcyjnej dystrybuanty) bez odwoływania się do liczbowych wartości parametru, mówimy o hipotezach **nieparametrycznych**. Na przykład hipoteza głosząca, że średnia wieku w pewnej populacji wynosi 35 lat jest hipotezą parametryczną, natomiast stwierdzenie, że populacja ta ma rozkład zbliżony do normalnego, jest hipotezą nieparametryczną. Na początek zajmiemy się sposobem budowy i weryfikacji hipotez parametrycznych, a później w dalszych podrozdziałach omówimy wybrane testy dla hipotez nieparametrycznych. Podobnie jak we wcześniejszym rozdziale, przyjmiemy, że nieznany parametr populacji generalnej oznaczymy symbolem Q. Proces weryfikacji rozpoczyna się od sformułowania tzw. **hipotezy zerowej** H_0, głoszącej brak różnic pomiędzy weryfikowanym parametrem Q, a założoną *a priori* wartością Q_0, oraz przeczącej jej jednej lub kilku **hipotez alternatywnych** H_1. Powszechnie termin hipoteza zerowa, zarezerwowany jest dla hipotezy o braku różnicy, jednak nic nie stoi na przeszkodzie, aby rozszerzyć znaczenie tego terminu również na przypadki, gdy różnica pomiędzy Q i Q_0 jest mniejsza lub równa lub też większa lub równa zero. Jeżeli hipotezę zerową zapiszemy jako:

$$H_0: Q = Q_0 \text{ lub } H_0: Q - Q_0 = 0, \tag{6.1}$$

to taką hipotezę nazwiemy hipotezą zerową prostą.

Natomiast, jeżeli hipoteza zerowa będzie miała postać:

$$H_0: Q \leq Q_0 \text{ lub } H_0: Q - Q_0 \leq 0, \tag{6.2}$$

albo:

$$H_0: Q \geq Q_0 \text{ lub } H_0: Q - Q_0 > 0, \tag{6.3}$$

to wowczas nazywać się ją będzie hipotezą zerową złożoną.

W stosunku do tak postawionych hipotez zerowych należy sformułować odpowiednie hipotezy alternatywne H_1:

$$H_1: Q \neq Q_0 \text{ lub } H_1: Q - Q_0 \neq 0, \tag{6.4}$$

$$H_1: Q > Q_0 \text{ lub } H_1: Q - Q_0 > 0, \tag{6.5}$$

$$H_1: Q < Q_0 \text{ lub } H_1: Q - Q_0 < 0, \tag{6.6}$$

Zauważmy, że każda z tych trzech złożonych hipotez alternatywnych zaprzecza hipotezie zerowej prostej (6.1). Jeżeli natomiast hipoteza zerowa będzie złożona i będzie miała postać (6.2) lub (6.1), to wówczas można do niej dopasować tylko jedną hipotezą alternatywną. I tak hipotezie zerowej (6.2) odpowiada hipoteza alternatywna (6.5), natomiast hipotezie zerowej (6.3), hipoteza alternatywna (6.6).

W trakcie weryfikacji hipotez statystycznych dopuszcza się możliwość popełnienia dwóch rodzajów błędu (tzw. **błąd pierwszego rodzaju** i **błąd drugiego rodzaju**). Do błędu pierwszego rodzaju dochodzi wówczas, gdy przyjmiemy, że prawdziwa jest hipoteza alternatywna, gdy w rzeczywistości prawdziwą będzie hipoteza zerowa. Natomiast jeżeli przyjmiemy, że prawdziwa jest hipoteza zerowa, a w rzeczywistości prawdziwą jest alternatywna, to popełnimy błąd drugiego rodzaju. Oczywiście błędów nie popełnimy, gdy przyjmiemy hipotezę zerową, gdy ta będzie prawdziwa, a odrzucimy ją, gdy będzie ona fałszywa. Prawdopodobieństwo popełnienia błędu pierwszego rodzaju oznaczamy symbolem α, natomiast prawdopodobieństwo, że popełniony zostanie błąd drugiego rodzaju, oznaczymy symbolem β. Do oznaczenia prawdopodobieństwa braku pomyłek będziemy stosować odpowiednio symbole $1 - \alpha$ i $1 - \beta$. Podczas formułowania hipotezy zerowej należy pamiętać o tym, aby była ona tak skonstruowana, by z dużym prawdopodobieństwem dało się ją odrzucić, w przypadku, gdy jest ona fałszywa. Całość powyższego rozumowania dobrze ilustruje rysunek 32 w formie tabelki.

		Hipoteza zerowa H_0	
		Prawdziwa	**Fałszywa**
Decyzja	Przyjąć	Decyzja prawidłowa $(1 - \alpha)$	Błąd drugiego rodzaju (β)
	Odrzucić	Bład pierwszego rodzaju (α)	Decyzja prawidłowa $(1 - \beta)$

Rys. 32. Błędy popełniane przy weryfikacji hipotez oraz prawdopodobieństwo ich realizacji

Źródło: opracowanie własne.

Reguły postępowania (procedury) pozwalające na rozstrzygnięcie, która z hipotez jest prawdziwa, a która fałszywa, nazywane są **testami statystycznymi**. Podstawą każdego testu jest reprezentatywna próba losowa. W praktyce używane są najczęściej tzw. **testy istotności**.

W standardowych testach istotności jedynie wartość α, nazywana **poziomem** (współczynnikiem) **istotności**, jest przyjmowana *a priori*, natomiast β jest wartością wynikową, zależną od liczebności próby n oraz od rzeczywistej wartości parametru Q, który to parametr zwykle pozostaje nieznany. Skutkiem braku znajomości rzeczywistej wartości parametru Q jest również brak znajomości prawdopodobieństwa β. Ze względu na to, że w testach istotności uwzględnia się tylko ryzyko popełnienia błędu pierwszego rodzaju, pomijając możliwość popełnienia błędu drugiego rodzaju, to wynikiem tego testu możliwa jest decyzja odrzucenia hipotezy zerowej lub stwierdzenie braku podstaw do jej odrzucenia, co nie jest tym samym co jej przyjęcie. Jeżeli na podstawie testu istotności poddaje się weryfikacji hipotezy parametryczne, to wówczas taki test określany jest mianem **parametrycznego testu istotności**. Natomiast, jeżeli przedmiotem weryfikacji są hipotezy nieparametryczne, to wówczas mowi się o **nieparametrycznych testach istotności**.

Aby lepiej zrozumieć poruszane tu problemy, przeanalizujmy krótki przykład.

Przykład 6.1

Założmy, że w procesie sądowym, przeciwko podejrzanemu o popełnienie przestępstwa, formułowana jest następująca hipoteza zerowa oraz hipoteza alternatywna.

H_0: podejrzany jest niewinny,

H_1: podejrzany jest winny.

Forma hipotezy zerowej jest uwarunkowana tutaj podstawową zasadą prawa, głoszącą domniemaną niewinność podejrzanego. Aby ją odrzucić i przyjąć hipotezę alternatywną, należy dowieść winny podejrzanego. Do błędu pierwszego rodzaju, dojdzie wowczas, gdy podejrzany zostanie uznany za winnego, podczas gdy w rzeczywistości jest niewinny, natomiast sąd popełni błąd rodzaju drugiego, gdy uniewinni podejrzanego, nie dlatego, że był on niewinny, tylko że na przykład miał dobrych adwokatow. Sąd może wydać rowniez decyzję, że z braku dowodów należy umorzyć postępowanie jurysdykcyjne, decyzja ta ma takie samo znaczenie jak decyzja: brak powodów do odrzucenia H_0 w przypadku standardowych testów istotności. Chyba nikogo nie trzeba przekonywać, że decyzja sądu nakazująca umorzenie postępowania jurysdykcyjnego z braków dowodów, jest zupełnie odmienna od decyzji o całkowitym uniewinnieniu podejrzanego. {kp}

Intuicyjnie mogłoby się wydawać, że w trakcie przeprowadzania testów statystycznych powinniśmy dążyć do minimalizacji prawdopodobieństw popełnienia błędu pierwszego i drugiego rodzaju α i β. W rzeczywistości, nie jest to jednak możliwe, gdyż przy zadanej liczności próby n, zmniejszenie α zaowocuje zwiększeniem β. Teoretycznie można próbować minimalizować α i β poprzez zwiększanie liczebności próby, jednak w wielu przypadkach jest to działanie ekonomicznie nieuzasadnione, gdyż zwiększając liczebność próby zwiększamy również całkowite koszty badań.

Dlatego też w standardowych testach istotności obszar krytyczny konstruowany jest w taki sposób, aby zminimalizować prawdopodobieństwo popełnienia błędu drugiego rodzaju β przy ustalonym *a priori* prawdopodobieństwie popełnienia błędu pierwszego rodzaju α. Zbudowane w tej konwencji testy są testami **najmocniejszymi**, gdyż odpowiada im największa **moc**, interpretowana jako prawdopodobieństwo odrzucenia fałszywej hipotezy zerowej na korzyść prawdziwej hipotezy alternatywnej. Odwołując się do przykładu 6.1, opisującego postępowania sądowe, moc testu jest prawdopodobieństwem, że podejrzany zostanie uznany za winnego, gdy rzeczywiście jest winien.

Podstawą każdego testu jest reprezentatywna próba losowa. Pobranie i opis n-elementowej próby losowej stanowi drugi po konstrukcji hipotez etap procesu weryfikacji. Pobraną próbę należy następnie opisać, używając odpowiedniej **statystyki** (charakterystyki) **z próby** η. Dobór statystyki z próby odbywa się stosownie do treści weryfikowanych hipotez, oraz warunków takich jak: typ rozkładu populacji, liczebność próby, znajomość parametrów opisujących zbiorowość generalną itd. W wyniku pomiaru próby otrzymujemy zbiór realizacji, dla których obliczamy wartość statystyki z próby η_n.

W kolejnym kroku, wykorzystując znajomość rozkładu statystyki z próby, oraz zakładając *a priori* **poziom istotności** α wyznacza się granice tzw. **przedziału krytycznego** η_k, w taki sposób, aby spełnione były relacje:

$$P[\eta_n \in \eta_k \,|\, Q = Q_0] = \alpha, \tag{6.7}$$

$$P[\eta_n \in \eta_k \,|\, Q < Q_0] < \alpha, \tag{6.8}$$

$$P[\eta_n \in \eta_k \,|\, Q > Q_0] \leq \alpha, \tag{6.9}$$

Powyższe wyrażenia należy czytać: prawdopodobieństwo zdarzenia losowego, że wartość statystyki z próby η_n należy do przedziału krytycznego η_k, przy założeniu, że prawdziwa jest hipoteza zerowa, jest równe (mniejsze lub równe) dowolnie małej różnej od zera wartości α. Wartość poziomu istotności α ustala się zwykle na poziomie: 0,01; 0,05; 0,025; 0,1 itp. Dopełnieniem logicznym przedziału krytycznego jest **przedział** η_t[1]. Przedział ten może być przedziałem ograniczonym dwustronnie lub jednostronnie (prawostronnie lub lewostronnie). Sposób ograniczenia przedziału η_t

[1] W literaturze przedmiotu raczej unika się nazywania tego przedziału. Zgodnie z funkcją, jaką spełnia w procesie podejmowania decyzji, można by go nazwać przedziałem braku podstaw do odrzucenia hipotezy zerowej. Z uwagi jednak na długą nazwę, w dalszej części pozostaniemy przy symbolicznym oznaczeniu η_t.

zależy od konstrukcji hipotezy alternatywnej. Jeżeli w hipotezie alternatywnej występuje znak „≠", to wówczas przedział η_t jest ograniczony dwustronnie i ma postać:

$$\eta_t = (\eta_d, \eta_g), \tag{6.10}$$

gdzie: η_d, η_g są dolnym i górnym kresem tego przedziału.

W konsekwencji przedział krytyczny w rozważnym przypadku będzie przedstawiał się następująco:

$$\eta_k = (-\infty, \eta_d] \cup [\eta_g, +\infty). \tag{6.11}$$

Jeżeli w hipotezie alternatywnej występuje znak „>", to wówczas przedział η_t i przedział krytyczny będą przedstawiały się następująco:

$$\eta_t = (-\infty, \eta_g) \text{ oraz } \eta_k = [\eta_g, +\infty). \tag{6.12}$$

Jak łatwo się domyśleć, gdy w hipotezie alternatywnej postawimy znak „<", to wówczas przedział η_t i przedział krytyczny będą miały postać:

$$\eta_t = (\eta_d, +\infty) \text{ oraz } \eta_k = (-\infty, \eta_d]. \tag{6.13}$$

Granice η_d *i* η_g we wzorach (6.11) – (6.13) określa się często mianem **wartości krytycznych**.

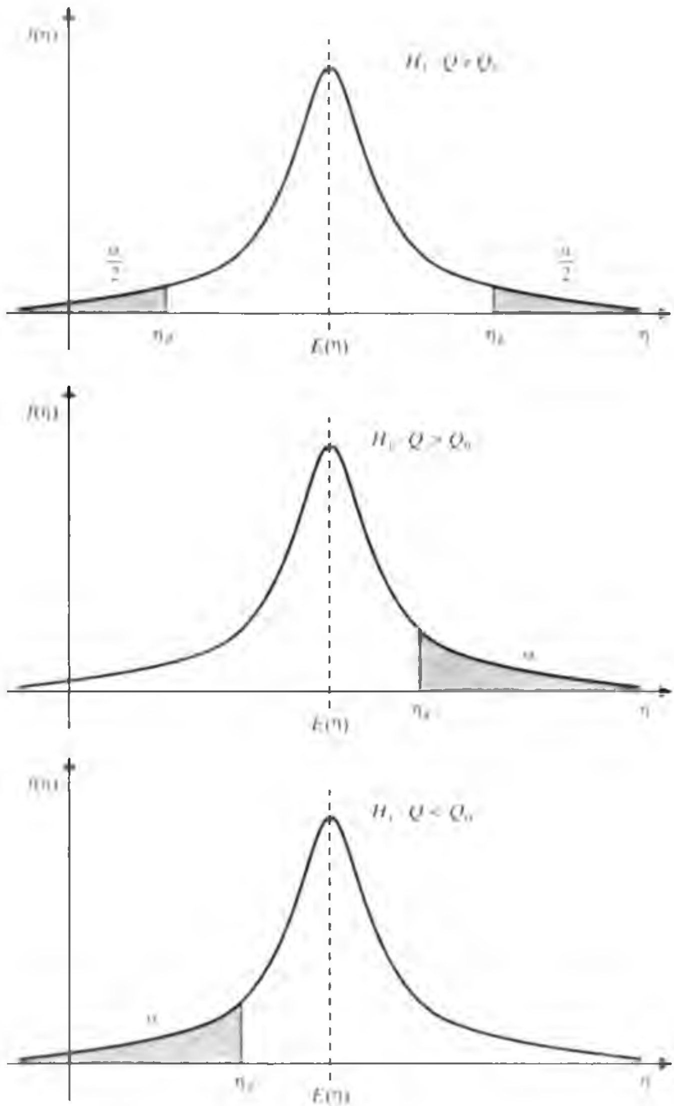
Na rysunku 33 przedstawiono graficzną prezentację omawianych przedziałów. Podczas tworzenia rysunku, założono, że statystyka z próby η posiada rozkład normalny o wartości oczekiwanej $E(\eta) = Q_0$. Zakreskowany obszar pod krzywą normalną ilustruje prawdopodobieństwa przypisane odpowiednim przedziałom krytycznym.

Końcowym etapem procesu weryfikacji jest podjęcie decyzji. W przypadku, gdy weryfikacja odbywa się przy użyciu testu istotności, podjęcie decyzji przebiega według następującego schematu. Jeżeli uzyskana wartość statystyki z próby η_n należy do przedziału krytycznego η_k ($\eta_n \in \eta_k$), to wówczas z prawdopodobieństwem równym lub nie większym niż α, odrzucamy hipotezę zerową na korzyść odpowiedniej hipotezy alternatywnej. Natomiast, jeżeli zachodzi relacja odwrotna ($\eta_n \notin \eta_k$), co oznacza, że wartość statystyki z próby należy do przedziału η_t, to wówczas brak jest powodów (nie ma podstaw) do odrzucenia hipotezy zerowej.

Nie możemy jednak w sposób jednoznaczny stwierdzić, że przyjmujemy hipotezę zerową jako hipotezę prawdziwą, gdyż nie zostało przyjęte *a priori* ryzyko, że decyzja taka będzie fałszywa. Ryzyko takie zakłada się w bardziej ogólnych rozważaniach dotyczących procedur weryfikacji hipotez statystycznych, w których uwzględnia się możliwość zarówno odrzucenia, jak i przyjęcia hipotezy zerowej.

Uogólniając powyższe rozważania, schemat postępowania weryfikacyjnego na podstawie testów istotności można w skrócie przedstawić w czterech punktach:

1) konstrukcja hipotez statystycznych,
2) pobranie i opis n-elementowej próby losowej,
3) konstrukcja tzw. przedziału krytycznego,
4) podjęcie decyzji.



Rys. 33. Zależność pomiędzy postacią hipotezy alternatywnej a przedziałem η_t i przedziałem krytycznym

Źródło: opracowanie własne.

Na koniec uwag ogólnych dotyczących weryfikacji hipotez statystycznych, rozważmy kilka innych ważnych problemów.

Zdefiniujmy obecnie w sposób formalny wprowadzone wcześniej pojęcie mocy testu. Moc testu $M(Q; n)$ jest funkcją opisywaną przez wartość parametru Q oraz liczność próby n. Korzystając z przyjętych wcześniej oznaczeń, można zapisać, że:

$$M(Q; n) = P[\eta_n \in \eta_k | Q]. \tag{6.14}$$

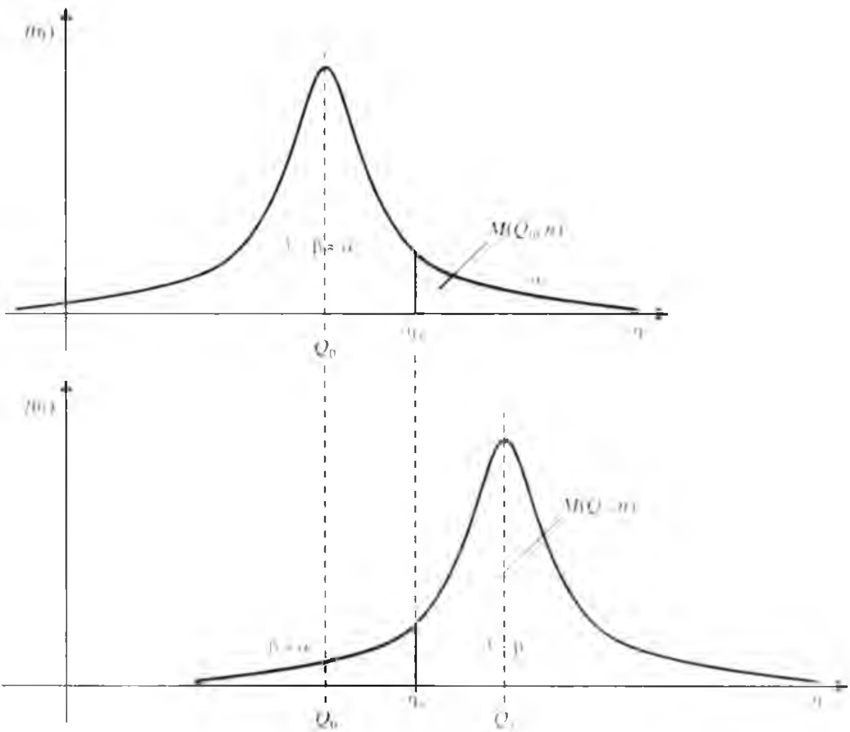
Załóżmy, że liczebność próby pozostaje na niezmienionym poziomie, natomiast zmianie ulega wartość parametru Q. Załóżmy, że przedział η_t jest ograniczony prawostronnie wartością η_g. Załóżmy także, że statystyka z próby η posiada rozkład normalny o wartosci oczekiwanej $E(\eta)$ wynoszącej kolejno Q_0 i Q_1, przy czym $Q_0 < Q_1$. Wartość funkcji mocy testu w punkcie Q_0 jest równa poziomowi istotności α:

$$M(Q_0; n) = \alpha, \tag{6.15}$$

natomiast w punkcie Q_1 wyniesie $1 - \beta$:

$$M(Q_1; n) = 1 - \beta. \tag{6.16}$$

W sytuacji, gdy moc testu będzie wynosić α, to wówczas $\beta = 1 - \alpha$. Natomiast, gdy moc zbliży się do wartości $1 - \beta$, to wówczas β będzie w przybliżeniu równa α ($\alpha \approx \beta$). W celu zilustrowania rozważanej sytuacji sporządzono rysunek 34, na którym funkcja mocy testu jest reprezentowana przez pole powierzchni zawarte pod krzywą gęstości prawdopodobieństwa w przedziale $(\eta_g, +\infty)$.



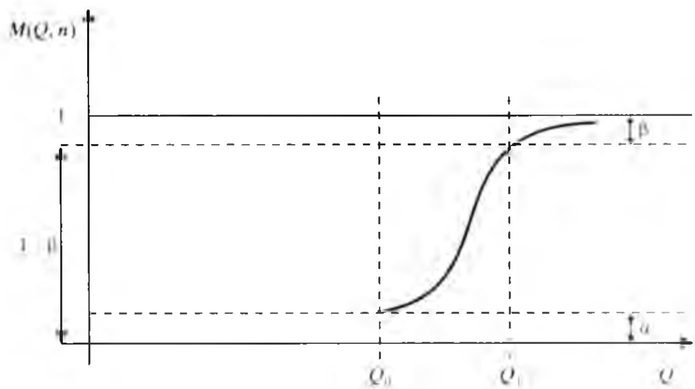
Rys. 34. Funkcja mocy testu dla Q_0 oraz Q_1 dla prawostronnie ograniczonego przedziału η_t

Źródło: opracowanie własne.

Funkcję mocy testu można również przedstawić za pomocą rzędnych, sporządzając wykres krzywej mocy testu (zob. rys. 35). W podobny sposób możemy skonstruować krzywą mocy testu dla przypadku, gdy przedział η_t jest ograniczony lewostronnie i dwustronnie (zob. rys. 36 i rys. 37). W przypadku testów lewostronnych (tzn. gdy przedział η_t jest ograniczony lewostronnie) należy uwzględnić dwie wartości parametru Q: Q_0 i Q_{-1} ($Q_0 > Q_{-1}$), natomiast gdy przedział η_t ograniczymy dwustronnie, to wówczas aż trzy: Q_0, Q_{-1}, Q_1 ($Q_{-1} < Q_0 < Q_1$).

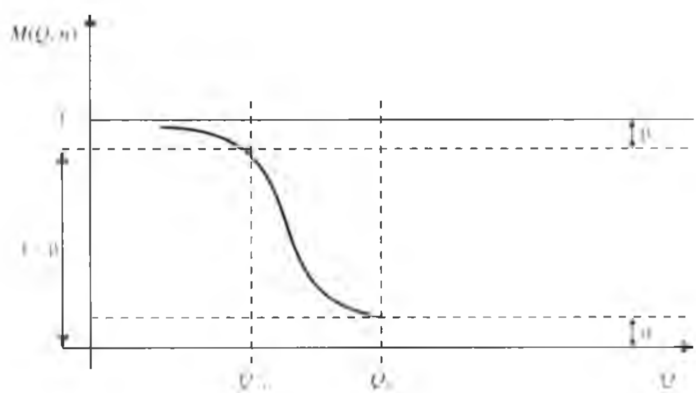
W praktyce, często zamiast funkcji mocy testu wykorzystuje się tzw. **funkcję operacyjno-charakterystyczną** (w skrócie funkcję OC), która określana jest wzorem:

$$O_c(Q; n) = P[\eta_n \in \eta_t | Q]. \tag{6.17}$$



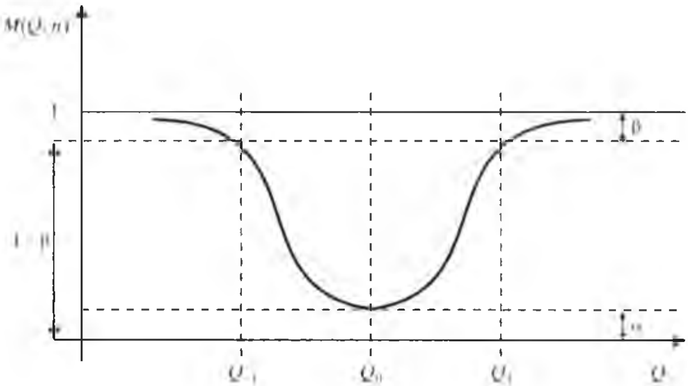
Rys. 35. Krzywa mocy testu dla Q_0 oraz Q_1 dla prawostronnie ograniczonego przedziału η_t (H_0: $Q \leq Q_0$ i H_1: $Q > Q_0$)

Źródło: opracowanie własne.



Rys. 36. Krzywa mocy testu dla Q_0 oraz Q_{-1} dla lewostronnie ograniczonego przedziału η_t (H_0: $Q > Q_0$ i H_1: $Q < Q_0$)

Źródło: opracowanie własne.



Rys. 37. Krzywa mocy testu dla Q_0, Q_1, Q_{-1} dla dwustronnie ograniczonego przedziału η_t (H_0: $Q = Q_0$ i H_1: $Q \neq Q_0$)

Źródło: opracowanie własne.

