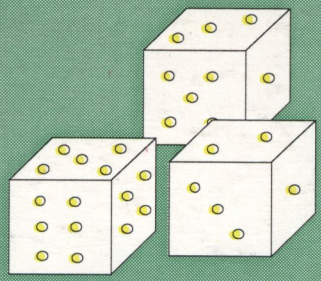
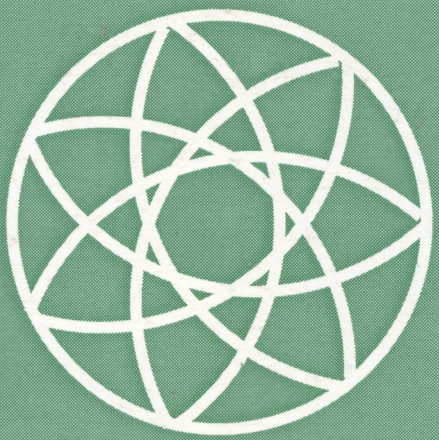


Jan Stankiewicz
Katarzyna Wilczek

Elementy rachunku prawdopodobieństwa i statystyki matematycznej

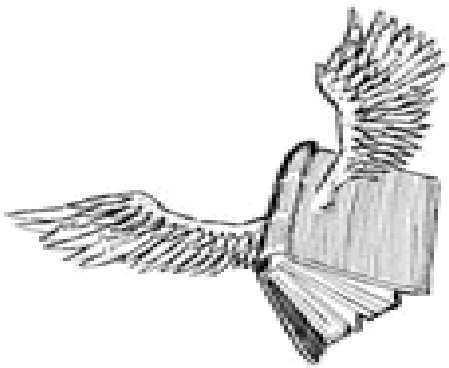
Teoria, przykłady, zadania



atematyka
a studentów
olitechniki Rzeszowskiej

ISBN 83-7199-146-0

Jan Stankiewicz
Katarzyna Wilczek



Elementy rachunku prawdopodobieństwa i statystyki matematycznej

Teoria, przykłady, zadania

Matematyka dla studentów Politechniki Rzeszowskiej

OFICYNA WYDAWNICZA POLITECHNIKI RZESZOWSKIEJ

Wydano za zgodą Rektora

Opiniodawca
prof. dr hab. Jerzy TOCKI

Redaktor
Genowefa SPÓLNIK

Redaktor techniczny
Anna MAZUR

Autor logo na okładce
Stanisława DUDA

Druk wykonano z matryc dostarczonych przez autorów

Skrypt uczelniany dla studentów Wydziału Budowy Maszyn i Lotnictwa
do wykładów z przedmiotu "matematyka"

Prawdopodobieństwo
Teoria korelacji
Statystyka
Weryfikacja hipotez
Tablice statystyczne

ISBN 83-7199-146-0

Oficyna Wydawnicza Politechniki Rzeszowskiej
ul. W. Pola 2, 35-959 Rzeszów

Nakład 400 + 35 egz. Ark. wyd. 10,1. Ark. druk. 9,5. Papier offset. kl. III 70g B1.
Oddano do druku w lipcu 2000 r. Wydrukowano w sierpniu 2000 r.
Zakład Poligrafii Politechniki Rzeszowskiej im. Ignacego Łukasiewicza, ul. W. Pola 2, 35-959 Rzeszów
Zam. nr 44/2000

Spis treści

Wstęp

W przedstawionej czytelnikom książce są omawiane podstawowe własności prawdopodobieństwa zdarzeń losowych, prawdopodobieństwo warunkowe, twierdzenia o prawdopodobieństwie całkowitym oraz schemat Bernoulliego. Następnie są wprowadzone pojęcia zmiennej losowej i jej rozkładu. Zaprezentowane są przykłady rozkładów zmiennej oraz podstawowe parametry rozkładu: wartość przeciętna, wariancja oraz momenty wyższych rzędów. Omawia się zagadnienia zależności i korelacji zmiennych losowych, prawa wielkich liczb oraz twierdzenia centralne.

W drugiej części, poświęconej statystyce matematycznej, są omawiane badania statystyczne, stawianie i weryfikowanie hipotez statystycznych różnych typów. Podane są wybrane testy statystyczne, a na zakończenie jest prezentowana metoda najmniejszych kwadratów.

Materiały te zostały opracowane na podstawie wykładów i ćwiczeń, jakie prowadziliśmy w Politechnice Rzeszowskiej na Wydziale Budowy Maszyn i Lotnictwa oraz Wydziale Elektrycznym w okresie 1982-1999. Nie stanowią one pełnego wykładu Rachunku Prawdopodobieństwa i Statystyki Matematycznej a jedynie, zgodnie z tytułem, wybrane elementy. Zostały one tak wybrane, aby stanowiły pewną całość możliwą do opanowania w czasie przewidzianym programem nauczania matematyki i, naszym zdaniem, potrzebną studentom studiów technicznych do dalszego kształcenia.

Oddajemy do rąk czytelników tę książkę mając nadzieję, że spełnia one następujące warunki:

a) prezentuje materiał możliwy do opanowania przez studentów drugiego roku studiów technicznych,

b) zawiera tylko konieczny materiał, bez zbytniego rozwijania go do pełnej, lecz zbyt obszernej teorii.

Odnosimy wrażenie, że na rynku podręczników akademickich brak jest zwięzłego opracowania prezentowanych przez nas zagadnień, opracowania, które spełniałoby te warunki.

Dziękujemy recenzentowi Panu dr. hab. Jerzemu Tockiemu za wnikliwe uwagi merytoryczne oraz Pani redaktor mgr Genowefie Spólnik za uwagi stylistyczno-redakcyjne. Przyczyniły się one istotnie do poprawienia jakości wydania.

Będziemy wdzięczni za wszelkie uwagi i sugestie czytelników, które nasuną się im w trakcie korzystania z niniejszego podręcznika. Zostaną one wykorzystane przy, ewentualnym, kolejnym wydaniu.

Jan Stankiewicz, Katarzyna Wilczek
Katedra Matematyki
Politechniki Rzeszowskiej
e-mail: jan.stankiewicz@man.rzeszow.pl
e-mail: wilczek@man.rzeszow.pl

Rozdział 1

Prawdopodobieństwo

1.1 Definicja i podstawowe własności prawdopodobieństwa

W rachunku prawdopodobieństwa podstawowe jest pojęcie **zdarzenia** – zdarzenia losowego. Zdarzenie łączymy zazwyczaj z wynikiem pewnej obserwacji lub doświadczenia. Wynik ten może być ilościowy lub jakościowy (otrzymanie "czwórki" podczas rzutu kostką, wylosowanie "czerwonej" kuli z urny czy asa z talii kart). Wsród zdarzeń wyróżniamy takie, których nie da się rozłożyć na zdarzenia prostsze. Takie zdarzenia są nazywane **zdarzeniami elementarnymi**. Na przykład otrzymanie "liczby parzystej" jest zdarzeniem losowym, ale nie jest zdarzeniem elementarnym, gdyż może być zrealizowane przez otrzymanie na kostce "dwójki", "czwórki" lub "szóstki". Te trzy ostatnie są zdarzeniami elementarnymi.

Przy rzucie klasyczną kostką mamy sześć zdarzeń elementarnych e_1, e_2, ..., e_6, polegających na otrzymaniu jednej z liczb naturalnych 1, 2, 3, 4, 5, 6.

Zbiór wszystkich możliwych zdarzeń elementarnych w danym doświadczeniu lub obserwacji nazywamy **przestrzenią zdarzeń elementarnych**, w skrócie PZE. Jest ona zwykle oznaczana przez Ω, a jej elementy literą e z indeksem jako e_k. W przypadku rzutu kostką sześcienną

$$\Omega = \{e_1, e_2, e_3, e_4, e_5, e_6\}.$$

Podzbiory przestrzeni Ω są **zdarzeniami losowymi** (ZL) lub krótko zdarzeniami i są oznaczane początkowymi literami alfabetu $A, B, \ldots$

Gdy przestrzeń Ω ma skończoną lub przeliczalną liczbę elementów, wówczas każdy podzbiór zbioru Ω jest zdarzeniem losowym. Tak nie musi być, gdy zbiór Ω nie jest przeliczalny.

Przed wprowadzeniem pojęcia prawdopodobieństwa potrzebne nam będą pewne pojęcia.

Zdarzeniem przeciwnym $\overline{A}$ do zdarzenia A nazywamy zdarzenie polegające na tym, że nie zachodzi zdarzenie A.

$$\overline{A} := \Omega \setminus A.$$

Zdarzeniem pewnym nazywamy zdarzenie, które zawiera wszystkie elementy PZE (które musi się zdarzyć). Oznaczamy go przez Ω.

Zdarzenie niemożliwe to takie, które nie może zajść. Odpowiada mu zbiór pusty $\emptyset$. Oznaczamy go więc przez $\emptyset$.

Sumą albo **alternatywą zdarzeń** A i B nazywamy zdarzenie odpowiadające sumie $A \cup B$ zbiorów. Podobnie określamy sumę skończonej lub przeliczalnej liczby zdarzeń

$$A_1 \cup A_2 \ldots \cup A_n = \bigcup_{k=1}^{n} A_k,$$

$$A_1 \cup A_2 \ldots = \bigcup_{k=1}^{\infty} A_k.$$

Iloczynem zdarzeń A i B nazywamy zdarzenie polegające na jednoczesnym zajściu obu zdarzeń i oznaczamy

$$A \cap B.$$

Dla większej liczby zdarzeń

$$A_1 \cap A_2 \ldots \cap A_n = \bigcap_{k=1}^{n} A_k,$$

$$A_1 \cap A_2 \ldots = \bigcap_{k=1}^{\infty} A_k.$$

Zdarzenia A i B nazywamy **wykluczającymi się**, jeżeli ich iloczyn jest zdarzeniem niemożliwym, to znaczy wtedy, gdy

$$A \cap B = \emptyset.$$

Wiadomo, że realizacja zdarzeń losowych jest różna. Na przykład: wyrzucenie kostką liczby parzystej jest trzy razy bardziej "prawdopodobne" niż wyrzucenie "dwójki". Dlatego wprowadza się liczbową miarę możliwości realizacji zdarzenia losowego, przypisującą zdarzeniu A pewną liczbę $P(A)$, zwaną prawdopodobieństwem zdarzenia A.

Jest wiele różnych definicji prawdopodobieństwa. Jedną z nich jest tak zwana klasyczna definicja pochodząca od P.S. Laplace'a, wielkiego francuskiego uczonego, żyjącego w latach 1749-1827. Definicja ta odnosi się do sytuacji, gdy PZE ma skończoną liczbę zdarzeń elementarnych.

Definicja 1.1 (Klasyczna definicja prawdopodobieństwa) *Niech przestrzeń zdarzeń elementarnych Ω składa się ze skończonej liczby zdarzeń elementarnych i zajście każdego z nich jest jednakowo możliwe. Jeżeli $n(A)$ jest liczbą zdarzeń elementarnych sprzyjających zdarzeniu A (należących do A) i $n(\Omega)$ jest liczbą wszystkich zdarzeń elementarnych, to prawdopodobieństwem zdarzenia A nazywamy liczbę*

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

Przykład 1.1 *Obliczyć prawdopodobieństwo wyciągnięcia kuli białej z urny zawierającej 10 kul białych, 15 kul czarnych i 5 kul czerwonych.*

Rozwiązanie. PZE ma w tym przypadku 30 elementów będących zdarzeniami elementarnymi. Zdarzeniu B polegającemu na wyciągnięciu kuli białej sprzyja 10 zdarzeń elementarnych, gdyż w urnie jest 10 kul białych. Zatem $n(\Omega) = 30$, $n(B) = 10$ i stąd

$$P(B) = \frac{10}{30} = \frac{1}{3}.$$

Z klasycznej definicji prawdopodobieństwa, przy ustalonej PZE, wynikają następujące własności prawdopodobieństwa – funkcji określonej na zdarzeniach – podzbiorach przestrzeni Ω.

Własność 1 Prawdopodobieństwo przyjmuje wartości nie mniejsze od zera i nie większe od jedności:

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

Własność 2 Prawdopodobieństwo zdarzenia pewnego $A = \Omega$ jest równe jedności:

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

Własność 3 Prawdopodobieństwo zdarzenia niemożliwego $A = \emptyset$ jest równe zeru:

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

Własność 4 Prawdopodobieństwo zdarzenia $\overline{A}$ przeciwnego do A jest równe różnicy prawdopodobieństwa zdarzenia pewnego i danego zdarzenia, to znaczy:

$$P(\overline{A}) = 1 - P(A)$$

lub inaczej

$$P(A) + P(\overline{A}) = 1.$$

Własność 5 Jeżeli A i B są zdarzeniami rozłącznymi, to prawdopodobieństwo sumy $A \cup B$ jest równe sumie prawdopodobieństw:

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

Przytoczona tutaj klasyczna definicja prawdopodobieństwa ma pewne mankamenty. Jest oparta na pojęciu zdarzeń elementarnych "jednakowo możliwych", to jest zawiera "błędne koło" w definiowaniu i dotyczy tylko przypadku, gdy zdarzeń tych jest skończona liczba.

Próbą usunięcia niektórych wad klasycznej definicji jest tak zwana definicja geometryczna, oparta na pojęciu miary zbioru (długość, pole, objętość).

Definicja 1.2 (Geometryczna definicja prawdopodobieństwa) *Niech G będzie ustalonym zbiorem (jednowymiarowym, płaskim, przestrzennym) o zadanej mierze $m(G)$. Niech $g \subset G$ ma miarę $m(g)$. Zakładamy, że wybranie każdego punktu ze zbioru G jest jednakowo możliwe. Niech zdarzenie A polega na tym, że wybrany losowo punkt należy do zbioru g. Prawdopodobieństwo zdarzenia A definiujemy następująco:*

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

Definicja geometryczna pozwala na rozpatrzenie przypadku nieskończonej, nieprzeliczalnej liczby zdarzeń elementarnych. Jest ona jednak trudna do zastosowania i nie usuwa dalej pojęcia "jednakowo prawdopodobne".

Dlatego została wprowadzona trzecia definicja pochodząca od A.N. Kołmogorowa (matematyk rosyjski, 1903-1987) i nazywana definicją aksjomatyczną.

Definicja 1.3 (Aksjomatyczna definicja prawdopodobieństwa) *Pojęciem pierwotnym jest pojęcie przestrzeni zdarzeń elementarnych Ω. Następnie zostaje wyróżniona pewna rodzina S podzbiorów przestrzeni Ω, nazywana σ-ciałem i spełniająca warunki:*

1° $\Omega \in S$,

2° $A \in S \Rightarrow \overline{A} = \Omega \setminus A \in S$,

3° $A_k \in S$, $k = 1, 2, \ldots \Rightarrow \bigcup_{k=1}^{\infty} A_k \in S$
(suma skończonej lub przeliczalnej rodziny zbiorów rodziny S należy do tej rodziny S),

4° $A_k \in S$, $k = 1, 2, \ldots \Rightarrow \bigcap_{k=1}^{\infty} A_k \in S$
(iloczyn skończonej lub przeliczalnej rodziny zbiorów rodziny S należy do tej rodziny S).

Rodzina S spełniająca te warunki stanowi σ-ciało, nazywamy ją **zbiorem zdarzeń losowych**, *a elementy zbioru S* **zdarzeniami losowymi**.

Prawdopodobieństwem P nazywamy każdą funkcję określoną na zbiorze S zdarzeń losowych spełniającą następujące trzy warunki, nazywane **aksjomatami prawdopodobieństwa**.

Aksjomat I. Każdemu zdarzeniu losowemu $A \in S$ jest przyporządkowana dokładnie jedna liczba $P(A) \in \langle 0, 1 \rangle$, nazywana prawdopodobieństwem zdarzenia A.

Aksjomat II. Prawdopodobieństwo zdarzenia pewnego jest równe jedności:

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

Aksjomat III. Prawdopodobieństwo sumy skończonej lub przeliczalnej liczby zdarzeń $A_1, A_2, \ldots$ parami się wykluczających równa się sumie prawdopodobieństw poszczególnych zdarzeń:

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

gdy $A_k \cap A_j = \emptyset$ dla $k \neq j$.

Aksjomaty te określają jakiego rodzaju funkcje można nazwać prawdopodobieństwem, wartości zaś liczbowe ustala się w poszczególnych przypadkach stosownie do rozpatrywanego zagadnienia. Definicja ta jest formalnie poprawna, jest jednak dość trudna do przełożenia na język praktyczny.

Łatwo sprawdzić, że funkcje $P(A)$, określone zarówno w klasycznej definicji, jak też w geometrycznej definicji, spełniają Aksjomaty I-III, a więc są prawdopodobieństwami w myśl aksjomatycznej definicji prawdopodobieństwa.

Wykorzystując definicję aksjomatyczną prawdopodobieństwa, można wykazać następujące twierdzenia.

Twierdzenie 1.1 $P(\emptyset) = 0$.

Dowód. $\emptyset \cap \Omega = \emptyset$ i dlatego na mocy Aksjomatów III i II mamy

$$1 = P(\Omega) = P(\emptyset \cup \Omega) = P(\emptyset) + P(\Omega) = P(\emptyset) + 1.$$

Stąd $P(\emptyset) = 0$.

Twierdzenie 1.2 $P(\overline{A}) = 1 - P(A)$.

Dowód. $\Omega = A \cup \overline{A}$, $A \cap \overline{A} = \emptyset$. Stąd

$$1 = P(\Omega) = P(A \cup \overline{A}) = P(A) + P(\overline{A}),$$

co daje potrzebną równość.

Twierdzenie 1.3 $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Dowód. Ze znanych własności działań na zbiorach mamy

$$A \cup B = A \cup (B \setminus (A \cap B)), \quad A \cap (B \setminus (A \cap B)) = \emptyset,$$

$$B = (A \cap B) \cup (B \setminus (A \cap B)), \quad (A \cap B) \cap (B \setminus (A \cap B)) = \emptyset.$$

Zatem z Aksjomatu III mamy

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

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

Odejmując stronami, otrzymujemy potrzebną równość.

Twierdzenie 1.4 *Jeżeli* $A \subset B$, *to* $P(A) \leq P(B)$.

Dowód. $B = A \cup (B \setminus A)$ i na mocy Aksjomatu III mamy

$$P(B) = P(A) + P(B \setminus A) \geq P(A),$$

gdyż $P(B \setminus A) \geq 0$.

1.2 Prawdopodobieństwo warunkowe i całkowite

Rozważmy dwa zdarzenia losowe A i B, $P(B) > 0$.

Definicja 1.4 *Prawdopodobieństwem zajścia zdarzenia A pod warunkiem zajścia zdarzenia B (oznaczane symbolem* $P(A|B)$*) nazywamy liczbę*

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

Liczba $P(A|B)$ jest istotnie prawdopodobieństwem (spełnia aksjomaty) i nazywamy ją **prawdopodobieństwem warunkowym**. Rzeczywiście

1° Dla każdego zdarzenia A mamy $A \cap B \subset B$ i dlatego

$$0 \leq P(A \cap B) \leq P(B).$$

Stąd

$$0 \leq \frac{P(A \cap B)}{P(B)} \leq 1$$

i w konsekwencji $0 \leq P(A|B) = \frac{P(A \cap B)}{P(B)} \leq 1$. Zatem $P(A|B)$ spełnia Aksjomat I.

2° $P(\Omega|B) = \frac{P(\Omega \cap B)}{P(B)} = \frac{P(B)}{P(B)} = 1.$
Jest więc spełniony Aksjomat II.

3° Jeżeli $A_1, A_2, \ldots$, są parami rozłączne, to również zbiory $A_1 \cap B$, $A_2 \cap B, \ldots$ są parami rozłączne i dlatego

$$P(A_1 \cup A_2 \cup \ldots | B) = \frac{P((A_1 \cup A_2 \ldots) \cap B)}{P(B)} =$$
$$= \frac{P(A_1 \cap B) \cup (A_2 \cap B) \cup \ldots)}{P(B)} = \frac{P(A_1 \cap B)}{P(B)} + \frac{P(A_2 \cap B)}{P(B)} + \ldots =$$
$$= P(A_1|B) + P(A_2|B) + \ldots$$

Jest zatem spełniony również Aksjomat III.

Definicja 1.5 *Mówimy, że* **zdarzenia** A *i* B **są niezależne**, *jeżeli* $P(B) = 0$ *lub* $P(B) \neq 0$ *i* $P(A|B) = P(A)$.

Jeżeli $P(A) > 0$, to również $P(B|A) = P(B)$.
Jeżeli zdarzenia A i B nie są niezależne, to mówimy, że są **zależne.**
Innymi słowy, A i B są niezależne, gdy

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

Używając ostatniej równości, możemy definicję niezależności rozszerzyć na przypadek większej liczby zdarzeń.

Definicja 1.6 *Mówimy, że zdarzenia* $A_1, A_2, \ldots, A_n$ *są niezależne, jeżeli dla dowolnego ciągu* $i_1, i_2, \ldots, i_k$, $1 \leq i_1 \leq i_2 \leq \ldots \leq i_k \leq n$ *ma miejsce równość*

$$P(A_{i_1} \cap A_{i_2} \cap \ldots A_{i_k}) = P(A_{i_1}) \cdot P(A_{i_2}) \cdot \ldots \cdot P(A_{i_k}).$$

Zatem w odniesieniu do zdarzeń niezależnych prawdopodobieństwo ich iloczynu jest równe iloczynowi prawdopodobieństw.
Dla iloczynu dowolnych zdarzeń mamy

Twierdzenie 1.5 *Niech A, B będą dowolnymi zdarzeniami losowymi. Wtedy:*

$$\begin{aligned} P(A \cap B) &= P(A) + P(B) - P(A \cup B), \\ P(A \cap B) &= P(A)P(B|A), \quad gdy\, P(A) > 0, \\ P(A \cap B) &= P(B)P(A|B), \quad gdy\, P(B) > 0 \end{aligned}$$

oraz ogólniej

$$P(A_1 \cap A_2 \cap \ldots \cap A_n) =$$

$$= P(A_1)P(A_2|A_1)P(A_3|A_1 \cup A_2)\ldots P(A_n|A_1 \cup \ldots \cup A_{n-1}).$$

Definicja 1.7 *Mówimy, że układ zdarzeń losowych $A_1, A_2, \ldots$* **jest zupełny,** *jeżeli zdarzenia te są parami wykluczające się (rozłączne), a ich suma jest zdarzeniem pewnym, to znaczy gdy*

$$\begin{aligned} A_k \cap A_l &= \emptyset \quad dla\, k \neq l, \\ \bigcup_{k=1} A_k &= \Omega. \end{aligned}$$

Układ zupełny jest wykorzystywany do wyznaczania w pewnych przypadkach prawdopodobieństwa danego zdarzenia, gdy znamy prawdopodobieństwa zdarzeń A_k układu zupełnego oraz prawdopodobieństwa zdarzeń warunkowych.

Twierdzenie 1.6 (Twierdzenie o prawdopodobieństwie całkowitym) *Jeżeli $A_1, A_2, \ldots$ jest układem zupełnym zdarzeń losowych i $P(A_k) > 0$, $k = 1, 2, \ldots$, to dla dowolnego zdarzenia losowego B zachodzi równość*

$$P(B) = \sum_{k=1}^{\infty} P(A_k)P(B|A_k).$$

Dowód. Z zupełności układu wynika, że

$$B = B \cap \Omega = B \cap \left(\bigcup_k A_k\right) = \bigcup_k (B \cap A_k).$$

Ale $B_k = B \cap A_k$ są parami rozłączne i dlatego

$$P(B) = P\left(\bigcup_k (B \cap A_k)\right) = \sum_k P(B \cap A_k) = \sum_k P(A_k)P(B|A_k),$$

co kończy dowód.

Wzór ten nosi nazwę **wzoru na prawdopodobieństwo całkowite.**

Następne twierdzenie podaje wzór na prawdopodobieństwo warunkowe poszczególnych zdarzeń A_k układu zupełnego zdarzeń. W tych i w następnych wzorach symbole $\bigcup_k$, $\sum_k$ oznaczają odpowiednio $\bigcup_{k=1}^{\infty}$, $\sum_{k=1}^{\infty}$, gdy układ jest przeliczalny, oraz $\bigcup_{k=1}^{n}$, $\sum_{k=1}^{n}$, gdy układ zupełny ma skończoną liczbę elementów.

Twierdzenie 1.7 (Twierdzenie Bayesa) *Jeżeli $P(B) > 0$ oraz $A_1, A_2, \ldots$ jest układem zupełnym zdarzeń losowych, $P(A_k) > 0$, $k = 1, 2, \ldots$, to zachodzi wzór*

$$P(A_k|B) = \frac{P(A_k)P(B|A_k)}{\sum_i P(A_i)P(B|A_i)} = \frac{P(A_k \cap B)}{P(B)}$$

nazywany **wzorem Bayesa** *na prawdopodobieństwo warunkowe a posteriori.*

Dowód. Ze wzoru na prawdopodobieństwo warunkowe oraz z Twierdzenia o prawdopodobieństwie całkowitym mamy

$$P(A_k|B) = \frac{P(A_k \cap B)}{P(B)} = \frac{P(A_k)P(B|A_k)}{\sum_i P(A_i)P(B|A_i)},$$

co kończy dowód.

Przykład 1.2 *Mamy n urn do losowania wypełnionych kulami białymi i czarnymi. Prawdopodobieństwo wylosowania kuli białej z k-tej urny jest równe p_k, a prawdopodobieństwo wylosowania k-tej urny wynosi $a_k > 0$, $k = 1, 2 \ldots, n$. Wyznaczyć prawdopodobieństwo tego, że wylosowana biała kula pochodzi z k-tej urny.*

Rozwiązanie. Niech A_k będzie zdarzeniem polegającym na wylosowaniu k-tej urny i niech B będzie zdarzeniem polegającym na wylosowaniu białej kuli. Widzimy, że $A_1, A_2, \ldots, A_n$ tworzą zupełny układ zdarzeń. Wiemy też, że $P(A_k) = a_k$, $P(B|A_k) = p_k$. Zatem prawdopodobieństwo zdarzenia $A_k|B$, a więc że wylosowano białą kulę z k-tej urny, wyznaczamy ze wzoru Bayesa

$$P(A_k|B) = \frac{P(A_k)P(B|A_k)}{\sum_i P(A_i)P(B|A_i)} = \frac{p_k a_k}{\sum_{i=1}^{n} p_i a_i}.$$

1.3 Schemat Bernoulliego

Omówimy teraz pewien schemat doświadczenia losowego polegającego na wielokrotnym powtarzaniu tego samego doświadczenia losowego. Pojedyncze doświadczenie polega na tym, że mogą zajść jedynie zdarzenia A lub $\overline{A}$. Innymi słowy, gdy zbiór zdarzeń losowych składa się ze zdarzeń $\emptyset$, A, $\overline{A}$, Ω i

$P(A) = p \in (0,1)$ oraz $P(\overline{A}) = q = 1 - p \in (0,1)$. Zakładamy, że wyniki kolejnych prób są zdarzeniami niezależnymi.