Pomiędzy funkcją mocy testu a funkcję OC zachodzi następujący związek:

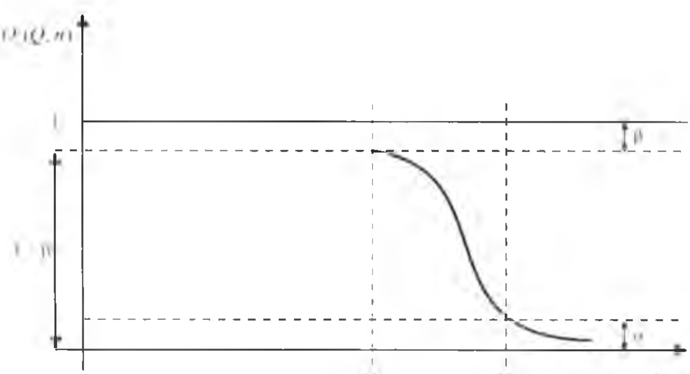
$$O_c(Q; n) = 1 - M(Q; n), \tag{6.18}$$

z czego wynika:

$$O_c(Q_0; n) = 1 - \alpha, \text{ oraz } O_c(Q_1; n) = \beta. \tag{6.19}$$

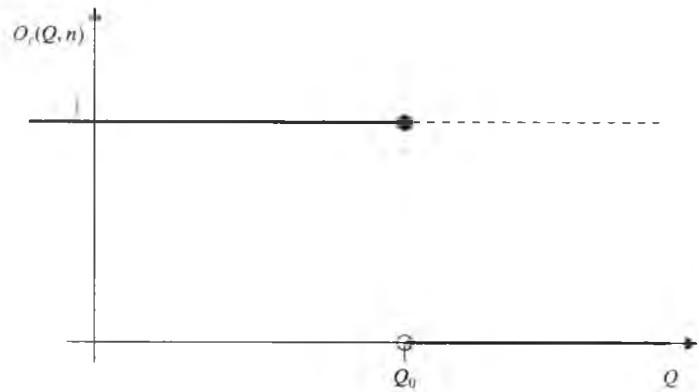
Idealna funkcja OC generuje wartość 1 w sytuacji, gdy prawdziwa jest hipoteza zerowa, i zero, gdy prawdziwa jest hipoteza alternatywna. Taka sytuacja jest możliwa tylko wówczas, gdy badania mają charakter wyczerpujący (tzn. gdy: $n = N$).

Wykres funkcji OC oraz idealnej funkcji OC, dla testu prawostronnego przedstawiono na rysunkach: 38 i 39.



Rys. 38. Krzywa operacyjno-charakterystyczna dla Q_0 oraz Q_1 dla prawostronnie ograniczonego przedziału η_t (H_0: $Q < Q_0$ i H_1: $Q > Q_0$)

Źródło: opracowanie własne.



Rys. 39. Idealna funkcja operacyjno-charakterystyczna dla H_0: $Q < Q_0$ i H_1: $Q > Q_0$

Źródło: opracowanie własne.

Z powyższych rozważań wynika, że każdy standardowy tekst można w prosty sposób przekształcić w test, w którym obok przyjętego *a priori* współczynnika α przyjmuje się również *a priori* wartość ryzyka β. Przekształceniu ulegają wówczas hipotezy alternatywne i tak na przykład w miejsce hipotezy (6.4) będziemy mieli obecnie aż dwie hipotezy alternatywne[2]:

$$H_1\colon Q = Q_0 + \Delta Q = Q_1 \text{ i } H_{-1}\colon Q = Q_0 - \Delta Q = Q_{-1} \tag{6.14}$$

gdzie:

$$Q_{-1} < Q_0 < Q_1.$$

Wówczas prawdopodobieństwo β' wyniesie:

$$P[\eta_n \notin \eta_k | Q = Q_{-1}] + P[\eta_n \notin \eta_k | Q = Q_1] = \beta', \tag{6.15}$$

Jeżeli $Q_0 - Q_{-1} = Q_1 - Q_0$, to wówczas:

$$P[\eta_n \notin \eta_k | Q = Q_{-1}] = \beta'/2, \tag{6.16}$$

$$P[\eta_n \notin \eta_k | Q = Q_1] = \beta'/2. \tag{6.17}$$

Jeżeli hipoteza zerowa ma postać (6.2), to wowczas odpowiednia hipoteza alternatywna będzie miała postać:

$$H_1\colon Q = Q_0 + \Delta Q = Q_1, \tag{6.18}$$

natomiast, gdy weryfikujemy hipotezę zerową postaci (6.3), odpowiednią hipotezę alternatywną należy zapisać jako:

$$H_{-1}\colon Q = Q_0 - \Delta Q = Q_{-1} \tag{6.19}.$$

Zob. A. Iwasiewicz, Z Paszek: *Statystyka...*, *op. cit.*, s. 212.

W powyższych przypadkach prawdopodobieństwo błędu drugiego rodzaju wynosi:

$$P[\eta_n \notin \eta_k | Q = Q_{-1}] = \beta', \tag{6.20}$$

$$P[\eta_n \notin \eta_k | Q = Q_1] = \beta'. \tag{6.21}$$

Parametr β', należy traktować jako malejącą funkcję różnicy ΔQ lub liczebności próby n. Zwiększanie liczebności próby przy stałej różnicy ΔQ lub zwiększanie wielkości różnicy ΔQ, przy założeniu stałej liczności próby n spowoduje, że prawdopodobieństwo błędu β' będzie zmierzać do zadanego *a priori* prawdopodobieństwa β. Zatem możemy zapisać:

$$\lim_{\substack{n \to N; \\ \Delta Q \to \infty}} P[\eta_n \notin \eta_k | Q = Q_{-1}] + P[\eta_n \notin \eta_k | Q = Q_1] = \beta' < \beta. \tag{6.22}$$

Z powyższego wynika, że można znaleźć (analitycznie lub na drodze symulacji) taką minimalną liczebność próby lub taką minimalną wartość różnicy ΔQ, przy której spełniona będzie nierówność $\beta' < \beta$.

6.2. Weryfikacja hipotez parametrycznych dotyczących istotności różnicy między wartością oczekiwaną zmiennej losowej a ustaloną wartością

Załóżmy, że zbiorowość generalna charakteryzowana jest za pomocą zmiennej losowej: $X \sim N(\mu; \sigma)$. Hipotezę zerową i odpowiednią hipotezę alternatywną konstruujemy dla parametru μ. Hipoteza zerowa może być hipotezą prostą typu:

$$H_0: \mu = \mu_0, \tag{6.23}$$

lub jedną z hipotez złożonych postaci:

$$H_0: \mu < \mu_0, \tag{6.24}$$

lub

$$H_0: \mu > \mu_0. \tag{6.25}$$

Zaprzeczeniem hipotezy zerowej jest hipoteza alternatywna, o postaci:

$$H_1: \mu \neq \mu_0, \tag{6.26}$$

lub

$$H_1: \mu > \mu_0, \tag{6.27}$$

lub

$$H_1: \mu < \mu_0. \tag{6.28}$$

Podobnie jak w rozważaniach ogólnych, każda z tych trzech hipotez alternatywnych może być zaprzeczeniem hipotezy zerowej prostej (6.23). Jeżeli natomiast hipoteza zerowa będzie złożona i będzie miała postać (6.24) lub (6.25), to wówczas można jej przyporządkować tylko jedną hipotezą alternatywną. I tak hipotezie zerowej (6.24) odpowiada hipoteza alternatywna (6.27), natomiast hipotezie zerowej (6.25), hipoteza alternatywna (6.28).

Podczas budowy testu istotności dla wartości oczekiwanej można wyróżnić trzy modele sytuacyjne (porównaj z estymacją przedziałową wartości oczekiwanej):

1) **model I**. Zmienna losowa X opisująca populację ma rozkład normalny lub zbliżony do normalnego, z nieznaną wartością oczekiwaną μ i znanym odchyleniem standardowym σ,
2) **model II**. Zmienna losowa X opisująca populację ma rozkład normalny lub zbliżony do normalnego, z nieznaną wartością oczekiwaną μ i nieznanym odchyleniem standardowym σ,
3) **model III**. Zmienna losowa X opisująca populację ma nieznany rozkład, z nieznaną wartością oczekiwaną μ i nieznanym odchyleniem standardowym σ.

Model I. Jeżeli zmienna losowa opisująca populację ma rozkład normalny o znanym i stałym odchyleniu standardowym σ, to wówczas bez względu na liczebność próby, do opisu próby losowej używa się statystyki z próby postaci (zob. wzór 5.61):

$$U = \frac{\overline{X}_n - \mu_0}{\sigma}\sqrt{n}, \tag{6.29}$$

o wartościach

$$u_0 = \frac{\overline{x}_n - \mu_0}{\sigma}\sqrt{n}. \tag{6.30}$$

Jeżeli prawdziwa jest hipoteza zerowa, to wówczas statystyka U ma rozkład normalny o parametrach $\mu = 0$ i $\sigma = 1$. Dzięki temu do konstrukcji przedziałów krytycznych i przedziałów η_t wykorzystujemy tablice dystrybuanty rozkładu normalnego standaryzowanego. Postać przedziału krytycznego jest zależna od postaci hipotezy alternatywnej. Jeżeli hipoteza alternatywna ma postać: (6.26), to wówczas przedział krytyczny będzie miał postać:

$$\eta_k = (-\infty; -u_{\alpha/2}] \cup [u_{\alpha/2}; +\infty). \tag{6.31}$$

Wartości $-u_{\alpha/2}$ i $u_{\alpha/2}$ wyznaczane są w taki sposób, aby

$$P(U < -u_{\alpha/2}) = \alpha/2, \tag{6.32}$$

oraz

$$P(U > u_{\alpha/2}) = \alpha/2. \tag{6.33}$$

W przypadku, gdy hipoteza alternatywna jest postaci (6.27), to wówczas przedział krytyczny przedstawia się następująco:

$$\eta_k = [u_\alpha; +\infty), \tag{6.34}$$

przy czym

$$P(U > u_\alpha) = \alpha. \tag{6.35}$$

Jeżeli hipoteza alternatywna przedstawiana jest wzorem (6.28), to wówczas przedział krytyczny jest określony jako:

$$\eta_k = (-\infty; -u_\alpha]. \tag{6.36}$$

Wartość graniczna $-u_\alpha$ jest określana w ten sposób, aby spełniona była relacja:

$$P(U < -u_\alpha) = \alpha. \tag{6.37}$$

Ogólna reguła decyzyjna jest taka, że w przypadku, gdy wartość obliczonej statystyki u_0 zawiera się w wyznaczonym przedziale krytycznym η_k, to wówczas należy odrzucić hipotezę zerową na korzyść hipotezy alternatywnej. Jeżeli natomiast obliczona wartość statystyki u_0 będzie leżeć poza przedziałem krytycznym, to wówczas, należy sformułować wniosek, że brak jest podstaw (powodów), aby odrzucić hipotezę zerową. Szczegółowe reguły postępowania zostały zamieszczone w tablicy 6.1.

Tablica 6.1. Zasady podejmowania decyzji[3]

Hipotezy / Decyzja	H_0: $\mu = \mu_0$ H_1: $\mu \neq \mu_0$	H_0: $\mu \leq \mu_0$ H_1: $\mu > \mu_0$	H_0: $\mu \geq \mu_0$ H_1: $\mu < \mu_0$
Odrzucić H_0 na korzyść H_1	$\lvert u_0 \rvert \geq u_{\alpha/2}$ ($u_0 > u_{\alpha/2}$ lub $u_0 < -u_{\alpha/2}$) (6.38)	$\lvert u_0 \rvert \geq u_\alpha$ ($u_0 \geq u_\alpha$) (6.40)	$\lvert u_0 \rvert \geq u_\alpha$ ($u_0 \leq -u_\alpha$) (6.42)
Brak podstaw do odrzucenia H_0	$\lvert u_0 \rvert < u_{\alpha/2}$ ($-u_{\alpha/2} < u_0 < u_{\alpha/2}$) (6.39)	$\lvert u_0 \rvert < u_\alpha$ ($u_0 < u_\alpha$) (6.41)	$\lvert u_0 \rvert < u_\alpha$ ($u_0 > -u_\alpha$) (6.43)

Źródło: opracowanie własne.

Zasady podejmowania decyzji przedstawione w tablicy 6.1, można transformować do nieco odmiennej postaci. Ich opis zamieszczono w tablicy 6.2[4].

[3] Dla większej przejrzystości w nagłówku tablicy umieszczono najczęściej spotykane postacie hipotez H_0 i H_1. (Zob. wzory 6.23–6.28).

[4] Forma podejmowania decyzji przedstawiona w tablicy 6.2 jest charakterystyczna dla tzw. kart kontrolnych Shewharta, które należą do podstawowych narzędzi wykorzystywanych w statystycznej kontroli jakości. Szerzej zob. np. A. Iwasiewicz: *Zarządzanie jakością*, PWN, Warszawa–Kraków 1999, rozdz. 8. Zob także: *idem*: *Zarządzanie jakością w przykładach i zadaniach*, Śląskie Wydawnictwo Naukowe Wyższej Szkoły Zarządzania i Nauk Społecznych, Tychy 2005, s. 273.

Tablica 6.2. Zmodyfikowane zasady podejmowania decyzji

Decyzja \ Hipotezy	H_0: $\mu = \mu_0$ H_1: $\mu \neq \mu_0$	H_0: $\mu < \mu_0$ H_1: $\mu > \mu_0$	H_0: $\mu > \mu_0$ H_1: $\mu < \mu_0$
Odrzucić H_0 na korzyść H_1	$\bar{x}_n < \bar{x}_d = \mu_0 - u_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ lub $\bar{x}_n > \bar{x}_g = \mu_0 + u_{\alpha}\frac{\sigma}{\sqrt{n}}$ (6.44)	$\bar{x}_n \geq \bar{x}_g = \mu_0 + u_{\alpha}\frac{\sigma}{\sqrt{n}}$ (6.46)	$\bar{x}_n \leq x_g = \mu_0 - u_{\alpha}\frac{\sigma}{\sqrt{n}}$ (6.48)
Brak podstaw do odrzucenia H_0	$\bar{x}_d < \bar{x}_n < \bar{x}_g$, gdzie $\bar{x}_g = \mu_0 + u_{\alpha/2}\frac{\sigma}{\sqrt{n}}$, $\bar{x}_d = \mu_0 - u_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ (6.45)	$\bar{x}_n < \bar{x}_g = \mu_0 + u_{\alpha}\frac{\sigma}{\sqrt{n}}$ (6.47)	$\bar{x}_n > \bar{x}_g = \mu_0 - u_{\alpha}\frac{\sigma}{\sqrt{n}}$ (6.49)

Źródło: opracowanie własne.

Zauważmy, że obecnie do podjęcia decyzji wystarczy porównać wartość średniej z próby z odpowiednią wartością krytyczna (wartościami krytycznymi). Wartości krytyczne $\bar{x}_d$ i $\bar{x}_g$ otrzymuje się poprzez przekształcenie funkcji (6.30) i ustalenie wartości u_0 na poziomie u_α lub $u_{\alpha/2}$. W przypadku wyznaczania dolnej granicy, przed przekształceniem wyrażenia (6.30) jego prawą stronę należy przemnożyć przez (–1).

Model II. Załóżmy, że populacja jest opisywana za pomocą zmiennej losowej $X \sim N(\mu, \sigma)$, przy czym nieznana jest rzeczywista wartość parametru σ.

Proces weryfikacji hipotez odbywa się z wykorzystaniem statystyki

$$t_r = \frac{\bar{X}_n - \mu}{S}\sqrt{n-1} = \frac{\bar{X}_n - \mu}{S^*}\sqrt{n}, \tag{6.50}$$

przyjmującej wartości:

$$t_r = \frac{\bar{x}_n - \mu}{s}\sqrt{n-1} = \frac{\bar{x}_n - \mu}{s^*}\sqrt{n}. \tag{6.51}$$

Statystyka t_r ma rozkład Studenta o $r = n - 1$ stopniach swobody.

W sytuacji, gdy prawdziwa jest hipoteza zerowa (6.23) – (6.25), to wówczas do jej weryfikacji wykorzystuje się sprawdzian $t_{r,0}$ postaci:

$$t_{r,0} = \frac{\bar{x}_n - \mu_0}{s}\sqrt{n-1} = \frac{\bar{x}_n - \mu_0}{s^*}\sqrt{n}. \tag{6.51a}$$

Przedział krytyczny konstruujemy, w podobny sposób jak w modelu I, korzystając z tablic rozkładu Studenta. W przypadku testu dwustronnego przedział krytyczny ma postać:

$$\eta_k = (-\infty; -t_{r,\alpha/2}] \cup [t_{r,\alpha/2}; +\infty), \tag{6.52}$$

przy czym

$$P(t_r < -t_{r,\alpha/2}) = P(t_r > t_{r,\alpha/2}) = \alpha/2. \tag{6.53}$$

Jeżeli test jest prawostronny, to wówczas przedział krytyczny będzie przedstawiał się następująco:

$$\eta_k = [t_{r,\alpha}; +\infty), \tag{6.54}$$

przy czym

$$P(U \geq t_{r,\alpha}) = \alpha.$$

Natomiast w przypadku testu lewostronnego przedział krytyczny będzie postaci:

$$\eta_k = (-\infty; -t_{r,\alpha}], \tag{6.55}$$

przy czym $P(t_r < -t_{r,\alpha}) = \alpha$.

Zasady podejmowania decyzji są analogiczne jak opisane w tablicy 6.1. Podobnie jak wcześniej ujmiemy je tabelarycznie (zob. tablica 6.3):

Tablica 6.3. Zasady podejmowania decyzji w modelu II

Hipotezy / Decyzja	$H_0: \mu = \mu_0$ $H_1: \mu \neq \mu_0$	$H_0: \mu \leq \mu_0$ $H_1: \mu > \mu_0$	$H_0: \mu \geq \mu_0$ $H_1: \mu < \mu_0$
Odrzucić H_0 na korzyść H_1	$\lvert t_{r,0}\rvert \geq t_{r,\alpha/2}$ $(t_{r,0} > t_{r,\alpha/2}$ lub $t_{r,0} < -t_{r,\alpha/2})$ (6.56)	$\lvert t_{r,0}\rvert \geq t_{r,\alpha}$ $(t_{r,0} > t_{r,\alpha})$ (6.58)	$\lvert t_{r,0}\rvert \geq t_{r,\alpha}$ $(t_{r,0} < -t_{r,\alpha})$ (6.60)
Brak podstaw do odrzucenia H_0	$\lvert t_{r,0}\rvert < t_{r,\alpha/2}$ $(-t_{r,\alpha/2} < t_{r,0} < t_{r,\alpha/2})$ (6.57)	$\lvert t_{r,0}\rvert < t_{r,\alpha}$ $(t_{r,0} < t_{r,\alpha})$ (6.59)	$\lvert t_{r,0}\rvert < t_{r,\alpha}$ $(t_{r,0} > -t_{r,\alpha})$ (6.61)

Źródło: opracowanie własne.

Oczywiście w podobny sposób można dokonać transformacji zasad podejmowania decyzji do formy przedstawionej w tablicy 6.1. Wystarczy tylko w miejsce odchylenia standardowego σ podstawić wartość estymatora S^* lub S. Budując tablicę 6.4. założono, że estymatorem odchylenia standardowego populacji jest odchylenie standardowe z próby S^*.

Tablica 6.4. Zmodyfikowane zasady podejmowania decyzji w modelu II

Hipotezy / Decyzja	H_0: $\mu = \mu_0$ H_1: $\mu \neq \mu_0$	H_0: $\mu < \mu_0$ H_1: $\mu > \mu_0$	H_0: $\mu > \mu_0$ H_1: $\mu < \mu_0$
Odrzucić H_0 na korzyść H_1	$\bar{x}_n \geq \bar{x}_g = \mu_0 + t_{r,\alpha/2}\frac{s^*}{\sqrt{n}}$ lub $\bar{x}_n \leq \bar{x}_d = \mu_0 - t_{r,\alpha/2}\frac{s^*}{\sqrt{n}}$, (6.62)	$\bar{x}_n \geq \bar{x}_g = \mu_0 + t_{r,\alpha}\frac{s^*}{\sqrt{n}}$ (6.64)	$\bar{x}_n < \bar{x}_g = \mu_0 - t_{r,\alpha}\frac{s^*}{\sqrt{n}}$ (6.66)
Brak podstaw do odrzucenia H_0	$\bar{x}_d < \bar{x}_n < \bar{x}_g$, gdzie $\bar{x}_g = \mu_0 + t_{r,\alpha/2}\frac{s^*}{\sqrt{n}}$, $\bar{x}_d = \mu_0 - t_{r,\alpha/2}\frac{s^*}{\sqrt{n}}$ (6.63)	$\bar{x}_n < \bar{x}_g = \mu_0 + t_{r,\alpha}\frac{s^*}{\sqrt{n}}$ (6.65)	$\bar{x}_n > \bar{x}_g = \mu_0 - t_{r,\alpha}\frac{s^*}{\sqrt{n}}$ (6.67)

Źródło: opracowanie własne.

Model III. W przypadku braku znajomości postaci rozkładu zbiorowości generalnej, do wiarogodnej weryfikacji wymagana jest duża liczebność próby. Do oszacowania wartości oczekiwanej populacji należy użyć średniej arytmetycznej z próby $\bar{X}_n$. Jeżeli próba losowa, wzrasta do nieskończoności, to zgodnie z centralnym twierdzeniem granicznym, rozkład z próby średnich zbliża się do rozkładu normalnego ze średnią μ i odchyleniem standardowym $\frac{\sigma}{\sqrt{n}}$. Gdy znana jest rzeczywista wartość odchylenia standardowego σ, to wówczas w procesie weryfikacji należy skorzystać ze statystyki U_0 (6.29) i postępować identycznie jak w modelu I. Jeżeli natomiast nie znamy wartości parametru σ, to przy dużej próbie można przyjąć, że w przybliżeniu jest on równy szacunkowi otrzymanemu przy użyciu estymatora S ($s \approx \sigma$) i skorzystać ze statystyki:

$$U = \frac{\bar{X}_n - \mu_0}{S}\sqrt{n}, \tag{6.68}$$

która w przypadku prawdziwości hipotezy zerowej ma rozkład $N(0,1)$.

Przykład 6.2

Wykorzystując dane i założenia z przykładu 5.5 oraz zakładając współczynnik ufności $\alpha = 0{,}05$, zweryfikować hipotezę zerową, głoszącą, że średnia liczba punktów otrzymanych z egzaminu z „Zarządzania jakością" równa się 30, wobec hipotezy alternatywnej głoszącej, że średnia liczba punków jest istotnie różna od 30.

Przypomnijmy, że rozkład punktów z egzaminu jest w przybliżeniu normalny z odchyleniem standardowym wynoszącym $\sigma = 6{,}4$, oraz średnia liczba punktów w badanej próbie ($n = 25$) wynosiła $\bar{x}_{25} = 30{,}5$.

Stawiamy hipotezę zerową:

$$H_0: \mu = 30,$$

wobec hipotezy alternatywnej:

$$H_1: \mu \neq 30.$$

Ponieważ znamy wartość odchylenia standardowego populacji σ oraz wiemy, że zmienna losowa X oznaczająca liczbę punktów ma w przybliżeniu rozkład normalny, dlatego do zweryfikowania hipotezy zerowej można wykorzystać statystykę (6.29) z modelu I.

Po podstawieniu danych do wzoru otrzymamy:

$$u_0 = \frac{30{,}5 - 30}{6{,}4}\sqrt{25} = 0{,}39$$

Przedział krytyczny jest postaci:

$$\eta_k = (-\infty; -1{,}96] \cup [1{,}96; +\infty).$$

Ponieważ $|0{,}39| < u_{\alpha/2} = 1{,}96$, dlatego brak jest podstaw do odrzucenia hipotezy zerowej, głoszącej, że średnia liczba punktów w całej populacji jest równa 30.

Powyższe zagadnienie można rozwiązać w inny sposób, wykorzystując relację (6.44).

Wartość średniej z próby porównujemy z wartościami $\bar{x}_d$ i $\bar{x}_g$. W naszym przypadku:

$$\bar{x}_d = \bar{x}_n - u_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 30{,}5 - 1{,}96\frac{6{,}4}{\sqrt{25}} \approx 28;$$

$$\bar{x}_g = \bar{x}_n + u_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 30{,}5 + 1{,}96\frac{6{,}4}{\sqrt{25}} \approx 33\,.$$

Ponieważ wartość średniej z próby leży pomiędzy wartościami $\bar{x}_d$ i $\bar{x}_g$, dlatego brak jest powodów, aby odrzucić hipotezę H_0. {kp}

Przykład 6.3

Załóżmy, obecnie, że utrzymując założenia z przykładu 6.2, sprawdzamy prawdziwość hipotezy zerowej postaci:

$$H_0: \mu = 30,$$

wobec hipotezy alternatywnej:

$$H_1: \mu = 35.$$

Naszym zadaniem jest ustalenie prawdopodobieństwa popełnienia błędu drugiego rodzaju oraz mocy testu, zakładając, że prawdopodobieństwo popełnienia błędu pierwszego rodzaju wynosi $\alpha = 0{,}05$.

Ponieważ $\mu_0 < \mu_1$ musimy zastosować przedział krytyczny jednostronny (prawostronny).