Taki schemat może być na przykład realizowany jako losowanie kuli z urny ze zwracaniem po każdym losowaniu danej kuli do urny. Zdarzenie A polegałoby na wylosowaniu kuli białej, a $\overline{A}$ na wylosowaniu kuli, która nie jest biała.

Zatem powtarzamy n-krotnie dane doświadczenie i chcemy wyznaczyć prawdopodobieństwo, że dokładnie k razy będzie zachodziło zdarzenie A i dokładnie $n-k$ razy zdarzenie $\overline{A}$. Oznaczmy to prawdopodobieństwo przez $P_n(k)$. Omówione doświadczenie może być realizowane na wiele sposobów, na przykład przez k początkowych losowań otrzymujemy A, przez pozostałe zaś – zdarzenie $\overline{A}$. Sposobów tych jest tyle, na ile sposobów możemy wybrać zbiór k-elementowy ze zbioru n elementowego, a więc jest ich

$$C_k^n = \binom{n}{k} = \frac{n!}{k!(n-k)!} = \frac{(k+1)(k+2)\dots n}{1\cdot 2\cdot \dots \cdot (n-k)} = \\ = \frac{(n-k+1)(n-k+2)\dots n}{1\cdot 2\cdot \dots \cdot k}.$$

Prawdopodobieństwo zajścia każdego takiego sposobu jest takie samo i jest równe iloczynowi

$$\underbrace{P(A)\cdot P(A)\dots P(A)}_{k}\cdot \underbrace{P(\overline{A})\cdot P(\overline{A})\dots P(\overline{A})}_{(n-k)} = p^k\cdot q^{n-k} = p^k\cdot(1-p)^{n-k}.$$

Jeżeli prawdopodobieństwo zajścia zdarzenia A nazwiemy **sukcesem**, a zdarzenia $\overline{A}$ **porażką**, to prawdopodobieństwo $P_n(k)$ osiągnięcia dokładnie k sukcesów w n próbach wyraża się wzorem

$$P_n(k) = \binom{n}{k}p^kq^{n-k} = \binom{n}{k}p^k(1-p)^{n-k}.$$

Wzór ten nosi nazwę **wzoru Bernoulliego**. (Jakub Bernoulli, matematyk i fizyk Szwajcarski, 1654-1705, ponadto brat Johan, 1667-1748, oraz syn Jakuba Daniel, 1700-1782.)

Na mocy wzoru Newtona mamy

$$\sum_{k=0}^{n} P_n(k) = \sum_{k=0}^{n}\binom{n}{k}p^kq^{n-k} = (p+q)^n = 1^n = 1.$$

Wzór Bernoulliego daje możliwość wyznaczania różnych innych zdarzeń związanych ze schematem Bernoulliego.

Przykład 1.3 *Połóżmy $n = 6$, $p = 1/3$, $q = 2/3$. Obliczyć prawdopodobieństwo, że uzyskamy nie więcej niż trzy sukcesy.*

Rozwiązanie. Prawdopodobieństwo uzyskania w sześciu próbach nie więcej niż trzech sukcesów $P_6(k \le 3)$ jest sumą prawdopodobieństw $P_6(0) + P_6(1) + P_6(2) + P_6(3)$. Zatem

$$P_6(k\le 3) = \binom{6}{0}\left(\frac{2}{3}\right)^6 + \binom{6}{1}\frac{1}{3}\left(\frac{2}{3}\right)^5 + \binom{6}{2}\left(\frac{1}{3}\right)^2\left(\frac{2}{3}\right)^4 + \\ + \binom{6}{3}\left(\frac{1}{3}\right)^3\left(\frac{2}{3}\right)^3 = \frac{2^6}{3^6} + 6\frac{2^5}{3^6} + 15\frac{2^4}{3^6} + 20\frac{2^3}{3^6} = \frac{659}{729} = 0,8998\dots$$

Przykład 1.4 *Rzucamy 4 razy kostką do gry. Obliczyć prawdopodobieństwo, że wyrzucimy 6 oczek:*

a) dokładnie 2 razy,

b) co najmniej 2 razy.

Rozwiązanie. Oznaczmy przez B zdarzenie polegające na otrzymaniu 6 oczek dokładnie dwa razy, a przez B zdarzenie polegające na otrzymaniu 6 oczek co najmniej dwa razy. Wtedy

$$P(B) = P_4(2) = \binom{4}{2}\left(\frac{1}{6}\right)^2\left(\frac{5}{6}\right)^2 = \frac{25}{216} = 0,1157\dots$$

$$P(C) = P_4(2) + P_4(3) + P_4(4) = 6\left(\frac{1}{6}\right)^2\left(\frac{5}{6}\right)^2 + 4\left(\frac{1}{6}\right)^3\frac{5}{6} + \left(\frac{1}{6}\right)^4 = \\ = \frac{150+20+1}{6^4} = \frac{171}{1296} = 0,1319\dots$$

Jednym z ciekawszych zagadnień związanych ze schematem Bernoulliego jest wyznaczenie **najbardziej prawdopodobnej liczby sukcesów** w serii o zadanej długości.

Niech p i n będą ustalone i niech $P_n(k)$ oznacza prawdopodobieństwo uzyskania dokładnie k sukcesów. Przy jakiej wartości k ta wartość będzie największa? W tym celu rozważmy iloraz

$$\frac{P_n(k+1)}{P_n(k)} = \frac{\binom{n}{k+1}p^{k+1}q^{n-k+1}}{\binom{n}{k}p^kq^{n-k}} = \frac{(n-k)p}{(k+1)q}, \quad q = 1-p.$$

Zatem $P_n(k)$ jest funkcją rosnąca zmiennej k, gdy

$$(n-k)p > (k+1)q$$

lub równoważnie, gdy

$$k < np - q = (n+1)p - 1.$$

Gdy zachodzi nierówność przeciwna

$$k > np - q = (n+1)p - 1,$$

wówczas $P_n(k)$ jest malejącą funkcją zmiennej k.

Zatem $P_n(k)$ osiąga wartość największą, gdy

$$P_n(k_0 - 1) \leq P_n(k_0) \leq P_n(k_0)$$

a więc, gdy

$$(n+1)p - 1 \leq k_0 \leq (n+1)p.$$

Jeżeli $(n+1)p$ nie jest liczbą naturalną, to k_0 jest wyznaczona jednoznacznie i

$$k_0 = [(n+1)p],$$

gdzie $[x]$ = całość (część całkowita) liczby x.

Jeżeli $(n+1)p$ jest liczbą naturalną, to

$$k_0 = (n+1)p \quad \text{lub} \quad k_0 = (n+1)p - 1$$

oraz

$$P_n(k_0) = P_n((n+1)p - 1) = P_n((n+1)p).$$

Przykład 1.5 *Znaleźć najbardziej prawdopodobną liczbę sukcesów w schemacie Bernoulliego:*

a) dla $n = 7$ i $p = 0,3$,

b) dla $n = 19$ i $p = 0,2$.

Rozwiązanie. a) $(n+1)p = 8 \cdot 0,3 = 2,4$. Stąd $k_0 = [2,4] = 2$. Istotnie:

$$P_7(0) = 0,0823\ldots, \quad P_7(1) = 0,2470\ldots, \quad P_7(2) = 0,3176\ldots,$$
$$P_7(3) = 0,2268\ldots, \quad P_7(4) = 0,0972\ldots, \quad P_7(5) = 0,0250\ldots,$$
$$P_7(6) = 0,0035\ldots, \quad P_7(7) = 0,0002\ldots$$

b) $(n+1)p = 20 \cdot 0,2 = 4$. Stąd $k_0 = 3$ lub $k_0 = 4$,

$$P_{19}(3) = \binom{19}{3}(0,2)^3(0,8)^{16} = \binom{19}{14}(0,2)^4(0,8)^{15} =$$
$$= P_{19}(4) = 0,2181\ldots$$

Dla porównania $P_{19}(0) = 0,0144\ldots$, $P_{19}(19) = 0,000000000000052\ldots$

Uwaga. Przy dużych n w schemacie Bernoulliego najbardziej prawdopodobna liczba sukcesów k_0 jest w przybliżeniu równa np.

Wzór Bernoulliego dla dużych wartości n jest uciążliwy rachunkowo. Jest to związane z symbolem $\binom{n}{k}$ występującym w tym wzorze. Są więc używane pewne przybliżone metody wyznaczania $P_n(k)$.

Jedna z nich jest zawarta w następującym twierdzeniu:

Twierdzenie 1.8 (Twierdzenie lokalne Moivre'a-Laplace'a) *Przy ustalonym $p \in (0,1)$ i wszystkich k, dla których*

$$x_k = \frac{k - np}{\sqrt{npq}}$$

leży w dowolnym skończonym przedziale $\langle a, b\rangle$, mamy

$$\lim_{n\to\infty} \frac{P_n(k)}{(2\pi npq)^{-\frac{1}{2}} e^{-\frac{1}{2}x_k^2}} = 1, \qquad \left(P_n(k) \approx \frac{e^{-\frac{1}{2}\left(\frac{np-k}{\sqrt{npq}}\right)^2}}{\sqrt{2\pi npq}},\ q = 1 - p \right).$$

Twierdzenie 1.9 (Twierdzenie integralne Moivre'a-Laplace'a) *Niech prawdopodobieństwo zajścia zdarzenia A w serii n niezależnych doświadczeń będzie stałe i równe p, $0 < p < 1$. Jeżeli k oznacza liczbę sukcesów (zaistnień zdarzenia A) w tych n doświadczeniach, to dla dowolnych a, b, takich że $a < b$, mamy*

$$\lim_{n\to\infty} P\left(a \leq \frac{k - np}{\sqrt{np(1-p)}} \leq b\right) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{1}{2}t^2} dt.$$

Liczba $P\left(a \leq \frac{k - np}{\sqrt{np(1-p)}} \leq b\right)$ jest prawdopodobieństwem tego, że wartość x_k leży w przedziale $\langle a, b\rangle$.

1.4 Zmienne losowe i ich rozkłady

W dalszych zastosowaniach prawdopodobieństwa istotną rolę odgrywa zmienna losowa i rozkład prawdopodobieństwa zmiennej losowej. Pozwalają one zapisać pewne realne – fizyczne zjawiska losowe – w postaci modeli matematycznych.

Niech Ω będzie ustaloną przestrzenią zdarzeń elementarnych ω, S zbiorem zdarzeń losowych (σ-ciałem zdarzeń losowych).

Definicja 1.8 Zmienną losową X *nazywamy każdą funkcję określoną na przestrzeni* Ω*, przyjmującą wartości rzeczywiste, i taką, że dla każdej liczby rzeczywistej* x *zbiór zdarzeń elementarnych* ω *spełniających nierówność* $X(\omega) < x$ *jest zdarzeniem losowym, czyli należy do* S.

Innymi słowy $X = X(\omega) : \Omega \to \mathbb{R}$ nazywamy zmienną losową, jeżeli

$$\bigwedge_{x \in \mathbb{R}} \{\omega \in \Omega : X(\omega) < x\} \in S.$$

W celu lepszego zrozumienia tego pojęcia rozważmy je na przykładach.

I. Niech doświadczenie polega na rzucie monetą. Wtedy Ω składa się z dwu zdarzeń elementarnych:

ω_1 – wyrzucenie orła, ω_2 – wyrzucenie reszki, to znaczy

$$\Omega = \{\omega_1, \omega_2\}, \quad S = \{\omega_1, \omega_2, \emptyset, \Omega\}.$$

Określamy funkcję $X : \Omega \to \mathbb{R}$ następująco:

$$X(\omega_1) = 1, \quad X(\omega_2) = 0.$$

Tak określona funkcja jest zmienną losową. Istotnie

$$\{\omega \in \Omega : X(\omega) < x\} = \begin{cases} \emptyset & \text{dla } x \leq 0 \\ \omega_2 & \text{dla } 0 < x \leq 1 \\ \Omega & \text{dla } 1 < x \end{cases}.$$

skąd widać, że w każdym przypadku zbiory te są zdarzeniami losowymi z rodziny S.

Zauważmy, że w tym przypadku mamy:

$$P(X = 0) = P(\{\omega : X(\omega) = 0\}) = P(\omega_2) = \frac{1}{2},$$
$$P(X = 1) = P(\{\omega : X(\omega) = 1\}) = P(\omega_1) = \frac{1}{2}.$$

Jeżeli $P(X < x) = P(\{\omega : X(\omega) < x\})$, to mamy

$$P(X < x) = \begin{cases} 0 & \text{dla } x \leq 0 \\ \frac{1}{2} & \text{dla } 0 < x \leq 1 \\ 1 & \text{dla } 1 < x \end{cases}.$$

Równości te opisują tak zwany rozkład zmiennej losowej X.

II. Niech będzie dany schemat Bernoulliego, polegający na n-krotnym powtórzeniu ustalonego doświadczenia. Niech funkcja X przyporządkowuje każdemu schematowi liczbę k równą liczbie uzyskanych w nim sukcesów. Funkcja ta jest zmienną losową, gdyż zbiór $\{\omega : X(\omega) < x\}$ jest zawsze zdarzeniem losowym, polegającym na uzyskaniu liczby sukcesów nie większej niż ustalona.

Ponadto

$$P(X = k) = P_n(k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, 2, \ldots, n-1, n.$$

Definicja 1.9 *Dla danej zmiennej losowej* X *funkcję*

$$F(x) = P(X < x) = P(\{\omega : X(\omega) < x\})$$

określającą prawdopodobieństwo tego, że zmienna losowa X *przyjmie wartości mniejsze od* x*, nazywamy* **dystrybuantą zmiennej losowej** X.

Dystrybuanta $F(x)$ określona dla każdej zmiennej losowej X ma następujące własności:

1° $F(x)$ jest funkcją niemalejącą dla $x \in \overline{\mathbb{R}} = \langle -\infty, +\infty \rangle$.

2° $0 \leq F(x) \leq 1$ dla $x \in \overline{\mathbb{R}}$.

3° $F(-\infty) = \lim_{x \to -\infty} F(x) = 0, \quad F(+\infty) = \lim_{x \to +\infty} F(x) = 1.$

4° $F(x)$ jest funkcją lewostronnie ciągłą na $\overline{\mathbb{R}}$, to znaczy $\lim_{x \to x_0^-} F(x) = F(x_0)$.

5° $P(x_1 \leq X < x_2) = F(x_2) - F(x_1), \quad x_1, x_2 \in \overline{\mathbb{R}}.$

Zazwyczaj są wyróżniane dwa rodzaje zmiennych losowych:

- **zmienne losowe typu skokowego**, gdy zbiór wartości funkcji $X(\omega)$ jest zbiorem skończonym lub przeliczalnym,
- **zmienne losowe typu ciągłego**, gdy zbiór wartości funkcji $X(\omega)$ zawiera pewien przedział właściwy lub niewłaściwy.

Nie wyczerpują one jednak wszystkich możliwości, rozważa się również tak zwane

– **zmienne losowe typu mieszanego**, wówczas zbiorem wartości funkcji $X(\omega)$ jest suma zbioru skończonego lub przeliczalnego oraz przedziału właściwego albo niewłaściwego.

Różnice pomiędzy poszczególnymi typami zmiennych uwidaczniają się zarówno podczas opisywania rozkładów samych zmiennych, jak i tworzenia ich dystrybuant.

Zmienne typu skokowego

Definicja 1.10 Definicja rozkładu zmiennej losowej typu skokowego
Niech $x_1, x_2, \ldots, x_k, \ldots$ będzie skończonym lub przeliczalnym zbiorem wartości zmiennej losowej X. Funkcję

$$p_i = P(X = x_i) = P(\{\omega : X(\omega) = x_i\})$$

przyporządkowującą wartościom $x_1, x_2, \ldots, x_k, \ldots$ zmiennej losowej X odpowiednie prawdopodobieństwa $p_1, p_2, \ldots, p_k, \ldots$ nazywamy **rozkładem zmiennej losowej typu skokowego**. *Mamy przy tym $\sum_i p_i = 1$.*

Dla zmiennej losowej typu skokowego dystrybuanta wyraża sie wzorem

$$F(x) = \sum_{x_i < x} P(X = x_i) = \sum_{x_i < x} p_i \ .$$

Suma ta rozciąga się na te wskaźniki "i", dla których wartości zmiennej x_i są mniejsze od liczby x.

Najczęściej rozważanymi rozkładami zmiennej losowej typu skokowego są: rozkład zero-jedynkowy, rozkład Bernoulliego, rozkład Poissona.

Rozkład zero-jedynkowy

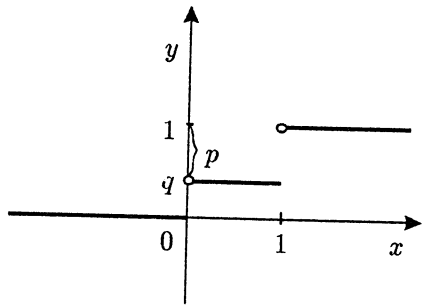
Rozkład zero-jedynkowy ma miejsce wtedy, gdy zmienna losowa X przyjmuje tylko dwie wartości: wartość 1 z prawdopodobieństwem p oraz wartość 0 z prawdopodobieństwem $q = 1-p$, $p \in (0,1)$. Możemy to zapisać następująco:

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

Dystrybuanta ma postać

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

Wykres funkcji $y = F(x)$ przedstawiony jest na Rys. 1.1.



Rys. 1.1. Wykres dystrybuanty rozkładu zero-jedynkowego

Rozkład Bernoulliego (rozkład dwumianowy)

Rozkład Bernoulliego (rozkład dwumianowy) ma miejsce wtedy, kiedy zmienna losowa X przyjmuje wartości $k \in \{0, 1, 2, \ldots, n-1, n\}$ odpowiednio z prawdopodobieństwem

$$p_k = P_n(k) = \binom{n}{k} p^k q^{n-k} = \binom{n}{k} p^k (1-p)^{n-k}, \ k = 0, 1, \ldots, n,$$

gdzie $p \in (0,1)$ jest ustalone i $q = 1 - p$, $(p + q = 1)$. Prawdopodobieństwo p_k jest równe k-temu wyrazowi dwumianu Newtona

$$(p+q)^n = \sum_{k=0}^{\infty} \binom{n}{k} p^k q^{n-k}$$

i stąd używana nazwa: rozkład dwumianowy.

Dystrybuanta tej zmiennej X wyraża się wzorem

$$F(x) = \sum_{k<x} P_n(k) = \sum_{k<x} \binom{n}{k} p^k q^{n-k}.$$

Gdy $\tilde{x} > n$, wówczas

$$F(\tilde{x}) = \sum_{k \leq n} P_n(k) = \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} = (p+q)^n = 1.$$

W szczególnym przypadku, gdy $n = 3$, $p = \frac{1}{2}$, mamy rozkład

$$p_0 = P_3(0) = \frac{1}{8}, \ p_1 = P_3(1) = \frac{3}{8}, \ p_2 = P_3(2) = \frac{3}{8}, \ p_3 = P_3(3) = \frac{1}{8}$$

oraz dystrybuantę

$$F(x) = \begin{cases} 0 & \text{dla } x \leq 0 \\ 1/8 & \text{dla } 0 < x \leq 1 \\ 1/2 & \text{dla } 1 < x \leq 2 \\ 7/8 & \text{dla } 2 < x \leq 3 \\ 1 & \text{dla } 3 < x \end{cases}.$$

Rozkład Poissona

Rozkład Poissona ma miejsce wtedy, kiedy zmienna X przyjmuje odpowiednio wartości będące liczbami naturalnymi $k \in \{0, 1, 2, \ldots\}$ z prawdopodobieństwami

$$p_k = \frac{\lambda^k}{k!} e^{-\lambda},$$

gdzie λ jest ustalonym dodatnim parametrem charakteryzującym ten rozkład.

Rozkład ten jest granicznym przypadkiem rozkładu Bernoulliego przy $n \to \infty$ i zmiennym p, takim że $p = \lambda/n$. Istotnie:

$$\lim_{n\to\infty} P_n(k) = \lim_{n\to\infty} \frac{n!}{k!(n-k)!} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} =$$

$$= \lim_{n\to\infty} \left\{ \left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right) \ldots \left(1 - \frac{k-1}{n}\right)\left(1 - \frac{\lambda}{n}\right)^{-k} \cdot \frac{\lambda^k}{k!} \left(1 - \frac{\lambda}{n}\right)^n \right\} =$$

$$= \frac{\lambda^k}{k!} e^{-\lambda}.$$

Dystrybuanta ma postać

$$F(x) = \sum_{k<x} \frac{\lambda^k}{k!} e^{-\lambda} = e^{-\lambda} \sum_{k<x} \frac{\lambda^k}{k!}$$

i stąd

$$F(+\infty) = e^{-\lambda} \sum_{k=x} \frac{\lambda^k}{k!} = e^{-\lambda} e^{\lambda} = 1.$$

W rozkładzie Poissona zmiena losowa X przyjmuje nieskończenie wiele, a dokładnie – przeliczalnie wiele wartości.

Zmienne typu ciągłego

Dla zmiennej losowej typu ciągłego nie można wprowadzić pojęcia rozkładu podobnego do rozkładu zmiennej typu skokowego. Dlatego w tym przypadku zamiast rozkładu wprowadza się pojęcie "gęstości prawdopodobieństwa" zmiennej losowej typu ciagłego.

Załóżmy, że X jest zmienną losową typu ciągłego i $F(x)$ jest dystrybuantą tej zmiennej. Wtedy przy $\Delta x > 0$ mamy

$$\lim_{\Delta x \to 0+} \frac{P(x \leq X < x + \Delta x)}{\Delta x} = \lim_{\Delta x \to 0} \frac{F(x + \Delta x) - F(x)}{\Delta x} = F'(x)$$

jeżeli dystrybuanta F jest funkcją różniczkowalną. Wówczas funkcję

$$f(x) \stackrel{\text{def}}{=} F'(x) \geq 0$$

nazywamy **gęstością prawdopodobieństwa** zmiennej losowej typu ciągłego.

Możemy teraz podać nieco ściślejszą definicję zmiennej losowej typu ciągłego.

Zmienna losowa X jest zmienną losową ciagłą (ciągłą na danym przedziale lub przedziałami ciągłą), jeżeli ma w danym obszarze gęstość prawdopodobieństwa $f(x)$ i $f(x)$ jest funkcją ciągłą (na danym przedziale lub przedziałami ciągłą).

Pomiędzy dystrybuantą F a gęstością f zachodzą zależności:

$$F'(x) = \frac{d}{dx} F(x) = f(x), \quad \text{gdy } F \text{ różniczkowalna},$$

$$F(x) = \int_{-\infty}^{x} f(t)dt = \int_{x_0}^{x} f(t)dt + F(x_0),$$

$$P(a \leq x < b) = \int_a^b f(t)dt = F(b) - F(a),$$

$$\int_{-\infty}^{+\infty} f(t)dt = F(+\infty) - F(-\infty) = 1.$$

Dla zmiennej losowej ciągłej gęstość prawdopodobieństwa, lub krótko gęstość, jest pojęciem analogicznym do pojęcia rozkładu zmiennej losowej typu skokowego. Mówimy więc, że gęstość określa rozkład zmiennej losowej typu ciągłego.

Rozkład normalny

Najpopularniejszym rozkładem zmiennej losowej typu ciągłego jest rozkład normalny, zwany również **rozkładem Gaussa.**

Gęstość rozkładu normalnego jest określona wzorem

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-m)^2}{2\sigma^2}} = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2},$$

gdzie m, σ są ustalonymi parametrami charakteryzującymi ten rozkład. Oznaczamy go symbolem $N(m,\sigma)$. Dystrybuanta $F(x)$ ma postać

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}(\frac{t-m}{\sigma})^2} dt.$$

Dla $m = 0$, $\sigma = 1$ rozkład $N(0,1)$ nazywamy rozkładem normalnym **znormalizowanym** lub rozkładem normalnym **standaryzowanym**. Wtedy

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} = \phi(x)$$

jest funkcją Laplace'a i podobnie

$$F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}t^2} dt = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}t^2} dt = \Phi(x),$$

gdzie $\Phi(x)$ jest inną funkcją Laplace'a.

Istnieją tablice funkcji $\Phi(x)$ i $\phi(x)$ (zamieszczone w Rozdziale 5) podobnie jak funkcji trygonometrycznych. Wspomniana funkcja Φ ma następujące własności:

1° $\Phi(-x) + \Phi(x) = 1$.

2° $\Phi(0) = \frac{1}{2}$.

3° $\Phi(x)$ rosnąca, gdyż $\Phi'(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2} > 0$.

4° $\Phi(+\infty) = \lim_{x\to+\infty} \Phi(x) = 1$.

5° $\frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{1}{2}t^2} dt = \Phi(b) - \Phi(a)$.

Zmienne typu mieszanego

Zmienną losową typu mieszanego najłatwiej określić podając jej dystrybuantę, która jest w postaci

$$F(x) = \alpha F_1(x) + \beta F_2(x),$$

gdzie $F_1(x)$ jest dystrybuantą pewnej zmiennej typu skokowego, $F_2(x)$ dystrybuantą pewnej zmiennej typu ciągłego oraz

$$\alpha + \beta = 1.$$

Przykład 1.6 *Doświadczenie losowe polega na rzucie kostką sześcienną do gry na kwadratową planszę o boku 4. Zmienna losowa X przyjmuje wartości równe:*
- liczbie wyrzuconych oczek, jeżeli liczba ta jest podzielna przez trzy,
- odległości kostki od ustalonego boku tarczy, gdy liczba wyrzuconych oczek na kostce jest niepodzielna przez trzy (kostkę na tarczy traktujemy jak punkt). Podać dystrybuantę zmiennej X.

Rozwiązanie. Dystrybuanta ma postać

$$F(x) = \frac{1}{3}F_1(x) + \frac{2}{3}F_2(x),$$

gdzie

$$F_1(x) = \begin{cases} 0 & \text{dla } x < 3 \\ 1/2 & \text{dla } 3 \le x < 6 \\ 1 & \text{dla } 6 \le x \end{cases}$$

jest dystrybuantą zmiennej skokowej przyjmującej wartości 3 lub 6 z prawdopodobieństwem 1/2 oraz

$$F_2(x) = \begin{cases} 0 & \text{dla } x < 0 \\ x/4 & \text{dla } 0 \le x < 4 \\ 1 & \text{dla } 4 \le x \end{cases}$$

jest dystrybuantą rozkładu ciągłego. Zatem

$$F_2(x) = \begin{cases} 0 & \text{dla } x < 0 \\ x/6 & \text{dla } 0 \le x < 3 \\ (1+x)/6 & \text{dla } 3 \le x < 4 \\ 5/6 & \text{dla } 4 \le x < 6 \\ 1 & \text{dla } 6 \le x \end{cases}.$$

1.5 Niezależność zmiennych losowych

Omówimy to pojęcie wpierw dla zmiennych losowych typu skokowego. W tym celu jednak musimy wprowadzić pojęcie zmiennej losowej dwuwymiarowej.

Definicja 1.11 *Zmienną losową $Z = (X,Y)$ przyjmującą skończoną bądź przeliczalną liczbę wartości (x_i, y_j), każdą z prawdopodobieństwem odpowiednio równym $p_{i,j}$, $i,j \in \mathbb{N}$, przy czym*

$$\sum_i \sum_j p_{i,j} = 1,$$

nazywamy **dwuwymiarową zmienną losową** *typu skokowego. Piszemy*

$$P(X = x_i, Y = y_i) = p_{i,j}.$$

Niech X będzie zmienną losową przyjmującą wartości $x_1, x_2, \ldots$ odpowiednio z prawdopodobieństwami

$$p_i = p_{i,\bullet} = P(X = x_i) = P(X = x_i | Y \text{ dowolne }), \quad i = 1, 2, \ldots$$

i podobnie Y zmienną losową przyjmującą wartości $y_1, y_2, \ldots$ odpowiednio z prawdopodobieństwami

$$p_j = p_{\bullet,j} = P(Y = y_j) = P(Y = y_j | X \text{ dowolne }), \quad j = 1, 2, \ldots$$

Zmienne te nazywamy **rozkładami brzegowymi** zmiennej losowej (X, Y).

Definicja 1.12 *Mówimy, że zmienne X, Y są* **niezależne**, *jeżeli dla wszystkich naturalnych i, j mamy*

$$\begin{aligned} P(X = x_i | Y = y_j) &= P(X = x_i) = p_{i,\bullet}\,, \\ P(Y = y_j | X = x_i) &= P(Y = y_j) = p_{\bullet,j}\,. \end{aligned}$$

Wykorzystując prawdopodobieństwo warunkowe

$$P(X = x_i | Y = y_j) = \frac{P(X = x_i, Y = y_j)}{P(Y = y_j)},$$

widzimy, że niezależność zmiennych X, Y można określić warunkiem: zmienne X, Y są niezależne, jeżeli

$$P(X = x_i, Y = y_j) = P(X = x_i) P(Y = y_j)$$

lub równoważnie, gdy

$$p_{i,j} = p_i p_j$$

dla wszystkich dopuszczalnych i, j.

Rozkład

$$P(X = x_i | Y = y_j), \quad i = 1, 2 \ldots, \quad y_j \text{ ustalone}$$

nazywamy **rozkładem warunkowym** zmiennej X przy warunku $Y = y_j$, podobnie

$$P(Y = y_j | X = x_i), \quad j = 1, 2 \ldots, \quad x_i \text{ ustalone}$$

nazywamy rozkładem warunkowym zmiennej Y przy warunku $X = x_i$.

Zmienną losową dwuwymiarową typu ciągłego nazywamy zmienną o dystrybuancie

$$F(x, y) = P(X < x, Y < y), \quad \text{dla } (x, y) \in \mathbb{R}^2.$$

Zmienne losowe o dystrybuantach

$$\begin{aligned} F_1(x) &= P(X < x) = P(X < x | Y \text{ dowolne}), \\ F_2(y) &= P(Y < y) = P(Y < y | X \text{ dowolne}) \end{aligned}$$

nazywamy zmiennymi losowymi o rozkładach brzegowych.

Niezależność zmiennych losowych typu dowolnego możemy określić następująco.

Mówimy, że zmienne losowe X, Y o dystrybuantach odpowiednio F_1, F_2 są niezależne, jeżeli dystrybuanta F zmiennej dwuwymiarowej (X, Y) spełnia warunek

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

dla wszystkich x, y.

Podobnie zmienne X, Y są niezależne, jeżeli

$$P(x_1 \le X \le x_2, y_1 \le Y \le y_2) = P(x_1 \le X \le x_2) P(y_1 \le Y \le y_2).$$

Dla zmiennych losowych typu ciągłego możemy wprowadzić pojęcie gęstości zmiennych dwuwymiarowych (X, Y)

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

Gdy zmienne są niezależne, wówczas

$$\begin{aligned} f(x, y) &= \lim_{(\Delta x, \Delta y) \to (0,0)} \frac{P(x \le X \le x + \Delta x, y \le Y \le y + \Delta y)}{\Delta x \Delta y} = \\ &= \lim_{\Delta x \to 0} \frac{P(x \le X \le x + \Delta x)}{\Delta x} \cdot \lim_{\Delta y \to 0} \frac{P(y \le Y \le y + \Delta y)}{\Delta y} = \\ &= f_1(x) \cdot f_2(y), \end{aligned}$$

gdzie $F_1'(x) = f_1(x)$, $F_2'(x) = f_2(x)$.

Wtedy mamy

$$\frac{\partial^2 F(x, y)}{\partial x \partial y} = \frac{\partial^2}{\partial x \partial y} (F_1(x) F_2(y)) = F_1'(x) F_2'(y) = f_1(x) f_2(y) = f(x, y).$$

Zatem dla zmiennej dwuwymiarowej (X,Y), gdy X, Y są niezależne, gęstość dwuwymiarowa jest równa iloczynowi gęstości poszczególnych zmiennych (rozkładów brzegowych):

$$f(x,y) = f_1(x)f_2(y).$$

Jeżeli $f(x,y)$ jest gęstością rozkładu zmiennej $Z = (X,Y)$, to funkcje

$$f_1(x) = \int_{-\infty}^{+\infty} f(x,y)\mathrm{d}y$$

oraz

$$f_2(y) = \int_{-\infty}^{+\infty} f(x,y)\mathrm{d}x$$

są gęstością rozkładów brzegowych.

Podobnie można wyznaczyć dystrybuanty rozkładów warunkowych:

$$F(x|y) = P(X < x|Y = y) = \int_{-\infty}^{x} f(u|y)\mathrm{d}u$$

oraz

$$F(y|x) = P(Y < y|X = x) = \int_{-\infty}^{y} f(v|x)\mathrm{d}v,$$

gdzie funkcje

$$f(x|y) = \frac{f(x,y)}{f_2(y)}, \quad f(y|x) = \frac{f(x,y)}{f_1(x)}$$

są gęstościami rozkładów warunkowych.

Pojęcie zmiennej losowej dwuwymiarowej można uogólnić na wyższe wymiary i określić niezależność takich zmiennych.

Dla zmiennych niezależnych (X,Y,Z) o zadanej dystrybuancie

$$F(x,y,z) = P(X < x, Y < y, Z < z)$$

gęstość prawdopodobieństwa $f(x,y,z)$ spełnia warunek

$$f(x,y,z) = \frac{\partial^3 F(x,y,z)}{\partial x \partial y \partial z}.$$

Przykład 1.7 *Niech zmienna (X,Y) ma gęstość prawdopodobieństwa $f(x,y)$ zadaną wzorem*

$$f(x,y) = \begin{cases} \frac{1}{2}\sin(x+y) & \text{dla } 0 \le x \le \frac{\pi}{2}, \quad 0 \le y \le \frac{\pi}{2}, \\ 0 & \text{dla pozostałych } (x,y). \end{cases}$$

Wyznaczyć dystrybuantę zmiennej (X,Y) oraz gęstości rozkładów brzegowych i warunkowych.

Rozwiązanie. Dystrybuantę $F(x,y)$ wyznaczamy ze wzoru

$$F(x,y) = P(X < x, Y < y) = \int_{-\infty}^{x}\left(\int_{-\infty}^{y} f(u,v)du\right)dv.$$

Otrzymujemy kolejno:

$$\begin{aligned} F(x,y) &= \frac{1}{2}\int_0^x\int_0^y \sin(u+v)dudv = \\ &= -\frac{1}{2}\int_0^x (\cos(u+v))\Big|_0^y du = \frac{1}{2}\int_0^x(-\cos(u+y)+\cos(u))du = \\ &= \frac{1}{2}(-\sin(u+y)+\sin(u))\Big|_0^x = \frac{1}{2}(-\sin(x+y)+\sin x+\sin y) \end{aligned}$$

dla $0 \le x \le \pi/2$, $0 \le y \le \pi/2$;

$$F(x,y) = \frac{1}{2}\int_0^x\int_0^{\pi/2}\sin(u+v)dudv = \frac{1}{2}(1+sinx-\cos x)$$

dla $0 \le x \le \pi/2$, $y > \pi/2$;

$$F(x,y) = \frac{1}{2}\int_0^{\pi/2}\int_0^y \sin(u+v)dudv = \frac{1}{2}(1+siny-\cos y)$$

dla $x > \pi/2$, $0 \le y \le \pi/2$;

$$F(x,y) = \frac{1}{2}\int_0^{\pi/2}\int_0^{\pi/2}\sin(u+v)dudv = 1$$

dla $x > \pi/2$, $y > \pi/2$.

Ogólnie dla dowolnych x, y mamy

$$F(x,y) = \begin{cases} 0 & \text{dla } x \le 0 \quad \text{lub} \quad y \le 0 \\ \frac{1}{2}(\sin x+\sin y-\sin(x+y)) & \text{dla } 0 \le x \le \frac{\pi}{2} \text{ i } 0 \le y \le \frac{\pi}{2} \\ \frac{1}{2}(1+\sin x-\cos x) & \text{dla } 0 \le x \le \frac{\pi}{2} \text{ i } \frac{\pi}{2} \le y \\ \frac{1}{2}(1+\sin y-\cos y) & \text{dla } \frac{\pi}{2} \le x \text{ i } 0 \le y \le \frac{\pi}{2} \\ 1 & \text{dla } \frac{\pi}{2} \le x \text{ i } \frac{\pi}{2} \le y \end{cases}.$$

Wyznaczamy funkcję gęstości rozkładów brzegowych:

$$\begin{aligned} f_1(x) &= \int_{-\infty}^{+\infty} f(x,y)\mathrm{d}y = \begin{cases} \int_0^{\pi/2}\frac{1}{2}\sin(x+y)\mathrm{d}y & \text{dla } 0 \le x \le 1 \\ 0 & \text{dla pozostałych } x \end{cases} = \\ &= \begin{cases} \frac{1}{2}(\sin x+\cos x) & \text{dla } 0 \le x \le 1 \\ 0 & \text{dla pozostałych } x \end{cases} \end{aligned}$$

$$f_2(y) = \int_{-\infty}^{+\infty} f(x,y)\mathrm{d}x = \begin{cases} \frac{1}{2}(\sin y + \cos y) & \text{dla } 0 \le y \le 1 \\ 0 & \text{dla pozostałych } y \end{cases}$$

oraz rozkładów warunkowych

$$f(x|y) = \frac{f(x,y)}{f_2(y)} = \begin{cases} \dfrac{\sin(x+y)}{\sin y + \cos y} & \text{dla } 0 \le x \le 1 \\ 0 & \text{dla pozostałych } x \end{cases}$$

$$f(y|x) = \frac{f(x,y)}{f_2(x)} = \begin{cases} \dfrac{\sin(x+y)}{\sin x + \cos x} & \text{dla } 0 \le y \le 1 \\ 0 & \text{dla pozostałych } y \end{cases}$$

1.6 Parametry rozkładu zmiennej losowej

Zmienną losową można scharakteryzować przez podanie pewnych parametrów liczbowych, które będą dawały informacje dotyczące rozkładu tej zmiennej lub gęstości prawdopodobieństwa. Najczęściej rozważanymi parametrami są **wartość średnia** oraz **wariancja** albo **odchylenie standardowe**.

Pierwszy z tych parametrów ma wiele nazw:

wartość średnia,
wartość przeciętna zmiennej losowej,
wartość oczekiwana zmiennej losowej,
nadzieja matematyczna,
wartość średnia zmiennej losowej

i jest oznaczany symbolem $E(X)$.

Definicja 1.13 *Wartością przeciętną (wartością oczekiwaną) zmiennej losowej X typu skokowego o rozkładzie*

$$p_i = P(X = x_i), \quad i = 1, 2, \ldots$$

nazywamy liczbę

$$E(X) := \sum_i p_i x_i = \sum_i x_i P(X = x_i)$$

przy założeniu, że suma $\sum_i p_i x_i$ jest skończona albo szereg nieskończony $\sum_{i=1}^{\infty} p_i|x_i|$ jest zbieżny.

Wartością przeciętną (wartością oczekiwaną) zmiennej losowej X typu ciągłego o gęstości rozkładu $f(x)$ nazywamy liczbę

$$E(X) \stackrel{\text{def}}{=} \int_{-\infty}^{+\infty} x f(x)dx$$

przy założeniu, że całka $\int_{-\infty}^{+\infty} |x|f(x)dx$ jest zbieżna.

Jeżeli $Y = aX + b$, to

$$E(Y) = E(aX+b) = \sum_i p_i(ax_i + b) = a\sum_i p_i x_i + b\sum_i p_i = aE(X) + b.$$

Przykład 1.8 *Obliczyć wartość oczekiwaną $E(X)$ zmiennej losowej X przyjmującej wartości $x_1 = 1$, $x_2 = 2$, $x_3 = 4$ z prawdopodobieństwami $p_1 = \frac{1}{2}$, $p_2 = \frac{1}{4}$, $p_3 = \frac{1}{4}$.*

Rozwiązanie. $E(X) = \sum_{i=1}^{3} p_i x_i = \frac{1}{2}\cdot 1 + \frac{1}{4}\cdot 2 + \frac{1}{4}\cdot 4 = 2.$

Przykład 1.9 *Obliczyć wartość oczekiwaną zmiennej losowej X o gęstości prawdopodobieństwa $f(x)$ zadanej wzorem*

$$f(x) = \begin{cases} 1 - |x-1| & \textit{dla } 0 \le x \le 2 \\ 0 & \textit{dla } x < 0 \textit{ lub } x > 2 \end{cases}.$$

Rozwiązanie.

$$\begin{aligned} E(X) &= \int_{-\infty}^{+\infty} x f(x)dx = \int_0^1 x(1-1+x)dx + \int_1^2 x(1+1-x)dx = \\ &= \int_0^1 x^2 dx + \int_1^2 (2x - x^2)dx = \frac{1}{3}x^3\Big|_0^1 + \left(x^2 - \frac{1}{3}x^3\right)\Big|_1^2 = \\ &= \frac{1}{3} + 4 - \frac{8}{3} - 1 + \frac{1}{3} = 1. \end{aligned}$$

Zatem

$$E(X) = 1.$$

Przykład 1.10 *Obliczyć $E(2X-3)$, gdzie X jest zmienną z* Przykładu 1.8.

Rozwiązanie.

$$E(2X-3) = \frac{1}{2}(2-3) + \frac{1}{4}(2\cdot 2 - 3) + \frac{1}{4}(2\cdot 4 - 3) = -\frac{1}{2} + \frac{1}{4} + \frac{5}{4} = 1$$

lub

$$E(2X-3) = 2E(X) - 3 = 2 \cdot 2 - 3 = 1.$$

Niech dana będzie dowolna zmienna losowa X i niech $g(x)$ będzie przekształceniem ciągłym wzajemnie jednoznacznym, $g'(x) \neq 0$, natomiast $h(x)$ przekształceniem do niego odwrotnym. Rozważmy nową zmienną losową $Y = g(X)$. Gęstość rozkładu tej zmiennej wyraża się wzorem

$$f_Y(y) = f_X(h(y))|h'(y)|,$$

gdzie f_X jest gęstością rozkładu zmiennej X.

W szczególnym przypadku, gdy $g(x) = ax + b$, piszemy

$$Y = aX + b$$

i rozkład f_Y ma postać

$$f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right).$$

Wartość oczekiwana zmiennej $Y = aX + b$ ma postać

$$E(Y) = aE(X) + b,$$

natomiast w ogólnym przypadku

$$E(Y) = E(g(X)) = \int_{-\infty}^{+\infty} f(x)g(x)dx = \int_{-\infty}^{+\infty} yf(h(y))h'(y)dy.$$

Dla zmiennej losowej X typu skokowego zmienna $Y = g(X)$ jest też typu skokowego o wartościach $y_i = g(x_i)$ i rozkładzie

$$g_i = P(Y = y_i) = p_i \ .$$

Wtedy

$$E(Y) = \sum_i p_i g(x_i) = \sum_i P(X = x_i) \cdot g(x_i).$$

Momentem rzędu k zmiennej losowej X nazywamy wartość oczekiwaną zmiennej $Y = X^k$, to znaczy moment ten obliczamy według wzoru

$$m_k = E\left(X^k\right) = \sum_i p_i x_i^k$$

lub

$$m_k = E\left(X^k\right) = \int_{-\infty}^{+\infty} x^k f(x)dx$$

w zależności od tego, czy zmienna losowa jest typu skokowego czy ciągłego.

Na szczególną uwagę zasługuje moment rzędu drugiego dla zmiennej $Y = X - E(X)$; nazywamy go **wariancją** i oznaczamy symbolem $V(X)$, to znaczy

$$V(X) := E\left([X - E(X)]^2\right).$$

Łatwo policzyć, że wariancja $V(X)$ może być zapisana w postaci

$$V(X) = E(X^2) - [E(X)]^2.$$

Istotnie:

$$\begin{aligned} E\left([X - E(X)]^2\right) &= E\left(X^2 - 2X \cdot E(X) + [E(X)]^2\right) = \\ &= E(X^2) - 2E(X) \cdot E(X) + [E(X)]^2 = E(X^2) - [E(X)]^2. \end{aligned}$$

Przykład 1.11 *Obliczyć wariancję* $V(X)$ *z* Przykładów 1.8 i 1.9.

Rozwiązanie. 1. Dla zmiennej X z Przykładu 1.8 mamy

$$E(X^2) = \frac{1}{2} \cdot 1^2 + \frac{1}{4} \cdot 2^2 + \frac{1}{4} \cdot 4^2 = \frac{1}{2} + 1 + 4 = 5\frac{1}{2} = \frac{11}{2},$$

$$V(X) = E(X^2) - [E(X)]^2 = \frac{11}{2} - 2^2 = \frac{3}{2}.$$

Drugim sposobem (bezpośrednio z definicji):

$$V(X) = E\left([X - E(X)]^2\right) = \frac{1}{2}(1-2)^2 + \frac{1}{4}(2-2)^2 + \frac{1}{4}(4-2)^2 = \frac{1}{2} + 1 = \frac{3}{2}.$$

2. Zmienna X z Przykładu 1.9. Korzystamy z przytoczonego wcześniej wzoru

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

oraz z tego, że $E(X) = 1$, i otrzymujemy:

$$\begin{aligned} V(X) &= E[(X - E(X))^2] = \int_0^2 (x-1)^2(1 - |x-1|)dx = \\ &= \int_0^1 (x-1)^2 x dx + \int_1^2 (x-1)^2(2-x)dx = \\ &= \int_0^1 (x^3 - 2x^2 + x)dx + \int_1^2 (2 - 5x + 4x^2 - x^3)dx = \\ &= \left(\frac{1}{4}x^4 - \frac{2}{3}x^3 + \frac{1}{2}x^2\right)\Big|_0^1 + \left(2x - \frac{5}{2}x^2 + \frac{4}{3}x^3 - \frac{1}{4}x^4\right)\Big|_1^2 = \frac{1}{6}. \end{aligned}$$

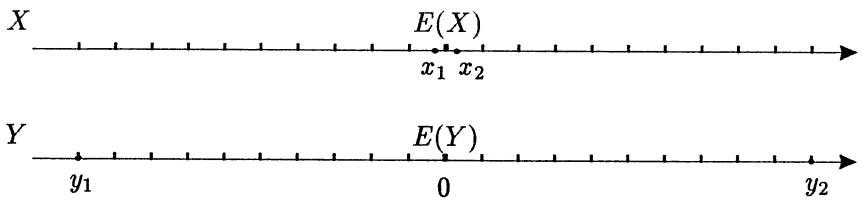
Wariancja $V(X)$ jest miarą rozproszenia wartości zmiennej losowej X od jej wartości średniej $E(X)$. Mała wariancja informuje, że wszystkie wartości zmiennej losowej o "dużym prawdopodobieństwie" leżą blisko wartości średniej $E(X)$. W celu ilustracji tego faktu rozważmy dwie zmienne losowe X, Y określone następująco:

$$P\left(X=\frac{1}{10}\right)=\frac{1}{2}, \qquad P\left(X=-\frac{1}{10}\right)=\frac{1}{2},$$
$$P(Y=10)=\frac{1}{2}, \qquad P(Y=-10)=\frac{1}{2}.$$

Wtedy

$$\begin{aligned} E(X) &= E(Y)=0, \\ V(X) &= E(X^2)-[E(X)]^2=\frac{1}{2}\frac{1}{100}+\frac{1}{2}\frac{1}{100}=\frac{1}{100}, \\ V(Y) &= E(Y^2)-[E(Y)]^2=\frac{1}{2}100+\frac{1}{2}100=100. \end{aligned}$$

W obu przypadkach wartość średnia jest równa zeru, a odchylenia standardowe – wariancje – są bardzo różne. Rozproszenie zmiennej Y jest wielokrotnie większe od rozproszenia zmiennej X, co można zaobserwować na wykresach przedstawionych na Rys. 1.2.



Rys. 1.2. Zmienne losowe X i Y

Wariancja jest momentem rzędu drugiego zmiennej $Y = X - E(X)$. W ogólnym przypadku wyrażenia

$$\mu_k = E\left([X-E(X)]^k\right)$$

nazywany **momentami centralnymi rzędu** k.

Przypomnijmy jeszcze raz wzory na wartość przeciętną i wariancję zmiennej losowej X:

$$E(X)=\sum_k x_k p_k,$$

gdy zmienna X jest typu skokowego, oraz:

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

$$V(X)=\int_{-\infty}^{+\infty}\left(x-\int_{-\infty}^{+\infty} t f(t)dt\right)^2 f(x)dx$$

dla zmiennej typu ciagłego.

Przedstawimy teraz pewne własności wartości przeciętnej:

1° $E(C)=c$, gdzie $X=C$ oznacza zmienną losową przyjmującą wartość c z prawdopodobieństwem 1,

2° $E(c\cdot X)=c\cdot E(X)$,

3° $E(X+Y)=E(X)+E(Y)$,

4° $E(X\cdot Y)=E(X)\cdot E(Y)$,

} gdy X, Y – zmienne niezależne.

We wzorach 3° i 4° przez sumę $X+Y$ oraz iloczyn $X\cdot Y$ rozumiemy zmienne przyjmujące wszystkie wartości będące, odpowiednio, sumami lub iloczynami wartości przyjmowanych przez poszczególne zmienne z prawdopodobieństwem (gęstością prawdopodobieństwa) będącym iloczynem prawdopodobieństw lub sumą iloczynów. Przedstawmy to na prostym przykładzie:

$$X : P(X=1)=\tfrac{1}{2}, \quad P(X=2)=\tfrac{1}{4}, \quad P(X=3)=\tfrac{1}{4},$$
$$Y : P(Y=0)=\tfrac{1}{2}, \quad P(Y=10)=\tfrac{1}{2},$$

wtedy mamy

$$Z=X+Y : P(Z=1)=\frac{1}{4}, \quad P(Z=2)=\frac{1}{8}, \quad P(Z=3)=\frac{1}{8},$$
$$P(Z=11)=\frac{1}{4}, \quad P(Z=12)=\frac{1}{8}, \quad P(Z=13)=\frac{1}{8}.$$

Stąd

$$\begin{aligned} E(X) &= \frac{1}{2}+\frac{1}{2}+\frac{3}{4}=\frac{7}{4}, \\ E(Y) &= 0+5=5, \\ E(Z) &= \frac{1}{4}+\frac{1}{4}+\frac{3}{8}+\frac{11}{4}+\frac{12}{8}+\frac{13}{8}=\frac{27}{4}=\frac{7}{4}+5. \end{aligned}$$

Dla iloczynu

$$U=X\cdot Y$$

mamy

$$P(U=0)=\frac{1}{4}+\frac{1}{8}+\frac{1}{8}=\frac{1}{2}, \qquad P(U=10)=\frac{1}{4},$$
$$P(U=20)=\frac{1}{8}, \qquad P(U=30)=\frac{1}{8},$$

$$E(U)=0\frac{1}{2}+10\frac{1}{4}+20\frac{1}{8}+30\frac{1}{8}=\frac{10+10+15}{4}=\frac{35}{4}=\frac{7}{4}\cdot 5.$$

Własności wariancji:

1° $V(C)=0$,

2° $V(c\cdot X)=c^2V(X)$,

$$\left.\begin{array}{lll} 3^\circ & V(X+Y) & = V(X)+V(Y), \\ 4^\circ & V(X-Y) & = V(X)+V(Y), \end{array}\right\} \text{ gdy } X, Y \text{ – zmienne niezależne.}$$

Twierdzenie 1.10 *Dla każdej liczby rzeczywistej $c \neq E(X)$ ma miejsce nierówność*

$$E\left([X-c]^2\right) > V(X) = E\left([X-E(X)]^2\right).$$

Nierówność ta mówi, że wartość oczekiwana zmiennej $[X-c]^2$ osiąga minimum właściwe, gdy c jest wartością przeciętną danej zmiennej.

Dowód.

$$\begin{aligned} E\left([X-c]^2\right) &= E\left(X^2-2cX+c^2\right) = E(X^2)-2cE(X)+c^2 = \\ &= E(X^2)-[E(X)]^2+\left\{c^2-2cE(X)+[E(X)]^2\right\} = \\ &= V(X)-[c-E(X)]^2 > V(X), \end{aligned}$$

gdy $c \neq E(X)$.