Zakładając poziom istotności $\alpha = 0{,}05$, obliczamy wartość krytyczną:

$$x_g = \mu_0 + u_\alpha \frac{\sigma}{\sqrt{n}} = 30 + 1{,}64 \frac{6{,}4}{\sqrt{25}} \approx 32{,}1.$$

Błąd pierwszego rodzaju popełniamy, gdy prawdziwa będzie hipoteza zerowa i $\bar{x}_{25} > x_g$. Prawdopodobieństwo tego błędu zostało ustalone *a priori* na poziomie $\alpha = 0{,}01$ $(P(\bar{x}_{25} > x_g | H_0 : \mu = \mu_0) = 0{,}01)$. Błąd drugiego rodzaju popełnimy, gdy $\bar{x}_{25} < x_g$, a prawdziwa będzie hipoteza alternatywna. Prawdopodobieństwo popełnienia tego błędu wynosi:

$$\beta = P(\bar{x}_{25} < x_g | H_1 : \mu = \mu_1) = P\left(\frac{\bar{x}_{25} - \mu_1}{\sigma/\sqrt{n}} < \frac{x_g - \mu_1}{\sigma/\sqrt{n}}\right) = P\left(U < \frac{32{,}1 - 35}{6{,}4/\sqrt{25}}\right) =$$

$$= P(U < -2{,}266) = \Phi(-2{,}266) = 1 - \Phi(2{,}266) = 1 - 0{,}9884 = 0{,}0116.$$

Zatem, moc testu wyniesie $1 - \beta = 1 - 0{,}0116 = 0{,}9884$. {kp}

Przykład 6.4

Wykorzystując dane z przykładu 6.2, uchylmy założenie o znajomosci odchylenia standardowego populacji σ. Załóżmy, natomiast, że wartość odchylenia standardowego została oszacowana, na podstawie próby za pomocą statystyki S i wynosi $s = 6{,}5$. Przyjmując ryzyko popełnienia błędu pierwszego rodzaju na poziomie 0,01, zweryfikujmy hipotezę zerową postaci:

$$H_0: \mu \leq 30, \text{ wobec hipotez alternatywnej } H_1: \mu > 30.$$

Z uwagi na brak znajomości odchylenia standardowego oraz małą liczebnosć proby, do weryfikacji należy wykorzystać sprawdzian (6.50). Po podstawieniu wartości otrzymamy:

$$t_{24\,0} = \frac{30{,}5 - 30}{6{,}5}\sqrt{24} = 0{,}377.$$

Wartość obliczonej statystyki porównujemy z przedziałem krytycznym:

$$\eta_k = [t_{r = 24, \alpha = 0{,}01};\ +\infty) = [2{,}492;\ +\infty).$$

Ponieważ $t_{24{,}0} = 0{,}377 < 2{,}492$, to nie ma powodów do odrzucenia hipotezy zerowej mówiącej, że średnia liczba punktów nie przekracza 30. {kp}

Przykład 6.5

W przykładzie tym wykorzystamy dane, z przykładu 5.7 (zobacz także M. Major, J. Niezgoda: *Elementy...*, *op. cit.* przykład 2.2), przedstawiające liczbę osób korzystających z Biblioteki Miejskiej w jednym z miast województwa podkarpackiego. Dane te potraktujemy jako losowy podzbiór większej zbiorowości generalnej. Przypomnijmy, że obserwację liczby osób odwiedzających bibliotekę prowadzono w ciągu $n = 100$ dni roboczych i otrzymano średnią arytmetyczną $\bar{x}_{100} = 76,1$ oraz odchylenie standardowe z próby $s = 18,74$. Naszym zadaniem jest weryfikacja hipotezy zerowej głoszącej, że średnia dzienna liczba osób odwiedzających bibliotekę jest większa lub równa 81 osób (H_0: $\mu > 81$). Podczas weryfikacji założymy, że poziom istotności $\alpha = 0,01$. W stosunku do tak sformułowanej hipotezy zerowej stawiamy hipotezę alternatywną: H_1: $\mu < 81$. Ponieważ próba losowa jest duża i nie znana jest rzeczywista wartość odchylenia standardowego populacji, dlatego do wyznaczenia przedziału ufności należy wykorzystać procedury opisane w modelu III (wzór 6.62). Wartość obliczonej statystyki wyniesie:

$$u_0 = \frac{76,1-81}{18,76}\sqrt{100} \approx -2,6$$

Przedział krytyczny ma postać $(-\infty; -2,33]$. Ponieważ wartość obliczonej statystyki $u_0 = -2,6 < -2,33$ (lub inaczej $|u_0| = 2,08 > u_\alpha = 2,33$), dlatego z prawdopodobieństwem błędu nie większym niż 0,01, należy odrzucić hipotezę zerową i przyjąć hipotezę alternatywną, głoszącą, że średnia dzienna liczna osób odwiedzających bibliotekę jest mniejsza niż 81 osób. {kp}

6.3. Weryfikacja hipotez parametrycznych dotyczących różnicy pomiędzy wariancją a ustaloną wartością

Załóżmy, że badana zmienna losowa X opisująca populację ma rozkład $N(\mu; \sigma)$. Z populacji tej pobrano próbę prostą o liczności n, na podstawie której należy zweryfikować jedną z następujących hipotez zerowych:

$$H_0: \sigma^2 = \sigma_0^2, \tag{6.69}$$

$$H_0: \sigma^2 \le \sigma_0^2, \tag{6.70}$$

$$H_0: \sigma^2 \ge \sigma_0^2, \tag{6.71}$$

wobec jednej z poniższych hipotez alternatywnych:

$$H_1: \sigma^2 \neq \sigma_0^2, \tag{6.72}$$

$$H_1: \sigma^2 > \sigma_0^2, \tag{6.73}$$

$$H_1: \sigma^2 < \sigma_0^2, \tag{6.74}$$

gdzie σ_0^2 oznacza ustaloną na podstawie pozastatystycznych przesłanek hipotetyczną wartość wariancji w zbiorowości generalnej.

Sposób przyporządkowania hipotezy alternatywnej odpowiednim hipotezom zerowym jest taki sam, jak w przypadku weryfikacji różnicy pomiędzy wartością oczekiwaną a założoną wartością. Hipotezie zerowej (6.69) można przyporządkować każdą z trzech hipotez alternatywnych, natomiast hipotezie zerowej (6.70) odpowiada hipoteza alternatywna (6.73), a hipotezie zerowej (6.71) – hipoteza alternatywna (6.74). W praktyce najczęściej weryfikuje się pary hipotez (6.70) i (6.73), gdyż zwykle nie jest korzystna sytuacja, podczas której wariancja zmiennej jest duża.

Sposób postępowania (konstrukcja testu) jest uzależniony od dwóch czynników: sposobu szacunku wariancji z próby (S^{*2} i S^2) i od liczebności próby (porównaj przedziałowa estymacja wariancji).

Jeżeli próba jest statystycznie mała ($n \leq 30$) i wariancja z próby została oszacowana za pomocą estymatora S^{*2}, to wówczas w trakcie procesu sprawdzającego należy wykorzystać charakterystykę testową χ^2 o wartościach uzyskiwanych według wzoru:

$$\chi^2_{emp} = \frac{(n-1)s^{*2}}{\sigma_0^2}. \tag{6.75}$$

Natomiast, jeśli próba jest mała, a wariancja została oszacowana przy użyciu estymatora S^2, to wowczas w procesie weryfikacji należy skorzystać ze statystyki χ^2, której wartości uzyskujemy ze wzoru:

$$\chi^2_{emp} = \frac{ns^2}{\sigma_0^2}. \tag{6.76}$$

W obydwu przypadkach statystyka z próby χ^2, ma rozkład chi-kwadrat, o $r = n - 1$ stopniach swobody. Korzystając z tablic rozkładu chi-kwadrat, przyjmując odpowiedni poziom istotności α, konstruujemy odpowiednie przedziały krytyczne. Sposób budowy przedziału krytycznego zależy od postaci hipotezy alternatywnej i wpływającej na nią hipotezy zerowej. Jeżeli hipoteza alternatywna jest postaci (6.72) $H_1: \sigma^2 \neq \sigma_0^2$, to wówczas przedział krytyczny zdefiniujemy:

$$\eta_k = (0; \chi^2_{r,1-\alpha/2}] \cup [\chi^2_{r,\alpha/2}; +\infty), \tag{6.77}$$

W przypadku, gdy hipotezę alternatywną zapiszemy jako: $H_1: \sigma^2 < \sigma_0^2$, to wówczas przedział krytyczny będzie postaci:

$$\eta_k = [\chi^2_{r,\alpha/2}; +\infty), \tag{6.78}$$

natomiast, jeżeli stawiamy hipotezę alternatywną: $H_1 : \sigma^2 > \sigma_0^2$, to wówczas przedział krytyczny będzie miał postać:

$$\eta_k = (0; \chi^2_{r,1-\alpha/2}], \tag{6.79}$$

Podobnie jak poprzednio, reguły decyzyjne przedstawiono w ujęciu tabelarycznym (zob. tablica 6.5).

Przy dużej liczności próby ($n > 30$), rozkład statystyki χ^2 dąży do rozkładu normalnego i wówczas można skorzystać ze statystyki:

$$U = \sqrt{2\chi^2_{r,0}} - \sqrt{2r-1}, \tag{6.80}$$

gdzie r jest liczbą stopni swobody.

Statystyka (6.80) ma rozkład asymptotycznie normalny o parametrach $\mu = 0$ i $\sigma = 1$. Sposób konstruowania przedziałów krytycznych i podejmowania decyzji jest analogiczny, jak podczas weryfikacji hipotez dotyczących wartości oczekiwanej (zob. podrozdział 6.2, oraz tabl. 6.1).

Tablica 6.5. Zasady podejmowania decyzji podczas weryfikacji hipotez dotyczących wariancji, w przypadku małej liczebności próby

Hipotezy / Decyzja	$H_0 : \sigma^2 = \sigma_0^2$ $H_1 : \sigma^2 \neq \sigma_0^2$	$H_0 : \sigma^2 \leq \mu_0^2$ $H_1 : \sigma^2 > \sigma_0^2$	$H_0 : \sigma^2 > \sigma_0^2$ $H_1 : \sigma^2 < \sigma_0^2$
Odrzucić H_0 na korzyść H_1	$\chi^2_{r,0} \leq \chi^2_{r,1-\alpha/2}$ lub $\chi^2_{r,0} \geq \chi^2_{r,\alpha/2}$ (6.81)	$\chi^2_{r,0} \geq \chi^2_{r,\alpha}$ (6.83)	$\chi^2_{r,0} \leq \chi^2_{r,1-\alpha}$ (6.85)
Brak podstaw do odrzucenia H_0	$\chi^2_{r,1-\alpha/2} < \chi^2_{r,0} < \chi^2_{r,\alpha/2}$ (6.82)	$\chi_{r,0} < \chi^2_{r,\alpha}$ (6.84)	$\chi^2_{r,0} > \chi^2_{r,1-\alpha}$ (6.86)

Źródło: opracowanie własne.

Przykład 6.6

Wykorzystajmy dane z przykładu 6.4, aby zweryfikować hipotezę zerową głoszącą, że wariancja liczby punktów uzyskanych z egzaminu z „Zarządzania jakością" nie przekracza 25. Podczas weryfikacji przyjmiemy współczynnik ufności na poziomie 0,05.

Hipoteza zerowa i alternatywna będą miały postać:

$$H_0 : \sigma^2 \leq 25,$$

$$H_0 : \sigma^2 > 25.$$

Przypomnijmy, z przykładu 6.4, że średnia liczba punktów w badanej próbie (n = 25) wynosiła $\bar{x}_{25} = 30{,}5$, natomiast odchylenie standardowe z próby wynosiło s = 6,5.

Z uwagi na małą liczebność próby, do weryfikacji hipotezy zerowej należy zastosować statystykę chi-kwadrat o wartościach według wzoru (6.70). Wartość obliczonej statystyki wyniesie:

$$\chi^2_{24,0} = \frac{25 \cdot 6{,}5^2}{25} = \frac{1056{,}25}{25} = 42{,}25.$$

Zakładając że α = 0,05, oraz że liczba stopni swobody r = 25 – 1 = 24, z tablic rozkładu chi-kwadrat (zob. tablica III) odczytujemy, taką wartość $\chi^2_{r,\alpha}$, dla której spełniona jest zależność $P(\chi^2_{r,0} > \chi^2_{r,\alpha}) = \alpha$. W naszym przypadku $\chi^2_{r,\alpha}$ = 36,415. Ponieważ obliczona wartość $\chi_{24,0}$ jest większa od krytycznej $\chi_{24,0.05}$, dlatego z prawdopodobieństwem błędu mniejszym bądź równym 0,05, należy odrzucić hipotezę zerową na korzyść hipotezy alternatywnej. {kp}

Przykład 6.7

Wykorzystując dane z przykładu 6.5, zweryfikować na poziomie istotności 0,01, hipotezę zerową, że odchylenie standardowe w populacji dla dziennej liczby osób odwiedzających bibliotekę nie przekracza 15 (tj. $H_0 : \sigma^2 \leq 15$) wobec hipotezy alternatywnej, że odchylenie standardowe w badanej populacji jest większe niż 15 (tj. $H_1 : \sigma^2 > 15$).

Przypomnijmy dane wejściowe: n = 100; $\bar{x}_{100} = 76{,}1$, oraz s = 18,74.

Z uwagi na dużą liczność próby, do weryfikacji można wykorzystać statystykę U postaci (6.80) o wartościach wyznaczonych według wzoru:

$$u_0 = \sqrt{2\chi^2_{r,0}} - \sqrt{2r-1}.$$

Do obliczenia wartości statystyki U jest niezbędne wcześniejsze wyznaczenie wartości statystyki $\chi^2_{r,0}$. W analizowanym przykładzie wyniesie ona:

$$\chi_{99,0} = \frac{100 \cdot 18{,}74^2}{15^2} = \frac{35118{,}76}{225} \approx 156{,}08.$$

Wartość statystyki U wyniesie:

$$u_0 = \sqrt{2 \cdot 156{,}08} - \sqrt{2 \cdot 99 - 1} \approx 17{,}668 - 14{,}036 \approx 3{,}6.$$

Korzystając z tablic dystrybuanty rozkładu normalnego standaryzowanego, odczytujemy dla α = 0,01, wartość graniczną przedziału krytycznego u_α = 2,33. Ponieważ wartość obliczonej statystyki u_0 jest większa niż wartość graniczna u_α = 2,33, dlatego z prawdopodobieństwem błędu nie większym niż 0,01, należy odrzucić weryfikowaną hipotezę zerową. {kp}

6.4. Weryfikacja hipotez parametrycznych dotyczących różnicy pomiędzy wskaźnikiem struktury a ustaloną wartością

Założmy, że zbiorowość generalna jest opisywana przez zmienną losową X o rozkładzie zero-jedynkowym, dla którego $P(X = 1) = p$. Naszym zadaniem jest porównanie rzeczywistej wartości parametru p z hipotetyczną wartością p_0. W pierwszym kroku należy postawić jedną z hipotez zerowych postaci:

$$H_0\colon p = p_0, \tag{6.87}$$

$$H_0\colon p < p_0, \tag{6.88}$$

$$H_0\colon p > p_0, \tag{6.89}$$

której przeciwstawimy odpowiednią hipotezę alternatywną:

$$H_1\colon p \neq p_0, \tag{6.90}$$

$$H_1\colon p > p_0, \tag{6.91}$$

$$H_1\colon p < p_0. \tag{6.92}$$

Sposób zestawienia hipotezy zerowej z hipotezą alternatywą odbywa się analogicznie do opisanego w punkcie 6.2 i 6.3.

Jeżeli próba jest dostatecznie liczna, to wówczas w procesie weryfikacji można wykorzystać statystykę $U{\sim}N(0,1)$, przyjmującą wartości według wzoru (zob. punkt 5.6.4):

$$u = \frac{w - p}{\sqrt{\dfrac{w(1-w)}{n}}}, \tag{6.93}$$

gdzie: $w = z/n$ jest wartością statystyki $p = W_n$ opisanej wzorem (5.89).

Jeżeli $p = p_0$, to również wyrażenie:

$$u_0 = \frac{w - p_o}{\sqrt{\dfrac{p_o(1-p_o)}{n}}}, \tag{6.94}$$

jest realizacją zmiennej losowej $U{\sim}N(0,1)$.

Sposób budowy przedziału krytycznego oraz zasady podejmowania decyzji są analogiczne jak w przypadku weryfikacji różnicy pomiędzy wartością oczekiwaną μ a ustaloną wartością μ_0.

Przykład 6.8

Z populacji studentów zdających egzamin z „Zarządzania jakością" pobrano losową próbę liczącą $n = 100$ osób i stwierdzono, że $z = 12$ osób otrzymało ocenę niedostateczną. Przyjmując poziom istotności $\alpha = 0{,}1$, zweryfikować hipotezę zerową, że faktyczna frakcja studentów, którzy nie zdali egzaminu, jest mniejsza lub równa 10% tj. $H_0: p \leq 0{,}1$, wobec hipotezy alternatywnej, że rzeczywista frakcja studentów z ocenami niedostatecznymi jest większa niż 10% tj. $H_1: p > 0{,}1$.

Frakcję studentów, którzy w badanej próbie nie zdali egzaminu, wyniesie: $w = z/n = 12/100 = 0{,}12$. Wartość statystyki z próby wyznaczamy, korzystając ze wzoru (6.94). Otrzymamy wówczas:

$$u_0 = \frac{0{,}12 - 0{,}1}{\sqrt{\dfrac{0{,}1 \cdot 0{,}9}{100}}} = \frac{0{,}02}{0{,}03} \approx 0{,}67.$$

Z tablic dystrybuanty rozkładu normalnego odczytamy, dla $\alpha = 0{,}1$ że $u_\alpha = 1{,}28$. Przedział krytyczny będzie miał postać $[1{,}28; +\infty)$. Ponieważ wartość u_0 nie leży w przedziale krytycznym, dlatego brak jest podstaw do odrzucenia hipotezy zerowej, głoszącej, że rzeczywista frakcja ocen negatywnych nie przekracza 10%. {kp}

6.5. Weryfikacja hipotez parametrycznych dotyczących dwóch wartości oczekiwanych

Załóżmy, że rozważamy dwie zbiorowości generalne o rozkładach normalnych $N(\mu_1, \sigma_1)$; $N(\mu_2, \sigma_2)$, przy czym wartości μ_1 i μ_2, pozostają nieznane. Załóżmy na początek, że znane pozostają wartości odchyleń standardowych σ_1 i σ_2. Załóżmy dalej, że formułujemy jedną z następujących hipotez zerowych:

$$H_0: \mu_1 = \mu_2, \tag{6.95}$$

$$H_0: \mu_1 \leq \mu_2, \tag{6.96}$$

$$H_0: \mu_1 \geq \mu_2, \tag{6.97}$$

oraz jedną z odpowiednich hipotez alternatywnych:

$$H_1: \mu_1 \neq \mu_2, \tag{6.98}$$

$$H_1: \mu_1 > \mu_2, \tag{6.99}$$

$$H_1: \mu_1 < \mu_2. \tag{6.100}$$

Sposób przyporządkowania hipotezy alternatywnej do odpowiedniej hipotezy zerowej odbywa się według analogicznych zasad, jak w przypadku weryfikacji jednej wartości oczekiwanej.

W dalszym ciągu rozważań założymy, że rozpatrujemy hipotezę zerową (6.95) wobec hipotezy alternatywnej (6.98). Z badanych populacji pobiera się dwie losowe próby o liczności wynoszącej odpowiednio n_1 i n_2, z których wyznacza się średnie $\overline{X}_1$ i $\overline{X}_2$.

Jak pamiętamy (zob. punkt 5.6.5), różnica $\overline{X}_1 - \overline{X}_2$ ma rozkład normalny o parametrach: $\mu_1 - \mu_2$ i $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$.

Dokonując standaryzacji zmiennej losowej $\overline{X}_1 - \overline{X}_2$, otrzymujemy następujące wyrażenie:

$$U = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}, \tag{6.101}$$

gdzie: U jest zmienną losową o rozkładzie normalnym zero-jedynkowym.

Jeżeli założymy, że hipoteza zerowa jest prawdziwa, to wówczas jej weryfikacja odbywa się na podstawie wyrażenia:

$$u_0 = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}, \tag{6.102}$$

które jest realizacją zmiennej losowej U o rozkładzie normalnym standaryzowanym.

W sytuacji braku znajomosci parametrów σ_1 i σ_2, należy je wcześniej oszacować, używając jednego z estymatorów S lub $\underline{S}$. Założmy, że odchylenia standardowe rozważanych populacji są jednakowe, tzn. $\sigma_1 = \sigma_2$. Załóżmy w dalszym ciągu, że do oszacowania odchyleń standardowych wykorzystano statystkę S^*. Odpowiednim sprawdzianem hipotezy zerowej jest wówczas (zob. wzór 5.95)[5]:

$$t_{r0} = \frac{\overline{x}_1 - \overline{x}_2}{s_p\sqrt{\frac{n_1 + n_2}{n_1 n_2}}}, \tag{6.103}$$

gdzie: $r = n_1 + n_2 - 2$, natomiast s_p jest tzw. uogólnionym oszacowaniem odchylenia standardowego $\sqrt{\sigma_1^2 + \sigma_2^2}$ i wynosi (zob. wzór 5.99):

[5] Zob. A. Iwasiewicz, Z. Paszek: *Statystyka...*, *op. cit.*, s. 225.

$$s_p = \sqrt{\frac{(n_1-1)s_1^{*2}+(n_2-1)s_2^{*2}}{(n_1-1)+(n_2-1)}}. \tag{6.104}$$

Jeżeli liczebność próby $n_1 = n_2$, to wówczas wzór (6.103) można przedstawić jako:

$$t_{r,0} = \frac{\bar{x}_1 - \bar{x}_2}{s_p}\sqrt{\frac{n}{2}}, \tag{6.105}$$

przy czym:

$$s_p = \sqrt{\frac{s_1^{*2}+s_2^{*2}}{2}}. \tag{6.106}$$

Sposób budowy przedziałów krytycznych oraz reguły decyzyjne są analogiczne, jak w przypadku testowania hipotez dotyczących różnicy pomiędzy wartością oczekiwaną, a z góry ustaloną wartością (zob. model II).

Stosowanie testu (6.103) ograniczone jest jednak tylko do przypadków, w których badane zmienne losowe są niepowiązane. Jeżeli pomiędzy zmiennymi występuje zależność, to wówczas, aby zweryfikować hipotezę o równości dwóch średnich, należy dokonać przekształcenia dwóch jednowymiarowych zmiennych losowych X_1 i X_2 w jedną zmienną dwuwymiarową (X_1, X_2) o realizacjach $(x_{1.i}, x_{2.i})$ dla $i = 1, ..., n$. W takim przypadku wyniki obu prób są traktowane jako wyniki pomiarów dla tego samego elementu zbiorowości generalnej. Typowym przypadkiem jest tu model: wynik $x_{1.i}$ przed zajściem określonego zdarzenia i wynik $x_{2.i}$, gdy zaszło określone zdarzenie dla tej samej i-tej jednostki populacji. Na przykład liczba wypadków drogowych na pewnym odcinku drogi przed wprowadzeniem znaków ograniczających prędkość i po ich wprowadzeniu.

W takiej sytuacji należy postawić hipotezę zerową:

$$H_0 : \mu_Y = 0, \tag{6.107}$$

wobec hipotezy alternatywnej:

$$H_1 : \mu_Y \neq 0, \tag{6.108}$$

gdzie: μ_Y jest średnią zmiennej losowej $Y = X_1 - X_2$, o realizacjach: $y_i = x_{1.i} - x_{2.i}$, dla $i = 1,2, \ldots, n$.

Sprawdzianem tak sformułowanej hipotezy zerowej jest wyrażenie:

$$t_{r,0} = \frac{\bar{y}_n}{s_Y^*}\sqrt{n}, \tag{6.109}$$

lub

$$t_{r,0} = \frac{\bar{y}_n}{s_Y}\sqrt{n-1}, \tag{6.110}$$

będące realizacją zmiennej losowej t-Studenta o $r = n - 1$ stopniach swobody.

Symbole we wzorach (6.109) i (6.110) oznaczają:

$\bar{y}_n$ jest wartością średniej arytmetycznej z próby, natomiast s_Y^* i s_Y oszacowaniem odchylenia standardowego zmiennej Y, uzyskanym za pomocą estymatorów S^* i S.

Podczas konstrukcji przedziału krytycznego korzysta się z tablic rozkładu Studenta, przy założeniu, że liczba stopni swobody $r = n - 1$. Zasady podejmowania decyzji są analogiczne do opisanych w tablicy 6.3.

W sytuacji, gdy dwie populacje mają rozkłady normalne o nieznanych odchyleniach standardowych, a pobrane próby n_1 i n_2 są odpowiednio liczne, to wówczas, można przyjąć w przybliżeniu, że wariancje oszacowane na podstawie prób z wykorzystaniem estymatora S^2, są w przybliżeniu równe rzeczywistym wartościom wariancji w populacji (tzn. $s_1^2 \approx \sigma_1^2$ oraz $s_2^2 \approx \sigma_2^2$). Oczywiście w takim przypadku podczas procesu weryfikacji możemy wykorzystać sprawdzian u_0 postaci (6.102).

Przykład 6.9

Na dwóch specjalnościach studiów magisterskich: Rachunkowość i Zarządzanie firmą, przeprowadzono egzamin z „Zarządzania jakością". Zakładając, że rozkłady liczby punktów w obydwu próbach są normalne, postanowiono sprawdzić hipotezę, że średnia liczba punktów na obydwu specjalnościach jest taka sama. Podczas weryfikacji założono poziom istotności $\alpha = 0{,}05$. W celu weryfikacji tak postawionej hipotezy, pobrano niezależną próbę losową 121 osób ze specjalności Rachunkowość, oraz 100 osób ze specjalności Zarządzanie firmą.

Średnia liczba punktów w poszczególnych próbach wyniosła: $\bar{x}_1 = 25{,}5$ oraz $\bar{x}_2 = 62{,}2$.

Odchylenia standardowe z badanych prób wyniosły: $s_1 = 5{,}5$ oraz $s_2 = 4{,}4$.