1.7 Przykłady najczęściej stosowanych rozkładów i ich parametry

Przedstawimy teraz pewne szczególne rozkłady zmiennej losowej oraz wyliczymy parametry $E(X)$ i $V(X)$ rozkładu tych zmiennych

I. Rozkład zero-jedynkowy

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

Wtedy

$$\begin{aligned} E(X) &= 0\cdot q+1\cdot p=p, \\ E(X^2) &= 0\cdot q+1\cdot p=p, \\ V(X) &= E(X^2)-[E(X)]^2=p-p^2=p(1-p)=p\cdot q. \end{aligned}$$

II. Rozkład równomierny skokowy

Zmienna losowa przyjmuje skończoną liczbę wartości, $n<\infty$, z jednakowym prawdopodobieństwem:

$$P(X=x_i)=\frac{1}{n}, \quad i=1,2\ldots,n.$$

Zatem

$$\begin{aligned} E(X) &= \frac{1}{n}\sum_1^n x_i, \\ V(X) &= \frac{1}{n}\sum_i^n (x_i-E(X))^2. \end{aligned}$$

III. Rozkład równomierny ciągły (rozkład jednostajny)

Zmienna losowa przyjmuje wartości z przedziału (a,b), z jednakowym prawdopodobieństwem równym $1/(b-a)$. Funkcja gęstości ma postać

$$f(x)=\begin{cases} \frac{1}{b-a} & \text{dla } a<x<b \\ 0 & \text{dla pozostałych } x \end{cases}.$$

Zatem

$$\begin{aligned} E(X) &= \frac{a+b}{2}, \\ V(X) &= \frac{(b-a)^2}{12}. \end{aligned}$$

IV. Rozkład liniowy

Funkcja gęstości jest fragmentem funkcji liniowej

$$f(x)=\begin{cases} \frac{2(x-a)}{(b-a)^2} & \text{dla } a<x<b \\ 0 & \text{dla pozostałych } x \end{cases}.$$

V. Rozkład normalny

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2}.$$

Wtedy

$$\begin{aligned} E(X) &= \int_{-\infty}^{+\infty} \frac{x}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2} dx = \left| \begin{array}{l} \frac{x-m}{\sigma} = u \\ \frac{dx}{\sigma} = du \end{array} \right| = \\ &= \int_{-\infty}^{+\infty} \frac{m+\sigma u}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} du = \\ &= m\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{1}{2}u^2} du + \frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} u e^{-\frac{1}{2}u^2} du = m, \end{aligned}$$

gdyż druga całka jest równa zeru jako całka z funkcji nieparzystej, a całka pierwsza jest równa $\sqrt{2\pi}$.

Zatem wartość oczekiwana zmiennej losowej o rozkładzie normalnym jest równa parametrowi m:

$$E(X) = m.$$

Wówczas

$$\begin{aligned} V(X) &= E([x - E(X)]^2) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} (x-m)^2 e^{-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2} dx = \\ &= \left| \begin{array}{l} \frac{x-m}{\sigma} = u \\ \frac{dx}{\sigma} = du \end{array} \right| = \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} u^2 e^{-\frac{1}{2}u^2} du = \\ &= \left| \begin{array}{ll} f = u & f' = 1 \\ g' = ue^{-\frac{1}{2}u^2} & g = -e^{-\frac{1}{2}u^2} \end{array} \right| = \\ &= \frac{-\sigma^2}{\sqrt{2\pi}} u e^{-\frac{1}{2}u^2} \Big|_{-\infty}^{+\infty} + \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{1}{2}u^2} du = \\ &= \sigma^2. \end{aligned}$$

Zatem wariancja $V(X) = \sigma^2$ jest kwadratem drugiego parametru zmiennej losowej o rozkładzie normalnym.

Liczbę $\sigma = \sqrt{V(X)}$ nazywa się **odchyleniem standardowym**.

Reasumując, jeżeli X ma rozkład normalny $N(m, \sigma)$, to

$$E(X) = m, \qquad V(X) = \sigma^2.$$

Jeżeli X ma rozkład normalny, to

$$Y = aX + b$$

ma też rozkład normalny $N(am + b, |a|\sigma)$

$$\begin{aligned} E(Y) &= aE(X) + b = am + b, \\ V(Y) &= a^2 V(X) = a^2\sigma^2 = (|a|\sigma)^2. \end{aligned}$$

Zmienna

$$Z = \frac{X - m}{\sigma} = \frac{1}{\sigma}X - \frac{m}{\sigma}$$

ma rozkład $N(0,1)$ nazywany **rozkładem normalnym standaryzowanym**. Wtedy

$$m = E(Z) = 0, \quad \sigma = \sqrt{V(Z)} = 1.$$

Dla zmiennej $N(0,\sigma)$ możemy policzyć momenty (centralne) dowolnego rzędu:

$$\mu_k(X) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} x^k e^{-\frac{1}{2}\frac{x^2}{\sigma^2}} dx = \begin{cases} 0, & \text{gdy } k = 2l-1 \\ 1 \cdot 3 \ldots (2l-1)\sigma^2, & \text{gdy } k = 2l. \end{cases}$$

Pierwsza część wzoru wynika z tego, że funkcja podcałkowa jest nieparzysta. Druga część wynika ze wzoru rekurencyjnego otrzymanego z całkowania przez części:

$$E(X^{2l}) = (2l-1)\sigma^2 E\left(X^{2(l-1)}\right).$$

Rozkład normalny możemy przedstawić graficznie, Rys. 1.3.

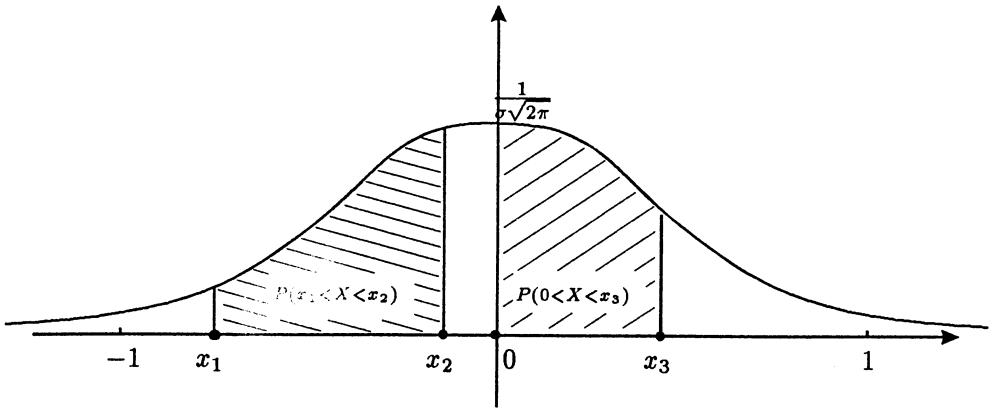
VI. Rozkład Bernoulliego

Niech X będzie zmienną w schemacie Bernoulliego

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

Wtedy

$$\begin{aligned} E(X) &= \sum_{k=0}^{n} k\binom{n}{k} p^k q^{n-k} = np \sum_{k=1}^{n} \binom{n-1}{k-1} p^{k-1} q^{n-k} = \\ &= np \sum_{l=0}^{n-1} \binom{n-1}{l} p^l q^{n-1-l} = n \cdot p. \end{aligned}$$



Rys. 1.3. Funkcja gęstości rozkładu normalnego

$$\begin{aligned} E(X^2) &= \sum_{k=0}^{n} k^2 \binom{n}{k} p^k q^{n-k} = np \sum_{k=1}^{n-1} \binom{n-1}{k-1} p^k q^{n-k} = \\ &= np\left(\sum_{l=0}^{n-1} l \binom{n-1}{l} p^l q^{n-1-l} + \sum_{l=0}^{n-1} \binom{n-1}{l} p^l q^{n-1-l}\right) = \\ &= np\,((n-1)p+1) = np(np+q). \end{aligned}$$

Zatem

$$V(X) = E(X^2) - E^2(X) = np(np+q) - (np)^2 = npq = np(1-p).$$

VII. Rozkład Poissona

Niech X będzie zmienną losową o rozkładzie Poissona

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

Wtedy

$$E(X) = \sum_{k=0}^{\infty} k \frac{\lambda^k}{k!} e^{-\lambda} = \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} = \lambda,$$

$$E(X^2) = \sum_{k=0}^{\infty} k^2 \frac{\lambda^k}{k!} e^{-\lambda} = \lambda e^{-\lambda} \sum_{k=1}^{\infty} (k-1+1) \frac{\lambda^{k-1}}{(k-1)!} =$$

$$\begin{aligned} &= \lambda e^{-\lambda} \left[\sum_{k=1}^{\infty} (k-1) \frac{\lambda^{k-1}}{(k-1)!} + \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!}\right] = \\ &= \lambda e^{-\lambda} \left[\lambda e^{-\lambda} + e^{-\lambda}\right] = \lambda^2 + \lambda, \end{aligned}$$

$$V(X) = E(X^2) - [E(X)]^2 = \lambda^2 + \lambda - \lambda^2 = \lambda.$$

Ostatecznie

$$E(X) = \lambda, \quad V(X) = \lambda, \quad \lambda > 0.$$

VIII. Rozkład wykładniczy

Jeżeli funkcja gęstości wyraża się wzorem

$$f(x) = \begin{cases} \frac{1}{\lambda} e^{-x/\lambda} & \text{dla } x \geq 0 \\ 0 & \text{dla } x < 0 \end{cases},$$

to mówimy, że zmienna X ma rozkład wykładniczy. Dystrybuanta ma wtedy postać

$$F(x) = \begin{cases} 1 - e^{-x/\lambda} & \text{dla } x \geq 0 \\ 0 & \text{dla } x < 0 \end{cases}.$$

Ponadto

$$E(X) = \lambda, \qquad V(X) = \lambda^2.$$

IX. Rozkład χ^2 (chi kwadrat)

Można dowieść, że jeżeli zmienne losowe $X_1, X_2, \ldots, X_k$ są niezależne i mają rozkłady normalne standaryzowane $N(0,1)$, to zmienna

$$\chi^2 := \sum_{i=1}^{k} X_i^2$$

ma rozkład o gęstości prawdopodobieństwa danej wzorem

$$f(x) = \begin{cases} 0 & \text{dla } x \leq 0 \\ \dfrac{x^{\frac{1}{2}k-1} e^{-\frac{1}{2}x}}{\Gamma\left(\frac{k}{2}\right) 2^{\frac{k}{2}}} & \text{dla } x > 0 \end{cases}.$$

Rozkład ten nazywamy rozkładem chi kwadrat i oznaczamy symbolem χ^2 (χ – grecka litera chi).

Występująca tu funkcja $\Gamma(x)$ nazywa się funkcją **gamma Eulera** i jest określona wzorem

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

Funkcja $\Gamma(x)$ jest uogólnieniem funkcji "silnia" i zachodzą wzory

$$\Gamma(n+1) = n!, \quad n \in \mathbb{N}; \quad \Gamma(x+1) = x\Gamma(x), \quad x > 0.$$

Liczbę k nazywamy **liczbą stopni swobody** rozkładu χ^2.

Załóżmy, że populacja generalna X ma rozkład normalny $N(m,\sigma)$, gdzie m, σ jest skończone, ale nieznane. Wybieramy losowo próbę o liczności n oraz każdą i-tą obserwację traktujemy jako niezależną zmienną X_i o wartości odpowiednio x_i. Definiujemy nową zmienną losową $\overline{X}$ (średnią z próby)

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Wtedy zmienna losowa

$$U = n\frac{S^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n^2\sigma^2}$$

ma rozkład χ^2 o liczbie stopni swobody $k = n-1$.

Rozkład ten jest wykorzystywany do weryfikacji hipotez statystycznych. Funkcja gęstości rozkładu prawdopodobieństwa i dystrybuanta rozkładu χ^2 są stablicowane i ich wartości można odczytać z tablic statystycznych.

X. Rozkład t Studenta

Można dowieść, że jeżeli zmienne $X_0, X_1, X_2, \ldots, X_n$ są niezależne i mają rozkład normalny $N(0,1)$, to zmienna

$$t = \frac{\sqrt{n}X_0}{\sqrt{\sum_{i=1}^{n} X_i^2}} = \frac{X_0}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} X_i^2}}$$

ma rozkład o gęstości prawdopodobieństwa zadanej wzorem

$$f(t) = \frac{1}{\sqrt{n}B\left(\frac{1}{2},\frac{n}{2}\right)}\left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}},$$

gdzie $B(x,y)$ jest znaną **funkcją Beta**, taką że

$$B\left(\frac{1}{2},\frac{n}{2}\right) = \frac{1}{\sqrt{n}}\int_{-\infty}^{+\infty}\left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}}dt.$$

Rozkład ten nosi nazwę rozkładu t Studenta i jest wykorzystywany do weryfikowania hipotez statystycznych.

Uwaga. Przy $n \to \infty$ rozkład t Studenta dąży do rozkładu normalnego standaryzowanego $N(0,1)$, gdyż

$$\lim_{n\to\infty}\left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}} = \lim_{n\to\infty}\left[\left(1+\frac{t^2}{n}\right)^{\frac{n}{t^2}}\right]^{-\frac{n+1}{n}t^2} = e^{-\frac{1}{2}t^2},$$

$$\lim_{n\to\infty}\left(\sqrt{n}B\left(\frac{1}{2},\frac{n}{2}\right)\right) = \lim_{n\to\infty}\int_{-\infty}^{+\infty}\left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}} = \int_{-\infty}^{+\infty} e^{-\frac{1}{2}t^2}dt = \sqrt{2\pi}.$$

1.8 Momenty dwuwymiarowej zmiennej losowej, współczynnik korelacji

Niech dana będzie zmienna losowa (X,Y) o rozkładzie $p_{i,j}$ w przypadku zmiennej skokowej lub gęstości $f(x,y)$ w przypadku zmiennej ciągłej. Jeżeli $h = h(X,Y)$ jest funkcją zmiennych X i Y (np. $X+Y$, $X\cdot Y$) prowadzącą do nowej zmiennej, to wartość oczekiwana tej zmiennej może być wyliczona ze wzorów

$$E(h(X,Y)) = \sum_{i,j} h(x_i,y_j)p_{i,j},$$

gdzie sumowanie rozciąga się na wszystkie dopuszczalne wartości i,j;

$$E(h(X,Y)) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} h(x,y)f(x,y)dxdy.$$

Dla funkcji $h(x,y) = X^kY^l$ wartości oczekiwane zmiennej $h(X,Y)$ nazywamy **momentami rzędu** $k+l$, oznaczamy symbolami

$$m_{k,l} = E\left(X^kY^l\right)$$

i obliczamy odpowiednio ze wzorów

$$m_{k,l} = \sum_{i,j} x_i^k y_j^l p_{i,j} \qquad \text{dla } (X,Y) \text{ typu skokowego,}$$

$$m_{k,l} = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} x^k y^l f(x,y)dxdy \quad \text{dla } (X,Y) \text{ typu ciągłego.}$$

Rozważa się też **momenty centralne**

$$\mu_{k,l} = E\left([X - m_{1,0}]^k \cdot [Y - m_{0,1}]^l\right).$$

Dla momentów centralnych mamy:

$$\begin{aligned}
\mu_{1,0} &= E(X - E(X)),\\
\mu_{0,1} &= E(Y - E(Y)),\\
\mu_{2,0} &= E\left([X - E(X)]^2\right) = V(X) = \sigma_1^2,\\
\mu_{0,2} &= E\left([Y - E(Y)]^2\right) = V(Y) = \sigma_2^2.
\end{aligned}$$

W tych wzorach σ_1, σ_2 są **odchyleniami standardowymi** zmiennych X i Y.

Moment centralny mieszany rzędu drugiego

$$\mu_{1,1} = E([X - E(X)][Y - E(Y)]) = E([X - m_{1,0}][Y - m_{0,1}])$$

nazywamy **kowariancją** zmiennej losowej dwuwymiarowej (X,Y).

Dla momentów centralnych mamy:

$$\begin{aligned}
\mu_{2,0} &= m_{2,0} - (m_{1,0})^2,\\
\mu_{0,2} &= m_{0,2} - (m_{0,1})^2,\\
\mu_{1,1} &= m_{1,1} - m_{1,0} \cdot m_{0,1}.
\end{aligned}$$

Kowariancja $\mu_{1,1}$ jest pewnym miernikiem niezależności zmiennych losowych. Mówi o tym następujące twierdzenie:

Twierdzenie 1.11 *Jeżeli zmienne losowe X i Y są niezależne, to ich kowariancja $\mu_{1,1}$ jest równa zeru.*

Dowód.

$$\mu_{1,1} = E([X - m_{1,0}][Y - m_{0,1}]) = E([X - m_{1,0}]) \cdot E([Y - m_{0,1}]) = 0.$$

Uwaga. Twierdzenie odwrotne nie jest prawdziwe. Zerowanie się kowariancji nie zawsze pociąga niezależność zmiennych.

Jeżeli odchylenia standardowe σ_1, σ_2 są różne od zera, to definiujemy **współczynnik korelacji** $\rho = \rho(X,Y)$ zmiennych X i Y w następujacy sposób:

$$\begin{aligned}
\rho &= \frac{\mu_{1,1}}{\sigma_1\sigma_2} = \frac{\mu_{1,1}}{\sqrt{\mu_{2,0}\mu_{0,2}}} = \frac{E([X - E(X)][Y - E(Y)])}{\sqrt{V(X)V(Y)}} =\\
&= \frac{E([X - m_{1,0}][Y - m_{0,1}])}{\sqrt{E([X - m_{1,0}]^2)E([Y - m_{0,1}]^2)}}.
\end{aligned}$$

Twierdzenie 1.12 *Dla dowolnych zmiennych losowych X i Y współczynnik korelacji ρ spełnia nierówność*

$$-1 \le \rho \le 1.$$

Dowód. Rozważmy funkcję zmiennej rzeczywistej t:

$$\begin{aligned}
g(t) &= E\left\{([X - m_{1,0}]t + [Y - m_{0,1}])^2\right\} =\\
&= t^2 E\left([X - m_{1,0}]^2\right) + 2tE([X - m_{1,0}][Y - m_{0,1}]) +\\
&\quad + E\left([Y - m_{0,1}]^2\right) = t^2\sigma_1^2 + 2t\mu_{1,1} + \sigma_2^2.
\end{aligned}$$

Funkcja $g(t)$ jako wartość oczekiwana kwadratu pewnej zmiennej nie może przyjmować wartości ujemnych. Jak widzimy, funkcja

$$g(t) = \sigma_1^2 \cdot t^2 + 2\mu_{1,1}t + \sigma_2^2$$

jest trójmianem kwadratowym względem t i dlatego wyróżnik Δ tego trójmianu musi być ujemny

$$\frac{1}{4}\Delta = \mu_{1,1}^2 - \sigma_1^2\sigma_2^2 \le 0$$

lub równoważnie

$$\left(\frac{\mu_{1,1}}{\sigma_1 \cdot \sigma_2}\right)^2 \le 1,$$

co daje tezę twierdzenia, $|\rho| \le 1$.

Definicja 1.14 *Zmienne losowe X, Y nazywamy zmiennymi* **nieskorelowanymi**, *jeżeli ich współczynnik korelacji jest równy zeru, to znaczy $\rho = 0$.*

Wiadomo, że kowariancja zmiennych niezależnych jest równa zeru, co daje rezultat

$$X, Y \text{ niezależne} \implies X, Y \text{ nieskorelowane.}$$

Zerowanie się współczynnika korelacji nie pociąga niezależności zmiennych, ale im większy – bliższy jedności – jest jego moduł, tym zmienne są "bardziej zależne", a przy $\rho = 1$ zależność jest liniowa.

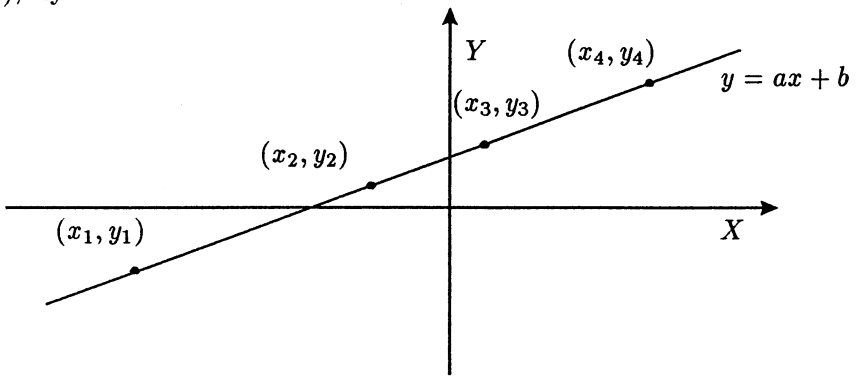
Twierdzenie 1.13 *Warunkiem koniecznym i wystarczającym na to, by zmienne X i Y spełniały warunek*

$$P(Y = aX + b) = 1,$$

jest, żeby ich współczynnik korelacji spełniał równość $|\rho| = 1$.

Oznacza to, że prawdopodobieństwo, iż zmienna dwuwymiarowa (X,Y) przybiera wartości leżące na pewnej prostej $y = ax + b$, jest równe jedności.

Przypuśćmy, że (X,Y) jest zmienną typu skokowego o skończonej liczbie wartości (x_i, y_j). Wtedy dla $|\rho| = 1$ ($\rho = -1$ lub $\rho = 1$) wszystkie wartości (x_i, y_j) leżą na jednej prostej (przy założeniu, że $p_{i,j} \neq 0$ dla dopuszczalnych i, j), Rys. 1.4.



Rys. 1.4 Rozkład zmiennej (X,Y) dla $|\rho| = 1$

1.9 Funkcja charakterystyczna

Definicja 1.15 Funkcją charakterystyczną *zmiennej losowej X nazywamy funkcję zespoloną zmiennej rzeczywistej t określoną wzorem*

$$\varphi(t) = E\left(e^{itX}\right).$$

Dla zmiennej losowej typu skokowego

$$\varphi(t) = \sum_k e^{itx_k} \cdot p_k = \sum_k [\cos(tx_k) + i\sin(tx_k)]\, p_k,$$

dla zmiennej losowej typu ciągłego

$$\varphi(t) = \int_{-\infty}^{+\infty} e^{itx} f(x)dx = \int_{-\infty}^{+\infty} \cos(tx) f(x)dx + i\int_{-\infty}^{+\infty} \sin(tx) f(x)dx.$$

Przykład 1.12 *Wyznaczyć funkcję charakterystyczną $\varphi(t)$ dla zmiennej losowej X o rozkładzie Poissona*

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

Rozwiązanie.

$$\varphi(t) = \sum_{k=1}^{+\infty} e^{itk} e^{-\lambda}\frac{\lambda^k}{k!} = e^{-\lambda}\sum_{k=1}^{+\infty}\frac{(\lambda e^{it})^k}{k!} = e^{-\lambda}e^{\lambda e^{it}} = e^{\lambda(e^{it}-1)}.$$

Przykład 1.13 *Wyznaczyć funkcję charakterystyczną $\varphi(t)$ dla zmiennej losowej X o rozkładzie Bernoulliego (dwumianowym).*

$$P(X = k) = \binom{n}{k} p^k q^{n-k}, \quad k = 0, 1, \ldots, n, \quad q = 1 - p, \; p \in (0,1).$$

Rozwiązanie.

$$\varphi(t) = \sum_{k=0}^{n}\binom{n}{k} p^k q^{n-k} e^{itk} = \sum_{k=0}^{n}\binom{n}{k}(pe^{it})^k q^{n-k} = (pe^{it} + q)^n.$$

Przykład 1.14 *Wyznaczyć funkcję charakterystyczną $\varphi(t)$ dla zmiennej losowej X o rozkładzie normalnym standaryzowanym.*

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}.$$

Rozwiązanie.

$$\begin{aligned}\varphi(t) &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} e^{itx} e^{-\frac{1}{2}x^2} dx = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} e^{-\frac{1}{2}(x^2 - 2itx + (it)^2) - \frac{1}{2}t^2} dt = \\ &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} e^{-\frac{1}{2}(x-it)^2} dx \cdot e^{-\frac{1}{2}t^2} = e^{-\frac{1}{2}t^2}.\end{aligned}$$

Funkcja charakterystyczna ma następujące własności:

1° $\varphi(0) = 1$. Istotnie

$$\varphi(0) = E(e^0) = E(1) = 1.$$

2° $|\varphi(t)| \leq 1$. Istotnie

$$|\varphi(t)| = \left|\int_{-\infty}^{+\infty} e^{itx} f(x)dx\right| \leq \int_{-\infty}^{+\infty} \left|e^{itx}\right| f(x)dx \leq \int_{-\infty}^{+\infty} f(x)dx = 1.$$

3° $\varphi(-t) = \overline{\varphi(t)}$. Wynika to z równości $e^{-itx} = \overline{(e^{itx})}$.

4° Jeżeli $\varphi(t)$ jest funkcją charakterystyczną zmiennej losowej X oraz zmienna $Y = aX + b$, to funkcja charakterystyczna $\varphi_1(t)$ zmiennej Y ma postać

$$\varphi_1(t) = e^{itb}\varphi(at).$$

Rzeczywiście

$$\varphi_1(t) = E\left(e^{itY}\right) = E\left(e^{it(aX+b)}\right) = E\left(e^{i(ta)X} \cdot e^{itb}\right) = e^{itb}\varphi(at).$$

Zatem, gdy zmienna Y ma rozkład normalny $N(m, \sigma)$, wówczas jej funkcja charakterystyczna ma postać

$$\varphi(t) = e^{itm} \cdot e^{-\frac{1}{2}\sigma^2 t^2} = e^{itm - \frac{1}{2}\sigma^2 t^2}.$$

5° Funkcja charakterystyczna sumy dwu zmiennych niezależnych X, Y jest równa iloczynowi funkcji charakterystycznych poszczególnych zmiennych

$$\varphi_{X+Y}(t) = \varphi_X(t) \cdot \varphi_Y(t).$$

Istotnie

$$\begin{aligned} \varphi_{X+Y}(t) &= E\left(e^{it(X+Y)}\right) = E\left(e^{itX}e^{itY}\right) = \\ &= E\left(e^{itX}\right) E\left(e^{itY}\right) = \varphi_X(t) \cdot \varphi_Y(t). \end{aligned}$$

Zastosowania funkcji charakterystycznych

Jeżeli znamy funkcję charakterystyczną $\varphi(t)$ zmiennej losowej X, to możemy między innymi obliczyć wartość oczekiwaną oraz wariancję zmiennej X.

Zauważmy, że dla zmiennej typu ciągłego mamy

$$\begin{aligned} \varphi(t) &= \int_{-\infty}^{+\infty} e^{itx} f(x)dx, \\ \varphi'(t) &= i\int_{-\infty}^{+\infty} xe^{itx} f(x)dx, \\ \varphi''(t) &= -\int_{-\infty}^{+\infty} x^2 e^{itx} f(x)dx. \end{aligned}$$

Zatem

$$\varphi'(0) = iE(X), \quad \varphi''(0) = -E(X^2).$$

Stąd mamy

$$\begin{aligned} E(X) &= \frac{1}{i}\varphi'(0) = -i\varphi'(0), \\ V(X) &= [\varphi'(0)]^2 - \varphi''(0). \end{aligned}$$

Podobnie mamy dla funkcji typu skokowego. Dla przykładu, gdy X ma rozkład dwumienny – Bernoulliego

$$P(X = k) = \binom{n}{k} p^k q^{n-k},$$

wówczas

$$\varphi(t) = \left(pe^{it} + q\right)^n.$$

Stąd

$$\begin{aligned} \varphi'(t) &= n\left(pe^{it} + q\right)^{n-1} pe^{it} \cdot i, \\ \varphi'(0) &= n \cdot p \cdot i, \\ \varphi''(t) &= n \cdot p \cdot i\left\{(n-1)\left(pe^{it} + q\right)^{n-2} pie^{2it} + \left(pe^{it} + q\right)^{n-1} ie^{it}\right\}, \\ \varphi''(0) &= -np(np + q). \end{aligned}$$

Zatem

$$\begin{aligned} E(X) &= \frac{1}{i}\varphi'(0) = n \cdot p, \\ V(X) &= [\varphi'(0)]^2 - \varphi''(0) = -n^2p^2 + np(np + q) = n \cdot p \cdot q. \end{aligned}$$

Uwaga. Z definicji funkcji charakterystycznej $\varphi(t)$ widać, że jest ona wyznaczona jednoznacznie przez jej rozkład – gęstość rozkładu.