Stawiamy hipotezę zerową:

$$H_0\colon \mu_1 = \mu_2,$$

oraz hipotezę alternatywną:

$$H_0\colon \mu_1 \neq \mu_2.$$

Ponieważ próba losowa jest duża, można w przybliżeniu przyjąć, że:

$$s_1^2 \approx \sigma_1^2 \text{ i } s_2^2 \approx \sigma_2^2.$$

Do sprawdzenia hipotezy zerowej można użyć wzoru (6.102). Po podstawieniu wartości do wzoru otrzymamy:

$$u_0 = \frac{25{,}5 - 26{,}2}{\sqrt{\frac{5{,}5^2}{121} + \frac{4{,}4^2}{100}}} = \frac{-0{,}7}{0{,}67} \approx -1{,}045.$$

Przedział krytyczny jest w tym przypadku postaci:

$$\eta_k = (-\infty; -u_{\alpha/2}] \cup [u_{\alpha/2}; +\infty).$$

Wartości $-u_{\alpha/2}$ i $u_{\alpha/2}$ wyznaczane są z tablic dystrybuanty rozkładu $N(0,1)$, w taki sposób, aby:

$$P(U \leq -u_{\alpha/2}) = \alpha/2 = 0{,}025,$$

oraz

$$P(U \geq u_{\alpha/2}) = \alpha/2 = 0{,}025.$$

W analizowanym przykładzie przedział krytyczny będzie miał postać:

$$\eta_k = (-\infty; -1{,}96] \cup [1{,}96; +\infty).$$

Ponieważ wartość obliczonej statystyki $u_0 \in \eta_k$, dlatego brak jest powodów, aby odrzucić hipotezę zerową głoszącą równość średniej liczby punktów w badanych grupach. {kp}

Przykład 6.10

Celem badania była weryfikacja wpływu leku na obniżenie ciśnienia tętniczego krwi pacjenta. Badanie ciśnienia prowadzono przez 20 kolejnych dni, przy czym przez dziesięć pierwszych dni nie podawano lekarstwa, a przez dziesięć kolejnych dni lekarstwo było podawane. Niech X_1 oznacza zmienną losową będącą poziomem ciśnienia skurczowego krwi przed podaniem leku, natomiast X_2 oznacza zmienną losową będącą ciśnieniem skurczowym krwi po podaniu leku. Wyniki doświadczenia oraz pomocnicze obliczenia ujęto w tablicy 6.6.

Tablica 6.6. Obliczenia pomocnicze do przykładu 6.10

Dzień i	Wynik pomiaru		$y_i = x_{1.i} - x_{2.i}$	y_i
	1	2		
	Ciśnienie przed podaniem leku $x_{1.i}$ [mmHg]	Ciśnienie po podaniem leku $x_{2.i}$ [mmHg]		
1	2	3	4	5
1	141	130	11	121
2	150	138	12	144
3	160	132	28	784
4	142	121	21	441
5	166	120	46	2116
6	145	133	12	144
7	160	125	35	1225
8	180	110	70	4900
9	170	130	40	1600
10	150	132	18	324
Suma	xxx	xxx	297	11799

Zródło: badania własne.

Naszym zadaniem jest weryfikacja hipotezy zerowej, że lek nie miał istotnego wpływu na zmianę poziomu ciśnienia skurczowego krwi (czyli różnica poziomu ciśnienia przed i po podaniu leku jest równa zero, (H_0: $\mu_Y = 0$), wobec hipotezy alternatywnej, że podanie leku spowodowało istotne zmniejszenie poziomu ciśnienia skurczowego krwi (H_1: $\mu_Y \neq 0$).

Ponieważ każda para pomiarów dotyczy tego samego pacjenta, wyniki obu prób możemy potraktować jako wyniki jednej próby, której elementami będą różnice $y_i = x_{1,i} - x_{2,i}$ (zob. kolumna 4, tablica 6.6). Hipotezę zerową zweryfikujemy za pomocą testu (6.109). Do obliczenia wartości statystyki z próby niezbędne jest obliczenie średniej z próby $\bar{y}$ oraz odchylenia standardowego s_Y^*. W analizowanym przykładzie będą one wynosiły:

$$\bar{y}_n = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{297}{10} = 29.7,$$

$$s_Y^* = \sqrt{\frac{1}{n-1}\left[\sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2\right]} = \sqrt{\frac{1}{9}\left[11799 - \frac{1}{10}(29{,}7)^2\right]} \approx 36{,}1.$$

Stąd wartość statystyki z próby wyniesie:

$$t_{9,0} = \frac{29{,}7}{36{,}1}\sqrt{10} \approx 2{,}6.$$

Załóżmy, że poziom istotności testu $\alpha = 0{,}05$. Z tablic rozkładu Studenta odczytamy wartość $t_{9,0,05} = 1{,}833$. Ponieważ $t_{9,0} = 2{,}6 > t_{9,0,05} = 1{,}833$, wnioskujemy – na poziomie istotności 0,05 – odrzucenie hipotezy zerowej i przyjęcie hipotezy alternatywnej, głoszącej że średnie ciśnienie krwi przed podaniem leku jest większe niż średnie ciśnienie krwi po jego podaniu. {kp}

6.6. Weryfikacja hipotez parametrycznych dotyczących dwóch wariancji

Załóżmy, że rozważamy dwie zbiorowości opisywane przez dwie zmienne X_1 i X_2 o rozkładach normalnych z parametrami równymi odpowiednio: $N(\mu_1, \sigma_1)$ oraz $N(\mu_2, \sigma_2)$. Żaden z parametrów tych rozkładów nie jest znany.

Formułujemy hipotezę zerową postaci:

$$H_0: \sigma_1^2 = \sigma_2^2, \tag{6.111}$$

wobec hipotezy alternatywnej:

$$H_1: \sigma_1^2 \neq \sigma_2^2. \tag{6.112}$$

Najczęściej do weryfikacji tak sformułowanej hipotezy zerowej używa się wariancji z próby obliczonej przy użyciu statystyki S^{*2} o wartościach wynoszących odpowiednio: s_1^{*2} i s_2^{*2}.

Jeżeli prawdziwa jest hipoteza zerowa, to wówczas statystyka

$$F_{r_1, r_2, 0} = \frac{S_1^{*2}}{\sigma_1^2} : \frac{S_2^{*2}}{\sigma_2^2} = \frac{S_1^{*2}}{S_2^{*2}} \tag{6.113}$$

ma rozkład F – Snedecora o $r_1 = n - 1$ i $r_2 = n - 1$ stopniach swobody.

Obszar krytyczny należałoby wyznaczyć, korzystając z relacji:

$$P(F_{r_1,r_2} \leq F_{r_1,r_2,1-\alpha/2}) = \frac{\alpha}{2}, \tag{6.114}$$

oraz

$$P(F_{r_1,r_2} \geq F_{r_1,r_2,\alpha/2}) = \frac{\alpha}{2}, \tag{6.115}$$

przy czym

$$F_{r_1,r_2,1-\alpha/2} < 1, \text{ gdy } s_1^{*2} < s_2^{*2}, \text{ oraz } F_{r_1,r_2,\alpha/2} > 1, \text{ gdy } s_1^{*2} < s_2^{*2}, \tag{6.116}$$

Otrzymany wówczas przedział krytyczny będzie miał postać:

$$\eta_k = (0; F_{r_1,r_2,-\alpha/2}] \cup [F_{r_1,r_2,\alpha/2}; +\infty), \tag{6.117}$$

przy czym

$$F_{r_1,r_2,1-\alpha/2} = \frac{1}{F_{r_1,r_2,\alpha/2}}. \tag{6.118}$$

Ponieważ tablice F – Snedecora, skonstruowane są zwykle tak, że pozwalają na odczytanie wartości $F_{r_1,r_2,\alpha/2}$, określającej jedynie prawą część obszaru krytycznego, podczas wyznaczania wartości statystyki (6.113) w liczniku umieszcza się większą z wariancji z obu prób. Statystykę (6.113) przyjmującą wartości w przedziale $[0; +\infty)$ można zastąpić statystyką przyjmującą wartości z przedziału $[1; +\infty)$ postaci:

$$F_{r_1,r_2,0} = \frac{S_1^{*2}}{S_2^{*2}}, \tag{6.113a}$$

przy czym zachodzi relacja[6]:

$$S_1^{*2} > S_2^{*2}. \tag{6.119}$$

Obszar krytyczny będzie miał wówczas postać:

$$\eta_k = [F_{r_1,r_2,\alpha/2}, +\infty). \tag{6.120}$$

Reguły podejmowania decyzji są podobne do testów prawostronnych.

---

[6] Jeżeli relacja jest odwrotna, to wówczas można przenumerować zmienne opisujące badane populacje.

Jeżeli

$$F_{r_1,r_2,0} \geq F_{r_1,r_2,\alpha/2}, \tag{6.121}$$

to wówczas należy odrzucić hipotezę H_0 na korzyść hipotezy H_1, natomiast jeśli:

$$F_{r_1,r_2,0} < F_{r_1,r_2,\alpha/2}, \tag{6.122}$$

to brak jest podstaw, aby odrzucić H_0.

Statystykę postaci (6.118) można również stosować w przypadku testów jednostronnych. Wówczas hipoteza zerowa może być postaci:

$$H_0 \colon \sigma_1^2 = c\sigma_2^2, \tag{6.123}$$

lub postaci

$$H_0 \colon \sigma_1^2 \leq c\sigma_2^2, \tag{6.124}$$

gdzie[7] $c > 0$.

W stosunku do tak sformułowanej hipotezy zerowej konstruujemy hipotezę alternatywną:

$$H_1 \colon \sigma_1^2 > c\sigma_2^2. \tag{6.125}$$

Parametr σ_1^2 koresponduje z większą wariancją z próby s_1^2, natomiast σ_2^2 odpowiada tej populacji, dla której wyznaczono mniejszą wariancję z próby s_2^2.

Odpowiedni przedział krytyczny będzie miał wówczas postać:

$$\eta_k = [F_{r_1,r_2,\alpha}; +\infty). \tag{6.126}$$

Oczywiście reguły decyzyjne są następujące:
jeżeli

$$F_{r_1,r_2,0} \geq F_{r_1,r_2,\alpha}, \tag{6.127}$$

to wówczas, z prawdopodobieństwem błędu rzędu α, należy odrzucić H_0 i przyjąć H_1, natomiast w przypadku, gdy spełniona jest relacja:

$$F_{r_1,r_2,0} < F_{r_1,r_2,\alpha}, \tag{6.128}$$

wówczas brak jest podstaw do odrzucenia hipotezy H_0.

Oczywiście w stosunku do hipotezy zerowej (6.123), można zbudować hipotezę alternatywną postaci:

$$H_1 \colon \sigma_1^2 > c\sigma_2^2. \tag{6.129}$$

Zaleca się wówczas przenumerować zmienne opisujące badane populacje i zastąpić test lewostronny testem prawostronnym.

[7] Najczęściej przyjmuje się $c = 1$.

Przykład 6.11

Z dwóch populacji studentów studiów magisterskich specjalności Rachunkowość i Zarządzanie firmą, którzy przystąpili do egzaminu z „Zarządzania jakością” pobrano losowo po 25 studentów. Odchylenie standardowe liczby otrzymanych punktów wynosiło w grupie pierwszej (Rachunkowość) $s_1^* = 6{,}5$, a w grupie drugiej (Zarządzanie firmą) $s_2^* = 5{,}4$. Zakładając, że rozkład liczby punktów na obydwu specjalnościach jest normalny, zweryfikować hipotezę, że wariancje liczby punktów na obydwu specjalnościach są jednakowe (H_0: $\sigma_1^2 = \sigma_2^2$), wobec hipotezy alternatywnej, według której wariancja liczby punktów na specjalności Rachunkowość jest wyższa niż wariancja liczby punktów na specjalności Zarządzanie firmą (H_1: $\sigma_1^2 > \sigma_2^2$). Podczas weryfikacji przyjąć $\alpha = 0{,}05$.

Hipotezę zerową weryfikujemy, wykorzystując statystykę F (6.118), a następnie jej wartość porównujemy z wartością krytyczną $F_{r_1 r_2 \alpha}$, odczytaną z tablic F – Snedecora, w ten sposób, aby spełniona była zależność $P(F_{r_1 r_2 0} > F_{r_1 r_2 \alpha}) = \alpha = 0{,}05$. W analizowanym przykładzie wartość statystyki

$$F_{24,24,0} = \frac{6{,}5^2}{5{,}4^2} = \frac{42{,}25}{29{,}16} \approx 1{,}45.$$

Wartość krytyczna (zob. tablica IVa) $F_{24,24,0,05} = 1{,}98$. Ponieważ $F_{24,24,0} = 1{,}45 < F_{24,24,0,05} = 1{,}98$, dlatego nie ma powodów do odrzucenia hipotezy zerowej, głoszącej, że wariancje liczby punktów na obydwu kierunkach studiów nie różnią się w sposób znaczący. {kp}

6.7. Weryfikacja hipotez parametrycznych dotyczących dwóch wskaźników struktury (dwóch frakcji)

Przyjmijmy obecnie, że rozważamy dwie populacje opisywane przez dwie zmienne losowe X_1 i X_2, o rozkładach zero-jedynkowych z wartościami oczekiwanymi równymi odpowiednio p_1 i p_2. Wartości tych parametrów są nieznane i zostały oszacowane na podstawie dwóch licznych próbek za pomocą estymatora W o wartościach równych $w_1 = z_1/n_1$ oraz $w_2 = z_2/n_2$.

Weryfikacji poddajemy najczęściej parę hipotez postaci:

$$H_0: p_1 = p_2, \tag{6.130}$$

$$H_1: p_1 \neq p_2. \tag{6.131}$$

Jak wiemy z podrozdziału 5.6.5, w przypadku dużych liczebności próby wyrażenie:

$$u = \frac{(w_1 - w_2) - (p_1 - p_2)}{\sqrt{\frac{w_1(1-w_1)}{n_1} + \frac{w_2(1-w_2)}{n_2}}}. \tag{6.132}$$

jest realizacją zmiennej losowej $U \sim N(0,1)$.

Jeżeli prawdziwa jest hipoteza zerowa i $p_1 - p_2 = 0$, to wówczas wyrażenie (6.132) zredukowane do postaci:

$$u_0 = \frac{(w_1 - w_2)}{\sqrt{\frac{w_1(1-w_1)}{n_1} + \frac{w_2(1-w_2)}{n_2}}} \tag{6.133}$$

jest realizacją zmiennej losowej U o rozkładzie normalnym standaryzowanym.

Sposób budowy przedziału krytycznego oraz reguły decyzyjne są analogiczne, jak przy weryfikacji różnicy pomiędzy wskaźnikiem struktury p, a ustaloną z góry wartością p_0. W pewnych sytuacjach istnieją merytoryczne przesłanki, aby prostą hipotezę zerową postaci (6.130) zastąpić jedną z hipotez złożonych:

$$H_0: p_1 \leq p_2, \tag{6.134}$$

lub

$$H_0: p_1 \geq p_2. \tag{6.135}$$

Wówczas w stosunku do hipotezy (6.134), formułujemy hipotezę alternatywną postaci:

$$H_1: p_1 > p_2, \tag{6.136}$$

natomiast w stosunku do hipotezy zerowej (6.135), hipotezę alternatywną:

$$H_1: p_1 < p_2, \tag{6.137}$$

We wszystkich przypadkach, podczas weryfikacji tak postawionej hipotezy zerowej należy skorzystać ze sprawdzianu (6.133). Jeżeli hipoteza alternatywna jest postaci (6.136), to wówczas przedział krytyczny jest postaci (6.34); natomiast jeżeli hipoteza alternatywna ma postać (6.134), to przedział krytyczny będzie postaci (6.36).

Oczywiście zasady podejmowania decyzji pozostają niezmienione tzn. jeśli wartość wyznaczonej statystki z próby będzie należeć do przedziału krytycznego, to wówczas z prawdopodobieństwem błędu α, należy odrzucić H_0, natomiast, gdy wartość obliczonej statystyki z próby nie będzie należała do przedziału krytycznego, to wówczas brak jest podstaw do odrzucenia H_0.

Przykład 6.12

Postanowiono sprawdzić, czy frakcja studentów, którzy nie zdali egzaminu ze Statystyki na kierunku Administracja publiczna, nie uległa zmianie w dwóch kolejnych latach akademickich 03/04 oraz 04/05. W tym celu pobrano losowo z każdej populacji dwie próby o liczność $n_1 = n_2 = 200$ osób. W pierwszym roku akademickim liczba ocen niedostatecznych $z_1 = 36$, natomiast w drugim $z_2 = 20$. Postawiono hipotezę zerową H_0: $p_1 = p_2$, oraz hipotezę alternatywną H_1: $p_1 \neq p_2$. Weryfikację przeprowadzono na poziomie istotności $\alpha = 0{,}05$.

Po podstawieniu wartości do wzoru (6.133) otrzymamy:

$$u_0 = \frac{\left(\frac{36}{200} - \frac{20}{200}\right)}{\sqrt{\frac{\frac{36}{200}\left(1-\frac{36}{200}\right)}{200} + \frac{\frac{20}{200}\left(1-\frac{20}{200}\right)}{200}}} =$$

$$= \frac{0{,}08}{\sqrt{0{,}000738 + 0{,}00045}} = \frac{0{,}08}{0{,}034467} = 2{,}32$$

Wartość krytyczna $u_{\alpha/2} = u_{0,025} = 1{,}96$ jest mniejsza od wartości obliczonej statystyki u_0, zatem hipotezę zerową należy odrzucić na korzyść hipotezy alternatywnej i stwierdzić, że frakcje ocen niedostatecznych w poszczególnych latach istotnie różnią się między sobą. {kp}

6.8. Ocena istotności współczynnika korelacji

Załóżmy obecnie, że badamy dwuwymiarową zmienną losową (X,Y) zbiorowości generalnej, która ma dwuwymiarowy rozkład normalny z nieznanym współczynnikiem korelacji liniowej ρ_{xy}. Ze zbiorowości generalnej pobrano n-elementową próbę losową, na podstawie której wyznaczono współczynnik korelacji liniowej r_{xy} (zob. wzór (5.108)). Stawiamy hipotezę zerową głoszącą, że zmienne w populacji są nieskorelowane:

$$H_0\colon \rho_{xy} = 0, \tag{6.138}$$

wobec jednej z hipotez alternatywnych[8]:

$$H_1\colon \rho_{xy} \neq 0, \tag{6.139}$$

[8] Najczęściej hipoteza alternatywna ma postać (6.139).

$$H_1: \rho_{xy} > 0, \tag{6.140}$$

$$H_1: \rho_{xy} < 0. \tag{6.141}$$

Przypomnijmy (zob. punkt 5.6.7), że jeżeli liczebność próby jest dostatecznie liczna (kilkaset elementów), to wówczas rozkład współczynnika korelacji z próby jest asymptotycznie zbieżny do rozkładu normalnego $N\left(\rho_{xy}; \frac{1-\rho_{xy}^2}{\sqrt{n}}\right)$. Dokonując standaryzacji współczynnika korelacji z próby, otrzymamy wyrażenie:

$$u = \frac{r_{xy} - \rho_{xy}}{\frac{1-\rho_{xy}^2}{\sqrt{n}}} \tag{6.142}$$

będące wartością zmiennej $U \sim N(0; 1)$.

Jeżeli prawdziwa jest hipoteza zerowa (6.138) tzn. $\rho_{xy} = 0$, to wówczas:

$$t_{r=n-2,0} = \frac{r_{xy}}{\sqrt{1-r_{xy}^2}}\sqrt{n-2}, \tag{6.143}$$

jest jedną z możliwych realizacji zmiennej losowej o rozkładzie Studenta, o $r = n - 2$ stopniach swobody. Przedział krytyczny w zależności o konstrukcji hipotezy alternatywnej jest opisywany wzorami: (6.52) – w przypadku hipotezy dwustronnej (6.139), (6.54) – w przypadku hipotezy alternatywnej prawostronnej (6.140) lub (6.55), gdy hipoteza alternatywna jest lewostronna postaci (6.141). Reguły decyzyjne są analogiczne do opisanych w tablicy 6.3.

Jeżeli jedna ze zmiennych losowych dwuwymiarowych (X lub Y), opisująca zbiorowość generalną jest zmienną zero-jedynkową, a druga zmienną ciągłą, to wówczas zależność pomiędzy nimi jest opisywana za pomocą współczynnika korelacji dwuseryjnej ρ_{dxy}. W takiej sytuacji do oszacowania współczynnika korelacji ρ_{dxy} należy użyć współczynnik korelacji dwuseryjnej z próby, którego wartości uzyskuje się według wzoru (zob. zob. M. Major, J. Niezgoda: *Elementy statystyki...*, *op. cit.*, wzór 4.27):

$$r_{dxy} = \frac{\bar{y}_1 - \bar{y}_0}{s_y^*}\sqrt{\frac{n_1 n_0}{n(n-1)}}, \tag{6.144}$$

gdzie:

Y – zmienna ciągła, X – zmienna zero-jedynkowa,

$\bar{y}_0$ – średnia arytmetyczna realizacji zmiennej Y, skojarzonych z realizacjami zmiennej X o wartości 0,

$\bar{y}_1$ – średnia arytmetyczna realizacji zmiennej Y, skojarzonych z realizacjami zmiennej X o wartości 1,

s^* – oszacowanie odchylenia standardowego ciągłej zmiennej losowej Y,

n_0 – liczebność podzbioru zer,
n_1 – liczebność podzbioru jedynek,
$n = n_0 + n_1$.

Do weryfikacji hipotezy zerowej (6.138) wobec jednej z hipotez alternatywnych (6.139) – (6.141) można użyć sprawdzianu postaci:

$$t_{r=n-2\,0} = \frac{r_{d.xy}}{\sqrt{1-r_{d.xy}^2}}\sqrt{n-2}. \tag{6.145}$$

Oczywiście sposób budowy przedziału krytycznego i reguły decyzyjne są analogiczne do tych, jakie stosuje się w przypadku oceny istotności współczynnika korelacji liniowej ρ_{xy}.

Przykład 6.13

Wykorzystując informacje zawarte w przykładzie 5.11, zweryfikować hipotezę zerową, że korelacja pomiędzy ocenami ze Statystyki na koniec I i II semestru nieistotnie różni się od zera. Wnioskowanie należy przeprowadzić na poziomie istotności $\alpha = 0{,}01$. Przypomnijmy, że współczynnik korelacji liniowej pomiędzy ocenami na koniec I i na koniec II semestru wynosił $r_{xy} = 0{,}46$. Wartość tego współczynnika została wyznaczona na bazie próby losowej liczącej $n = 103$ osoby.

Stawiamy hipotezę zerową:

$$H_0\colon \rho_{xy} = 0,$$

wobec hipotezy alternatywnej:

$$H_0\colon \rho_{xy} \neq 0.$$

Korzystając ze wzoru (6.143), otrzymamy:

$$t_{r=n-2\,0} = \frac{0{,}46}{\sqrt{1-0{,}46^2}}\sqrt{103-2} = \frac{4{,}63}{0{,}8879} = 5{,}2.$$

Z tablic rozkładu Studenta odczytamy, że $t_{101,\alpha/2=0{,}005} = 2{,}576$. Ponieważ wartość obliczonej statystyki jest $t_{101,0} = 5{,}2 > t_{101,0,0{,}005} = 2{,}576$, dlatego z prawdopodobieństwem błędu 0,01, należy odrzucić hipotezę zerową i przyjąć hipotezę alternatywną. Pomiędzy wynikami na koniec I i II semestru istnieje istotna nielosowa współzależność. {kp}

6.9. Ocena istotności współczynnika regresji liniowej[9]

Badanie istotności współczynnika regresji liniowej sprowadza się do zweryfikowania hipotezy zerowej, głoszącej, że współczynnik regresji liniowej jest równy zero, co przekłada się na stwierdzenie, że funkcja regresji w zbiorowości generalnej jest stała, co wskazuje na brak zależności pomiędzy badanymi zmiennymi. Jeżeli natomiast współczynnik regresji okaże się istotnie różny od zera, to oznacza to, że istnieje wyraźna zależność pomiędzy badanymi zmiennymi, a funkcję regresji można użyć jako narzędzia prognozy wartości jednej badanej cechy na podstawie wartości drugiej skorelowanej z nią cechy.

Podstawowe założenie, przy którym konstruuje się omawiany test istotności, dotyczy normalności rozkładu dwuwymiarowej zmiennej losowej (X,Y). Z badanej zbiorowości generalnej pobiera się n-elementową próbę losową o wartościach (x_i, y_i) dla $i = 1,2, \ldots, n$.

Na podstawie otrzymanych wyników z próby stosując metodę najmniejszych kwadratów, szacuje się parametry funkcji regresji i otrzymuje się funkcję $y=ax+b$.

Na podstawie badanej próby weryfikujemy hipotezę zerową, że współczynnik regresji A liniowej funkcji regresji w populacji ma określoną wartość a_0. Najczęściej zakłada się, że $a_0 = 0$, co oznacza brak zależności między badanymi zmiennymi. Zatem można zapisać:

$$H_0{:}\, A = 0. \tag{6.146}$$

Wobec tak sformułowanej hipotezy zerowej stawia się jedną z hipotez alternatywnych postaci:

$$H_1{:}\, A \neq 0, \tag{6.147}$$

$$H_1{:}\, A > 0, \tag{6.148}$$

$$H_1{:}\, A < 0, \tag{6.149}$$

przy czym najczęściej stosuje się hipotezę alternatywną (6.147).

W trakcie weryfikacji tak sformułowanych hipotez wykorzystuje się statystykę t o wartościach wyznaczonych ze wzoru:

$$t_{r=n-2,0} = \frac{a-a_0}{s_\varepsilon}\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}, \tag{6.150}$$

gdzie:

a – jest oszacowanym na podstawie próby współczynnikiem regresji,

a_0 – jest hipotetycznym współczynnikiem regresji,

s_ε – jest wartością błędu standardowego estymacji funkcji regresji wyznaczoną ze wzoru:

---

[9] Przed przeczytaniem tego podrozdziału zaleca się czytelnikowi przestudiowanie zagadnień związanych z liniową funkcją regresji dwóch zmiennych losowych (zob. np. par. 4.1.2 w pracy: M. Major, J. Niezgoda: *Elementy statystyki..., op. cit.*).

$$s_e = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{n-2}}. \tag{6.151}$$

Wzór (6.150) można również zapisać w krótszej postaci:

$$t_{r=n-2,0} = \frac{a-a_0}{s_a}, \tag{6.152}$$

gdzie s_a jest wartością tzw. średniego błędu szacunku współczynnika regresji obliczoną ze wzoru:

$$s_a = \frac{s_e}{\sum_{i=1}^{n}(x_i-\bar{x})^2}. \tag{6.153}$$

Wartość średniego błędu szacunku współczynnika regresji informuje nas, o ile przeciętnie (*in plus lub in minus*) popełniono błąd, szacując wartość współczynnika regresji w populacji na podstawie wyników otrzymanych z próby.

Statystyka t o wartościach wyznaczonych ze wzoru (6.150) lub wzoru (6.152), przy założeniu prawdziwości hipotezy zerowej, posiada rozkład t Studenta o $r = n - 2$ stopniach swobody. Jeżeli hipoteza alternatywna ma postać (6.147), to wówczas korzystając z tablic rozkładu Studenta, należy odczytać dla ustalonego poziomu ufności α oraz liczby stopni swobody $r = n - 2$, taką wartości $t_{r,\alpha/2}$, aby zachodziło [zob. (6.53)]: $P(t_r < -t_{r,\alpha/2}) = P(t_r \geq t_{r,\alpha/2}) = \alpha/2$. Dwustronny przedział krytyczny ma wówczas postać (6.52): $\eta_k = (-\infty; -t_{r,\alpha/2}] \cup [t_{r,\alpha/2}; +\infty)$. Jeżeli hipoteza alternatywna jest postaci (6.148) lub (6.149), to wówczas w teście zamiast dwustronnego przedziału krytycznego należy użyć odpowiednio prawostronnego lub lewostronnego przedziału krytycznego [zob. wzory (6.54) i (6.55)]. Jeżeli wartość obliczonej statystyki $t_{r,0}$ zawiera się w przedziale krytycznym, to wówczas należy odrzucić hipotezę zerową, w przeciwnym wypadku należy stwierdzić brak podstaw do odrzucenia hipotezy H_0.

Przykład 6.14

W celu oszacowania zależności pomiędzy powierzchnią mieszkania (typu M1) (x_i [m^2]) i jego ceną (y_i), w Krakowie, pobrano losowo 10 ofert i otrzymano następujące dane:

Tablica 6.7. Powierzchnia i cena mieszkania typu M1 w Krakowie

i	**Powierzchnia mieszkania x_i [m^2]**	**Cena y_i [tys. zł]**
1	16	69
2	18	73
3	20	59
4	24	77
5	25	74
6	29	82
7	35	81
8	39	83
9	40	103
10	50	129

Źródło: badania własne, listopad 1999 r.

Na podstawie powyższych danych wyznaczyć równanie regresji liniowej i zweryfikować hipotezę zerową, że współczynnik regresji w populacji mieszkań nieistotnie różni się od zera. Podczas weryfikacji założyć, że poziom ufności $\alpha = 0{,}05$.

Korzystając ze wzorów i zależności opisanych w podręczniku M. Major, J. Niezgoda: *Elementy statystyki...*, *op. cit.*, rozdział 6.1.2, wzory: (4.15) i (4.16) szacowane wartości parametrów a i b wynoszą odpowiednio (całość niezbędnych obliczeń zawiera tablica 6.8):

$$a = \frac{\sum_{i=1}^{10}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{10}(x_i - \bar{x})^2} = \frac{622{,}2}{236{,}9} = 2{,}626425,$$

$$b = \bar{y} - a\bar{x} = 61{,}8 - 2{,}626425 \cdot 25{,}1 = -4{,}12326.$$

Zatem funkcja regresji ma postać: $y_i = 2{,}626425x_i - 4{,}12326$.

Parametr a nazywany jest współczynnikiem regresji. W analizowanym przykładzie parametr ten informuje nas, że jeżeli powierzchnia mieszkania wzrośnie o jednostkę (1 m^2), to wówczas cena mieszkania wzrośnie średnio o 2,63 tys. zł.

Natomiast wyraz wolny b wskazuje, ile będzie wynosić średnia wartość zmiennej Y, jeżeli zmienna $X = 0$. W analizowanym przypadku parametr ten nie ma logicznej interpretacji, gdyż nie można żądać zgodnie z prawem pieniędzy za mieszkanie, które nie istnieje (powierzchnia mieszkania nie może wynosić 0 m^2).

Błąd standardowy estymacji funkcji regresji wyniesie:

$$s_{\varepsilon} = \sqrt{\frac{573{,}4386}{10-2}} \approx 8{,}4664,$$

natomiast średni błąd szacunku współczynnika regresji:

$$s_{a} = \frac{8{,}4664}{\sqrt{236{,}9}} = 0{,}550067 \approx 0{,}55.$$

Tablica 6.8. Obliczenia pomocnicze do przykładu 6.14

i	pow. w m² x_i	cena w mln zł y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i-\bar{x})\cdot(y_i-\bar{y})$	$(x_i-\bar{x})^2$	$\hat{y}_i$	$(y_i-\hat{y}_i)^2$
1	20	60	–5,1	–1,8	9,18	26,01	48,40523	134,4386
2	33	85	7,9	23,2	183,28	62,41	82,54875	6,008603
3	27	50	1,9	–11,8	–22,42	3,61	66,79021	281,911
4	27	73	1,9	11,2	21,28	3,61	66,79021	38,56153
5	33	90	7,9	28,2	222,78	62,41	82,54875	55,52106
6	25	58	–0,1	–3,8	0,38	0,01	61,53736	12,5129
7	18	46	–7,1	–15,8	112,18	50,41	43,15238	8,108911
8	24	55	–1,1	–6,8	7,48	1,21	58,91093	15,2954
9	20	46	–5,1	–15,8	80,58	26,01	48,40523	5,785152
10	24	55	–1,1	–6,8	7,48	1,21	58,91093	15,2954
Σ	**251**	**618**	**0,00**	**0,00**	**622,2**	**236,9**	xxx	**573,4386**

Źródło: obliczenia własne.

Stawiamy hipotezę zerową: H_0: $A = 0$, oraz hipotezę alternatywną H_1: $A \neq 0$, Wartość charakterystyki z próby po podstawieniu do wzoru (6.152) wyniesie:

$$t_{8,0} = \frac{2{,}63 - 0}{0{,}550067} \approx 4{,}77.$$

Wartość krytyczna $t_{8;0,025} = 2{,}306$. Ponieważ $|t_{8,0}| = 4{,}77 > 2{,}306 = t_{8;0,025}$, należy odrzucić hipotezę zerową. Zatem można stwierdzić, że współczynnik regresji istotnie różni się od zera, co świadczy o występowaniu zależności pomiędzy powierzchnią mieszkania (X) i ceną mieszkania (Y). {kp}

6.10. Wybrane nieparametryczne testy istotności

6.10.1. Test niezależności dwóch cech

Test niezależności stosowany jest najczęściej w przypadku badania niezależności dwóch cech niemierzalnych (jakościowych) lub w przypadku badania niezależności cechy niemierzalnej i cechy ilościowej. Oznaczmy badane cechy symbolami B_1 i B_2, przy czym dla cechy B_1 można wyróżnić k_1, a dla cechy B_2, k_2 rozróżnialnych stanów. Stany te oznaczymy odpowiednio:

$$b_{1i} \text{ oraz } b_{2j}, \text{ przy czym } i = 1, ..., k_1 \text{ oraz } j = 1, ..., k_2.$$

Ze zbiorowości generalnej należy pobrać n-elementową próbę losową (przy czym ważne jest, aby liczebność próby wynosiła co najmniej 30 elementów). Wyniki próby klasyfikuje się w postaci tzw. tablicy wielodzielczej (kontygencji, niezależności) według jednej cechy (B_1) w k_1 wierszach i według drugiej cechy (B_2) w k_2 kolumnach. Wnętrze tablicy wielodzielczej stanowią liczebności n_{ij} elementów próby, które spełniają jednocześnie kryteria zawarte w i-tym wierszu i j-tej kolumnie. Przykład takiej tablicy zaprezentowano na rysunku 40.

Podział na kategorie powinien być taki, aby $n_{ij} > 5$. Po zsumowaniu wierszy i kolumn otrzymujemy odpowiednie liczebności brzegowe $n_{i\cdot}$ dla cechy B_1 i $n_{\cdot j}$ dla cechy B_2.

Zachodzą następujące zależności:

$$n_{i\cdot} = \sum_{j=1}^{k_2} n_{ij}, \tag{6.154}$$

$$n_{\cdot j} = \sum_{i=1}^{k_1} n_{ij}, \tag{6.155}$$

$$\sum_{i=1}^{k_1}\sum_{j=1}^{k_2} n_{ij} = \sum_{i=1}^{k_1} n_{i\cdot} = \sum_{j=1}^{k_2} n_{\cdot j} = n. \tag{6.156}$$

Tablica wielodzielcza jest podstawą weryfikacji nieparametrycznej hipotezy zerowej głoszącej, że w populacji nie ma zależności między cechami B_1 i B_2. Możemy więc zapisać:

$$H_0\text{: cechy } B_1 \text{ i } B_2 \text{ są niezależne,} \tag{6.157}$$

$$H_1\text{: cechy } B_1 \text{ i } B_2 \text{ są zależne.} \tag{6.158}$$

Powyższe hipotezy można również zapisać w kategoriach prawdopodobieństwa. Niezależność cech gwarantuje, że prawdziwa jest hipoteza zerowa:

$$H_0\text{: } P(B_1 = b_{1i}, B_2 = b_{2j}) = P(B_1 = b_{1i})\ P(B_2 = b_{2j}), \tag{6.159}$$

wobec której formułowana jest hipoteza alternatywna:

$$H_1\text{: } P(B_1 = b_{1i}, B_2 = b_{2j}) \neq P(B_1 = b_{1i})\ P(B_2 = b_{2j}). \tag{6.160}$$

		B_2						
		b_{21}	...	b_{2j}	...	b_{2k}	Σ	
B_1	b_{11}	n_{11} / $\hat{n}_{11}$	...	n_{1j} / $\hat{n}_{1j}$	...	n_{1k_2} / $\hat{n}_{1k}$	$n_{1.}$	$p_{1.}$
	$\vdots$	$\vdots$	...	$\vdots$	...	$\vdots$	$\vdots$	$\vdots$
	b_{1i}	n_{i1} / $\hat{n}_{i1}$	...	n_{ij} / $\hat{n}_{ij}$	...	n_{ik_2} / $\hat{n}_{ik_2}$	$n_{i.}$	$p_{i.}$
	$\vdots$	$\vdots$	...	$\vdots$	...	$\vdots$	$\vdots$	$\vdots$
	b_{1k_1}	$n_{k_1 1}$ / $\hat{n}_{k_1 1}$	...	$n_{k_1 j}$ / $\hat{n}_{k_1 j}$	...	$n_{k_1 k_2}$ / $\hat{n}_{k_1 k_2}$	$n_{k_1.}$	$p_{k_1.}$
	Σ	$n_{.1}$	...	$n_{.j}$	...	$n_{.k_2}$	n	
		$p_{.1}$	...	$p_{.j}$	...	$p_{.k_2}$		1,00

Rys. 40. Tablica wielodzielcza

Źródło: opracowanie własne.

Na podstawie liczebności brzegowych $n_{i.}$ i $n_{.j}$ wyznacza się szacunkowe wielkości prawdopodobieństw brzegowych i umieszcza w skrajnej prawej kolumnie i w dolnym wierszu tablicy wielodzielczej (zob. rys. 40). Prawdopodobieństwa brzegowe wyznacza się przez podzielenie liczebności brzegowych przez liczebność próby n.

Zatem możemy zapisać:

$$p_{i.} = \frac{n_{i.}}{n}, \tag{6.161}$$

oraz

$$p_{.j} = \frac{n_{.j}}{n}. \tag{6.162}$$

Jeżeli założymy, że prawdziwa jest hipoteza zerowa (6.157) i w konsekwencji (6.159), to wówczas zachodzi zależność:

$$p_{i.} \cdot p_{.j} = p_{ij}, \tag{6.163}$$

przy czym

$$\sum_{i=1}^{k_1} \sum_{j=1}^{k_2} p_{ij} = 1. \tag{6.164}$$

Otrzymane w ten sposób prawdopodobieństwo nazwiemy hipotetycznym prawdopodobieństwem zajścia kombinacji ($B_1 = b_{1i}$, $B_2 = b_{2j}$). Po przemnożeniu hipotetycznych prawdopodobieństw przez liczebność próby otrzymamy tzw. liczebności teoretyczne (oczekiwane):

$$\hat{n}_{ij} = np_{ij}, \tag{6.165}$$

które umieszczamy w prawych dolnych rogach odpowiedniej ij – tej komórki tablicy wielodzielczej (zob. rys. 40).

Jeżeli prawdziwa jest hipoteza zerowa (6.157), to wówczas, statystyka:

$$\chi^2_{r,0} = \sum_{i=1}^{k_1} \sum_{j=1}^{k_2} \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}}, \tag{6.166}$$

ma rozkład chi-kwadrat $r = (k_1 - 1)(k_2 - 1)$ stopniach swobody. Obszar krytyczny ma postać (zob. 6.78):

$$\eta_k = [\chi^2_{r,\alpha}; +\infty), \tag{6.167}$$

gdzie $\chi^2_{r,\alpha}$ jest wartością odczytaną z tablic rozkładu chi-kwadrat w ten sposób, aby spełniona była zależność:

$$P(\chi^2_{r,0} \geq \chi^2_{r,\alpha}) = \alpha. \tag{6.168}$$

Reguły decyzyjne są analogiczne jak przy prawostronnych parametrycznych testach istotności (zob. tab. 6.5 (6.83) i (6.84)).

Przykład 6.15

Oceny ze statystyki na koniec semestru, na II roku studiów dziennych, kierunku towaroznawstwa kształtowały się następująco:

Tablica 6.9. Dane do przykładu 6.14

Ocena / Płeć	ndst	dst	+dst	db	+db	bdb
Kobiety	8	19	16	24	18	17
Mężczyźni	10	13	9	14	13	10

Źródło: badania własne.

Wykorzystując test niezależności chi-kwadrat, zbadać, czy rozkład stopni ze statystyki jest zależny od płci studenta. Podczas weryfikacji założyć, że poziom istotności $\alpha = 0{,}01$.

Hipoteza zerowa ma postać:

H_0: ocena ze statystyki nie zależy od płci studenta,

natomiast hipoteza alternatywna brzmi:

H_1: ocena ze statystyki zależy od płci zdającego.

W kolejnych krokach wyznaczamy i umieszczamy w tablicy wielodzielczej (zob. tab. 6.10): liczebności brzegowe $n_{i.}$ i $n_{.j}$ oraz prawdopodobieństwa brzegowe $p_{i.}$ i $p_{.j}$. Na podstawie prawdopodobieństw brzegowych wyznaczamy następnie prawdopodobieństwa hipotetyczne p_{ij}.

Wartości obliczonych prawdopodobieństw hipotetycznych umieszczono w tablicy roboczej 6.11.

Mnożąc prawdopodobieństwa hipotetyczne przez liczebność próby n otrzymujemy liczebności teoretyczne $\hat{n}_{ij}$.

Tablica 6.10. Obliczenia pomocnicze do przykładu 6.14

Płeć \ Ocena			ndst	dst	+dst	db	+db	bdb	$n_{i.}$	$p_{i.}$
			j							
			1	2	3	4	5	6		
Kobiety	i	1	8	19	16	24	18	17	102	0,4
Mężczyźni		2	10	13	9	14	13	10	69	0,6
$n_{.j}$			18	32	25	38	31	27	171	xxx
$p_{.j}$			0,105	0,187	0,146	0,222	0,181	0,158	xxx	1,00

Źródło: obliczenia własne.

Tablica 6.11. Wartości prawdopodobieństw hipotetycznych

Płeć \ Ocena			ndst	dst	+dst	db	+db	bdb	$n_{i.}$	$p_{i.}$
			j							
			1	2	3	4	5	6		
Kobiety	i	1	0,047	0,129	0,088	0,164	0,123	0,099	102	0,4
Mężczyźni		2	0,025	0,070	0,047	0,089	0,066	0,054	69	0,6
$n_{.j}$			18	32	25	38	31	27	171	xxx
$p_{.j}$			0,105	0,187	0,146	0,222	0,181	0,158	xxx	1,00

Źródło: obliczenia własne.

Mnożąc prawdopodobieństwa hipotetyczne przez liczebność próby n, otrzymujemy liczebności oczekiwane (teoretyczne) $\hat{n}_{ij}$, które umieszczamy w prawym dolnym rogu każdej komórki ij, tablicy wielodzielczej (zob. tab. 6.12). Podczas obliczeń wartości zaokrąglono do dwóch miejsc po przecinku.

Tablica 6.12. Liczebności empiryczne n_{ij} i oczekiwane $\hat{n}_{ij}$.

Płeć \ Ocena		ndst	dst	+dst	db	+db	bdb	$n_{i.}$
		j						
		1	2	3	4	5	6	
Kobiety	i = 1	8 / 10,74	19 / 19,09	16 / 14,91	24 / 22,67	18 / 18,49	17 / 16,11	72
Mężczyźni	i = 2	10 / 7,26	13 / 12,91	9 / 10,09	14 / 15,33	13 / 12,51	10 / 10,89	39
$n_{.j}$		8	22	15	28	21	17	111

Źródło: obliczenia własne.

Wartość statystyki wyniesie:

$$\chi^2_{5,0} = \frac{(8-10{,}74)^2}{10{,}74} + \frac{(19-19{,}09)^2}{19{,}09} + \frac{(16-14{,}91)^2}{14{,}91} + \frac{(24-22{,}67)^2}{22{,}67} +$$

$$+ \frac{(18-18{,}49)^2}{18{,}49} + \frac{(17-16{,}11)^2}{16{,}11} + \frac{(10-7{,}26)^2}{7{,}26} + \frac{(13-12{,}91)^2}{12{,}91} +$$

$$+ \frac{(9-10{,}09)^2}{10{,}09} + \frac{(14-15{,}33)^2}{15{,}33} + \frac{(13-12{,}51)^2}{12{,}51} + \frac{(10-10{,}89)^2}{10{,}89} =$$

$$= 0{,}6976 + 0{,}0004 + 0{,}0793 + 0{,}0784 + 0{,}0130 + 0{,}0497 + 1{,}0313 +$$

$$+ 0{,}0006 + 0{,}1173 + 0{,}1159 + 0{,}0193 + 0{,}0735 = 2{,}28.$$

Zakładając $\alpha = 0{,}01$ oraz liczbę stopni swobody $r = (2-1)\cdot(6-1) = 5$, z tablic rozkładu chi-kwadrat odczytujemy: $\chi^2_{5,0,01} = 15{,}086$. Ponieważ zachodzi nierówność $\chi^2_{5,0} < \chi^2_{5,0,01}$, dlatego brak jest powodów, aby odrzucić hipotezę zerową głoszącą brak zależności pomiędzy płcią studenta i stopniem ze statystyki. {kp}

6.10.2. Test zgodności chi-kwadrat

Do najstarszych testów zgodności należy test zgodności chi-kwadrat, który pozwala zweryfikować hipotezę zerową, że populacja ma określony typ rozkładu tzn. określoną postać dystrybuanty.

Hipoteza zerowa oraz hipoteza alternatywna mają postać:

$$H_0\text{: } F(x) = F, \tag{6.169}$$

$$H_1\text{: } F(x) \neq F, \tag{6.170}$$

gdzie: $F(x)$ oznacza nieznaną dystrybuantę zmiennej losowej X, natomiast F oznacza postulowaną klasę dystrybuant.

Wybór hipotetycznego rozkładu oraz postaci dystrybuanty odbywa się na drodze analizy rodzaju zmiennej losowej (zmienna ciągła lub zmienna skokowa) oraz graficznej prezentacji rozkładu (kształt histogramu lub wieloboku liczności)[10]. I tak, np. jeżeli zmienna losowa jest ciągła, a jej rozkład jest symetryczny i ma jedną modalną, to wówczas można założyć, że hipotetyczny rozkład to rozkład normalny. Natomiast, jeżeli rozkład jest skrajnie asymetryczny lewostronnie, to prawdopodobnie w przypadku ciągłej zmiennej losowej, mamy do czynienia z rozkładem wykładniczym, a gdy zmienna jest dyskretna i jej rozkład jest asymetryczny prawostronnie, z rozkładem Poissona.

Do sprawdzenia hipotezy zerowej (6.169) stosuje się statystykę postaci:

$$\chi^2_{r,\alpha} = \sum_{j=1}^{k} \frac{\left(f_j - \hat{f}_j\right)}{\hat{f}_j} = \sum_{j=1}^{k} \frac{(f_j - np_j)^2}{np_j} \tag{6.171}$$

gdzie:

- f_j – liczebność przedziału klasowego $(x_{d.j}; x_{g.j}]$ w przypadku ciągłej zmiennej losowej X, lub liczebność z jaką występowała j – ta wartość zmiennej losowej dyskretnej X,
- $\hat{f}_j$ – hipotetyczna (oczekiwana) liczebność przedziału klasowego $(x_{d.j}; x_{g.j}]$ w przypadku ciągłej zmiennej losowej X, lub hipotetyczna (oczekiwana) liczebność, z jaką powinna występować j – ta wartość zmiennej losowej dyskretnej X,
- p_j – prawdopodobieństwo, że badana ciągła zmienna losowa X przyjmie wartość z przedziału klasowego $(x_{d.j}; x_{g.j}]$, lub prawdopodobieństwo, że dyskretna zmienna losowa X przyjmie wartość x_j,
- n – liczebność próbki, przy czym $\sum_{j=1}^{k} f_j = n$,
- k – liczba przedziałów klasowych na które został rozcięty zbiór wartości zmiennej losowej ciągłej X, lub liczba wariantów wartości dyskretnej zmiennej losowej X.

Omawiany test można stosować, gdy liczebności empiryczne poszczególnych przedziałów klasowych szeregu rozdzielczego przedziałowego lub liczebności przypisane j – tym wartościom zmiennej losowej w szeregu rozdzielczym punktowym, wynoszą, co najmniej 5. W sytuacji gdy nie jest spełniony ten warunek, liczebności te należy połączyć z licznościami sąsiednimi. Oczywiście wpłynie to na zmniejszenie liczby stopni swobody.

Statystyka (6.171) – w przypadku, gdy prawdziwa jest hipoteza zerowa (6.169) – ma rozkład chi-kwadrat o $r = k - l_p - 1$ stopniach swobody. Symbol l_p oznacza liczbę parametrów rozkładu, które zostały oszacowane na podstawie próby.

[10] Zob. M. Major, J. Niezgoda: *Elementy statystyki...*, *op. cit.*, rys. 3.2 i 3.3.

Wartości statystyki z próby (6.171) są porównywane z wartością krytyczną odczytaną z tablicy III, a zasady podejmowania decyzji są analogiczne do reguł opisanych w tab. 6.5 wzory (6.83) i (6.84).

Przykład 6.16

Skorzystajmy z danych z przykładu 2.2, z: M. Major, J. Niezgoda: *Elementy statystyki..., op. cit.*, przedstawiających strukturę liczby osób korzystających z biblioteki miejskiej w 100 dniach pracy. Dane te potraktujemy obecnie jako próbę losową większej zbiorowości generalnej. Podczas obliczeń wykorzystamy dane zestawione w szeregu rozdzielczym przedziałowym (zob. M. Major, J. Niezgoda: *Elementy statystyki..., op. cit.*, tab. 4.11 oraz. tab. 2.2). Rozkład liczebności w przedziałach klasowych oraz jedna wartość modalnej, pozwała na wysunięcie tezy, że może to być rozkład normalny. Stawiamy hipotezę zerową H_0: $F(x) = F_N$, oraz hipotezę alternatywną H_1: $F(x) \neq F_N$, przy czym symbol F_N oznacza w tym przypadku klasę dystrybuant normalnych. Podczas weryfikacji hipotezy zerowej załóżmy, że poziom istotności $\alpha = 0{,}05$.

Parametry rozkładu μ i σ szacujemy, obliczając średnią z próby ($\bar{x}_n$) i odchylenie standardowe z próby (s^*).

Tablica 6.13. Obliczanie pomocnicze do przykładu 6.16.

j	$(x_{d,j}; x_{g,j}]$	f_j	x_j^o	$f_j x_j^o$	$f_j (x^o - \bar{x}_n)^2$
1	2	3	4	5	6
1	(36; 48]	6	42	252	6822,23
2	(48; 60]	14	54	756	6604,618
3	(60; 72]	26	66	1716	2456,438
4	(72; 84]	24	78	1872	124,7616
5	(84; 96]	15	90	1350	3058,776
6	(96; 108]	9	102	918	6215,746
7	(108; 120]	4	114	456	5861,434
8	(120; 132]	2	126	252	5056,157
xxx	SUMA	100	xxx	7572	36200,16

Źródło: obliczenia własne.

Korzystając z obliczeń pomocniczych w tablicy 6.13, otrzymamy:

$$\bar{x}_n = \frac{\sum_{j=1}^{k} f_j x_j^o}{n} = \frac{7572}{100} = 75{,}72.$$

$$s^* = \sqrt{\frac{\sum_{j=1}^{k} f_j (x_j^o - \bar{x}_n)^2}{n-1}} = \sqrt{\frac{36200{,}16}{99}} \approx 19{,}12$$

W kolejnym kroku standaryzujemy górne krańce przedziałów klasowych, a następnie dla otrzymanych wartości wyznaczamy dystrybuantę rozkładu $N(0,1)$. Podczas standaryzacji górnych krańców przedziałów klasowych kres górny ostatniego przedziału klasowego przesuwa się do $+\infty$. Na podstawie otrzymanych dystrybuant wyznaczamy p_j oraz $\hat{f}_j = np_j$ a następnie podstawiamy do wzoru (6.171). Z uwagi na małą liczebność połączone zostały dwa ostatnie przedziały szeregu rozdzielczego. Liczba przedziałów klasowych wyniesie obecnie 7. Wpłynie to na zmianę stopni swobody. Całość obliczeń pomocniczych zestawiono w tablicy 6.14.