Można pokazać też, że funkcja charakterystyczna $\varphi(t)$ wyznacza jednoznacznie rozkład – dystrybuantę danego rozkładu.

Rozważmy zmienną losową X o gęstości $f(x) = \frac{1}{2}e^{-|x|}$. Funkcja charakterystyczna φ wyraża się wzorem

$$\begin{aligned} \varphi(t) &= \frac{1}{2}\int_{-\infty}^{+\infty} e^{itx}e^{-|x|}dx = \frac{1}{2}\int_{-\infty}^{0} e^{(it+1)x}dx + \frac{1}{2}\int_{0}^{+\infty} e^{(it-1)x}dx = \\ &= \frac{e^{(it+1)x}}{2(it+1)}\Big|_{-\infty}^{0} + \frac{e^{(it-1)x}}{2(it-1)}\Big|_{0}^{+\infty} = \frac{1}{2}\left(\frac{1}{it+1} - \frac{1}{it-1}\right) = \\ &= \frac{1}{t^2+1}. \end{aligned}$$

Stąd

$$\varphi'(t) = \frac{-2t}{(t^2+1)^2}, \qquad \varphi''(t) = \frac{2(t^2-1)}{(t^2+1)^3}.$$

Zatem

$$E(X) = 0, \qquad V(X) = 2.$$

1.10 Prawa wielkich liczb

Zdarzenia losowe wykazują pewną prawidłowość w częstości występowania, gdy mamy dużą liczbę doświadczeń.

Jeżeli rzucamy monetą 100 razy, to możemy oczekiwać, że częstość (prawdopodobieństwo) otrzymania "orła" będzie bliska 1/2. Jeżeli rozważać będziemy 1000 rzutów, to z jeszcze wiekszą pewnością możemy oczekiwać, że prawdopodobieństwo praktyczne będzie 1/2.

Prawidłowość ta jest wyrażona przez twierdzenia znane jako **prawa wielkich liczb.**

Wpierw rozpatrzmy tak zwaną nierówność Czebyszewa, na której opierają się dowody tych twierdzeń.

Twierdzenie 1.14 (Nierówność Czebyszewa) *Jeżeli X jest zmienną losową o skończonej wariancji $V(X)$, to dla dowolnego $\varepsilon > 0$ zachodzi nierówność*

$$P(|X - E(X)| \geq \varepsilon) \leq \frac{V(X)}{\varepsilon^2},$$

nazywana nierównością Czebyszewa.

Dowód. Dowód przedstawimy dla zmiennej X typu ciągłego o dystrybuancie $F(x)$. Mamy wtedy

$$\begin{aligned} P(|X - E(X)| \geq \varepsilon) &= P(X \leq E(X) - \varepsilon) + P(X \geq E(X) + \varepsilon) = \\ &= \int_{-\infty}^{E(X)-\varepsilon} dF(x) + \int_{E(X)+\varepsilon}^{+\infty} dF(x). \end{aligned}$$

Dla x należących do przedziałów całkowania mamy $|x - E(X)| \geq \varepsilon$ lub inaczej $\frac{|x-E(X)|}{\varepsilon} \geq 1$ i dlatego

$$\begin{aligned} &\int_{-\infty}^{E(X)-\varepsilon} dF(x) + \int_{E(X)+\varepsilon}^{+\infty} dF(x) = \int_{-\infty}^{E(X)-\varepsilon} f(x)dx + \int_{E(X)+\varepsilon}^{+\infty} f(x)dx \leq \\ &\leq \int_{-\infty}^{E(X)-\varepsilon} \frac{[x - E(X)]^2}{\varepsilon^2} f(x)dx + \int_{E(X)+\varepsilon}^{+\infty} \frac{[x - E(X)]^2}{\varepsilon^2} f(x)dx \leq \\ &\leq \frac{1}{\varepsilon^2} \int_{-\infty}^{+\infty} [x - E(X)]^2 f(x)dx = \frac{V(X)}{\varepsilon^2}. \end{aligned}$$

Stąd otrzymujemy żądaną nierówność. To kończy dowód.

Z nierówności tej wynika, że przy ustalonym $\varepsilon > 0$, im mniejsza jest wariancja $V(X)$, tym mniejsze jest prawdopodobieństwo, że zmienna losowa X różni się od swej wartości przeciętnej $E(X)$ więcej niż o ε. Jest to zgodne z faktem, że wariancja $V(X)$ jest miarą odchylenia wartości zmiennej losowej X od jej wartości przeciętnej.

Możemy teraz przedstawić i udowodnić jedno z twierdzeń znanych jako prawa wielkich liczb.

Twierdzenie 1.15 (Twierdzenie Czebyszewa) *Jeżeli $X_1, X_2, \ldots$ jest ciągiem zmiennych losowych parami niezależnych, których wariancje są wspólnie ograniczone, to znaczy istnieje stała C, taka że*

$$V(X_i) < C, \quad i = 1, 2, \ldots,$$

to dla dowolnej liczby $\varepsilon > 0$ zachodzi równość:

$$\lim_{n\to\infty} P\left(\left|\frac{X_1 + X_2 + \ldots + X_n}{n} - \frac{E(X_1) + E(X_2) + \ldots + E(X_n)}{n}\right| < \varepsilon\right) = 1.$$

Dowód. Położmy

$$\overline{X_n} = \frac{X_1 + X_2 + \ldots + X_n}{n}.$$

Z nierówności Czebyszewa dla $\overline{X_n}$ wynika, że dla każdego $\varepsilon > 0$ mamy

$$P\left(|\overline{X_n} - E(\overline{X_n}| \geq \varepsilon\right) \leq \frac{V(\overline{X_n})}{\varepsilon^2}$$

lub

$$P\left(|\overline{X_n} - E(\overline{X_n})| < \varepsilon\right) \geq 1 - \frac{V(\overline{X_n})}{\varepsilon^2}.$$

Wykorzystując własności wartości oczekiwanej $E(X)$ i wariancji $V(X)$ dla sumy zmiennych niezależnych, mamy

$$V(\overline{X_n}) = \frac{V(X_1) + V(X_2) + \ldots + V(X_n)}{n^2} \leq \frac{n \cdot C}{n^2} = \frac{C}{n}.$$

Zatem

$$P\left(|\overline{X_n} - E(\overline{X_n})| < \varepsilon\right) \geq 1 - \frac{V(\overline{X_n})}{\varepsilon^2} \geq 1 - \frac{C}{n\varepsilon^2}.$$

Przechodząc do granicy przy $n \to \infty$, otrzymujemy:

$$\lim_{n\to\infty} P\left(|\overline{X_n} - E(\overline{X_n})| < \varepsilon\right) \geq \lim_{n\to\infty} \left(1 - \frac{C}{n\varepsilon^2}\right) = 1.$$

Ponieważ prawdopodobieństwo jest zawsze mniejsze od jedności, więc z tych nierówności otrzymujemy równość

$$\lim_{n\to\infty} P\left(|\overline{X_n} - E(\overline{X_n})| < \varepsilon\right) = 1.$$

Uwaga. Zwróćmy uwagę, że o zmiennych losowych X_n nie zakładamy nic poza ograniczonością wariancji $V(X_n)$.

Twierdzenie 1.16 (Twierdzenie Bernoulliego) *Niech X będzie liczbą sukcesów w serii n-doświadczeń niezależnych (w schemacie Bernoulliego) i p jest prawdopodobieństwem sukcesu w każdym doświadczeniu. Wtedy dla dowolnego $\varepsilon > 0$ mamy*

$$\lim_{n\to\infty} P\left(\left|\frac{X}{n} - p\right| < \varepsilon\right) = 1.$$

Dowód. Rozważmy zmienną $Y_n = \frac{X}{n}$. Wtedy

$$\begin{aligned} E(Y_n) &= \frac{1}{n}E(X) = \frac{1}{n}np = p, \\ V(Y_n) &= \frac{1}{n^2}V(X) = \frac{npq}{n^2} = \frac{pq}{n} < 1. \end{aligned}$$

Zatem na mocy Twierdzenia Czebyszewa

$$\lim_{n\to\infty} P\left(|Y_n - E(Y_n)| < \varepsilon\right) = \lim_{n\to\infty} P\left(\left|\frac{X}{n} - p\right| < \varepsilon\right) = 1.$$

1.11 Twierdzenia centralne

W wielu praktycznych zagadnieniach techniki, fizyki, biologii, demografii itp. bardzo często mamy zjawiska o rozkładzie normalnym. Do pewnego czasu uważano, że rozkład normalny jest jedynym rozkładem dającym dobre przybliżenie rozkładów empirycznych (doświadczalnych). Obecnie wiadomo, że istnieją też inne rozkłady teoretyczne przybliżające pewne rozkłady empiryczne. Rozkład normalny występuje jednak najczęściej.

Wyjaśnienie tego faktu można znaleźć przez tak zwane twierdzenia centralne, które podają warunki na to, aby zmienna losowa miała rozkład normalny lub bliski rozkładowi normalnemu.

Rozważmy ciąg X_k zmiennych losowych o jednakowym rozkładzie. Niech

$$E(X_k) = m, \quad V(X_k) = \sigma^2, \quad k = 1, 2, \ldots$$

Dla zmiennej losowej

$$Y_n = X_1 + X_2 + \ldots + X_n$$

mamy

$$E(Y_n) = n \cdot m, \quad V(Y_n) = n \cdot \sigma^2.$$

Oznaczmy zmienną standaryzowaną zmiennej Y_n przez Z_n, kładąc

$$Z_n = \frac{Y_n - n \cdot m}{\sigma\sqrt{n}} = \frac{1}{\sigma\sqrt{n}} \sum_{k=1}^{n} (X_k - m).$$

Wtedy

$$E(Z_n) = 0, \quad V(Z_n) = \frac{1}{n\sigma^2} \sum_{k=1}^{n} V(X_k) = 1.$$

Twierdzenie 1.17 (Twierdzenia centralne Lindeberga-Levy'ego) *Jeżeli $X_1, X_2, \ldots$ są zmiennymi losowymi niezależnymi o tym samym rozkładzie ($E(X_k) = m$, $V(X_k) = \sigma^2$), to rozkład zmiennej losowej*

$$Z_n = \frac{1}{\sigma\sqrt{n}} \sum_{k=1}^{n} (X_k - m) = \frac{Y_n - n \cdot m}{\sigma\sqrt{n}}$$

dąży do rozkładu normalnego standaryzowanego, to znaczy

$$\lim_{n\to\infty} P(a < Z_n < b) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{1}{2}z^2} dz.$$

Dowód. Wszystkie zmienne losowe $(X_k - m)$ mają tę samą (wspólną) funkcję charakterystyczną $\varphi_n(t)$, a zatem funkcja charakterystyczna $\varphi_n(t)$ zmiennej losowej Z_n ma postać

$$\varphi_n(t) = \left[\varphi_1\left(\frac{t}{\sigma\sqrt{n}}\right)\right]^n.$$

Wyliczmy początkowe wyrazy rozwinięcia funkcji $\varphi_1(t)$ w szereg potęgowy. Zauważmy, że

$$\begin{aligned} \varphi_1(0) &= E(1) = E(e^0) = 1, \\ \varphi_1'(0) &= iE(X_k - m) = 0, \\ \varphi_1''(0) &= -E([X_k - m]^2) = -V(X_k) = -\sigma^2. \end{aligned}$$

Zatem

$$\varphi_1(t) = \varphi_1(0) + \varphi_1'(0)t + \frac{1}{2}\varphi_1''(0)t^2 + O(t^2) = 1 - \frac{1}{2}\sigma^2 t^2 + O(t^2),$$

gdzie $\lim_{t\to 0}\frac{O(t^2)}{t^2}=0$, wobec tego

$$\varphi_1\left(\frac{t}{\sigma\sqrt{n}}\right)=1-\frac{1}{2}\frac{t^2}{n}+O\left(\frac{t^2}{n}\right),$$

i dla $\varphi_n(t)$ mamy

$$\begin{aligned}\varphi_n(t) &= \left[1-\frac{t^2}{2n}+O\left(\frac{t^2}{n}\right)\right]^n,\\ \ln\varphi_n(t) &= n\ln\left[1-\frac{t^2}{2n}+O\left(\frac{t^2}{n}\right)\right].\end{aligned}$$

Korzystając z rozwinięcia funkcji $\ln(1-u)$ w szereg potęgowy

$$\ln(1-u)=-u-\frac{1}{2}u^2-\ldots,$$

przy $u=\frac{t^2}{2n}+O\left(\frac{t^2}{n}\right)$ otrzymujemy:

$$\ln\varphi_n(t)=-n\left[\frac{t^2}{2n}+O\left(\frac{t^2}{n}\right)\right]=-\frac{1}{2}t^2-n\cdot O\left(\frac{t^2}{n}\right).$$

Zauważmy, że dla ustalonego t zachodzi równość

$$\lim_{n\to\infty}\left(n\cdot O\left(\frac{t^2}{n}\right)\right)=\lim_{n\to\infty}\frac{O\left(\frac{t^2}{n}\right)t^2}{\frac{t^2}{n}}=0,$$

a zatem

$$\lim_{n\to\infty}[\ln\varphi_n(t)]=-\frac{1}{2}t^2$$

lub równoważnie

$$\varphi(t)=\lim_{n\to\infty}\varphi_n(t)=e^{-\frac{1}{2}t^2}.$$

Pokazaliśmy, że granica zmiennej losowej Z_n ma funkcję charakterystyczną $\varphi(t)$ równą funkcji charakterystycznej rozkładu normalnego standaryzowanego, co kończy dowód.

Twierdzenie 1.18 (Twierdzenie centralne Lapunowa) *Niech $X_1, X_2, \ldots$ będzie ciągiem zmiennych losowych niezależnych o dowolnych rozkładach mających skończone momenty do rzędu trzeciego włącznie. Niech m_k, σ_k będą*

odpowiednio wartością oczekiwaną i odchyleniem standardowym zmiennej X_k. Oznaczmy

$$\begin{aligned}B_k^3 &= E\left(|X_k-m_k|^3\right),\quad k=1,2,\ldots,\\ W_n &= \sqrt[3]{B_1^3+B_2^3+\ldots+B_n^3},\quad n=1,2,\ldots,\\ C_n &= \sqrt{\sigma_1^2+\sigma_2^2+\ldots+\sigma_n^2},\quad n=1,2,\ldots\end{aligned}$$

Jeżeli

$$\lim_{n\to\infty}\frac{W_n}{C_n}=0,$$

to dla zmiennej

$$Z_n=\frac{1}{C_n}\sum_{k=1}^{n}(X_k-m_k)$$

ma miejsce równość

$$\lim_{n\to\infty}P(a<Z_n<b)=\frac{1}{\sqrt{2\pi}}\int_a^b e^{-\frac{1}{2}z^2}dz.$$

Twierdzenie to jest uogólnieniem poprzedniego twierdzenia na przypadek, gdy zmienne X_k mają różne rozkłady.

1.12 Elementy teorii korelacji

Regresja typu pierwszego

W wielu zagadnieniach praktycznych zachodzi konieczność jednoczesnego badania danej populacji generalnej ze względu na dwie cechy X oraz Y. Zależności pomiędzy X i Y na ogół nie są funkcyjne i ulegają wpływom o charakterze losowym. W celu uchwycenia zależności pomiędzy X i Y wprowadza się tak zwane **linie regresji.**

Rozważmy dwuwymiarową zmienną losową (X,Y). Załóżmy, że dla każdej wartości zmiennej losowej Y istnieje warunkowa wartość przeciętna $E(X|Y=y)$. Jest ona pewną funkcją zmiennej y:

$$E(X|Y=y)=m_1(y). \tag{1.1}$$

Równanie

$$x=m_1(y)$$

nazywamy równaniem regresji typu pierwszego zmiennej X względem zmiennej Y.

Analogicznie załóżmy, że dla każdej wartości zmiennej X istnieje warunkowa wartość przeciętna $E(Y|X=x)$. Jest ona pewną funkcją zmiennej x:

$$E(Y|X=x) = m_2(x). \tag{1.2}$$

Równanie

$$y = m_2(x)$$

nazywamy równaniem regresji typu pierwszego zmiennej Y względem zmiennej X.

Obrazem geometrycznym zależności (1.1) i (1.2) są pewne krzywe na płaszczyźnie xOy. Nazywamy je **krzywymi regresji pierwszego rodzaju.** Z przyjętego określenia widać, że na krzywej regresji zmiennej X względem zmiennej Y "leżą" warunkowe wartości przeciętne zmiennej losowej X, gdy Y przybiera daną ustaloną wartość y. Analogicznie na krzywej regresji zmiennej Y względem zmiennej X "leżą" warunkowe wartości przeciętne zmiennej Y, gdy X przybiera daną ustaloną wartość x.

Na ogół linia regresji typu pierwszego zmiennej X względem zmiennej Y nie pokrywa się z linią regresji zmiennej Y względem zmiennej X. Jeżeli zmienne X, Y spełniają ustaloną zależność liniową

$$Y = aX + b,$$

to wartościom zmiennej podwójnej (X,Y) odpowiadają punkty prostej

$$y = ax + b.$$

Wtedy obie krzywe regresji pokrywają się z tą prostą.

Jeżeli zmienne X, Y są niezależne, to

$$m_1(y) = E(X|Y=y) = E(X); \quad m_1(x) = E(Y|X=x) = E(Y).$$

Równania linii regresji mają postać

$$x = E(X), \quad y = E(Y),$$

są więc odpowiednio prostymi równoległymi do osi układu xOy przecinającymi się w punkcie $(E(X), E(Y))$.

Krzywe regresji mają pewną bardzo ważną **własność minimalności.** Można ją przedstawić w formie następującego twierdzenia:

Twierdzenie 1.19 *Wartość przeciętna kwadratu odchylenia zmiennej losowej Y od zmiennej $u(X)$ jest najmniejsza, gdy funkcja $u(x)$ pokrywa się z funkcją $m_2(x)$ i podobnie wartość przeciętna kwadratu odchylenia zmiennej*

losowej X od zmiennej $v(Y)$ jest najmniejsza, gdy funkcja $v(y)$ pokrywa się z funkcją $m_1(y)$. Fakty te można zapisać następująco:

$$\min_{u(x)} E\left[(Y-u(X))^2\right] = E\left[(Y-m_2(X))^2\right],$$
$$\min_{v(y)} E\left[(X-v(Y))^2\right] = E\left[(X-m_1(Y))^2\right].$$

Szkic dowodu. Niech (X,Y) będzie zmienną losową typu ciągłego o gęstości rozkładu $f(x,y)$. Wtedy

$$E\left[(Y-u(X))^2\right] = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}(y-u(x))^2 f(x,y)dxdy =$$
$$= \int_{-\infty}^{+\infty}\left\{\int_{-\infty}^{+\infty}(y-u(x))^2 f(x,y)dy\right\}dx.$$

Dla każdego ustalonego x funkcja $f(x,y) = f(y|X=x)$ jest (rozkładem) gęstością prawdopodobieństwa zmiennej $Y|X=x$. Dlatego całka w nawiasie "klamrowym" dla każdego ustalonego x jest najmniejsza, ale nieujemna, gdy jest równa wariancji zmiennej warunkowej $Y|X=x$, a więc gdy

$$u(x) = m_2(x) = E(Y|Y=x).$$

To kończy dowód pierwszej równości. Drugą dowodzimy analogicznie.

Regresja typu drugiego

Zostało więc wykazane, że krzywe regresji pierwszego typu, mające równania $y = m_2(x)$ oraz $x = m_1(y)$, czynią zadość pewnym warunkom minimalności, a więc, że funkcjonał

$$E\left[(Y-u(X))^2\right] \tag{1.3}$$

określony na zbiorze wszystkich funkcji $u(x)$ osiąga minimum dla $u(x) = m_2(x)$.

W praktyce często poszukujemy minimum funkcjonału (1.3) nie na zbiorze wszystkich funkcji, lecz wśród funkcji należących do pewnej ustalonej klasy funkcji, na przykład dla funkcji liniowych lub wielomianów ustalonego stopnia lub inaczej wyróżnionych. Wtedy krzywą

$$y = p(x),$$

dla której funkcjonał (1.3) osiąga minimum w wybranej klasie funkcji, nazywamy **linią regresji drugiego typu** zmiennej Y względem zmiennej X.

Szczególnie ważny przypadek to ten, gdy ograniczymy się do funkcji liniowych. Wtedy krzywe regresji drugiego typu są prostymi.

Wśród wszystkich prostych na płaszczyźnie xOy znajdziemy taką prostą

$$y = \alpha x + \beta, \tag{1.4}$$

że

$$E\left[(Y - aX - b)^2\right] \geq E\left[(Y - \alpha X - \beta)^2\right],$$

dla dowolnej prostej $y = ax + b$. Prostą (1.4) nazywa się **prostą regresji drugiego rodzaju** zmiennej Y względem zmiennej X.

Podobnie prostą

$$x = \gamma y + \delta \tag{1.5}$$

realizującą warunek

$$E\left[(X - cY - d)^2\right] \geq E\left[(X - \gamma Y - \delta)^2\right],$$

dla dowolnej prostej $y = ax + b$, nazywa się prostą regresji drugiego rodzaju zmiennej X względem zmiennej Y.

Współczynniki prostych regresji wyznaczmy za pomocą momentów zmiennej (X, Y).

Połóżmy

$$m_{1,0} = E(X), \qquad \sigma_1^2 = \mu_{2,0} = E[(X - m_{1,0})^2],$$

$$m_{0,1} = E(Y), \qquad \sigma_2^2 = \mu_{0,2} = E[(Y - m_{0,1})^2],$$

$$\mu_{1,1} = E[(X - m_{1,0})(Y - m_{0,1})]$$

i załóżmy, że odchylenia standardowe $\sigma_1 = \sqrt{\mu_{2,0}} > 0$, $\sigma_2 = \sqrt{\mu_{0,2}} > 0$ są skończone. Wtedy mamy

$$\begin{aligned} E\left[(Y - \alpha X - \beta)^2\right] &= \\ &= E\left[(Y - m_{0,1} - \alpha(X - m_{1,0}) + m_{0,1} - \alpha m_{1,0} - \beta)^2\right] = \\ &= E\left[(Y - m_{0,1})^2 + \alpha^2(X - m_{1,0})^2 + (m_{0,1} - \alpha m_{1,0} - \beta)^2 + \right. \\ &\quad +2\alpha(X - m_{1,0})(Y - m_{0,1}) + (m_{0,1} - \alpha m_{1,0} - \beta)(Y - m_{0,1}) - \\ &\quad \left. -2\alpha(m_{0,1} - \alpha m_{1,0} - \beta)(X - m_{1,0})\right] = \\ &= \mu_{0,2} + \alpha^2\mu_{2,0} - 2\alpha\mu_{1,1} + (m_{0,1} - \alpha m_{1,0} - \beta)^2 =: q(\alpha, \beta). \end{aligned} \tag{1.6}$$

Aby wyznaczyć minimum ostatniego wyrażenia, potraktujemy je jako funkcję dwóch zmiennych α, β i zastosujemy standardową metodę wyznaczania ekstremów. Punkt stacjonarny wyznaczamy z równań:

$$\begin{cases} \dfrac{\partial q(\alpha,\beta)}{\partial \alpha} &= 2\alpha\mu_{2,0} - 2\mu_{1,1} - 2m_{1,0}(m_{0,1} - \alpha m_{1,0} - \beta) &= 0, \\ \dfrac{\partial q(\alpha,\beta)}{\partial \beta} &= -2(m_{0,1} - \alpha m_{1,0} - \beta) &= 0. \end{cases}$$

Stąd

$$\begin{cases} \alpha &= \dfrac{\mu_{1,1}}{\mu_{2,0}} = \rho\dfrac{\sigma_2}{\sigma_1}, \\ \beta &= m_{0,1} - \dfrac{m_{1,0}\mu_{1,1}}{\mu_{2,0}} = m_{0,1} - \rho\dfrac{\sigma_2}{\sigma_1}m_{1,0}, \end{cases} \tag{1.7}$$

gdzie

$$\rho = \frac{\mu_{1,1}}{\sigma_1\sigma_2} \tag{1.8}$$

nazywamy **współczynnikiem korelacji** zmiennych X, Y. Dla ρ danego przez (1.8) α, β są wartościami, przy których w (1.6) otrzymujemy minimum.

Podobnie współczynniki prostej regresji typu drugiego zmiennej X względem zmiennej Y są odpowiednio równe:

$$\begin{cases} \gamma &= \rho\dfrac{\sigma_1}{\sigma_2}, \\ \delta &= m_{1,0} - \rho\dfrac{\sigma_1}{\sigma_2}m_{0,1}. \end{cases} \tag{1.9}$$

Zatem na mocy zależności (1.7), (1.8) i (1.9) proste regresji (1.4) i (1.5) mają równania:

$$\begin{cases} y - m_{0,1} &= \rho\dfrac{\sigma_2}{\sigma_1}(x - m_{1,0}), \\ y - m_{0,1} &= \dfrac{1}{\rho}\dfrac{\sigma_2}{\sigma_1}(x - m_{1,0}). \end{cases} \tag{1.10}$$

Proste regresji – jak widać z (1.10) – przechodzą zawsze przez punkt $(m_{1,0}, m_{0,1})$. Jeżeli $\rho^2 = 1$, to $\rho = 1/\rho$ i proste regresji pokrywają się.

Jeżeli $\rho = 0$, to z (1.10) wynika, że proste regresji są prostopadłe względem siebie i mają postać

$$y = m_{0,1}, \qquad x = m_{1,0}.$$

Zauważmy, że jeżeli linia regresji typu pierwszego jest prostą, to pokrywa się z prostą regresji typu drugiego. Nie zawsze prosta regresji typu drugiego musi być linią regresji typu pierwszego.