Tablica 6.14. Obliczanie pomocnicze do przykładu 6.16 cd.

j	$(x_{dj}; x_{gj}]$	$\frac{x_{gj}-\bar{x}_n}{s^*}$	$\Phi\left(\frac{x_{gj}-\bar{x}_n}{s^*}\right)$	f_j	p_j	$\hat{f}_j = np_j$	$\frac{(f_j-\hat{f}_j)^2}{\hat{f}_j}$
1	2	3	4	5	6	7	8
1	(36; 48]	–1,45	0,0735	6	0,0735	7,3500	0,2480
2	(48; 60]	–0,82	0,2061	14	0,1326	13,2600	0,0413
3	(60; 72]	–0,19	0,4247	26	0,2186	21,8600	0,7841
4	(72; 84]	0,43	0,6664	24	0,2417	24,1700	0,0012
5	(84; 96]	1,06	0,8554	15	0,1890	18,9000	0,8048
6	(96; 108]	1,69	0,9545	9	0,0991	9,9100	0,0836
7	(108; 120]	2,32	0,9898	4 } 6	0,0353 } 0,0455	4,5500	0,4621
8	(120; 132]	*	1,0000	2	0,0102		
SUMA		xxx	xxx	100	1,0000	xxx	**2,4249**

Źródło: obliczenia własne.

Suma wartości kolumny (7) jest szukaną wartością statystyki $\chi^2_{r\,0}$.
Liczba stopni swobody wynosi $r = k - l_p - 1 = 7 - 2 - 1 = 4$.

Korzystając z tablic rozkładu chi-kwadrat (zob. tab. III) oraz przyjmując $\alpha = 0{,}05$, można odczytać, że $\chi^2_{4;0,05} = 9{,}488$.

Ponieważ zachodzi nierówność $\chi^2_{r\,0} = 2{,}4249 < \chi^2_{4;0,05} = 9{,}488$, brak jest podstaw, aby odrzucić hipotezę zerową, że badana zmienna losowa ma rozkład normalny. {kp}

6.10.3. Test serii

Test serii najczęściej stosuje się do weryfikacji hipotezy, że próba ma charakter losowy. Na podstawie testu serii można również sprawdzić przypuszczenia, że dwie próby pochodzą z tej samej populacji (mają takie same rozkłady).

6.10.3.1. Weryfikacja hipotezy o losowości próby

Weryfikacji podawana jest następująca nieparametryczna hipoteza zerowa oraz hipoteza alternatywna:

$$H_0\text{: próba ma charakter losowy,} \tag{6.172}$$

$$H_1\text{: próba jest nielosowa.} \tag{6.173}$$

Sprawdzenie hipotezy zerowej (6.172) rozpoczyna się od uporządkowania wszystkich wyników pobranej próby w szereg szczegółowy uporządkowany i wyznaczenia z tego szeregu wartości mediany ($x_{me,n}$) (zob. M. Major, J. Niezgoda: *Elementy statystyki*..., *op. cit.*, 3.2.1). Po wyznaczeniu mediany powraca się do pierwotnego uporządkowania wyników (do szeregu szczegółowego nieuporządkowanego) i poszczególne obserwacje próby x_i zastępuje się symbolami a lub b, stosując zasadę: jeżeli $x_i < x_{me,n}$, to x_i należy zastąpić symbolem a, natomiast, jeżeli $x_i > x_{me,n}$, to x_i należy zastąpić symbolem b.

Każdy podciąg symboli jednego rodzaju nazywany jest serią. Liczbę liter a oznaczymy przez n_a, natomiast liczbę liter b – n_b. Liczbę serii symboli a i b oznaczymy przez k_{n_a,n_b}. Liczba serii ma stablicowany rozkład zależny od liczby liter a – n_a i liczby liter b – n_b.

Przyjmując poziom istotności α, z tablic serii (zob. tab. Va–Vb) odczytuje się dwie wartości krytyczne $k_{n_a,n_b;\alpha/2}$, oraz $k_{n_a,n_b;1-\alpha/2}$ dla których spełnione są własności:

$$P(k_{n_a,n_b} \leq k_{n_a,n_b;\alpha/2}) = \alpha/2, \tag{6.174}$$

oraz

$$P(k_{n_a,n_b} \leq k_{n_a,n_b;\alpha/2}) = 1 - \alpha/2. \tag{6.175}$$

Jeżeli liczba serii k_{n_a,n_b}, spełnia własność:

$$k_{n_a,n_b;\alpha/2} < k_{n_a,n_b} \leq k_{n_a,n_b;1-\alpha/2}, \tag{6.176}$$

to wówczas brak jest podstaw do odrzucenia hipotezy zerowej H_0.

Natomiast, jeżeli spełniona będzie nierówność:

$$k_{n_a,n_b} > k_{n_a,n_b;1-\alpha/2}, \tag{6.177}$$

lub:

$$k_{n_a,n_b} \leq k_{n_a,n_b;\alpha/2}, \tag{6.178}$$

to wówczas z prawdopodobieństwem błędu rzędu α, należy odrzucić hipotezę zerową i przyjąć hipotezę alternatywną, głoszącą, że pobrana próba nie jest losowa.

Opisany powyżej sposób postępowania jest właściwy, gdy próba losowa jest mała (tzn. gdy $n_a \leq 20$ i $n_b \leq 20$). W przypadku próby o dużej liczności, zaleca się skorzystanie ze sprawdzianu przyjmującego wartości według wzoru[11]:

$$u = \frac{k_{n_a,n_b} - \bar{k}_{n_a,n_b}}{s_{k_{n_a,n_b}}} \tag{6.179}$$

gdzie:

$$\bar{k}_{n_a,n_b} = \frac{2n_a n_b}{n_a + n_b} + 1, \tag{6.180}$$

$$s_{k_{n_a,n_b}} = \sqrt{\frac{2n_a n_b (2n_a n_b - n_a - n_b)}{(n_a + n_b)^2 (n_b + n_a - 1)}}. \tag{6.181}$$

Statystyka przyjmująca wartości według wzoru (6.179) ma rozkład asymptotycznie normalny $N(0,1)$. Przedział krytyczny jest w tym przypadku przedziałem dwustronnym typu (6.31), o wartościach krytycznych odczytanych z tablic dystrybuanty rozkładu normalnego standaryzowanego. Zasady podejmowania decyzji są analogiczne jak te opisane w tablicy 6.1 – wzory (6.38) i (6.39).

Przykład 6.17

Ze zbiorowości studentów piszących pracę kontrolną ze statystyki pobrano losową próbę o liczności $n = 18$ studentów i zbadano liczbę uzyskanych punktów. Wyniki badania prezentuje tablica 6.15.

Tablica 6.15. Dane do przykładu 6.17

i	Liczba punków x_i	i	Liczba punków x_i
1	**2**	**1**	**2**
1	23	10	22
2	27	11	22
3	27	12	25
4	28	13	26
5	26	14	22
6	16	15	28
7	29	16	27
8	29	17	26
9	29	18	19

Źródło: badania własne.

[11] Zob. np. M. Sobczyk: *Statystyka*, *op. cit.*, s. 184.

Zweryfikować hipotezę zerową, że pobrana próba ma charakter losowy. Podczas weryfikacji założyć poziom istotności $\alpha = 0{,}01$.

W pierwszym kroku postępowania wyznaczamy wartość mediany z próby. W tym celu przekształcamy badany szereg pierwotny w szereg pozycyjny (zob. tablica 6.16).

Ponieważ liczba obserwacji w próbie jest parzysta, dlatego medianę wyznaczymy ze wzoru (zob. zob. M. Major, J. Niezgoda: *Elementy statystyki…*, *op. cit.*, wzór (3.10)):

$$x_{me,n} = \frac{1}{2}\left[x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right] = \frac{1}{2}\left[x_{(9)} + x_{(10)}\right] = \frac{26+26}{2} = 26.$$

Tablica 6.16. Dane do przykładu 6.17, przedstawione w szeregu pozycyjnym

(i)	i	$x_{(i)}$	(i)	i	$x_{(i)}$
1	**2**	**3**	**1**	**2**	**3**
1	6	16	10	17	26
2	18	19	11	2	27
3	10	22	12	3	27
4	11	22	13	16	27
5	14	22	14	4	28
6	1	23	15	15	28
7	12	25	16	7	29
8	5	26	17	8	29
9	13	26	18	9	29

Źródło: opracowanie własne.

Wykorzystując wartość mediany, poszczególne obserwacje próby x_t w szeregu pierwotnym (tab. 6.15) zastępuje się symbolami a lub b, przy czym x_t kodujemy jako a gdy $x_t < x_{me,n}$, natomiast, gdy $x_t > x_{me,n}$, x_t kodujemy jako b.

Wyniki kodowania zostały przedstawione w tablicy 6.17.

Tablica 6.17. Kodowanie wartości x_i na symbole *a* i *b*

i	Liczba punków (x_i)	*a*; *b*	serie
1	**2**	**3**	**4**
1	23	*a*	1
2	27	*b*	2
3	27	*b*	
4	28	*b*	
5	26	*a*	3
6	16	*a*	
7	29	*b*	4
8	29	*b*	
9	29	*b*	
10	22	*a*	5
11	22	*a*	
12	25	*a*	
13	26	*a*	
14	22	*a*	
15	28	*b*	6
16	27	*b*	
17	26	*a*	7
18	19	*a*	

Źródło: opracowanie własne.

W omawianym przykładzie $n_a = 10$, $n_b = 8$. Liczba serii (zob. tab. 6.17, kol. 4) $k_{10,8} = 7$. Z tablicy Vb odczytujemy, że wartości krytyczne $k_{10,8;\,0,05} = 6$ oraz $k_{10,8;\,0,95} = 13$. Ponieważ spełniona jest zależność $k_{10,8;\,0,05} = 6 < k_{10,8} = 7 < k_{10,8;\,0,95} = 13$, dlatego nie ma powodów, aby odrzucić hipotezę zerową głoszącą, że pobrana próba ma charakter losowy. {kp}

6.10.3.2. Weryfikacja hipotezy, że dwie próby mają ten sam rozkład

Test serii można również wykorzystać do sprawdzenia hipotezy zerowej postaci:

H_0: dwie próby pochodzą z tej samej zbiorowości generalnej, (rozkłady obydwu populacji, z których pochodzą próby są identyczne), (6.182) wobec hipotezy alternatywnej:

H_1: dwie próby nie pochodzą z tej samej zbiorowości generalnej, (rozkłady obydwu populacji, z których pochodzą próby są różne). (6.183)

W pierwszym kroku weryfikacji dwie próby o licznościach n_a i n_b, pobrane z dwóch zbiorowości o dowolnych rozkładach, łączy się i porządkuje niemalejąco. W tak uporządkowanym szeregu literą a oznacza się obserwacje należące do pierwszej próby, natomiast literą b, obserwacje należące do drugiej próby. Efektem takiego postępowania jest k_{n_a, n_b} serii symboli a i b. Jeżeli liczba otrzymanych serii spełnia nierówność:

$$k_{n_a, n_b} \leq k_{n_a, n_b; \alpha}, \tag{6.184}$$

to wówczas należy odrzucić hipotezę zerową (6.182).

W przypadku, gdy:

$$k_{n_a, n_b} > k_{n_a, n_b; \alpha}, \tag{6.185}$$

stwierdzamy brak podstaw do odrzucenia hipotezy zerowej głoszącej, że obie próby pochodzą z populacji o jednakowych rozkładach.

Przykład 6.18

Z dwóch różnych grup dziekańskich piszących pracę kontrolną ze statystyki pobrano dwie próby losowe o liczebnościach $n_a = n_b = 10$ osób i zbadano liczbę otrzymanych punktów. Wyniki doświadczenia przedstawia tablica 6.18.

Tablica 6.18. Wyniki z pracy kontrolnej dwóch prób losowych

Próba „*a*”			Próba „*b*”		
i	*a*	x_i	*i*	*b*	x_i
1	**2**	**3**	**4**	**5**	**6**
1	a	21	1	*b*	25
2	a	15	2	*b*	18
3	a	26	3	*b*	27
4	a	29	4	*b*	29
5	a	13	5	*b*	12
6	a	24	6	*b*	26
7	a	16	7	*b*	16
8	a	28	8	*b*	23
9	a	24	9	*b*	28
10	a	16	10	*b*	19

Źródło: opracowanie własne.

Wszystkim obserwacjom pochodzącym z pierwszej zbiorowości (z pierwszej grupy dziekańskiej) przypisano symbol „*a*", natomiast obserwacje pochodzące z drugiej zbiorowości (drugiej grupy dziekańskiej) zakodowano, używając symbolu „*b*".

Naszym zadaniem jest zbadanie, czy prawdziwa jest hipoteza zerowa, głosząca, że rozkłady obydwu populacji, z których pochodzą próby, są identyczne. Weryfikację przeprowadzimy przy założonym poziomie istotności $\alpha = 0,05$.

Po połączeniu prób i uporządkowaniu niemalejąco otrzymujemy zbiór 20 obserwacji przedstawionych w tablicy 6.19.

Liczba $n_a = 10$, $n_b = 10$, liczba serii $k_{10; 10} = 13$. Odczytana z tablicy Vc, wartość krytyczna, przy $\alpha = 0,05$, oraz $n_a = n_b = 10$, wynosi $k_{10; 10; 0,05} = 6$.

Tablica 6.19. Szereg pozycyjny z połączonej próby

a; *b*	x_i	Serie
1	2	3
b	12	1
a	13	2
a	15	
a	16	
a	16	
b	16	3
b	18	
b	19	
a	21	4
b	23	5
a	24	6
a	24	
b	25	7
a	26	8
b	26	9
b	27	
a	28	10
b	28	11
a	29	12
b	29	13

Źródło: opracowanie własne.

Ponieważ $k_{10; 10} = 13 > k_{10; 10; 0,05} = 6$, brak jest powodów do odrzucenia hipotezy zerowej głoszącej, że rozkłady obu populacji, z których pobrano próby nieistotnie różnią się od siebie. {kp}

6.11. Podsumowanie

Przedstawione procedury nie obejmują wszystkich, lecz jedynie ważniejsze parametryczne i nieparametryczne testy istotności. Obszerną dawkę wiadomości na temat możliwości i różnorodności testów istotności można znaleźć np. w cytowanej już pozycji J. Grenia, *Statystyka matematyczna. Modele i zadania*, która pomimo upływu dość długiego czasu od momentu jej wydania jest jedną z ważniejszych i godnych uwagi pozycji poświęconej omawianym tu zagadnieniom. Odrębną grupę w teorii weryfikacji hipotez statystycznych spełniają także tzw. sekwencyjne testy oraz oparte na nich tzw. procedury sum skumulowanych, które są procedurami sekwencyjnymi realizowanymi wstecznie. Teoria analizy sekwencyjnej sprowadza się do losowego pobierania pojedynczych lub małych zespołów zbiorowości generalnej, oraz każdorazowym rozstrzyganiu, czy zdobyty dotąd zasób informacji pozwala na podjęcie określonej decyzji. Przy sformułowanej hipotezie zerowej H_0 i alternatywnej H_1 decyzje takie mogą polegać na:

- przyjęciu hipotezy H_0,
- odrzuceniu hipotezy H_0 i przyjęcia hipotezy H_1,
- odłożeniu decyzji do czasu pobrania następnej jednostki (próbki n) do próby m.

Próba m jest w tym ujęciu sumą wszystkich próbek n pobieranych w kolejnych krokach k, czyli:

$$m = n_1 + n_2 + \ldots + n_k. \qquad (6.186)$$

Jeżeli w każdym k-tym przedziale próbkowania liczebność próbki wynosi jeden, to wówczas liczność całej skumulowanej próby wyniesie k ($m = k$).

Podczas każdego etapu badań obliczana jest wartość sekwencyjnego testu ilorazowego postaci:

$$Q_k = \frac{\prod_{i=1}^{n} p_1(x_i)}{\prod_{i=1}^{n} p_0(x_i)}, \qquad (6.187)$$

gdzie:

$p_1(x_i)$ – jest prawdopodobieństwem zdarzenia losowego $X = x_i$, gdy założona jest prawdziwość hipotezy H_1,

$p_0(x_i)$ – jest prawdopodobieństwem zdarzenia losowego $X = x_i$, gdy założona jest prawdziwość hipotezy H_0,

i sprawdza się, czy zachodzi nierówność

$$A < Q_k < B, \qquad (6.188)$$

gdzie:

$$A = \frac{\beta}{1-\alpha}, \text{ natomiast } B = \frac{1-\beta}{\alpha}$$

Jeżeli $Q_k \leq A$, to przyjmuje się hipotezę H_0 z prawdopodobieństwem błędu nie większym niż β.

Gdy $Q_k \geq B$, to należy przyjąć hipotezę H_1 z prawdopodobieństwem błędu nie większym niż α.

Natomiast, kiedy $A \leq Q_k \leq B$, to brak jest podstaw do podjęcia jednej z dwóch wymienionych powyżej decyzji i należy kontynuować badanie.

Techniki sekwencyjne oraz powiązane z nimi techniki sum skumulowanych są szeroko wykorzystywane w praktyce w trakcie monitorowania procesów, na przykład procesów produkcyjnych ze względu na poziom jakości wytwarzanego strumienia produktów[12].

[12] Techniki te zostały szeroko opisane np. w: A. Iwsiewicz, Z. Paszek, J. Steczkowski: *Sekwencyjne metody kontroli jakości*, AE, Kraków 1988 oraz w 9 rozdziale książki: A. Iwasiewicz: *Zarządzanie jakością*, PWN, Warszawa–Kraków 1999.

Cytowana literatura

Ferguson G.A., Takane Y.: *Analiza statystyczna w psychologii i pedagogice*, PWN, Warszawa 2002.

Freund J.E.: *Podstawy nowoczesnej statystyki*, PWE, Warszawa 1968.

Gerstenkorn T., Środka T.: *Kombinatoryka i rachunek prawdopodobieństwa*, wyd. 5, PWN, Warszawa 1979.

Greń J.: *Statystyka matematyczna, modele i zadania*, PWN, Warszawa 1974.

Helwig Z.: *Elementy rachunku prawdopodobieństwa i statystyki matematycznej*, PWN, Warszawa 1967.

Iwasiewicz A., Paszek Z., Steczkowski J.: *Sekwencyjne metody kontroli jakości*, Akademia Ekonomiczna, Kraków 1988.

Iwasiewicz A.: *Zarządzanie jakością*, PWN, Warszawa–Kraków 1999.

Iwasiewicz A., Paszek Z.: *Statystyka z elementami statystycznych metod monitorowania procesów*, Akademia Ekonomiczna, Kraków 2004.

Iwasiewicz A.: *Zarządzanie jakością w przykładach i zadaniach*, Śląskie Wydawnictwo Naukowe Wyższej Szkoły Zarządzania i Nauk Społecznych, Tychy 2005.

Jóźwiak J., Podgórski J.: *Statystyka od podstaw*, PWE, Warszawa 1994.

Krysicki W. i inni: *Rachunek prawdopodobieństwa i statystyka matematyczna w zadaniach*, cz I., PWN, Warszawa 1986.

Krzyśko M.: *Wykłady z teorii prawdopodobieństwa*, WNT, Warszawa 2000.

Major M., Niezgoda J.: *Elementy statystyki, cz I. Statystyka opisowa*, Krakowskie Towarzystwo Edukacyjne sp. z o.o., Kraków 2003.

Sobczyk M.: *Statystyka*, PWN, Warszawa1998.

Steczkowski J.: *Metoda reprezentacyjna w badaniach zjawisk ekonomiczno-społecznych*, PWN, Warszawa–Kraków 1995.

PN N-03010; Statystyczna kontrola jakości. Losowy wybór sztuk do próbek; PKN 1951.

Statystyka ogólna, red. M. Woźniak, Akademia Ekonomiczna, Kraków 1997.

Tablice statystyczne

Tablica I. Dystrybuanta $\Phi(u) = P(U < u)$ rozkładu normalnego standaryzowanego

	$\Phi(u)$									
u	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,4960	0,4920	0,4880	0,4840	0,4801	0,4761	0,4721	0,4681	0,4641
–0,1	0,4602	0,4562	0,4522	0,4483	0,4443	0,4404	0,4364	0,4325	0,4286	0,4247
–0,2	0,4207	0,4168	0,4129	0,4090	0,4052	0,4013	0,3974	0,3936	0,3897	0,3859
–0,3	0,3821	0,3783	0,3745	0,3707	0,3669	0,3632	0,3594	0,3557	0,3520	0,3483
–0,4	0,3446	0,3409	0,3372	0,3336	0,3300	0,3264	0,3228	0,3192	0,3156	0,3121
–0,5	0,3085	0,3050	0,3015	0,2981	0,2946	0,2912	0,2877	0,2843	0,2810	0,2776
–0,6	0,2743	0,2709	0,2676	0,2643	0,2611	0,2578	0,2546	0,2514	0,2483	0,2451
–0,7	0,2420	0,2389	0,2358	0,2327	0,2296	0,2266	0,2236	0,2206	0,2177	0,2148
–0,8	0,2119	0,2090	0,2061	0,2033	0,2005	0,1977	0,1949	0,1922	0,1894	0,1867
–0,9	0,1841	0,1814	0,1788	0,1762	0,1736	0,1711	0,1685	0,1660	0,1635	0,1611
–1,0	0,1587	0,1562	0,1539	0,1515	0,1492	0,1469	0,1446	0,1423	0,1401	0,1379
–1,1	0,1357	0,1335	0,1314	0,1292	0,1271	0,1251	0,1230	0,1210	0,1190	0,1170
–1,2	0,1151	0,1131	0,1112	0,1093	0,1075	0,1056	0,1038	0,1020	0,1003	0,0985
–1,3	0,0968	0,0951	0,0934	0,0918	0,0901	0,0885	0,0869	0,0853	0,0838	0,0823
-1,4	0,0808	0,0793	0,0778	0,0764	0,0749	0,0735	0,0721	0,0708	0,0694	0,0681
–1,5	0,0668	0,0655	0,0643	0,0630	0,0618	0,0606	0,0594	0,0582	0,0571	0,0559
–1,6	0,0548	0,0537	0,0526	0,0516	0,0505	0,0495	0,0485	0,0475	0,0465	0,0455
–1,7	0,0446	0,0436	0,0427	0,0418	0,0409	0,0401	0,0392	0,0384	0,0375	0,0367
-1,8	0,0359	0,0351	0,0344	0,0336	0,0329	0,0322	0,0314	0,0307	0,0301	0,0294
–1,9	0,0287	0,0281	0,0274	0,0268	0,0262	0,0256	0,0250	0,0244	0,0239	0,0233
–2,0	0,0228	0,0222	0,0217	0,0212	0,0207	0,0202	0,0197	0,0192	0,0188	0,0183
–2,1	0,0179	0,0174	0,0170	0,0166	0,0162	0,0158	0,0154	0,0150	0,0146	0,0143
–2,2	0,0139	0,0136	0,0132	0,0129	0,0125	0,0122	0,0119	0,0116	0,0113	0,0110
–2,3	0,0107	0,0104	0,0102	0,0099	0,0096	0,0094	0,0091	0,0089	0,0087	0,0084
–2,4	0,0082	0,0080	0,0078	0,0075	0,0073	0,0071	0,0069	0,0068	0,0066	0,0064
–2,5	0,0062	0,0060	0,0059	0,0057	0,0055	0,0054	0,0052	0,0051	0,0049	0,0048
–2,6	0,0047	0,0045	0,0044	0,0043	0,0041	0,0040	0,0039	0,0038	0,0037	0,0036
–2,7	0,0035	0,0034	0,0033	0,0032	0,0031	0,0030	0,0029	0,0028	0,0027	0,0026
–2,8	0,0026	0,0025	0,0024	0,0023	0,0023	0,0022	0,0021	0,0021	0,0020	0,0019
–2,9	0,0019	0,0018	0,0018	0,0017	0,0016	0,0016	0,0015	0,0015	0,0014	0,0014
–3,0	0,0013	0,0013	0,0013	0,0012	0,0012	0,0011	0,0011	0,0011	0,0010	0,0010

Tablica I. ciąg dalszy

	$\Phi(u)$									
u	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3,0	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990

Źródło: opracowanie własne z wykorzystaniem funkcji Microsoft Excel.