Wykorzystując (1.7) i (1.9), znajdujemy wartości minimów:

$$\begin{aligned} \min E\left[(Y-\alpha X-\beta)^2\right] &= E\left[(Y-\alpha_{2,1}X-\beta_{2,1})^2\right]=\sigma_2^2(1-\rho^2), \\ \min E\left[(X-\alpha Y-\beta)^2\right] &= E\left[(X-\alpha_{1,2}Y-\beta_{1,2})^2\right]=\sigma_1^2(1-\rho^2). \end{aligned} \tag{1.11}$$

Współczynnik korelacji

Współczynnik korelacji ρ, o którym już była mowa, wyraża się wzorem

$$\rho=\frac{\mu_{1,1}}{\sigma_1\sigma_2}=\frac{E[(X-m_{1,0})(Y-m_{0,1})]}{\sqrt{E[(X-m_{1,0})^2]}\sqrt{E[(Y-m_{0,1})^2]}}.$$

Łatwo zauważyć, że

$$-1<\rho<1.$$

Własności współczynnika korelacji ρ są związane z charakterystyką rozkładu zmiennych (X,Y) i ściśle się łączą z prostymi regresji typu drugiego, co było pokazane przed chwilą.

Liczby $\sigma_2^2(1-\rho^2)$, $\sigma_1^2(1-\rho^2)$, wartości minimów (1.11) są nazywane **wariancjami resztkowymi** odpowiednio zmiennych Y i X.

Zaważmy, wykorzystując (1.11), (1.7) i (1.9), że stosunek wariancji zmiennej

$$Y-\alpha X-\beta$$

do wariancji zmiennej Y jest równy $1-\rho^2$ i podobnie stosunek wariancji zmiennej

$$X-\gamma Y-\delta$$

do wariancji zmiennej X jest też równy $1-\rho^2$.

Wynika stąd, że:

1° Kiedy współczynnik korelacji $\rho=0$, to znaczy gdy X, Y są nieskorelowane, wówczas wariancji zmiennej Y nie można pomniejszyć przez odjęcie od Y funkcji liniowej zmiennej X, i odwrotnie: wariancji zmiennej X nie można pomniejszyć przez odjęcie funkcji liniowej zmiennej Y. W szczególności ma to miejsce, gdy zmienne X, Y są niezależne.

2° Jeżeli $\rho\neq 0$, to różną od zera wariancję zmiennej Y można zmniejszyć przez odjęcie od Y pewnej funkcji liniowej zmiennej X, i odwrotnie: wariancję zmiennej X można pomniejszyć przez odjęcie od X pewnej funkcji liniowej zmiennej Y. Mówimy w tym przypadku, że zmienne X, Y są skorelowane, dodatnio skorelowane dla $\rho\in(0,1\rangle$, i ujemnie skorelowane dla $\rho\in\langle -1,0)$.

Zmniejszenie wariancji może być tym większe, im większa jest wartość $|\rho|$.

3° Jeżeli $\rho^2=1$, to wariancje resztkowe są równe zeru. Wtedy prawdopodobieństwo, że zmienne X, Y łączy pewna zależność liniowa, to znaczy istnieją liczby a, b, takie że $X=aY+b$ lub $Y=aX+b$, jest równe jedności:

$$\exists_{a,b}P(Y=aX+b)=1.$$

Widać stąd, że współczynnik ρ korelacji można traktować jako pewną (probabilistyczną) miarę stopnia liniowości zmiennych X i Y. Im wartość $|\rho|\in(0,1)$ jest większa, tym stopień liniowości jest większy.

Szacowanie parametrów prostej regresji

Wyznaczenie dokładnych wartości parametrów prostych regresji typu drugiego jest możliwe, gdy znamy funkcję gęstości prawdopodobieństwa $f(x,y)$ populacji generalnej (X,Y). W praktyce możemy zazwyczaj dysponować tylko danymi z próby. Znamy więc zwykle układy (x_i,y_j), $p_{i,j}$ lub tylko (x_i,y_i), $i=1,2,\ldots,n$, i przypuszczamy, że wszystkie wybory są jednakowo prawdopodobne. Jeżeli na podstawie tych danych chcemy przybliżyć parametry α, β prostej regresji $y=\alpha x+\beta$, to potrzebne nam wielkości możemy przybliżyć, zastąpić następującymi wielkościami obliczonymi z próby:

$$\begin{aligned} m_{1,0}\approx\overline{x} &= \frac{1}{n}\sum_{i=1}^{n}x_i\,, \\ m_{0,1}\approx\overline{y} &= \frac{1}{n}\sum_{i=1}^{n}y_i\,, \\ \sigma_1^2\approx s_1^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2=\frac{1}{n}\sum_{i=1}^{n}x_i^2-\overline{x}^2\,, \\ \sigma_2^2\approx s_2^2 &= \frac{1}{n}\sum_{i=1}^{n}(y_i-\overline{y})^2=\frac{1}{n}\sum_{i=1}^{n}y_i^2-\overline{y}^2\,, \\ \sigma_{1,1}\approx s_{1,1} &= \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})=\frac{1}{n}\sum_{i=1}^{n}x_iy_i-\overline{x}\,\overline{y}\,. \end{aligned}$$

Przybliżenie współczynnika korelacji ma postać

$$\rho \approx r = \frac{s_{1,1}}{s_1 s_2} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}.$$

Parametry przybliżonej prostej regresji możemy wyznaczyć ze wzorów:

$$\alpha \approx r\frac{s_2}{s_1}, \qquad \beta \approx \overline{y} - r\frac{s_2}{s_1}\overline{x},$$
$$\gamma \approx r\frac{s_1}{s_2}, \qquad \delta \approx \overline{x} - r\frac{s_1}{s_2}\overline{y},$$

same zaś proste mają postać

$$y - \overline{y} = r\frac{s_2}{s_1}(x - \overline{x}),$$
$$y - \overline{y} = \frac{1}{r}\frac{s_2}{s_1}(x - \overline{x}).$$

Proste te mają w przybliżeniu własności podobne do własności dokładnych prostych regresji.

Rozdział 2

Statystyka

2.1 Elementy statystyki opisowej

Statystyka matematyczna zajmuje się opisywaniem i analizą zjawisk masowych za pomocą metod rachunku prawdopodobieństwa. Celem badań statystycznych jest poznanie prawidłowości ilościowych i jakościowych w masowych zjawiskach losowych (przypadkowych) i opisywanie ich za pomocą liczb.

Badane zbiory nazywamy **populacjami statystycznymi**. Badać można wszystkie elementy danej populacji statystycznej, zwanej też **populacją generalną**, albo tylko ich część, zwaną **próbką statystyczną** lub krótko **próbką**.

W pierwszym przypadku **badanie** jest **kompletne** i nie ma potrzeb używania elementów rachunku prawdopodobieństwa. W drugim mówimy, że **badania** są **częściowe**. Zadaniem statystyki jest wnioskowanie o własnościach całej populacji na podstawie informacji o tych własnościach elementów pewnego skończonego podzbioru tej populacji, zwanego **próbką**. Próbka Z_1 powinna stanowić reprezentację populacji Z w tym sensie, że częstość występowania w próbce każdej z badanych cech nie powinna znacznie różnić się od częstości występowania tych cech w populacji generalnej. Aby to osiągnać, elementy próbki Z_1 zwykle losujemy spośród elementów zbioru Z. Otrzymany w ten sposób zbiór nazywamy **próbką losową**. **Próbka losowa prosta** n-elementowa to taka, w której każdy element populacji generalnej ma takie same szanse wylosowania. **Statystyka opisowa** zajmuje się wstępnym opracowaniem próbki.

Szeregi rozdzielcze

Niech $x_1, x_2, \ldots, x_n$ będzie n-elementową próbką prostą o zadanych wartościach. **Rozstępem** badanej cechy X w tej próbce nazywamy liczbę

$$R = x_{\max} - x_{\min},$$

gdzie $x_{\max}, x_{\min}$ oznaczają, odpowiednio, największą i najmniejszą liczbę w ciągu x_k. Przy większej liczności próbki (zwykle od 30 wzwyż), w celu ułatwienia analizy, wartości próbki grupuje się w **klasach**, przedziałach zwykle jednakowej długości, przyjmując upraszczające założenia, że wszystkie wartości w danej klasie są identyczne ze **środkiem klasy**. Jeżeli R jest rozstępem próbki, k zaś liczbą klas, to jako **długość klasy** przyjmujemy liczbę $b \approx \frac{R}{k}$, tak jednak, żeby $b \cdot k \geq R$. Punkty podziału na klasy ustala się zwykle z dokładnością nieco wiekszą niż ta, z którą wyznacza się wartości w próbce. Formalnie można przyjąć podział na klasy punktami

$$x_{\min} \approx a, a + b, \ldots, a + kb \approx x_{\max},$$

gdzie $b \approx \frac{R}{k}$ jest tak dobrane, żeby $a \leq x_{\min}$, $b \geq x_{\max}$ i żeby środki klas $\overline{x}_i$ były w miarę proste (dane z pewną ustaloną dokładnością), $\overline{x}_i \approx a + (i - 1/2)b$. Liczbę wartości próbki zawartych w i-tej klasie nazywamy **licznością, (liczebnością) i-tej klasy** i oznaczamy symbolem n_i. Oczywiście

$$\sum_{i=1}^{k} n_i = n.$$

Jeżeli liczność n próbki $x_1, \ldots, x_n$ kwalifikuje się do podziału na klasy, to dokonuje się grupowania, w wyniku czego otrzymuje się **szereg rozdzielczy**, który stanowią pary liczb: środki kolejnych klas oraz ich liczności n_i, $i = 1, 2, \ldots, k$.

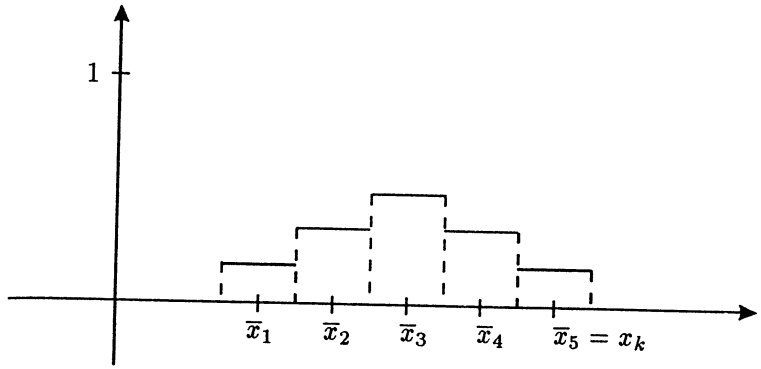
Sposób, w jaki liczności n_i są rozłożone w poszczególnych klasach, nazywamy **rozkładem liczności badanej cechy** przy danej liczbie k klas.

Otrzymany szereg można przedstawić graficznie w postaci tzw. **histogramu**. Na osi poziomej zaznacza się środki klas, na osi pionowej liczności n_i. Na osi pionowej można, również, odkładać inne wielkości, np. **częstości** n_i/n, wyrażane często w procentach $\frac{n_i}{n}100\%$;

$$n = 45, \quad n_1 = 5, \quad n_2 = 10, \quad n_3 = 15, \quad n_4 = 10, \quad n_5 = 5,$$

$$\sum \frac{n_i}{n} = \frac{1}{n} \sum n_i = 1.$$

Łącząc punkty $(\overline{x}_1 - b, 0), (\overline{x}_i, \frac{n_i}{n})$, $i = 1, 2, \ldots, k, (x_k + b, 0)$, otrzymamy tak zwaną **łamaną częstości**.



Rys. 2.1. Przykład graficznego przedstawienia szeregu rozdzielczego

Na Rys. 2.1 zostały zaznaczone środki klas $\overline{x}_i$ oraz częstości $\frac{n_i}{n}$.

Częstość skumulowaną otrzymujemy, dodając do częstości danej klasy częstości wszystkich poprzednich klas. Częstość skumulowana jest odpowiednikiem dystrybuanty.

Przykład szeregu rozdzielczego, $n = 150$:

Numer klasy	Przedział klasowy	Środek klasy $\overline{x}_i$	Liczność klasy n_i	Częstość klasy n_i/n	Częstość skumulowana $\sum_{l=1}^{i} x_l/n$
1	120 – 124	122	20	2/15	2/15
2	124 – 128	126	30	3/15	1/3
3	128 – 132	130	50	1/3	2/3
4	132 – 136	134	10	1/15	11/15
5	136 – 140	138	15	1/10	5/6
6	140 – 144	142	25	1/6	1

Liczby charakteryzujące szereg rozdzielczy

Szereg rozdzielczy można opisać przez podanie pewnych liczb, z których najważniejsze to **średnia arytmetyczna wartości danej cechy** oraz **wariancja**.

Jeżeli badana cecha przyjmuje wartości

$$\overline{x}_1, \overline{x}_2, \ldots, \overline{x}_k$$

z odpowiadającymi im licznościami

$$n_1, n_2, \ldots, n_k$$

oraz

$$n = n_1 + n_2 + \ldots + n_k,$$

to średnią arytmetyczną (wartość oczekiwaną) określamy następująco:

$$\overline{x} = \frac{n_1\overline{x}_1 + n_2\overline{x}_2 + \ldots + n_n\overline{x}_n}{n} = \frac{1}{n}\sum_{i=1}^{k} n_i\overline{x}_i.$$

Łatwo pokazać, że średnia $\overline{x}$ może być wyrażona za pomocą następującego, wygodnego wzoru:

$$\overline{x} = x_0 + \frac{1}{n}\sum_{i=1}^{k} n_i(\overline{x}_i - x_0),$$

gdzie x_0 jest dowolną liczbą rzeczywistą.

Występujące tu wielkości $\overline{x}_i$ mogą być wartościami zmiennej losowej X lub **środkami klas**, a n_i **licznościami klas** danego **szeregu rozdzielczego**.

Wariancję s^2 danej cechy X określamy wzorami:

$$s^2 = \frac{1}{n}\sum_{i=1}^{k} n_i\,(\overline{x}_i - \overline{x})^2 = \frac{1}{n}\sum_{i=1}^{k} n_i\overline{x}_i^2 - \overline{x}^2.$$

Liczbę s nazywamy **odchyleniem standardowym.**

Przykład 2.1 *Obserwowano przez tydzień (7 dni) ciśnienie atmosferyczne w danej miejscowości i zapisywano wyniki w postaci tabeli ($n = 7$, $k = 4$).*

Ciśnienie x_i	753	755	758	762
Liczność n_i	1	2	3	1

$$\begin{aligned}
\overline{x} &= \frac{1}{n}\sum_{i=1}^{k} n_i x_i = \frac{1}{7}(753 + 2\cdot 755 + 3\cdot 758 + 762) = 757,\\
s^2 &= \frac{1}{N}\sum_{i=1}^{n} n_i(x_i - \overline{x})^2 = \frac{1}{7}\left((-4)^2 + 2\cdot(-2)^2 + 3\cdot 1^2 + 5^2\right) =\\
&= \frac{52}{7} = 7,428\ldots,\\
s &= 2,7255\ldots.
\end{aligned}$$

Jeżeli przyjmiemy $x_0 = 755$, to

$$\overline{x} = 755 + \frac{1}{7}(-2 + 2\cdot 0 + 3\cdot 3 + 7) = 755 + \frac{14}{7} = 757,$$

a przy $x_0 = 757$ mamy

$$\overline{x} = 757 + \frac{1}{7}(-4 - 2\cdot 2 + 3\cdot 1 + 5) = 757 + 0 = 757.$$

Przykład 2.2 *Niech szereg rozdzielczy będzie zadany tabelą na stronie 69 ($n = 150$, $k = 6$). Wtedy*

$$\begin{aligned}
\overline{x} &= \frac{1}{n}\sum_{i=1}^{6} n_i\overline{x}_i = \frac{1}{150}(20\cdot 122 + 30\cdot 126 + 50\cdot 130 + 10\cdot 134 +\\
&\quad +15\cdot 138 + 25\cdot 142) = \frac{19680}{150} = \frac{1968}{15} = \frac{656}{5} = 131,2.
\end{aligned}$$

Kładąc $x_0 = 130$, mamy

$$\begin{aligned}
\overline{x} &= x_0 + \frac{1}{N}\sum_{i=1}^{6} n_i(\overline{x}_i - x_0) = 130 + \frac{1}{150}(20\cdot(-8) + 30\cdot(-4) +\\
&\quad +50\cdot 0 + 10\cdot 4 + 15\cdot 8 + 25\cdot 12) = 130 + \frac{180}{150} = 130 + \frac{6}{5} = 131,2,\\
s^2 &= \frac{1}{150}(20\cdot(-9,2)^2 + 30\cdot(-5,2)^2 + 50\cdot(-1,2)^2 + 10\cdot(2,8)^2 +\\
&\quad +15\cdot(6,8)^2 + 25\cdot(10,8)^2) = \frac{6264}{150} = 41,76\\
s &= 6,462\ldots
\end{aligned}$$

Elementy teorii prób, pojęcie próby

Zbiór wszystkich elementów, które poddajemy obserwacjom statystycznym, nazywamy **populacją generalną.** Elementy tej populacji badamy ze względu na interesującą nas **cechę** X, którą traktujemy **jako zmienną losową.**

Zwykle trudne lub wręcz niemożliwe jest pełne zbadanie całej populacji generalnej. Dlatego wykonuje się badania częściowe. Z populacji generalnej losujemy więc n-elementów i obserwujemy wartości $x_1, x_2, \ldots, x_n$ badanej cechy X. Losowanie odbywa się tak, że dla każdego elementu danej populacji generalnej istnieje ta sama szansa wylosowania (po wybraniu danego elementu i odczytaniu wartości x_i danej cechy X zwracamy element z powrotem do ponownego losowania lub liczba elementów populacji generalnej jest tak duża, że losowanie nawet bez zwracania nie zmienia szansy wylosowania pozostałych elementów).

Układ liczb $(x_1, x_2, \ldots, x_n)$ stanowi **próbę losową.**

Jest ona realizacją zmiennej losowej n-wymiarowej $(X_1, X_2, \ldots, X_n)$. Wszystkie zmienne X_i mają ten sam rozkład co cecha X i są niezależne. Zbiór wartości zmiennej losowej X_i to zbiór wszystkich możliwych wartości obserwacji i-tego elementu. Liczbę n nazywamy **licznością próby.**

Pobieranie próby może być liczne i wtedy opis doświadczenia musi być podany szeregiem rozdzielczym.

Przedział i poziom ufności

Załóżmy, że badamy populację generalną ze względu na pewną cechę ilościową X. Załóżmy też, że wartość oczekiwana (przeciętna) cechy X, nazywana **średnią generalną**, jest znana i równa m, że rozkład jest normalny oraz ma odchylenie standardowe σ.

Rozważmy próbę o liczności n, która jest realizacją n-wymiarowej zmiennej losowej $(X_1, X_2, \ldots, X_n)$, i nową zmienną losową

$$\overline{X} = \frac{1}{n}(X_1, X_2, \ldots, X_n),$$

nazywaną **średnią z próby**.

Rozkład tej nowej zmienej $\overline{X}$ jest też normalny, o tej samej średniej m i odchyleniu standardowym

$$s = \sigma/\sqrt{n}.$$

Wyznaczmy prawdopodobieństwo α zdarzenia losowego polegającego na tym, że średnia $\overline{X}$ leży w przedziale (a_1, a_2), gdzie a_1, a_2 to dowolne liczby rzeczywiste:

$$\begin{aligned} P(a_1 < \overline{X} < a_2) &= \alpha = P(t_1 < T < t_2) = \\ &= \frac{1}{\sqrt{2\pi}} \int_{t_1}^{t_2} e^{-\frac{1}{2}v^2} dv = \Phi(t_2) - \Phi(t_1), \end{aligned}$$

gdzie

$$T = \frac{\overline{X} - m}{\sigma}\sqrt{n}$$

jest zmienną losową o rozkładzie normalnym standaryzowanym,

$$t_1 = \frac{a_1 - m}{\sigma}\sqrt{n}, \qquad t_2 = \frac{a_2 - m}{\sigma}\sqrt{n}$$

oraz Φ jest funkcją Laplace'a $\Phi(t) = \int_{-\infty}^{t} e^{-\frac{1}{2}v^2} dv$, której wartości można odczytać z tablic.

Widać z tych przeliczeń, że liczby a_1, a_2 wyznaczają jednoznacznie prawdopodobieństwo α. Przedstawione rozumowania można wykorzystać do wyznaczania przedziału (a_1, a_2), w którym z zadanym z góry prawdopodobieństwem α leży (nieznana) średnia generalna m.

Znajomość prawdopodobieństwa α nie wyznacza jednoznacznie liczb t_1, t_2, gdyż równanie $\Phi(t_2) - \Phi(t_1) = \alpha$ ma wiele rozwiązań. Jeżeli jednak przyjąć dodatkowo, że $t_2 = -t_1 = t$ i ponieważ $\Phi(t) + \Phi(-t) = 1$, więc równanie

$$\Phi(t) - \Phi(-t) = 2\Phi(t) - 1 = \alpha, \quad \Phi(t) = \frac{1}{2} + \frac{\alpha}{2},$$

ma jednoznaczne rozwiązanie, które można odczytać z tablic funkcji Laplace'a. Dla takiej wartości t mamy więc

$$P\left(-t < \frac{\overline{X} - m}{\sigma/\sqrt{n}} < t\right) = P\left(-t < \frac{m - \overline{X}}{\sigma/\sqrt{n}} < t\right) = \alpha$$

lub równoważnie

$$P\left(\overline{X} - \frac{t\sigma}{\sqrt{n}} < m < \overline{X} + \frac{t\sigma}{\sqrt{n}}\right) = \alpha.$$

Jeżeli $\overline{x}$ jest średnią z próby, to przedział

$$\left(\overline{x} - \frac{t\sigma}{\sqrt{n}}, \overline{x} + \frac{t\sigma}{\sqrt{n}}\right),$$

gdzie wartość t jest odczytywana z tablicy rozkładu normalnego (Tab. 5.3), nazywamy **przedziałem ufności** dla wartości średniej, liczbę zaś α, wyznaczającą ten przedział, nazywamy **poziomem ufności**.

Przykład 2.3 *Dana jest populacja generalna, o której wiemy, że ma rozkład normalny $N(m, \sigma)$ o znanym odchyleniu standardowym $\sigma = 30$. Na podstawie próby z liczności $n = 100$ wyznaczono średnią z próby $\overline{x} = 90$. Wyznaczyć przedział ufności, w którym z prawdopodobieństwem $\alpha = 0,96$ znajduje się nieznana średnia generalna m.*

Rozwiązanie. Mamy $\sigma/\sqrt{n} = 30/\sqrt{100} = 3$. Z tablic funkcji Laplace'a znajdujemy wartość t, dla której $\Phi(t) = \frac{1}{2} + \frac{\alpha}{2} = 0,98$. Wartość ta wynosi

$$t = 2,06.$$

Stąd $t \cdot \sigma/\sqrt{n} = (2,054) \cdot 3 = 6,18$. Zatem przedziałem ufności jest przedział

$$(90 - 6,18\,,\, 90 + 6,18) = (83,82\,,\, 96,18).$$

Możemy więc twierdzić z prawdopodobieństwem $\alpha = 0,96$, że średnia generalna m spełnia warunek

$$83,82 < m < 96,18$$

lub równoważnie, że należy do przedziału ufności o poziomie ufności $0,96$

$$m \in (83,82\,,\, 96,18).$$

Można to też zapisać następująco:

$$P(83,82 < m < 96,18) = 0,96.$$

Rozważmy podobny problem dla wariancji.

Załóżmy teraz, że badana populacja generalna X ma rozkład normalny $N(m,\sigma)$, gdzie m i σ nie są nam znane. Wybieramy losowo próbę o liczności n oraz każdą i-tą obserwację traktujemy jako wartość zmiennej X_i.

Wprowadzimy zmienną losową

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Wtedy zmienna losowa

$$U = \frac{n}{\sigma^2}S^2,$$

gdzie

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2,$$

ma rozkład χ^2 o $(n-1)$ stopniach swobody.

Znając więc rozkład zmiennej χ^2, możemy wyznaczyć **przedział ufności,** to znaczy przedział, w którym z zadanym z góry prawdopodobieństwem α znajdzie się (będzie leżała) nieznana nam wariancja $V(X) = \sigma^2$ dla populacji generalnej.

W celu wyznaczenia tego przedziału szukamy przedziału (x_1, x_2), takiego że

$$P(x_1 < \chi^2 < x_2) = \alpha$$

lub równoważnie

$$-P(\chi^2 < x_1) + P(\chi^2 > x_2) = \alpha$$

albo

$$P(\chi^2 < x_1) + P(\chi^2 > x_2) = 1 - \alpha.$$

Aby jednoznacznie wyznaczyć ten przedział ufności, należy dodać jeszcze dodatkowy warunek. Nie można teraz założyć symetrii względem "zera", gdyż rozkład χ^2 przyjmuje tylko wartości dodatnie (z prawdopodobieństwem różnym od zera). Zazwyczaj ten dodatkowy warunek ma postać

$$P(\chi^2 > x_2) = P(\chi^2 < x_1)$$

lub równoważnie

$$P(\chi^2 < x_1) + P(\chi^2 < x_2) = 1.$$

Łącząc ze sobą warunki na x_1 i x_2, otrzymujemy:

$$P(\chi^2 < x_1) = \frac{1-\alpha}{2}, \quad P(\chi^2 < x_2) = \frac{1+\alpha}{2}.$$

Z tablic rozkładu prawdopodobieństwa χ^2, przy $k = n-1$, odczytujemy odpowiednie wartości x_1 i x_2.

Ponieważ zmienna nS^2/σ^2 ma rozkład χ^2, więc warunek

$$P(x_1 < \chi^2 < x_2) = \alpha$$

ma postać

$$P\left(x_1 < \frac{nS^2}{\sigma^2} < x_2\right) = \alpha$$

lub, dla nieznanego parametru σ,

$$P\left(\frac{nS^2}{x_2} < \sigma^2 < \frac{nS^2}{x_1}\right) = \alpha.$$

W ten sposób wyznaczamy przedział ufności dla σ^2, a więc przedział, w którym wariancja σ^2 leży z prawdopodobieństwem α. Jest to przedział

$$\left(\frac{nS^2}{x_2}, \frac{nS^2}{x_1}\right),$$

gdzie x_1, x_2 są policzone poprzednio dla rozkładu χ^2 o $(n-1)$ stopniach swobody, n jest zadane, a S^2 policzone z wybranej n-elementowej próby.

Przykład 2.4 *Na podstawie badań próby 25-elementowej wyznaczono odchylenie standardowe próby* $s = \sqrt{3}$. *Wyznaczyć* **przedział ufności**, *w którym z prawdopodobieństwem* $\alpha = 0,98$ *znajduje się nieznana nam wariancja* σ^2 *rozkładu populacji generalnej.*

Rozwiązanie. Warunki na x_1, x_2 mają postać

$$P(\chi^2 < x_1) = \frac{1-0,98}{2} = 0,01 \quad P(\chi^2 < x_2) = \frac{1+0,98}{2} = 0,99.$$

W tablicy rozkładu χ^2 przy $k = 24$ stopni swobody znajdujemy interesujące nas wartości x_1, x_2, które wynoszą

$$x_1 = 10,856, \qquad x_2 = 42,980.$$

Zatem przedziałem ufności jest przedział

$$\left(\frac{25\cdot 3}{42,980}, \frac{25\cdot 3}{10,857}\right) \approx \left(\frac{1}{2}, 7\right).$$

Możemy to rozumieć następująco: prawdopodobieństwo tego, że zmienna σ^2 przyjmuje wartości z przedziału $(1/2, 7)$, jest równe $0,98$.