Tablica Ia. Wybrane kwantyle rozkładu normalnego $P(U > u_\alpha)$ dla $U \sim N(0,1)$

α	0,2	0,1	0,05	0,025	0,01	0,005
u_α	0,842	1,282	1,645	1,960	2,326	2,576

Tablica II. Kwantyle rozkładów zmiennej losowej t_r –Studenta
$P(t_r > t_{r,\varepsilon}) = \varepsilon$, gdzie: $\varepsilon = \alpha$ lub $\varepsilon = \alpha/2$, r – liczba stopni swobody

r	$t_{r,\varepsilon}$					r
	$t_{r;0,1}$	$t_{r;0,05}$	$t_{r;0,025}$	$t_{r;0,01}$	$t_{r;0,005}$	
1	3,078	6,314	12,706	31,821	63,657	1
2	1,886	2,920	4,303	6,965	9,925	2
3	1,638	2,353	3,182	4,541	5,841	3
4	1,533	2,132	2,776	3,747	4,604	4
5	1,476	2,015	2,571	3,365	4,032	5
6	1,440	1,943	2,447	3,143	3,707	6
7	1,415	1,895	2,365	2,998	3,499	7
8	1,397	1,860	2,306	2,896	3,355	8
9	1,383	1,833	2,262	2,821	3,250	9
10	1,372	1,812	2,228	2,764	3,169	10
11	1,363	1,796	2,201	2,718	3,106	11
12	1,356	1,782	2,179	2,681	3,055	12
13	1,350	1,771	2,160	2,650	3,012	13
14	1,345	1,761	2,145	2,624	2,977	14
15	1,341	1,753	2,131	2,602	2,947	15
16	1,337	1,746	2,120	2,583	2,921	16
17	1,333	1,740	2,110	2,567	2,898	17
18	1,330	1,734	2,101	2,552	2,878	18
19	1,328	1,729	2,093	2,539	2,861	19
20	1,325	1,725	2,086	2,528	2,845	20
21	1,323	1,721	2,080	2,518	2,831	21
22	1,321	1,717	2,074	2,508	2,819	22
23	1,319	1,714	2,069	2,500	2,807	23
24	1,318	1,711	2,064	2,492	2,797	24
25	1,316	1,708	2,060	2,485	2,787	25
26	1,315	1,706	2,056	2,479	2,779	26
27	1,314	1,703	2,052	2,473	2,771	27
28	1,313	1,701	2,048	2,467	2,763	28
29	1,311	1,699	2,045	2,462	2,756	29
∞	1,282	1,645	1,960	2,326	2,576	

Źródło: opracowanie własne z wykorzystaniem funkcji Microsoft Excel.

Tablica III. Kwantyle rozkładów zmiennej losowej χ_r^2

$P(\chi_r^2 > \chi_{r;\varepsilon}^2) = \varepsilon$, gdzie: r – liczba stopni swobody

r	$\chi_{r;\varepsilon}^2$								r
	$\chi_{r;0,995}^2$	$\chi_{r;0,99}^2$	$\chi_{r;0,975}^2$	$\chi_{r;0,95}^2$	$\chi_{r;0,05}^2$	$\chi_{r;0,025}^2$	$\chi_{r;0,01}^2$	$\chi_{r;0,005}^2$	
1	0,0000393	0,000157	0,000982	0,00393	3,841	5,024	6,635	7,879	1
2	0,0100	0,0201	0,0506	0,103	5,991	7,378	9,210	10,597	2
3	0,0717	0,115	0,216	0,352	7,815	9,348	11,345	12,838	3
4	0,207	0,297	0,484	0,711	9,488	11,143	13,277	14,860	4
5	0,412	0,554	0,831	1,145	11,070	12,832	15,086	16,750	5
6	0,676	0,872	1,237	1,635	12,592	14,449	16,812	18,548	6
7	0,989	1,239	1,690	2,167	14,067	16,013	18,475	20,278	7
8	1,344	1,646	2,180	2,733	15,507	17,535	20,090	21,955	8
9	1,735	2,088	2,700	3,325	16,919	19,023	21,666	23,589	9
10	2,156	2,558	3,247	3,940	18,307	20,483	23,209	25,188	10
11	2,603	3,053	3,816	4,575	19,675	21,920	24,725	26,757	11
12	3,074	3,571	4,404	5,226	21,026	23,337	26,217	28,300	12
13	3,565	4,107	5,009	5,892	22,362	24,736	27,688	29,819	13
14	4,075	4,660	5,629	6,571	23,685	26,119	29,141	31,319	14
15	4,601	5,229	6,262	7,261	24,996	27,488	30,578	32,801	15
16	5,142	5,812	6,908	7,962	26,296	28,845	32,000	34,267	16
17	5,697	6,408	7,564	8,672	27,587	30,191	33,409	35,718	17
18	6,265	7,015	8,231	9,390	28,869	31,526	34,805	37,156	18
19	6,844	7,633	8,907	10,117	30,144	32,852	36,191	38,582	19
20	7,434	8,260	9,591	10,851	31,410	34,170	37,566	39,997	20
21	8,034	8,897	10,283	11,591	32,671	35,497	38,932	41,401	21
22	8,643	9,542	10,982	12,338	33,924	36,781	40,289	42,796	22
23	9,260	10,196	11,689	13,091	35,172	38,076	41,638	44,181	23
24	9,886	10,856	12,401	13,848	36,415	39,364	42,980	45,558	24
25	10,520	11,524	13,120	14,611	37,652	40,646	44,314	46,928	25
26	11,160	12,198	13,844	15,379	38,885	41,923	45,642	48,290	26
27	11,808	12,879	14,573	16,151	40,113	43,194	46,963	49,645	27
28	12,461	13,565	15,308	16,928	41,337	44,461	48,278	50,993	28
29	13,121	14,256	16,047	17,708	42,557	45,722	49,588	52,336	29
30	13,787	14,953	16,791	18,493	43,773	46,979	50,892	53,672	30

Źródło: opracowanie własne z wykorzystaniem funkcji Microsoft Excel.

Tablica IVa. Kwantyle rozkładów zmiennej losowej F

$P(F_{r_1, r_2} > F_{r_1, r_2, \alpha}) = \alpha$, gdzie $r_1 = n_1 - 1$ oraz $r_2 = n_2 - 1$

stopnie swobody mianownika $(n_2 - 1)$	$F_{(n_1-1);(n_2-1);0,05}$ — stopnie swobody licznika $(n_1 - 1)$																		
	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	161	200	216	225	230	234	237	239	241	242	244	246	248	249	250	251	252	253	254
2	18,5	19,0	19,2	19,2	19,3	19,3	19,4	19,4	19,4	19,4	19,4	19,4	19,4	19,5	19,5	19,5	19,5	19,5	19,5
3	10,1	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79	8,74	8,70	8,66	8,64	8,62	8,59	8,57	8,55	8,53
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,91	5,86	5,80	5,77	5,75	5,72	5,69	5,66	5,63
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74	4,68	4,62	4,56	4,53	4,50	4,46	4,43	4,40	4,37
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	4,00	3,94	3,87	3,84	3,81	3,77	3,74	3,70	3,67
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	3,57	3,51	3,44	3,41	3,38	3,34	3,30	3,27	3,23
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35	3,28	3,22	3,15	3,12	3,08	3,04	3,01	2,97	2,93
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14	3,07	3,01	2,94	2,90	2,86	2,83	2,79	2,75	2,71
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98	2,91	2,85	2,77	2,74	2,70	2,66	2,62	2,58	2,54
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85	2,79	2,72	2,65	2,61	2,57	2,53	2,49	2,45	2,40
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75	2,69	2,62	2,54	2,51	2,47	2,43	2,38	2,34	2,30
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67	2,60	2,53	2,46	2,42	2,38	2,34	2,30	2,25	2,21
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60	2,53	2,46	2,39	2,35	2,31	2,27	2,22	2,18	2,13
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54	2,48	2,40	2,33	2,29	2,25	2,20	2,16	2,11	2,07
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,42	2,35	2,28	2,24	2,19	2,15	2,11	2,06	2,01
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45	2,38	2,31	2,23	2,19	2,15	2,10	2,06	2,01	1,96
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41	2,34	2,27	2,19	2,15	2,11	2,06	2,02	1,97	1,92
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38	2,31	2,23	2,16	2,11	2,07	2,03	1,98	1,93	1,88
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35	2,28	2,20	2,12	2,08	2,04	1,99	1,95	1,90	1,84
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32	2,25	2,18	2,10	2,05	2,01	1,96	1,92	1,87	1,81
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30	2,23	2,15	2,07	2,03	1,98	1,94	1,89	1,84	1,78
23	4,28	3,42	3,03	2,80	2,64	2,53	2,44	2,37	2,32	2,27	2,20	2,13	2,05	2,01	1,96	1,91	1,86	1,81	1,76
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25	2,18	2,11	2,03	1,98	1,94	1,89	1,84	1,79	1,73
25	4,24	3,39	2,99	2,76	2,60	2,49	2,40	2,34	2,28	2,24	2,16	2,09	2,01	1,96	1,92	1,87	1,82	1,77	1,71
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16	2,09	2,01	1,93	1,89	1,84	1,79	1,74	1,68	1,62
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08	2,00	1,92	1,84	1,79	1,74	1,69	1,64	1,58	1,51
60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99	1,92	1,84	1,75	1,70	1,65	1,59	1,53	1,47	1,39
120	3,92	3,07	2,68	2,45	2,29	2,18	2,09	2,02	1,96	1,91	1,83	1,75	1,66	1,61	1,55	1,50	1,43	1,35	1,25
∞	3,84	3,00	2,60	2,37	2,21	2,10	2,01	1,94	1,88	1,83	1,75	1,67	1,57	1,52	1,46	1,39	1,32	1,22	1,00

Źródło: opracowanie własne z wykorzystaniem funkcji Microsoft Excel.

Tablica IVb. Kwantyle rozkładów zmiennej losowej F

$P(F_{r_1,r_1} > F_{r_1,r_1,\alpha}) = \alpha$, gdzie $r_1 = n_1 - 1$ oraz $r_2 = n_2 - 1$

stopnie swobody mianownika $(n_2 - 1)$	$F_{(n_1-1);(n_2-1);0,05}$ stopnie swobody licznika $(n_1 - 1)$: 1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	4052	5000	5403	5625	5764	5859	5928	5982	6023	6056	6106	6157	6209	6235	6261	6287	6313	6339	6366
2	98,5	99,0	99,2	99,2	99,3	99,3	99,4	99,4	99,4	99,4	99,4	99,4	99,4	99,5	99,5	99,5	99,5	99,5	99.5
3	34,1	30,8	29,5	28,7	28,2	27,9	27,7	27,5	27,3	27,2	27,1	26,9	26,7	26,6	26,5	26,4	26,3	26,2	26,1
4	21,2	18,0	16,7	16,0	15,5	15,2	15,0	14,8	14,7	14,5	14,4	14,2	14,0	13,9	13,8	13,7	13,7	13,6	13,5
5	16,3	13,3	12,1	11,4	11,0	10,7	10,5	10,3	10,2	10,1	9,89	9,72	9,55	9,47	9,38	9,29	9,20	9,11	9,02
6	13,7	10,9	9,78	9,15	8,75	8,47	8,26	8,10	7,98	7,87	7,72	7,56	7,40	7,31	7,23	7,14	7,06	6,97	6.88
7	12,2	9,55	8,45	7,85	7,46	7,19	6,99	6,84	6,72	6,62	6,47	6,31	6,16	6,07	5,99	5,91	5,82	5,74	5,65
8	11,3	8,65	7,59	7,01	6,63	6,37	6,18	6,03	5,91	5,81	5,67	5,52	5,36	5,28	5,20	5,12	5,03	4,95	4,86
9	10,6	8,02	6,99	6,42	6,06	5,80	5,61	5,47	5,35	5,26	5,11	4,96	4,81	4,73	4,65	4,57	4,48	4,40	4,31
10	10,0	7,56	6,55	5,99	5.64	5,39	5,20	5,06	4,94	4,85	4,71	4,56	4,41	4,33	4,25	4,17	4,08	4,00	3,91
11	9,65	7,21	6,22	5,67	5,32	5,07	4,89	4,74	4,63	4,54	4,40	4,25	4,10	4,02	3,94	3,86	3,78	3,69	3,60
12	9,33	6,93	5,95	5,41	5.06	4,82	4,64	4.50	4,39	4,30	4.16	4,01	3,86	3,78	3,70	3,62	3,54	3,45	3,36
13	9,07	6,70	5,74	5,21	4,86	4,62	4,44	4,30	4,19	4,10	3,96	3,82	3,66	3,59	3,51	3,43	3,34	3,25	3,17
14	8,86	6,51	5,56	5,04	4,70	4,46	4,28	4.14	4,03	3,94	3,80	3,66	3,51	3.43	3,35	3,27	3,18	3,09	3.00
15	8,68	6,36	5,42	4,89	4,56	4,32	4,14	4,00	3.89	3,80	3,67	3,52	3,37	3,29	3,21	3,13	3,05	2,96	2,87
16	8,53	6,23	5.29	4,77	4,44	4,20	4,03	3,89	3,78	3,69	3,55	3,41	3,26	3,18	3,10	3,02	2,93	2,84	2,75
17	8,40	6,11	5,19	4,67	4,34	4,10	3,93	3,79	3,68	3,59	3,46	3,31	3,16	3,08	3,00	2,92	2,83	2.75	2,65
18	8,29	6,01	5,09	4,58	4,25	4,01	3,84	3,71	3,60	3,51	3,37	3,23	3,08	3,00	2,92	2,84	2,75	2,66	2,57
19	8,19	5,93	5,01	4,50	4,17	3,94	3,77	3,63	3,52	3,43	3,30	3,15	3,00	2,92	2,84	2,76	2,67	2,58	2,49
20	8,10	5 85	4,94	4,43	4,10	3,87	3,70	3,56	3 46	3,37	3,23	3,09	2,94	7 86	2,78	2,69	2,61	2,52	2,42
21	8.02	5,78	4,87	4,37	4,04	3,81	3,64	3,51	3,40	3,31	3,17	3,03	2,88	2,80	2,72	2,64	2,55	2,46	2,36
22	7.95	5,72	4.82	4,31	3,99	3,76	3.59	3.45	3,35	3,26	3,12	2,98	2,83	2,75	2,67	2,58	2,50	2,40	2,31
23	7.88	5,66	4,76	4,26	3,94	3,71	3,54	3,41	3,30	3,21	3,07	2,93	2,78	2,70	2,62	2,54	2,45	2,35	2,26
24	7,82	5 61	4,72	4,22	3 90	3,67	3,50	3,36	3?6	3,17	3,03	2,89	2,74	?66	2,58	2,49	2,40	2,31	2,21
25	7.77	5,57	4,68	4,18	3,86	3,63	3,46	3,32	3,22	3,13	2,99	2,85	2,70	2,62	2,53	2,45	2,36	2,27	2,17
30	7,56	5,39	4.51	4,02	3,70	3,47	3,30	3,17	3,07	2,98	2,84	2,70	2,55	2,47	2,39	2,30	2,21	2,11	2,01
40	7,31	5,18	4,31	3,83	3,51	3,29	3,12	2,99	2,89	2,80	2,66	2,52	2,37	2,29	2,20	2,11	2,02	1,92	1,80
60	7,08	4,98	4,13	3,65	3,34	3,12	2,95	2,82	2,72	2,63	2,50	2,35	2,20	2,12	2,03	1,94	1,84	1,73	1,60
120	6,85	4,79	3,95	3,48	3,17	2,96	2,79	2,66	2,56	2,47	2,34	2,19	2,03	1,95	1,86	1,76	1,66	1,53	1,38
∞	6,63	4,61	3,78	3.32	3,02	2,80	2,64	2,51	2,41	2,32	2,18	2,04	1,88	1,79	1,70	1,59	1,47	1,32	1,00

Źródło: opracowanie własne z wykorzystaniem funkcji Microsoft Excel.

Tablica Va.

Kwantyle rozkładu serii

$P(k_{n_a, n_b} \leq k_{n_a, n_b; 0,025}) = 0,025$ $\qquad\qquad$ $P(k_{n_a, n_b} \leq k_{n_a, n_b; 0,975}) = 0,975$

n_b \ n_a	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	n_b \ n_a
2																				4																			2
3																				5	6																		3
4																				5	7	8																	4
5			2	2																5	7	8	9																5
6		2	2	3	3															5	7	8	9	10															6
7		2	2	3	3	3														5	7	9	10	11	12														7
8		2	3	3	3	4	4													5	7	9	10	11	12	13													8
9		2	3	3	4	4	5	5												5	7	9	11	12	13	13	14												9
10		2	3	3	4	5	5	5	6											5	7	9	11	12	13	14	15	15											10
11		2	3	4	4	5	5	6	6	7										5	7	9	11	12	13	14	15	16	16										11
12	2	2	3	4	4	5	6	6	7	7	7									5	7	9	11	12	13	15	15	16	17	18									12
13	2	2	3	4	5	5	6	6	7	7	8	8								5	7	9	11	13	14	15	16	17	18	18	19								13
14	2	2	3	4	5	5	6	7	7	8	8	9	9							5	7	9	11	13	14	15	16	17	18	19	19	20							14
15	2	3	3	4	5	6	6	7	7	8	8	9	9	10						5	7	9	11	13	14	15	17	17	18	19	20	21	21						15
16	2	3	4	4	5	6	6	7	8	8	9	9	10	10	11					5	7	9	11	13	15	16	17	18	19	20	20	21	22	22					16
17	2	3	4	5	5	6	7	7	8	9	10	10	11	11	11	11				5	7	9	11	13	15	16	17	18	19	20	21	22	22	23	24				17
18	2	3	4	5	5	6	7	8	8	9	9	10	10	11	11	12	12			5	7	9	11	13	15	16	17	18	19	20	21	22	23	24	24	25			18
19	2	3	4	5	6	6	7	8	8	9	10	10	11	11	12	12	13	13		5	7	9	11	13	15	16	17	19	20	21	22	22	23	24	25	25	26		19
20	2	3	4	5	6	6	7	8	9	9	10	10	11	12	12	13	13	13	14	5	7	9	11	13	15	16	17	19	20	21	22	23	24	24	25	26	26	27	20

Źródło: A. Iwasiewicz; Z. Paszek: *Statystyka…, op. cit.* s. 356.

Tablica Vb.

Kwantyle rozkładu serii

$P(k_{n_a, n_b} < k_{n_a, n_b; 0,05}) = 0,05$ $\qquad$ $P(k_{n_a, n_b} \leq k_{n_a, n_b; 0,95}) = 0,95$

n_b \ n_a	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	n_b \ n_a
2																				4																			2
3																				5	6																		3
4			2																	5	6	7																	4
5		2	2	3																5	7	8	8																5
6		2	3	3	3															5	7	8	9	10															6
7		2	3	3	4	4														5	7	8	9	10	11														7
8	2	2	3	3	4	4	5													5	7	9	10	11	12	12													8
9	2	2	3	4	4	5	5	6												5	7	9	10	11	12	13	13												9
10	2	3	3	4	5	5	6	6	6											5	7	9	10	11	12	13	14	15											10
11	2	3	3	4	5	5	6	6	7	7										5	7	9	11	12	13	14	14	15	16										11
12	2	3	4	4	5	6	6	7	7	8	8									5	7	9	11	12	13	14	15	16	16	17									12
13	2	3	4	4	5	6	6	7	8	8	9	9								5	7	9	11	12	13	14	15	16	17	17	18								13
14	2	3	4	5	5	6	7	7	8	8	9	9	10							5	7	9	11	12	13	15	16	16	17	18	19	19							14
15	2	3	4	5	6	6	7	8	8	9	9	10	10	11						5	7	9	11	13	14	15	16	17	18	18	19	20	20						15
16	2	3	4	5	6	6	7	8	8	9	10	10	11	11	11					5	7	9	11	13	14	15	16	17	18	19	20	20	21	22					16
17	2	3	4	5	6	7	7	8	9	9	10	10	11	11	12	12				5	7	9	11	13	14	15	16	17	18	19	20	21	21	22	23				17
18	2	3	4	5	6	7	8	8	9	10	10	11	11	12	12	13	13			5	7	9	11	13	14	15	18	18	19	20	20	21	22	23	23	24			18
19	2	3	4	5	6	7	8	8	9	10	10	11	12	12	13	13	14	14		5	7	9	11	13	14	15	18	18	19	20	21	22	22	23	24	24	25		19
20	2	3	4	5	6	7	8	9	9	10	11	11	12	12	13	13	14	14	15	5	7	9	11	13	14	16	18	18	19	20	21	22	23	24	24	25	26	26	20

Źródło: A. Iwasiewicz; Z. Paszek: *Statystyka...*, *op. cit.* s. 356.

Tablica Vc.

Kwantyle rozkładu serii $P(k < k_1) = \alpha$

$n_a = n_b$ \ α	0,025	0,05	0,95	0,975
10	6	6	15	15
11	7	7	16	16
12	7	8	17	18
13	8	9	18	19
14	9	10	19	20
15	10	11	20	21
16	11	11	22	22
17	11	12	23	24
18	12	13	24	25
19	13	14	25	26
20	14	15	26	27
21	15	16	27	28
22	16	17	28	29
23	16	17	30	31
24	17	18	31	32
25	18	19-	32	33
26	19	20	33	34
27	20	21	34	35
28	21	[illegible]	35	36
29	22	23	36	37
30	[illegible]	24	37	39
31	23	25	38	40
32	24	25	40	41

$n_a = n_b$ \ α	0,025	0,05	0,95	0,975
33	25	26	41	42
34	26	27	42	43
35	27	28	43	44
36	28	29	44	45
37	29	30	45	46
38	30	31	46	47
39	30	32	47	49
40	31	33	48	50
41	32	34	49	51
42	33	35	50	52
43	34	35	52	53
44	35	36	53	54
45	36	37	54	55
46	37	38	55	56
47	38	39	56	57
48	38	40	57	59
49	39	41	58	60
50	40	42	59	61
51	41	43	60	62
52	42	44	61	63
53	43	45	62	64
54	44	45	64	65
55	45	46	65	66

$n_a = n_b$ \ α	0,025	0,05	0,95	0,975
56	46	47	66	67
57	47	48	67	68
58	47	49	68	70
59	48	50	69	71
60	49	51	70	72
61	50	52	71	73
62	51	53	72	74
63	52	54	73	75
64	53	55	74	76
65	54	56	75	77
66	55	57	76	78
67	56	58	77	79
68	57	58	79	80
69	58	59	80	81
70	58	60	81	83
71	59	61	82	84
72	60	62	83	85
73	61	63	84	86
74	62	64	85	87
75	63	65	86	88
76	64	66	87	89
77	65	67	88	90
78	66	68	89	91

$n_a = n_b$ \ α	0,025	0,05	0,95	0,975
79	67	69	90	92
80	68	70	91	93
81	69	71	92	94
82	69	71	94	96
83	70	72	95	97
84	71	73	96	98
85	72	74	97	99
86	73	75	98	100
87	74	76	99	101
88	75	77	100	102
89	76	78	101	103
90	77	79	102	104
91	78	80	103	105
92	79	81	104	106
93	80	82	105	107
94	81	83	106	108
95	82	84	107	109
96	82	85	108	111
97	83	86	109	112
98	84	87	110	113
99	85	87	112	114
100	86	88	113	115

Źródło: A. Iwasiewicz; Z. Paszek: *Statystyka.. . op. cit*. s. 358.

Tablica VIa.
Tablice liczb losowych
Blok 1

numer wiersza	*numer kolumny* 1	2	3	4	5	6	7	8	9	10
1	45 83	84 22	92 14	37 32	94 19	89 10	74 94	92 29	35 51	58 41
2	06 81	33 27	81 89	85 31	88 95	60 43	02 92	40 58	47 28	23 03
3	27 39	51 15	68 38	89 69	35 58	75 34	46 22	11 02	79 79	51 20
4	47 96	72 82	44 99	15 07	42 42	03 53	85 93	09 64	55 59	28 11
5	89 79	95 91	27 42	80 22	60 51	79 38	57 39	24 61	22 59	62 99
6	18 72	51 84	25 93	55 04	09 97	66 66	31 15	68 67	50 69	46 31
7	16 01	05 31	29 06	19 40	14 07	67 28	23 06	28 16	26 73	79 05
8	17 19	32 62	46 68	42 34	03 97	34 67	96 74	32 31	23 64	13 81
9	46 08	35 37	37 28	60 22	09 56	43 93	53 02	10 38	82 92	82 37
10	21 65	35 94	74 32	18 38	75 48	81 72	93 63	41 02	69 14	75 15
11	42 16	57 54	75 61	81 60	73 35	12 84	05 79	29 40	12 96	58 09
12	99 49	76 19	26 30	52 58	30 29	85 61	95 62	04 81	83 33	85 63
13	36 24	49 23	74 09	61 00	12 74	77 21	40 73	76 49	49 74	82 35
14	84 67	04 41	71 86	57 71	71 37	07 91	30 12	07 11	24 38	70 52
15	91 44	36 63	97 55	37 82	83 09	56 10	49 66	62 79	38 14	51 15
16	62 39	41 53	57 30	14 46	25 21	28 33	34 26	30 78	35 35	23 71
17	53 60	83 19	31 17	46 27	18 46	81 92	09 85	18 09	63 44	62 91
18	75 12	72 34	28 55	72 40	10 32	89 84	31 30	09 23	77 33	05 43
19	19 77	06 76	27 91	63 97	90 03	30 35	14 93	54 56	62 61	16 74
20	62 64	38 13	86 35	33 31	00 73	99 37	99 38	43 61	73 28	97 59
21	58 62	52 12	84 63	43 70	05 26	66 18	80 16	95 98	65 79	92 20
22	48 64	73 61	96 77	20 03	68 97	73 01	79 23	69 45	50 53	17 63
23	30 29	65 70	48 86	59 62	74 26	59 92	92 64	03 23	74 42	26 10
24	18 85	27 43	44 51	97 65	85 44	92 91	99 19	18 47	66 04	36 69
25	33 95	24 70	71 15	96 16	74 60	79 77	80 13	49 02	59 64	52 09

Źródło: opracowanie własne z wykorzystaniem funkcji Microsoft Excel.

Tablica VIb.
Tablice liczb losowych
Blok 2

	numer kolumny									
numer wiersza	1	2	3	4	5	6	7	8	9	10
1	55 47	77 34	78 29	34 41	38 55	29 01	25 15	78 50	63 41	81 92
2	81 18	27 36	31 07	84 00	29 51	24 46	47 93	88 13	77 54	05 56
3	98 64	51 01	06 59	02 83	83 86	49 45	47 52	12 18	34 21	76 24
4	96 16	38 79	91 82	18 21	61 44	19 83	04 44	87 77	28 20	21 5
5	35 58	43 30	39 33	53 30	06 31	49 40	10 26	77 30	42 65	46 57
6	62 06	73 05	91 91	71 50	96 8	54 95	72 64	34 94	45 46	82 58
7	16 69	29 95	69 66	07 48	37 75	21 04	98 91	24 21	64 74	37 89
8	49 08	70 90	60 91	20 29	45 83	07 75	64 11	29 05	55 99	39 02
9	49 69	89 84	02 11	41 27	96 26	36 46	49 47	02 76	97 28	30 08
10	35 91	78 06	60 99	94 99	65 65	74 41	62 99	09 98	56 54	84 26
11	12 98	90 64	64 11	96 78	86 14	81 57	52 42	24 43	27 37	68 90
12	41 94	73 25	14 32	34 94	16 40	43 44	58 62	48 94	90 57	56 25
13	11 60	68 98	85 61	18 17	95 23	52 09	55 90	01 43	34 95	30 81
14	74 43	14 82	94 49	37 50	44 94	21 72	54 78	56 97	36 08	28 58
15	70 30	11 08	24 81	92 18	95 94	37 87	97 94	02 80	32 90	91 94
16	09 12	55 14	97 12	57 48	46 08	02 00	38 19	92 38	51 49	82 81
17	29 23	05 10	59 69	02 78	95 95	61 39	99 57	98 49	31 51	03 64
18	44 21	75 25	72 31	69 56	39 10	59 57	52 50	23 21	88 42	83 72
19	06 23	93 35	36 95	76 90	65 13	61 19	25 64	09 28	17 25	57 60
20	47 58	99 28	59 99	89 07	48 01	75 49	80 60	86 75	49 02	98 50
21	63 07	89 50	36 66	97 55	98 56	35 22	84 86	59 32	82 98	77 21
22	21 66	18 28	08 00	45 12	91 68	19 42	21 52	43 65	07 38	49 30
23	91 67	76 99	84 06	93 93	50 96	75 89	58 87	90 64	43 57	23 69
24	30 90	37 34	21 49	04 91	85 27	44 05	40 99	26 22	27 38	25 39
25	22 98	71 31	02 82	04 27	73 07	68 42	51 89	52 41	38 88	97 02

Źródło: opracowanie własne z wykorzystaniem funkcji Microsoft Excel.

Tablica VIc.
Tablice liczb losowych
Blok 3

		numer kolumny																			
		1		2		3		4		5		6		7		8		9		10	
numer wiersza	1	70	73	27	49	04	15	09	11	41	15	61	36	43	49	56	78	42	92	36	20
	2	43	70	17	61	37	8	50	89	37	20	28	93	02	75	27	88	96	71	58	67
	3	97	87	98	08	10	43	10	19	70	58	30	34	02	34	48	81	33	64	07	82
	4	19	54	28	85	92	04	03	66	61	34	51	54	93	43	14	21	96	72	10	01
	5	33	99	44	90	57	91	42	55	98	71	07	71	27	53	14	23	67	47	62	24
	6	81	84	49	50	56	69	96	51	80	53	62	81	76	38	80	56	72	43	91	17
	7	76	75	22	27	74	45	48	59	64	79	99	46	82	50	19	92	51	82	00	65
	8	45	10	16	80	17	08	83	80	90	03	97	87	09	27	60	17	17	77	68	22
	9	93	15	84	18	19	47	42	40	69	44	86	51	93	43	70	77	00	34	96	45
	10	46	54	24	11	53	53	85	81	15	77	84	60	36	60	28	88	55	67	18	35
	11	79	51	58	95	93	78	50	76	60	77	59	52	93	94	25	40	13	54	05	04
	12	26	60	52	07	16	11	79	99	35	43	37	90	90	27	93	14	01	63	69	67
	13	85	14	22	57	38	72	51	26	93	88	85	92	01	16	60	21	59	82	34	42
	14	04	00	97	08	99	23	25	84	27	97	17	76	83	99	08	16	35	02	00	94
	15	01	18	82	52	86	14	91	06	71	88	31	80	41	05	48	71	04	63	90	01
	16	09	06	77	64	42	86	94	93	16	88	47	25	67	93	91	08	41	66	48	35
	17	75	33	36	80	16	17	78	95	31	72	89	24	31	74	79	07	99	48	83	43
	18	76	93	27	91	43	32	41	32	26	73	19	21	53	36	29	17	27	42	89	23
	19	62	90	89	35	35	15	94	08	73	30	36	98	85	16	15	79	24	91	03	31
	20	60	79	52	59	40	86	28	44	25	76	72	12	13	27	49	49	90	01	38	12
	21	32	42	09	22	43	24	54	61	84	60	84	26	32	72	74	31	32	28	25	00
	22	36	29	03	31	96	49	08	70	34	01	07	94	65	25	45	55	96	95	10	10
	23	76	87	43	21	90	55	27	74	55	50	51	39	96	66	01	86	71	96	75	99
	24	73	72	39	98	86	23	90	96	91	63	79	10	61	39	17	19	86	93	79	05
	25	40	02	66	74	70	11	11	18	25	78	65	09	22	34	41	36	97	62	43	65

Źródło: opracowanie własne z wykorzystaniem funkcji Microsoft Excel.

Tablica VId.
Tablice liczb losowych
Blok 4

	numer kolumny									
numer wiersza	1	2	3	4	5	6	7	8	9	10
1	04 97	97 07	85 84	62 75	19 47	70 80	74 22	14 82	89 58	76 04
2	18 54	22 94	47 22	88 59	05 58	60 20	91 36	37 90	72 85	43 29
3	27 02	38 41	74 41	02 92	93 63	99 50	43 41	86 13	63 06	06 54
4	87 13	08 66	25 25	78 00	46 28	13 60	52 25	27 01	82 77	68 67
5	06 32	49 81	49 16	06 92	40 10	95 32	44 74	58 34	36 17	01 08
6	11 21	83 12	40 96	14 80	64 26	53 17	09 10	12 95	20 39	72 44
7	59 47	90 20	05 84	17 71	40 35	40 90	55 55	92 69	95 32	38 19
8	76 88	28 48	38 00	18 57	73 58	15 22	07 86	81 98	82 47	45 65
9	36 12	77 85	76 45	07 19	14 65	50 23	51 20	48 07	56 70	15 81
10	72 40	10 99	28 24	81 13	51 47	56 59	14 76	44 62	19 42	28 58
11	84 41	81 18	20 64	92 17	54 83	42 10	57 32	92 89	11 63	87 16
12	54 4	17 26	36 90	44 77	88 66	92 90	24 35	57 11	20 91	87 84
13	25 43	75 93	76 97	72 38	96 47	71 29	13 73	79 67	51 43	39 28
14	78 99	45 46	26 67	58 98	06 24	72 36	96 97	36 66	05 54	10 97
15	12 81	35 93	90 84	99 03	17 20	49 83	82 03	31 06	91 94	43 49
16	06 32	20 99	29 70	52 50	67 74	46 94	97 47	49 91	51 02	70 61
17	78 19	13 78	86 73	40 45	32 75	82 30	96 85	29 38	36 62	85 11
18	07 88	79 28	62 65	25 28	12 47	39 10	28 68	34 46	93 30	44 84
19	76 53	50 68	93 97	96 85	71 33	97 42	85 82	11 96	39 87	27 05
20	66 31	21 38	79 03	90 99	89 80	48 59	83 55	20 69	84 86	93 32
21	03 36	37 93	27 77	61 33	67 18	03 71	42 92	11 08	72 50	42 57
22	56 99	98 35	93 08	13 74	70 34	08 06	54 14	13 96	70 13	31 20
23	40 44	11 04	65 34	80 10	45 10	34 18	05 10	52 56	09 08	73 52
24	96 18	05 86	36 49	04 50	99 69	36 73	86 90	39 47	20 03	55 86
25	28 99	10 48	37 85	42 43	09 62	14 24	49 66	08 64	58 73	44 22

Źródło: opracowanie własne z wykorzystaniem funkcji Microsoft Excel.

Tablica VIe.
Tablice liczb losowych
Blok 5

	numer kolumny									
numer wiersza	1	2	3	4	5	6	7	8	9	10
1	00 19	64 98	01 38	86 72	34 45	23 25	51 54	09 61	18 49	19 27
2	45 35	65 35	54 13	97 64	42 63	88 05	14 19	48 30	80 91	73 29
3	32 84	16 20	24 16	96 50	91 85	60 59	80 92	96 23	64 79	53 06
4	40 48	28 18	13 05	91 77	03 60	95 58	25 50	93 18	16 07	79 99
5	31 87	42 09	02 89	79 33	32 35	62 56	02 62	27 10	19 18	25 23
6	32 03	08 25	32 94	77 27	34 88	31 17	75 99	89 14	88 21	46 76
7	65 84	40 97	81 20	87 45	61 09	24 26	64 40	99 69	89 05	43 23
8	05 66	67 75	31 16	74 07	76 35	15 70	60 90	52 86	42 17	45 23
9	85 48	37 79	11 15	53 36	72 43	14 40	97 80	15 93	01 22	03 43
10	39 94	89 83	08 70	62 11	83 64	76 59	96 61	52 22	02 07	96 52
11	17 12	98 53	16 84	62 23	85 07	45 25	99 15	23 61	06 4	67 16
12	27 71	17 06	59 28	16 94	12 27	37 41	14 82	99 87	49 84	59 45
13	48 30	39 08	76 75	19 93	54 18	04 69	12 70	68 75	40 01	74 59
14	34 99	41 92	02 42	83 16	06 16	49 69	18 41	28 37	83 64	14 71
15	10 89	48 53	24 48	07 38	56 16	18 83	53 39	88 64	11 28	21 23
16	01 50	33 50	66 73	18 36	29 74	50 96	47 90	38 38	89 91	52 43
17	94 81	38 66	01 85	18 44	74 92	39 51	86 03	87 07	72 85	53 40
18	70 67	10 81	80 37	05 28	99 83	01 62	72 06	25 66	18 50	64 12
19	88 69	12 90	87 89	79 70	59 41	16 14	53 28	25 92	37 80	39 17
20	69 61	04 71	88 45	17 11	05 47	45 21	57 20	37 05	91 40	62 06
21	76 92	13 54	43 73	45 39	73 54	45 80	17 10	91 37	34 05	40 43
22	23 62	94 26	83 90	66 28	96 63	74 08	93 41	49 04	88 51	88 08
23	86 39	62 16	68 38	26 45	21 09	87 24	13 65	12 61	85 09	07 32
24	93 68	89 67	35 93	32 13	22 76	08 63	72 94	65 49	03 60	06 61
25	19 21	28 47	89 97	21 75	81 35	02 05	37 50	38 19	73 99	75 87

Źródło: opracowanie własne z wykorzystaniem funkcji Microsoft Excel.

Tablica VIf.
Tablice liczb losowych
Blok 6

numer wiersza		numer kolumny 1	2	3	4	5	6	7	8	9	10
	1	11 40	41 79	72 70	14 42	67 85	73 14	89 97	70 63	24 93	11 40
	2	50 15	94 38	10 12	05 25	73 43	19 07	58 08	32 55	07 31	50 15
	3	17 84	40 26	63 64	52 47	83 14	89 40	52 05	77 03	98 53	17 84
	4	70 04	08 22	94 54	32 13	13 94	45 93	75 55	92 85	68 22	70 04
	5	52 67	43 23	91 78	38 21	97 58	58 26	59 67	39 43	70 44	52 67
	6	62 63	30 01	38 86	10 16	81 11	65 97	11 09	62 23	53 99	62 63
	7	25 57	33 67	73 46	68 01	51 17	98 10	70 31	66 80	87 67	25 57
	8	12 64	37 41	75 40	54 41	63 20	29 98	25 81	10 48	10 53	12 64
	9	15 66	81 91	88 76	14 88	47 70	29 74	97 99	01 28	99 51	15 66
	10	97 88	81 88	87 38	20 93	48 89	59 68	61 52	13 76	81 40	97 88
	11	10 35	71 87	59 03	76 92	68 25	64 55	75 14	51 96	95 69	10 35
	12	02 93	69 75	89 55	61 58	28 14	52 07	61 37	82 90	37 20	02 93
	13	95 2	55 16	70 28	09 28	06 93	98 86	36 26	08 61	01 87	95 02
	14	53 92	31 70	58 37	05 60	45 29	90 13	70 37	04 88	00 38	53 92
	15	09 74	04 69	72 21	92 59	43 18	50 44	57 68	49 02	87 10	09 74
	16	12 84	02 07	48 23	84 73	49 99	20 78	76 12	27 24	28 23	12 84
	17	03 10	76 23	18 28	10 09	56 79	89 97	99 57	64 04	21 38	03 10
	18	11 22	41 61	58 54	99 75	62 27	29 80	15 24	10 86	18 78	11 22
	19	77 38	60 69	31 44	19 72	25 90	38 70	37 31	51 59	27 81	77 38
	20	76 75	54 32	70 47	24 15	93 66	01 45	32 06	40 65	80 74	76 75
	21	34 80	49 30	20 28	51 62	35 99	22 62	68 47	70 23	28 48	34 80
	22	11 40	41 79	72 70	14 42	67 85	73 14	89 97	70 63	24 93	11 40
	23	50 15	94 38	10 12	05 25	73 43	19 07	58 08	32 55	07 31	50 15
	24	17 84	40 26	63 64	52 47	83 14	89 40	52 05	77 03	98 53	17 84
	25	70 04	08 22	94 54	32 13	13 94	45 93	75 55	92 85	68 22	70 04

Źródło: opracowanie własne z wykorzystaniem funkcji Microsoft Excel.

Tablica VIg.
Tablice liczb losowych
Blok 7

	numer kolumny									
numer wiersza	1	2	3	4	5	6	7	8	9	10
1	47 57	72 36	78 97	70 93	34 70	00 72	09 10	32 33	13 31	19 91
2	05 10	01 33	20 77	72 53	47 99	79 41	71 65	12 19	35 18	69 70
3	75 12	36 99	79 62	82 48	03 58	98 14	77 21	71 40	10 70	24 40
4	88 32	72 23	06 47	38 45	16 85	30 54	18 37	23 23	54 66	30 19
5	88 74	46 51	60 15	19 50	58 56	13 39	92 28	07 81	84 80	90 03
6	33 86	33 25	73 44	76 80	84 81	57 91	14 45	62 40	11 52	02 88
7	18 48	19 89	17 31	97 83	31 83	82 91	60 28	21 37	84 25	22 38
8	23 46	08 36	89 42	00 02	99 48	76 94	83 50	84 65	10 95	16 64
9	97 14	94 64	69 62	85 31	05 23	79 58	15 91	85 48	60 16	44 01
10	17 81	91 78	49 18	92 66	54 11	41 92	98 33	12 91	82 85	10 26
11	47 84	38 56	03 28	99 07	35 84	70 17	30 58	22 90	48 70	45 26
12	54 73	91 08	03 73	21 32	69 64	71 71	24 45	62 21	75 60	08 39
13	36 25	99 57	33 89	49 77	04 39	28 58	97 39	46 67	54 20	10 17
14	09 47	46 73	87 06	96 44	42 25	36 22	19 64	93 83	35 03	41 32
15	71 03	73 03	23 20	62 56	03 42	66 73	53 52	20 96	14 63	32 42
16	17 83	43 75	71 54	54 97	16 55	66 11	70 35	71 51	88 11	72 77
17	29 51	14 23	59 66	09 05	51 06	84 77	70 10	54 48	52 91	26 55
18	07 16	38 25	31 82	84 61	95 35	15 68	82 59	98 46	28 10	83 62
19	62 13	92 25	73 39	83 40	39 89	33 45	01 01	51 83	84 44	02 42
20	21 74	21 50	65 25	60 75	29 20	95 10	61 35	70 35	21 17	97 15
21	35 13	78 54	82 75	85 17	87 31	84 54	25 77	57 39	43 99	63 26
22	91 89	69 60	52 33	59 10	31 75	02 42	79 04	04 75	66 56	61 68
23	70 93	01 19	01 85	76 29	39 68	37 50	27 65	42 58	70 42	76 05
24	85 55	92 51	48 13	79 24	61 47	60 97	83 42	45 68	93 73	75 40
25	57 83	06 10	99 26	47 76	24 44	92 59	93 19	62 84	75 98	24 93

Źródło: opracowanie własne z wykorzystaniem funkcji Microsoft Excel.

Tablica VIh.
Tablice liczb losowych
Blok 8

	numer kolumny									
numer wiersza	1	2	3	4	5	6	7	8	9	10
1	91 03	67 97	51 44	48 07	74 08	87 21	78 25	44 35	33 56	91 03
2	50 53	80 49	21 10	81 85	01 87	48 35	50 11	56 33	25 21	50 53
3	17 87	83 34	98 29	55 35	23 17	06 33	65 75	71 06	67 49	17 87
4	70 33	96 46	63 23	45 12	87 40	68 69	05 15	57 92	15 02	70 33
5	52 96	72 67	20 32	68 96	74 49	41 17	68 45	27 26	60 48	52 96
6	62 70	08 64	67 21	87 78	90 42	17 18	14 79	09 98	67 89	62 70
7	25 61	01 07	10 26	19 80	44 89	49 06	70 73	39 68	00 68	25 61
8	12 96	19 15	14 63	07 86	01 44	30 83	41 38	74 97	99 14	12 96
9	15 95	52 14	27 96	05 15	90 13	52 45	57 51	06 55	58 88	15 95
10	97 84	70 04	30 28	43 77	37 28	99 27	48 39	71 31	96 09	97 84
11	10 86	34 38	33 30	49 28	41 44	60 55	89 33	89 46	35 76	10 86
12	02 90	11 14	39 19	63 53	65 22	64 57	50 28	17 06	40 57	02 90
13	95 06	14 28	48 23	61 11	86 01	73 01	16 72	90 52	31 90	95 06
14	53 18	26 06	92 87	68 56	57 16	67 14	73 68	45 61	39 73	53 18
15	09 29	31 00	95 55	63 88	71 31	60 32	97 22	14 91	91 25	09 29
16	12 41	10 75	95 78	66 33	84 78	19 12	47 63	82 07	80 90	12 41
17	03 92	27 90	40 98	07 30	60 06	15 58	38 48	58 98	57 10	03 92
18	11 16	63 32	29 15	61 79	58 00	08 82	01 39	44 82	90 67	11 16
19	77 55	06 21	25 83	20 85	68 48	96 42	91 82	68 60	44 80	77 55
20	76 05	97 92	30 28	93 17	19 96	18 54	24 08	70 73	10 12	76 05
21	34 05	30 22	63 99	56 95	87 04	95 38	67 19	41 59	42 92	34 05
22	91 03	67 97	51 44	48 07	74 08	87 21	78 25	44 35	33 56	91 03
23	50 53	80 49	21 10	81 85	01 87	48 35	50 11	56 33	25 21	50 53
24	17 87	83 34	98 29	55 35	23 17	06 33	65 75	71 06	67 49	17 87
25	70 33	96 46	63 23	45 12	87 40	68 69	05 15	57 92	15 02	70 33

Źródło: opracowanie własne z wykorzystaniem funkcji Microsoft Excel.

Tablica VIi.

Tablice liczb losowych

Blok 9

numer wiersza	*numer kolumny* 1	2	3	4	5	6	7	8	9	10
1	57 75	34 74	49 35	30 83	10 96	33 42	65 86	10 42	93 83	57 75
2	50 92	70 09	03 36	01 12	10 37	14 38	86 95	99 96	29 10	50 92
3	17 11	37 53	92 84	95 38	27 80	53 85	98 58	23 76	58 02	17 11
4	70 28	70 49	16 41	51 51	08 00	01 70	53 07	63 58	84 42	70 28
5	52 51	26 20	22 66	31 76	41 35	66 05	00 09	22 57	33 85	52 51
6	62 92	07 13	60 21	12 13	84 74	02 01	89 21	06 74	63 80	62 92
7	25 15	25 20	72 24	07 02	74 37	40 40	56 77	61 51	40 92	25 15
8	12 97	40 02	58 37	45 87	30 06	47 79	21 06	26 21	07 86	12 97
9	15 19	65 56	38 71	07 15	03 25	76 22	28 47	18 06	01 24	15 19
10	97 79	38 48	17 32	57 85	38 83	40 39	81 48	20 88	24 33	97 79
11	10 73	41 33	24 20	15 90	06 85	63 59	48 60	53 75	22 18	10 73
12	02 27	64 07	63 81	27 16	92 65	34 97	71 92	70 37	70 74	02 27
13	95 11	82 89	20 75	88 69	84 70	42 14	69 46	31 44	33 21	95 11
14	53 20	97 39	16 59	04 74	67 35	74 01	74 61	67 69	57 03	53 20
15	09 55	93 22	53 41	63 61	27 46	74 01	87 52	85 82	01 51	09 55
16	12 18	50 91	46 37	51 90	10 42	52 55	82 30	12 40	10 10	12 18
17	03 67	10 15	83 97	45 71	44 20	36 94	84 01	10 70	17 98	03 67
18	11 04	91 36	23 50	44 62	34 21	24 02	75 82	67 05	50 90	11 04
19	77 74	32 09	14 57	84 32	60 45	63 03	09 96	76 86	50 38	77 74
20	76 01	89 85	67 41	03 71	51 69	43 14	64 95	45 48	22 09	76 01
21	34 47	54 73	24 12	61 29	41 09	72 25	47 07	40 24	33 10	34 47
22	57 75	34 74	49 35	30 83	10 96	33 42	65 86	10 42	93 83	57 75
23	50 92	70 09	03 36	01 12	10 37	14 38	86 95	99 96	29 10	50 92
24	17 11	37 53	92 84	95 38	27 80	53 85	98 58	23 76	58 02	17 11
25	70 28	70 49	16 41	51 51	08 00	01 70	53 07	63 58	84 42	70 28

Źródło: opracowanie własne z wykorzystaniem funkcji Microsoft Excel.

Tablica VIj.
Tablice liczb losowych
Blok 10

		numer kolumny																			
		1		2		3		4		5		6		7		8		9		10	
numer wiersza	1	34	25	7	66	31	3	13	13	32	37	13	66	44	38	29	44	83	04	34	25
	2	50	96	49	72	22	32	77	55	92	48	24	33	29	65	89	48	45	51	50	96
	3	17	11	40	71	87	56	89	14	49	83	19	46	64	37	95	71	05	51	17	11
	4	70	37	91	73	63	92	52	02	13	61	50	02	90	99	66	94	85	41	70	37
	5	52	49	50	31	85	53	31	30	21	56	10	93	78	12	71	09	30	40	52	49
	6	62	81	26	99	84	04	70	12	70	69	82	16	36	57	71	91	42	74	62	81
	7	25	56	09	84	21	23	52	96	09	64	46	65	04	27	31	82	17	08	25	56
	8	12	57	48	04	41	88	57	75	65	37	92	89	28	85	30	37	82	59	12	57
	9	15	27	23	66	14	54	92	35	55	40	20	63	66	26	02	75	72	50	15	27
	10	97	73	75	33	46	28	44	04	83	87	04	19	65	41	44	71	91	75	97	73
	11	10	87	05	34	93	25	77	95	91	51	43	80	52	03	48	36	90	77	10	87
	12	02	80	74	62	78	27	82	39	38	11	99	41	65	51	10	23	33	83	02	80
	13	95	14	25	10	98	61	31	73	11	16	35	72	41	96	44	69	19	20	95	14
	14	53	61	19	81	66	34	74	80	40	77	42	23	11	10	48	25	03	92	53	61
	15	09	14	97	48	36	50	63	40	96	91	38	22	82	05	47	52	30	97	09	14
	16	12	94	90	78	14	02	57	82	38	14	52	85	57	06	30	15	42	53	12	94
	17	03	27	76	66	17	88	67	45	07	92	71	56	11	79	54	79	37	16	03	27
	18	11	70	40	21	85	25	42	28	78	94	69	99	61	72	07	60	88	83	11	70
	19	77	67	96	75	89	07	47	22	71	23	11	38	82	06	48	07	75	59	77	67
	20	76	50	15	15	21	37	34	63	94	17	97	71	50	24	85	62	58	53	76	50
	21	34	59	81	94	65	67	69	65	88	24	03	50	56	00	75	93	45	06	34	59
	22	34	25	07	66	31	03	13	13	32	37	13	66	44	38	29	44	83	04	34	25
	23	50	96	49	72	22	32	77	55	92	48	24	33	29	65	89	48	45	51	50	96
	24	17	11	40	71	87	56	89	14	49	83	19	46	64	37	95	71	05	51	17	11
	25	70	37	91	73	63	92	52	02	13	61	50	02	90	99	66	94	85	41	70	37

Źródło: opracowanie własne z wykorzystaniem funkcji Microsoft Excel.

Dr Michał Major jest wykładowcą w Krakowskiej Szkole Wyższej im. Andrzeja Frycza Modrzewskiego oraz adiunktem w Katedrze Statystyki w Akademii Ekonomicznej w Krakowie. Zajmuje się zagadnieniami z zakresu statystyki, ekonometrii, zarządzania jakością oraz dziedzinami pokrewnymi.

Podręcznik został opracowany jako kontynuacja książki napisanej przez Michała Majora i Janusza Niezgodę pt. *Elementy statystyki. Cz. I. Statystyka opisowa* (Kraków 2003). Może on być wykorzystywany także samodzielnie przy założeniu, że czytelnik posiada podstawowy zasób wiedzy ze statystyki. Podręcznik charakteryzuje się wysokim poziomem wiedzy przy jednoczesnej próbie prostego jej przekazu. Za bardzo istotne uważam prezentowanie różnych wariantów zastosowania omawianych metod, zależnie od zmieniających się warunków.

(z recenzji prof. KSW dr hab. Barbary Podolec)

Książka jest wartościową pozycją dydaktyczną dla studentów studiujących statystykę, która nie jest przedmiotem łatwym. Autor podjął się bardzo ambitnego zadania dostarczenia studentom pomocy dydaktycznej.
Przedstawiony podręcznik napisany jest w sposób przystępny. Liczne przykłady ułatwiają zrozumienie przekazywanych treści.

(z recenzji prof. AE dr hab. Józefa Biolika)

OFICYNA WYDAWNICZA
AFM

ISBN 978-83-89823-27-