2.2 Badania statystyczne

Statystyka matematyczna jest działem probabilistyki ściśle związanym z rachunkiem prawdopodobieństwa. Punkt widzenia statystycznego jest jednak inny. W rachunku prawdopodobieństwa rozważając zmienną losową, zakładamy, że jej rozkład jest znany. Wykorzystując ten fakt, wyznacza się prawdopodobieństwo różnych zdarzeń.

W statystyce nie mamy (nie zakładamy) zazwyczaj pełnej znajomości rozkładu zmiennej losowej interpretowanej w praktycznych zastosowaniach jako "cecha" statystyczna elementów badanej (zbiorowości) populacji generalnej.

Punktem wyjścia w badaniach statystycznych jest wybór (wylosowanie) z całej populacji generalnej skończonej liczby n elementów i zbadanie ich ze względu na interesującą nas cechę – zmienną losową X. Uzyskane w ten sposób wartości $x_1, x_2, \ldots, x_n$ badanej cechy X są zaobserwowanymi wartościami n-elementowej próby.

W statystyce opisowej ograniczamy się do opisu wyników przeprowadzonej próby bez wyciągania wniosków o całej populacji. W statystyce matematycznej natomiast, na podstawie wyników badania próbnego, staramy się wyciągnąć wnioski dotyczące badanej cechy (zmiennej losowej) w całej populacji generalnej.

Do najważniejszych form wnioskowania statystycznego należą:

- **estymacja** (ocena) nieznanych **parametrów** (badź ich funkcji), które charakteryzują rozkład badanej cechy populacji generalnej,
- **weryfikacja** (badania prawdziwości) postawionych hipotez statystycznych,
- **wnioskowanie statystyczne**, jako oparte na częściowej informacji, dostarcza jedynie wniosków **wiarygodnych a nie absolutnie prawdziwych**; wiarygodnych – to jest prawdziwych z pewnym zadanym, może bliskim jedności, prawdopodobieństwem.

Estymacja punktowa

Niech rozkład badanej cechy X populacji generalnej zależy od nieznanego parametru θ. Parametr ten będziemy szacowali na podstawie n-elementowej próby prostej $X_1, X_2, \ldots, X_n$.

Funkcję

$$g(X_1, X_2, \ldots, X_n),$$

będącą funkcją próby losowej $X_1, X_2, \ldots, X_n$, nazywamy **statystyką**. Statystyka jest funkcją zmiennych losowych, jest też zmienną losową mającą swój

własny rozkład zależny od postaci funkcji g i od rozkładu zmiennych X_i. Każdą statystykę $\hat{\theta}_n(X_1, X_2, \ldots, X_n)$, której wartości przyjmujemy jako oceny (przybliżenia) nieznanego parametru θ, nazywamy **estymatorem parametru** θ. Otrzymaną na podstawie realizacji konkretnej próby **wartość estymatora** nazywamy **oceną** (przybliżeniem) tego **parametru**. Zrozumiałe jest, że ze wzrostem liczności próby zwiększa się dokładność oszacowania parametru θ. Zatem estymator $\hat{\theta}_n$ powinien spełniać warunek

$$\forall_{\varepsilon>0} \lim_{n\to\infty} P\left(|\hat{\theta}_n - \theta| < \varepsilon\right) = 1.$$

Taki estymator $\hat{\theta}_n$ nazywamy **estymatorem zgodnym** parametru θ.

Na przykład estymatorem nieznanej wartości przeciętnej $\theta = E(X)$ dla populacji generalnej, dla której $|E(X)| < \infty$, jest statystyka

$$\hat{\theta}_n = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Estymatorami zgodnymi są też wszystkie estymatory

$$\hat{\theta}_n = \alpha_n \overline{X}, \qquad \text{gdy } \lim_{n\to\infty} \alpha_n = 1.$$

Estymator $\hat{\theta}_n$ nazywamy **estymatorem nieobciążonym** parametru θ, jeżeli dla każdego n mamy

$$E(\hat{\theta}_n) = \theta.$$

Jeżeli istnieje $E(\hat{\theta}_n) \neq \theta$, to $\hat{\theta}_n$ nazywamy **estymatorem obciążonym** parametru θ, a różnicę

$$B_n(\theta) = E(\hat{\theta}_n) - \theta$$

obciążeniem estymatora. Jeśli $\lim_{n\to\infty} B_n(\theta) = 0$, to $\hat{\theta}_n$ nazywamy **estymatorem asymptotycznie nieobciążonym**.

Przykład 2.5 *Niech cecha X populacji generalnej ma wariancję σ^2 skończoną, różną od zera. Zbadać, czy wariancja empiryczna S^2 dla pobranej próbki $X_1, \ldots, X_n$ jest* **estymatorem** *obciążonym czy nieobciążonym nieznanej wariancji σ^2.*

Rozwiązanie. Wariancja empiryczna S^2 może być przedstawiona w postaci

$$\begin{aligned} S^2 &= \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left[(X_i - \mu) + (\mu - \overline{X})\right]^2 = \\ &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 + \frac{2(\mu - \overline{X})}{n}\sum_{i=1}^{n}(X_i - \mu) + (\mu - \overline{X})^2 = \\ &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - (\overline{X} - \mu)^2, \end{aligned}$$

gdzie $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ oraz $E(X_i) = \mu$ dla $i = 1, 2, \ldots, n$.

Ponieważ X_i są niezależnymi zmiennymi losowymi o tym samym rozkładzie co badana cecha X, więc

$$E(X_i) = \mu, \quad E\left((X_i - \mu)^2\right) = E\left((X - E(X))^2\right) = \sigma^2$$

dla $i = 1, 2, \ldots, n$. Stąd

$$E\left((\overline{X} - \mu)^2\right) = E\left(\frac{\sum_{i=1}^{n}(X_i - \mu)^2}{n^2}\right) = \frac{1}{n^2}\sum_{i=1}^{n}\left(E(X_i - \mu)^2\right) = \frac{n \cdot \sigma^2}{n^2} = \frac{\sigma^2}{n}$$

i dlatego

$$E(S^2) = \frac{1}{n}\sum_{i=1}^{n} E\left((X_i - \mu)^2\right) - E\left((\overline{X} - \mu)^2\right) = \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n}\sigma^2.$$

Zatem rozpatrywany estymator S^2 jest **estymatorem zgodnym obciążonym**, o obciążeniu $B_n(\sigma^2) = -\frac{1}{n}\sigma^2$.

$$\lim_{n\to\infty} B_n(\sigma^2) = -\lim_{n\to\infty}\frac{\sigma^2}{n} = 0,$$

więc estymator ten jest asymptotycznie nieobciążony.

Estymatorem nieobciążonym wariancji σ^2 jest estymator

$$S^{*2} = \frac{n}{n-1}S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2, \qquad n \geq 2.$$

Dla danego parametru θ może istnieć wiele estymatorów nieobciążonych. Jeżeli $\hat{\theta}_n^*$, $\hat{\theta}_n^{**}$ są dwoma takimi estymatorami nieobciążonymi parametru mającymi wariancje

$$V\left(\hat{\theta}_n^*\right) = D^2\left(\hat{\theta}_n^*\right), \qquad V\left(\hat{\theta}_n^{**}\right) = D^2\left(\hat{\theta}_n^{**}\right)$$

spełniające warunek

$$V\left(\hat{\theta}_n^*\right) < V\left(\hat{\theta}_n^{**}\right),$$

to mówimy, że estymator $\hat{\theta}_n^*$ jest **asymptotycznie efektywniejszym** estymatorem parametru θ niż estymator $\hat{\theta}_n^{**}$. Ten estymator, który ma najmniejszą wariancję, nazywamy **estymatorem efektywnym** lub **najefektywniejszym**.

Jeżeli $f = f(X, \theta)$ oznacza gęstość prawdopodobieństwa zmiennej losowej X (rozkładu parametru θ), to

$$nE\left[\frac{\partial}{\partial\theta}\ln f(X,\theta)\right]^2$$

nosi nazwę (miary) **informacji zawartej** w n próbce. Wariancja $V\left(\hat{\theta}_n^*\right)$ spełnia nierówność:

$$V\left(\hat{\theta}_n^*\right) \geq \frac{1}{nE\left[\frac{\partial}{\partial\theta}\ln f(X,\theta)\right]^2}.$$

Jeżeli istnieje estymator efektywny $\tilde{\theta}_n$ parametru θ, to **miarą efektywności nieobciążonego estymatora** $\hat{\theta}_n^*$ nazywamy liczbę

$$\text{ef}\,\hat{\theta}_n^* = \frac{V\left(\tilde{\theta}_n\right)}{V\left(\hat{\theta}_n^*\right)} = \frac{D^2\left(\tilde{\theta}_n\right)}{D^2\left(\hat{\theta}_n^*\right)}.$$

Mamy

$$0 < \text{ef}\,\hat{\theta}_n \leq 1$$

i równość ma miejsce jedynie dla **estymatora efektywnego**.

Estymator $\hat{\theta}_n$ nazywamy **estymatorem asymptotycznie efektywnym**, gdy

$$\lim_{n\to\infty}\text{ef}\,\hat{\theta}_n = 1.$$

Metody wyznaczania estymatorów

I. Metoda największej wiarygodności

Niech badana cecha X ma rozkład zależny od k nieznanych parametrów $\theta_1, \ldots, \theta_k$, które chcemy oszacować (przybliżyć) na podstawie n-elementowej próby prostej. Oznaczmy przez L tak zwaną **funkcję wiarygodności** daną wzorem

$$L = f(x_1, \theta_1, \ldots, \theta_k) \cdot f(x_2, \theta_1, \ldots, \theta_k) \cdot \ldots \cdot f(x_n, \theta_1, \ldots, \theta_k),$$

gdzie $f(x, \theta_1, \ldots, \theta_k)$ oznacza gęstość prawdopodobieństwa cechy typu ciągłego lub funkcję rozkładu prawdopodobieństwa w przypadku cechy typu skokowego.

Metoda największej wiarygodności polega na tym, że jako estymatory $\hat{\theta}_1, \ldots, \hat{\theta}_k$ parametrów $\theta_1, \ldots, \theta_k$ przyjmujemy te wartości, dla których funkcja wiarygodności L przyjmuje wartość największą. Ponieważ funkcja $\ln L$ przyjmuje wartość największą dla tych samych parametrów co funkcja L, więc zadanie sprowadza się do wyznaczenia maksimum funkcji $\ln L$. Jeśli f jest różniczkowalna, to wartości $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k$ maksymalizujące L muszą spełniać warunki konieczne na ekstremum, to znaczy spełniać układ równań

$$\frac{\partial \ln L}{\partial\theta_i} = 0, \quad i = 1, 2, \ldots, k$$

nazywany zwykle **równaniami wiarygodności.**

Aby to było maksimum, to forma kwadratowa

$$\left\{\sum_{i=1}^{k}\sum_{j=1}^{k}\left(\frac{\partial^2 \ln L}{\partial\theta_i\partial\theta_j}\right)h_ih_j\right\} \qquad \theta_i=\hat{\theta}_i,\ i=1,2,\ldots,k$$

musi być określona ujemnie (to znaczy przyjmować wartości ujemne dla wszystkich rzeczywistych h_i, h_j nie zerujących się jednocześnie). Dla $k=1$ forma jest określona ujemnie, gdy

$$\left(\frac{d^2\ln L}{d\theta^2}\right)_{\theta=\hat{\theta}} = (\ln L)''_{\theta=\hat{\theta}} < 0.$$

Warunek ten, przy $k=1$, gwarantuje, że jeżeli istnieje estymator efektywny $\hat{\theta}(X_1, X_2, \ldots, X_n)$ parametru θ_1, to jedynym rozwiązaniem równania wiarygodności $\frac{d}{d\theta_1}\ln L = 0$ jest właśnie ten estymator.

Przykład 2.6 *Badana cecha populacji generalnej ma rozkład Poissona*

$$p(x,\lambda) = P(X=x;\lambda) = \frac{\lambda^x}{x!}e^{-\lambda}, \quad x\in\mathbb{N}\cup\{0\}$$

o nieznanym parametrze λ. Na podstawie n-elementowej próby prostej wyznaczyć metodą największej wiarygodności estymator $\hat{\lambda}$ parametru λ tego rozkładu.

Rozwiązanie. Funkcja wiarygodności $\ln L$ przyjmuje postać

$$L = \prod_{i=1}^{n} p(X_i,\lambda) = \prod_{i=1}^{n}\frac{\lambda^{X_i}e^{-\lambda}}{X_i!} = \frac{\lambda^{\sum_{i=1}^n X_i}\cdot e^{-n\lambda}}{\prod_{i=1}^n X_i!},$$

więc

$$\ln L = \sum_{i=1}^{n} X_i\cdot\ln\lambda - n\lambda - \sum_{i=1}^{n}\ln X_i!\,.$$

Stąd

$$\begin{aligned}\frac{d\ln L}{d\lambda} &= \frac{1}{\lambda}\sum_{i=1}^{n}X_i - n\\ \frac{d^2\ln L}{d\lambda^2} &= -\frac{1}{\lambda^2}\sum_{i=1}^{n}X_i\,.\end{aligned}$$

Zatem estymatorem $\hat{\lambda}$ parametru λ uzyskanym metodą największej wiarygodności jest rozwiązanie (pierwiastek) równania wiarygodności

$$\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n}X_i = \overline{X}.$$

Estymator ten jest estymatorem nieobciążonym i efektywnym.

$$\begin{aligned}E(\hat{\lambda}) &= E\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right) = \frac{1}{n}\sum_{i=1}^{n}E(X_i) = \frac{1}{n}nE(X) = \lambda,\\ E(\hat{\lambda}^2) &= E\left(\frac{1}{n}\sum_{i=1}^{n}X_i^2\right) = E(X^2) = \lambda^2+\lambda,\\ V(\hat{\lambda}) &= E(\hat{\lambda}^2) - \left[E(\hat{\lambda})\right]^2 = \lambda^2+\lambda-\lambda^2 = \lambda = V(X).\end{aligned}$$

Zatem

$$\text{ef}\,\hat{\lambda} = \frac{V(\hat{\lambda})}{V(X)} = \frac{\lambda}{\lambda} = 1.$$

Przykład 2.7 *Dana jest populacja generalna, w której badana cecha X ma rozkład normalny $N(\mu,\sigma)$ o nieznanych μ i σ. Wyznaczyć metodą największej wiarygodności estymatory parametrów μ i σ^2 tego rozkładu na podstawie n-elementowej próbki.*

Rozwiązanie. Funkcja wiarygodności L ma postać

$$\begin{aligned}L &= \prod_{i=1}^{n}f(X_i,\mu,\sigma) = \prod_{i=1}^{n}\left[\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2\sigma^2}(X_i-\mu)}\right] =\\ &= \frac{1}{\sigma^n(2\pi)^{n/2}}\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2\right].\end{aligned}$$

Stąd

$$\ln L = \frac{-n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2.$$

Układ równań wiarygodności ma więc postać

$$\begin{cases}\dfrac{\partial}{\partial\mu}\ln L = \dfrac{1}{\sigma^2}\displaystyle\sum_{i=1}^{n}(X_i-\mu) = 0,\\ \dfrac{\partial}{\partial(\sigma^2)}\ln L = -\dfrac{n}{2\sigma^2} + \dfrac{1}{2\sigma^4}\displaystyle\sum_{i=1}^{n}(X_i-\mu)^2 = 0.\end{cases}$$

Rozwiązaniem tego układu jest para

$$\begin{aligned}\hat{\mu} &= \frac{1}{n}\sum_{i=1}^{n}X_i = \overline{X},\\ \hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2 = S^2.\end{aligned}$$

Zostało więc wykazane, że średnia arytmetyczna $\overline{X}$ wyników próby i wariancja S^2 wyników próby są estymatorami największej wiarygodności nieznanych parametrów: wartości przeciętnej μ i wariancji σ^2 rozkładu $N(\mu, \sigma)$, przy czym wiadomo, że $\hat{\mu} = \overline{X}$ jest estymatorem zgodnym, nieobciążonym i efektywnym, a wariancja σ^2 jest estymatorem zgodnym asymptotycznie nieobciążonym.

II. Metoda momentów

Metoda ta polega na porównaniu pewnej liczby momentów (zwykle kolejnych) rozkładu policzonego z próby z odpowiednimi momentami rozkładu populacji generalnej będącymi funkcjami nieznanych parametrów. Wykorzystujemy tyle momentów (zwykle początkowych), ile jest szacowanych (estymowanych) parametrów. Rozwiązując otrzymane układy równań, otrzymujemy estymatory tych parametrów.

Rozpatrzmy to na przykładzie.

Przykład 2.8 *Niech badana cecha X ma rozkład Γ (gamma).*

$$f(x;p,\beta) = \begin{cases} \frac{\beta^p}{\Gamma(p)} x^{p-1} e^{-\beta x} & dla \quad x > 0 \\ 0 & dla \quad x \leq 0 \end{cases},$$

gdzie $p > 0$, $\beta > 0$ są nieznanymi parametrami. Na podstawie n-elementowej próbki prostej, wybranej z populacji generalnej, w której cecha X ma dany rozkład, wyznaczyć metodą momentów estymatory $\hat{p}$, $\hat{\beta}$ parametrów p i β.

Rozwiązanie. Pierwsze dwa momenty zwykłe rozkładu gamma m_1, m_2 są równe

$$m_1 = E(X) = \frac{p}{\beta}, \qquad m_2 = E(X^2) = \frac{p(p+1)}{\beta^2}.$$

Stąd otrzymujemy równania

$$\frac{p}{\beta} = \overline{X}, \qquad \frac{p(p+1)}{\beta^2} = \frac{1}{n}\sum X_i^2.$$

Rozwiązując, otrzymujemy

$$\hat{p} = \frac{\overline{X}^2}{S^2}, \quad \hat{\beta} = \frac{\overline{X}}{S^2},$$

gdzie

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2.$$

Estymatory otrzymane metodą momentów nie mają zwykle dużej efektywności, ale stosujemy je ze względu na łatwość obliczeń.

Przegląd podstawowych estymatorów

Najczęściej stosowanymi estymatorami w badaniach statystycznych, ze względu na jedną cechę, są estymatory wartości przeciętnej i wariancji rozpatrywanej cechy populacji. Nie zawsze jednak badania są prowadzone ze względu na cechę mierzalną. Czasem badana cecha ma charakter jakościowy. Wtedy zamiast wartości liczbowej z badania próbnego uzyskujemy jedynie informację o tym, czy dany element ma wyróżnioną cechę, czy nie. Podstawowym parametrem szacowanym w tym przypadku jest prawdopodobieństwo sukcesu (tego, że element ma wyróżnioną cechę), nazywane **frakcją** θ, która po pomnożeniu przez 100 daje procent elementów wyróżnionych (mających badaną cechę) w populacji generalnej. Zadanie sprowadza się tutaj do estymacji parametru θ w rozkładzie dwumianowym

$$P(k;n,\theta) = \binom{n}{k}\theta^k(1-\theta)^{n-k}.$$

W przypadku szacowania parametru θ na podstawie n-elementowej próby prostej, estymatorem zgodnym, nieobciążonym i efektywnym jest częstość względna

$$\hat{\theta} = \frac{k_n}{n},$$

gdzie k_n jest liczbą elementów wyróżnionych (mających daną cechę) wśród n-elementowej próbki.

Przedstawimy w postaci tabeli (patrz str. 84) najczęściej występujące estymatory parametrów lub ich funkcje, utworzone na podstawie n-elementowej próbki prostej, zakładając ich istnienie w populacji generalnej.

Estymacja przedziałowa

Metody estymacji punktowej pozwalają uzyskiwać oceny – przybliżenia punktowe parametrów rozkładu, przy czym nie potrafimy dać odpowiedzi na pytania, jaka jest dokładność danej cechy – przybliżenia.

Innym sposobem estymacji, dającym możliwość oceny jej dokładności, jest **metoda przedziałowa** polegająca na podaniu (wyznaczeniu) tak zwanych **przedziałów ufności** dla nieznanych parametrów rozkładu (bądź funkcji tych parametrów).

Przypomnijmy, że przedziałem ufności α, $0 < \alpha < 1$, nazywamy przedział (θ_1, θ_2) spełniający warunki:

1° Jego końce $\theta_1 = \theta_1(X_1, \ldots, X_n)$, $\theta_2 = \theta_2(X_1, \ldots, X_n)$ są funkcjami próby losowej i nie zależą od szacowanego parametru θ.

Tabela wybranych estymatorów

Nieznany parametr populacji	Estymator	Własność	Dla jakiej rodziny rozkładów
Wartość przeciętna (oczekiwana) μ	$\overline{X} = \frac{1}{n}\sum_1^n X_i$	zgodny, nieobciążony	rozkład dowolny, dla rozkładu $N(\mu,\sigma)$ również efektywny
	mediana z próby	zgodny, asymptotycznie nieobciążony	rozkład dowolny, dla rozkładu $N(\mu,\sigma)$ efektywność równa $\approx 0,64$
Wariancja σ^2, gdy μ znane	$S_1^2 = \frac{1}{n}\sum_1^n (X_i-\mu)^2$	zgodny, nieobciążony	rozkład dowolny, dla rozkładu $N(\mu,\sigma)$ również efektywny
Wariancja σ^2, gdy μ nieznane	$S^2 = \frac{1}{n}\sum_1^n (X_i-\overline{X})^2$	zgodny, asymptotycznie nieobciążony	rozkład dowolny
	$S^{*2} = \frac{1}{n-1}\sum_1^n (X_i-\overline{X})^2$	zgodny, nieobciążony	rozkład dowolny, dla rozkładu $N(\mu,\sigma)$ efektywność równa $\frac{n-1}{n}$
Odchylenie standardowe	S_1, S, S^*	zgodny	rozkład dowolny
	$b_n S, c_n S^*(\alpha)$	zgodny, nieobciążony asymptotycznie efektywny	rozkład normalny
	$Rd_n = (X_{\max} - X_{\min})d_n(\alpha)$	zgodny, nieobciążony asymp. efekt.= 0	rozkład normalny

2° Prawdopodobieństwo pokrycia przez ten przedział nieznanego parametru θ jest równe α, to znaczy

$$P(\theta_1 < \theta < \theta_2) = \alpha.$$

Liczbę α nazywamy **poziomem ufności.**

Wyznaczenie przedziału ufności to jednocześnie wyznaczenie takiego przedziału, że każdy jego element jest estymatorem nieznanego – badanego parametru i to takim przybliżeniem – estymatorem, że różnica pomiędzy rzeczywistą wartością parametru θ a jego estymatorem $\hat{\theta}$ nie przekracza długości przedziału ufności $|\hat{\theta} - \theta| < \theta_2 - \theta_1$.

Konstrukcję przedziału ufności omawialiśmy już w poprzednich rozważaniach.

Jest on (przedział ufności) wykorzystywany do weryfikacji hipotez statystycznych.

2.3 Weryfikacja hipotez

Drugim, po estymacji, celem wnioskowania statystycznego jest weryfikacja hipotez statystycznych (podejmowanie decyzji o prawdziwości albo fałszywości hipotez).

Hipotezą statystyczną nazywamy każde przypuszczenie dotyczące nieznanego rozkładu badanej cechy populacji. O prawdziwości lub fałszywości hipotezy wnioskuje się na podstawie pobranej próbki. Przypuszczenia te najczęściej dotyczą postaci rozkładu badanej cechy lub wartości jego parametrów.

Rozważmy następujące sytuacje:

a) Wiadomo, że badana cecha X ma rozkład wykładniczy o nieznanej wartości przeciętnej. Wysuwamy hipotezę, że wartość przeciętna $E(X) = 5$.

b) Niech T oznacza czas pomiędzy dwoma kolejnymi zgłoszeniami do centrali pogotowia ratunkowego. Wysuwamy hipotezę, że rozkład tego czasu jest wykładniczy.

c) Wiadomo, że badana cecha X ma rozkład normalny $N(20,\sigma)$ o nieznanej wariancji σ^2. Wysuwamy hipotezę, że wariancja σ^2 nie przekracza σ_0^2, np. $\sigma^2 \leq 9$.

d) Dane są dwa zbiory obserwacji, na przykład wyniki pomiarów tej samej cechy elementów wyprodukowanych w dwu różnych (konkurencyjnych) fabrykach. Wysuwamy hipotezę, że oba zbiory można traktować jako pochodzące z populacji o tym samym rozkładzie badanej cechy.

Hipotezy, które dotyczą wyłącznie wartości parametru (bądź parametrów) określonej klasy rozkładów, nazywamy **hipotezami parametrycznymi** (hipotezy a) i c)); każdą hipotezę, która nie jest parametryczna, nazywamy **hipotezą nieparametryczną** (hipotezy b) i d)).

Jeżeli hipoteza parametryczna precyzuje dokładnie wartości nieznanych parametrów rozkładu badanej cechy, to nazywamy ją **hipotezą prostą** (hipoteza a)), w przeciwnym przypadku mamy do czynienia z **hipotezą złożoną** (hipoteza c)).

Weryfikację hipotez statystycznych przeprowadzamy na podstawie wyników próby losowej. Metodę postępowania, która każdej możliwej realizacji próbki $X_1, X_2, \ldots, X_n$ przyporządkowuje, z ustalonym prawdopodobieństwem, decyzję przyjęcia albo odrzucenia sprawdzanej hipotezy, nazywamy **testowaniem statystycznym.**

W praktyce zwykle, oprócz weryfikowanej hipotezy H, wyróżnia się jeszcze inną hipotezę, K zwaną **hipotezą alternatywną.** Hipoteza K jest taką hipo-

tezą, którą w danym zagadnieniu skłonni jesteśmy przyjąć, jeżeli weryfikowaną hipotezę H należy odrzucić.

Sytuację, wobec której stoimy weryfikując hipotezę, można przedstawić następująco:

Hipoteza H jest prawdziwa	Hipoteza H jest fałszywa
Przyjąć hipotezę H – – decyzja poprawna	Odrzucić hipotezę H – – decyzja poprawna
Odrzucić hipotezę H – – decyzja błędna **Błąd pierwszego rodzaju**	Przyjąć hipotezę H – – decyzja błędna **Błąd drugiego rodzaju**

Chcemy, żeby prawdopodobieństwa błędów obu rodzajów były możliwie małe. Nie można jednak zmniejszyć jednocześnie obu rodzajów błędów przy ustalonej liczności próby.

Ustalmy więc z góry jakieś małe prawdopodobieństwo α **błędu pierwszego rodzaju**, które nazywamy **poziomem istotności testu**. Zazwyczaj przyjmujemy $\alpha = 0,01$ albo $\alpha = 0,05$ Następnie, dla ustalonego poziomu istotności α, przy wykorzystaniu statystyki testowej $\delta(X_1, \ldots, X_n)$, wyznacza się **zbiór krytyczny** W, aby w przypadku, gdy prosta hipoteza H jest prawdziwa, był spełniony warunek

$$P(\delta(X_1, \ldots, X_n) \in W) = \alpha.$$

Gdy hipoteza H nie jest prosta lub gdy zmienna jest typu skokowego, wówczas nie zawsze jest możliwa równość i zbiór krytyczny określamy warunkiem

$$P(\delta(X_1, \ldots, X_n) \in W) \leq \alpha.$$

Uwaga. Wybór zbioru (testowego) krytycznego może być zazwyczaj dokonany na nieskończenie wiele sposobów. Wobec wielu możliwości wyboru zbioru krytycznego najbardzej celowy byłby wybór zbioru W w taki sposób, aby minimalizował prawdopodobieństwo błędu drugiego rodzaju.

Niech alternatywną – konkurencyjną hipotezą względem hipotezy H będzie prosta hipoteza K. Wtedy prawdopodobieństwo β błędu drugiego rodzaju wyraża się wzorem

$$\beta = P(\delta \in \overline{W}|K) = 1 - P(\delta \in W|K),$$

$$\delta = \delta(X) = \delta(X_1, \ldots, X_n), \quad \overline{W} = \mathbb{R} \backslash W,$$

$P(\delta \in W|K)$ = prawdopodobieństwo, że $\delta \in W$ przy założeniu, że prawdziwa jest hipoteza alternatywna K.

Zatem zbiór krytyczny W należy tak dobrać, żeby $P(\delta \in W|K)$ było maksymalne (wtedy β jest minimalne).

Test, który przy ustalonym prawdopodobieństwie błędu pierwszego rodzaju minimalizuje prawdopodobieństwo błędu drugiego rodzaju, nazywamy **testem najmocniejszym** dla hipotezy H względem hipotezy alternatywnej K.

Jeżeli natomiast test jest najmocniejszy względem każdej hipotezy alternatywnej ze zbioru hipotez K, a więc w przypadku hipotezy złożonej, to nazywamy go **testem jednostajnie najmocniejszym** względem K.

Do ilustracji poprzednich rozważań przyjmijmy następującą sytuację: przypuśćmy, że pobrano próbkę n-elementową o wynikach $x_1, x_2, \ldots, x_n$ i dla niej obliczono wartość statystyki testowej $\delta_0 = \delta(x_1, x_2, \ldots, x_n)$. Załóżmy, że próbka ta może pochodzić z rozkładu o gęstości $f_1(x)$ albo z rozkładu o gęstości $f_2(x) \neq f_1(x)$.

Możemy zatem rozważyć dwie hipotezy:

H: próbka pochodzi z rozkładu o gęstości $f_1(x)$,

K: próbka pochodzi z rozkładu o gęstości $f_2(x)$.

Kiedy prawdziwa jest hipoteza H, wtedy statystyka testowa $\delta(X_1, X_2, \ldots, X_n)$ ma rozkład $g_1(\delta)$, natomiast gdy prawdziwa jest hipoteza K, wówczas $\delta(X_1, X_2, \ldots, X_n)$ ma rozkład $g_2(\delta)$.

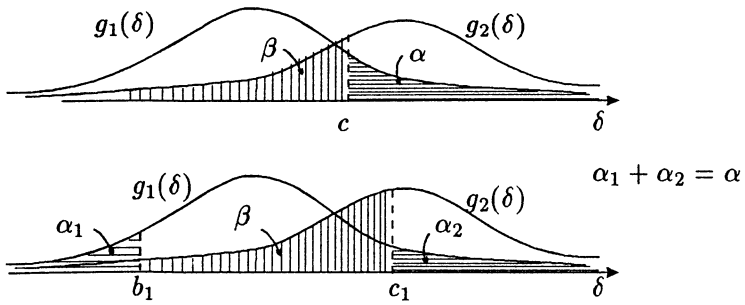
Rozpatrzmy teraz dwa sposoby konstrukcji zbioru krytycznego W.

1. W przypadku prawdziwości hipotezy H, przy założonym poziomie istotności α, za zbiór krytyczny W_1 przyjmujemy przedział $[c, +\infty)$, przy spełnieniu warunku

$$\int_c^{+\infty} g_1(\delta)d\delta = \alpha.$$

Jeśli hipoteza H jest prawdziwa (a tym samym hipoteza alternatywna K fałszywa), to jest możliwe otrzymanie z próby takiej wartości δ_0 statystyki δ, która "wpadnie" do zbioru W_1 odrzuceń hipotezy H (co prawda prawdopodobieństwo takiego zdarzenia jest małe i wynosi α). Uznajemy więc tę hipotezę za fałszywą, mimo że w rzeczywistości jest prawdziwa, a przyjmujemy tym samym hipotezę alternatywną K, mimo że jest ona fałszywa. Prawdopodobieństwo błędu drugiego rodzaju wynosi

$$\beta = \int_{-\infty}^{c} g_2(\delta)d\delta$$



Rys. 2.2. Prawdopodobieństwo β błędu drugiego rodzaju przy różnych zbiorach krytycznych testu uzyskanych dla tego samego poziomu istotności α

i jest równe polu obszaru zakreskowanego poziomo (Rys 2.2) Warto zwrócić uwagę na fakt, że gdybyśmy zmniejszyli prawdopodobieństwo błędu pierwszego rodzaju α (większe c), wówczas zwiększyłby się błąd drugiego rodzaju β.

2. Przy tym samym poziomie ufności α za zbiór krytyczny W_2 przyjmujemy sumę przedziałów $(-\infty, b_1] \cup [c_1, +\infty)$, takich że

$$\int_{-\infty}^{b_1} g_1(\delta)d\delta + \int_{c_1}^{+\infty} g_1(\delta)d\delta = \alpha_1 + \alpha_2 = \alpha.$$

Wtedy

$$\tilde{\beta} = \int_{b_1}^{c_1} g_2(\delta)d\delta.$$

I jak można zauważyć na Rys 2.2, jest ono większe niż w przypadku poprzednim.

Porównując oba przypadki, stwierdzamy, że test w przypadku pierwszym jest mocniejszy niż w przypadku drugim, bo przy tym samym poziomie istotności α daje mniejsze prawdopodobieństwo błędu drugiego rodzaju.

Aby skonstruować test statystyczny pozwalający weryfikować hipotezę H, należy zatem określić następujące elementy:

a) wybrać statystykę testową stosownie do treści postawionej hipotezy H,

b) ustalić dopuszczalne prawdopodobieństwo α błędu pierwszego rodzaju (poziom istotności testu),

c) określić hipotezę alternatywną,

d) wyznaczyć zbiór krytyczny W tak, aby przy wybranym poziomie istotności α zminimalizować prawdopodobieństwo błędu drugiego rodzaju.

Rozważany przypadek dotyczył konstrukcji testu, gdy hipoteza alternatywna była hipotezą prostą. Zagadnienie komplikuje się, gdy hipoteza alternatywna jest hipotezą złożoną.

Rozważmy przypadek hipotezy dotyczącej pewnego parametru θ w rozkładzie badanej cechy populacji. Niech zbiorem krytycznym testu dla hipotezy $H : \theta = \theta_0$ będzie zbiór W, taki że

$$P(\delta(X_1, \ldots, X_n) \in W|\theta_0) = \alpha.$$

Symbolem $M(\theta, W)$ oznaczamy prawdopodobieństwo tego, że funkcja testowa $\delta \in W$, przy założeniu, że nieznany parametr jest równy θ, czyli

$$M(\theta, W) = P(\delta \in W|\theta).$$

Funkcję tę, jako funkcję zmiennej θ, nazywamy **mocą testu.**

Przykład 2.9 *Badana cecha ma rozkład $N(\mu, \sigma)$ o znanej wariancji σ^2. Weryfikujemy hipotezę $H : \mu = \mu_0$ na podstawie 25-elementowej próby, przy wykorzystaniu jako funkcji testowej statystyki*

$$U = \frac{\overline{X} - \mu_0}{\sigma}\sqrt{n} = \frac{5(\overline{X} - \mu_0)}{\sigma}.$$

Dla poziomu istotności $\alpha = 0,05$ wyznaczyć moc testu w przypadku, gdy:

a) zbiorem krytycznym W jest przedział $[c, +\infty)$, gdzie c spełnia warunek $P(U \geq c|\mu_0) = \alpha$,

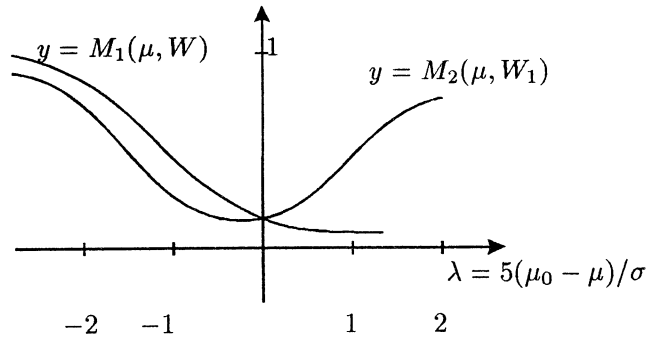
b) zbiorem krytycznym W_1 jest suma przedziałów $(-\infty, -c_1] \cup [c_1, +\infty)$, gdzie c_1 spełnia warunek

$$P(U \leq -c_1|\mu_0) + P(U \geq c_1|\mu_0) = P(|U| \geq c_1|\mu_0) = \alpha.$$

Rozwiązanie.

a) Ponieważ zmienna U ma rozkład zbliżony do rozkładu normalnego standaryzowanego (Twierdzenia graniczne), zatem z tablic rozkładu normalnego $N(0,1)$ przy $\alpha = 0,05$ znajdujemy liczbę c, taką że $\Phi(c) = 1 - \alpha$, $c = 1,64$. Zatem $W = [1,64; +\infty)$. Moc testu wyraża się następująco:

$$\begin{aligned} M_1(\mu, W) &= P(U \geq 1,64|\mu) = P\left(\frac{5(\overline{X} - \mu + \mu - \mu_0)}{\sigma} \geq 1,64|\mu\right) = \\ &= P\left(\frac{5(\overline{X} - \mu)}{\sigma} \geq 1,64 + \frac{5(\mu_0 - \mu)}{\sigma}\middle|\mu\right) = \\ &= P(U > 1,64 + \lambda|\mu) = 1 - \Phi(1,64 + \lambda), \end{aligned}$$



Rys. 2.3. Porównanie mocy testów z Przykładu 2.9

gdzie $\lambda = \dfrac{5(\mu_0 - \mu)}{\sigma}$, Φ zaś jest dystrybuantą rozkładu $N(0,1)$.

b) Z tablic rozkładu normalnego znajdujemy $c_1 = 1,96$ spełniające warunek $P(|U| > c_1|\mu_0) = 0,05$, $c_1 = 1,96$, i dlatego

$$W_1 = (-\infty, -1,96] \cup [1,96\,, +\infty).$$

Moc testu teraz wynosi

$$\begin{aligned} M_2(\mu, W_1) = P(|U| \geq 1,96|\mu) &= P\left(\left|\frac{5(\overline{X} - \mu + \mu - \mu_0)}{\sigma}\right| \geq 1,96\middle|\mu\right) = \\ &= P\left(\frac{5(\overline{X} - \mu)}{\sigma} \leq -1,96 + \lambda\right) + P\left(\frac{5(\overline{X} - \mu)}{\sigma} \geq 1,96 + \lambda\right) = \\ &= 1 - P\left(\lambda - 1,96 \leq \frac{5(\overline{X} - \mu)}{\sigma} \leq \lambda + 1,96\right) = \\ &= 1 - \Phi(\lambda + 1,96) + \Phi(\lambda - 1,96). \end{aligned}$$

Wykresy mocy testów dla obu przypadków przedstawiono na Rys. 2.3. Przy $\lambda < 0$ test z punktu a) ma moc większą niż test z punktu **b)**, natomiast gdy $\lambda > 0$, ma mniejszą moc.

Jeżeli zatem hipotezę H będziemy weryfikowali wobec hipotezy alternatywnej $K : \mu > \mu_0$, czyli $\lambda > 0$, to należy korzystać z pierwszego z tych testów. Jeżeli jednak hipotezą alternatywną będzie hipoteza $K : \mu \neq \mu_0$, to należy korzystać z drugiego z tych testów.

2.4 Weryfikacja hipotez dotyczących wartości przeciętnej

Model 1 – przy znanej wariancji σ^2

Rozważmy hipotezę o wartości przeciętnej:

$$H : \mu = \mu_0.$$

W testach rozpatrujemy tę hipotezę wobec hipotez alternatywnych:

a) K: $\mu = \mu_1 > \mu_0$,
b) K: $\mu = \mu_1 < \mu_0$,
c) K: $\mu = \mu_1 \neq \mu_0$

z ustalonym poziomem istotności α, $\alpha \in (0,1)$.

Do weryfikacji tej hipotezy, gdy cecha X populacji generalnej ma rozkład normalny $N(\mu, \sigma)$ przy nieznanym μ i znanym σ, wykorzystujemy statystykę testową U określoną wzorem

$$U = \frac{\overline{X} - \mu_0}{\sigma}\sqrt{n},$$

która przy założeniu prawdziwości hipotezy $H : \mu = \mu_0$ jest standaryzowaną zmienną losową $N(0,1)$.

Zbiór krytyczny testu, który przy danym poziomie istotności α minimalizuje prawdopodobieństwo popełnienia błędu, wyznacza się – jeżeli istnieje – w zależności od wartości μ_1 określonej hipotezą alternatywną. Można wykazać, że najmocniejszy test do zweryfikowania hipotezy $H : \mu = \mu_0$ przy alternatywnej hipotezie $K : \mu = \mu_1$ jest taki sam dla wszystkich hipotez K (jest ich nieskończenie wiele), w których $\mu_1 > \mu_0$, a najlepszym zbiorem krytycznym jest przedział $[u(1-\alpha), +\infty)$, gdzie $u(1-\alpha)$ jest **kwantylem rzędu** $(1-\alpha)$ rozkładu $N(0,1)$, to znaczy

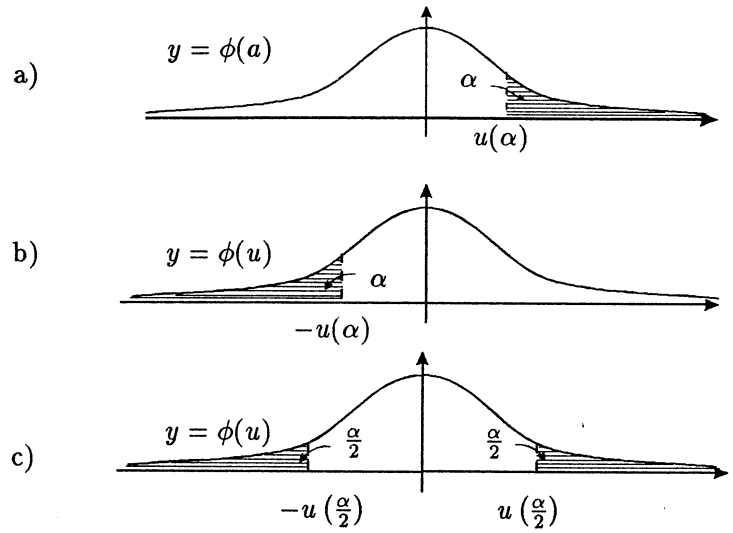
$$P(N > u(1-\alpha)) = \alpha, \quad N \text{ oznacza zmienną o rozkładzie } N(0,1).$$

Oczywiście $u(1-\alpha) + u(\alpha) = 1$.

Test ten nazywamy **testem jednostronnym z prawostronnym przedziałem krytycznym**, patrz Rys. 2.4 a).

Gdy hipotezą alternatywną jest $K : \mu_1 < \mu_0$, wtedy zbiorem krytycznym jest przedział $(-\infty, u(\alpha)]$. Test ten jest nazywany **testem jednostronnym z lewostronnym przedziałem krytycznym**, patrz Rys. 2.4 b).

Z symetrii rozkładu normalnego $N(0,1)$ wynika, że $-u(1-\alpha) = u(\alpha)$ i przedział lewostronny ma postać $(-\infty, -u(1-\alpha)]$.



Rys. 2.4. Testy do weryfikacji hipotezy o wartości przeciętnej $H : \mu = \mu_0$ wobec hipotezy alternatywnej a) $K : \mu > \mu_0$, b) $K : \mu < \mu_0$, c) $K : \mu \neq \mu_0$

Jeżeli hipotezą alternatywną jest $K : \mu = \mu_1 \neq \mu_0$, to zbiorem krytycznym jest suma przedziałów $(-\infty, -u(1-\alpha/2)] \cup [u(1-\alpha/2), +\infty)$. Test ten nosi nazwę **testu dwustronnego**, patrz Rys. 2.4 c).

Hipotezę $H : \mu = \mu_0$ odrzucamy, gdy obliczona z próby wartość statystyki testowej

$$U = \frac{\overline{X} - \mu_0}{\sigma}\sqrt{n}$$

należy do zbioru krytycznego (zależnie od wyboru testu i poziomu istotności). W przeciwnym przypadku twierdzimy tylko, że nie mamy podstaw do odrzucenia hipotezy $H : \mu = \mu_0$, co nie oznacza, że hipoteza H jest prawdziwa. Oznacza to tylko, że wyniki tej próby, na podstawie której weryfikujemy hipotezę H przy przyjętym "ryzyku błędu" α, nie przeczą tej hipotezie. Jeżeli hipoteza H jest prawdziwa, to błędna decyzja jej odrzucenia zdarza się w dużej liczbie takich decyzji przeciętnie $(n\alpha)$ razy, na przykład przy poziomie istotności $\alpha = 0,05$ zdarza się $100 \cdot 0,05 = 5$ razy na sto decyzji.

Do podjęcia decyzji "przyjęcie hipotezy H" nie wystarcza więc znajomość – przyjętego przed weryfikacją – poziomu istotności, ale konieczna jest informacja dotycząca prawdopodobieństwa przyjęcia weryfikowanej hipotezy za

prawdziwą, gdy w rzeczywistości jest ona fałszywa. Prawdopodobieństwo popełnienia tego błędu, jak już zostało wspomniane, nosi nazwę błędu drugiego rodzaju i jest zwykle oznaczane przez β, podczas gdy prawdopodobieństwo nieprzyjęcia za prawdziwą hipotezy, która jest w rzeczywistości prawdziwa, nosi nazwę błędu pierwszego rodzaju (poziomu istotności) i jest oznaczane zwykle przez α – patrz tabela na stronie 86.

Przykład 2.10 *Z populacji, w której badana cecha ma rozkład normalny* $N(\mu, 4)$, *wylosowano próbkę złożoną z* 16 *obserwacji i wyznaczono* $\overline{x} = 1,5$. *Na poziomie istotności* $\alpha = 0,05$ *zweryfikować hipotezę* $H : \mu = 2$ *przy hipotezie alternatywnej* $K : \mu = \mu_1 < 2$.

Rozwiązanie. Obliczamy wartość statystyki testowej

$$U_0 = \frac{\overline{x} - \mu_0}{\sigma}\sqrt{n} = \frac{1.5 - 2}{16}\sqrt{4} = -0,5.$$

Z tablic kwantyli rozkładu $N(0,1)$ odczytujemy wartość $u(1-\alpha) = u(0,95) = 1,64$. Zbiorem krytycznym lewostronnym jest więc przedział $(-\infty, -1,64] = W$. Zatem

$$U_0 \notin W$$

i na poziomie istotności 0,05 nie mamy podstaw do odrzucenia testowanej hipotezy.

Model 2 – przy nieznanej wariancji σ^2

Badana cecha X populacji generalnej ma rozkład $N(\mu, \sigma)$ o obu parametrach nieznanych. Weryfikujemy hipotezę $H : \mu = \mu_0$ wobec hipotezy alternatywnej

a) K: $\mu = \mu_1 > \mu_0$,
b) K: $\mu = \mu_1 < \mu_0$,
c) K: $\mu = \mu_1 \neq \mu_0$

z ustalonym poziomem istotności α, $\alpha \in (0,1)$.

Do weryfikacji tej hipotezy stosujemy test oparty na statystyce

$$t = \frac{\overline{X} - \mu_0}{S}\sqrt{n-1},$$

gdzie

$$S^2 = (S^*)^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \quad \text{lub} \quad S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

Wtedy zmienna t przy prawdziwości hipotezy H ma rozkład t Studenta o $(n-1)$ stopniach swobody.

Zbiór krytyczny testu, podobnie jak w poprzednim modelu, wyznacza się w zależności od położenia μ_1.

Gdy $\mu_1 > \mu_0$, wówczas zbiorem krytycznym prawostronnym jest przedział

$$[t(1-\alpha, n-1), +\infty),$$

gdzie $t(1-\alpha, n-1)$ jest kwantylem rzędu $(1-\alpha)$ rozkładu t Studenta przy $(n-1)$ stopniach swobody.

Gdy $\mu_1 < \mu_0$, wtedy zbiorem krytycznym lewostronnym jest przedział

$$[-\infty, -t(1-\alpha, n-1)],$$

gdzie $t(1-\alpha, n-1)$ jest kwantylem rzędu $(1-\alpha)$ rozkładu t Studenta przy $(n-1)$ stopniach swobody.

Gdy $\mu_1 \neq \mu_0$, wówczas zbiorem krytycznym dwustronnym jest suma przedziałów

$$\left(-\infty, -t\left(1-\frac{\alpha}{2}, n-1\right)\right] \cup \left[t\left(1-\frac{\alpha}{2}, n-1\right), +\infty\right),$$

gdzie $t\left(1-\frac{\alpha}{2}, n-1\right)$ jest kwantylem rzędu $\left(1-\frac{\alpha}{2}\right)$ rozkładu t Studenta przy $(n-1)$ stopniach swobody.

Jeśli t_{obl} wyliczone dla statystyki weryfikacyjnej należy do zbioru krytycznego, to hipotezę $H : \mu = \mu_0$ odrzucamy na korzyść odpowiedniej hipotezy alternatywnej K na poziomie istotności α. W przeciwnym przypadku nie ma podstaw do odrzucenia hipotezy H.

Przykład 2.11 *W celu ustalenia, czy dotychczasowa norma użytkowania pewnego narzędzia, wynosząca 150 dni, nie jest zbyt wysoka, zbadano faktyczny okres użytkowania na przykładzie losowo wybranych 65 sztuk tych narzędzi pracujących w normalnych warunkach. Otrzymamo średnią długość okresu użytkowania 139 dni oraz odchylenie standardowe $s = 9,8$ dni. Zakładając, że czas poprawnej pracy narzędzia ma rozkład normalny, stwierdzić na poziomie istotności $\alpha = 0,01$, czy uzyskane wyniki stanowią podstawę do zmiany, zmniejszenia normy.*

Rozwiązanie. Obliczając wartość statystyki testowej, otrzymujemy

$$t_{obl} = \frac{\overline{X} - \mu_0}{s}\sqrt{64} = \frac{139-150}{9,8} \cdot 8 = -8,98.$$

Ponieważ hipotezą alternatywną jest hipoteza $K : \mu < 150$, więc zbiorem krytycznym lewostronnym testu jest przedział

$$W = (-\infty, -t(1-\alpha, n-1)] = (-\infty, t(0,99,64)] = (-\infty, -2,385].$$

Mamy więc $t_{obl} = -8,98 \in (-\infty, -2,389] = W$. Zatem hipotezę $H : \mu = 150$ należy na poziomie istotności $\alpha = 0,01$ odrzucić na korzyść hipotezy alternatywnej $\mu < 150$, a zatem uzyskane rezultaty świadczą o tym, że norma była zbyt wysoka (można – należy ją obniżyć).

Model 3 – przy nieznanej wariancji σ^2 i dużej liczebności próby n

Badana cecha ma rozkład dowolny o nieznanej wartości przeciętnej μ i o skończonym, ale nieznanym, odchyleniu standardowym σ. Weryfikujemy hipotezę $H : \mu = \mu_0$ wobec hipotezy alternatywnej

a) K: $\mu = \mu_1 > \mu_0$,
b) K: $\mu = \mu_1 < \mu_0$,
c) K: $\mu = \mu_1 \neq \mu_0$.

Liczność próby $n \geq 100$.

Weryfikację hipotezy H w tym modelu przeprowadzamy testem analogicznym do testu w Modelu 1. Wiadomo, z tak zwanych praw wielkich liczb, że przy "dużych" n rozkład będzie bliski rozkładowi normalnemu, przy czym za σ możemy przyjąć odchylenie s obliczone z próby.

Przykład 2.12 *Zmierzono długość 196 włókien bawełny, a wyniki pomiarów zgrupowano w następującym szeregu rozdzielczym:*

Nr klasy	1	2	3	4	5	6	7
Środek przedziału	8	13	18	23	28	33	38
Liczność	2	9	18	70	75	19	3

Na poziomie istotności $\alpha = 0,01$ zweryfikować hipotezę, że średnia długość włókna dla całej badanej partii bawełny jest równa $\mu = 24$ wobec alternatywnej hipotezy $K : \mu \neq 24$.

Rozwiązanie. Obliczone wartości $\overline{x}$ i s^2 szeregu rozdzielczego odpowiednio wynoszą

$$\overline{x} = \frac{1}{196}(2 \cdot 8 + 9 \cdot 13 + \ldots + 3 \cdot 38) = 25,04,$$

$$s^2 \approx \frac{1}{196}\left(2 \cdot (17,04)^2 + 9 \ldots + 3 \cdot (12,96)^2\right) = 27,72.$$

Wartość statystyki U jest więc równa

$$u_{obl} = \frac{\overline{x} - \mu_0}{s}\sqrt{n} = \frac{25,04 - 24}{5,27}14 = 2,77.$$

Zbiorem krytycznym dwustronnym, wobec hipotezy H, jest zbiór

$$\begin{aligned} W &= \left(-\infty, -u\left(1-\frac{\alpha}{2}\right)\right] \cup \left[u\left(1-\frac{\alpha}{2}\right), +\infty\right) = \\ &= (-\infty, -u(0,895)] \cup [u(0,995), +\infty) = (-\infty; -2,58) \cup (2,58; +\infty). \end{aligned}$$

Zatem $u_{obl} = 2,77 \in W$ i odrzucamy weryfikowaną hipotezę $\mu = 24$ na rzecz hipotezy alternatywnej $\mu \neq 24$ na przyjętym poziomie istotności $\alpha = 0,01$.

2.5 Testy istotności dla wariancji

Analogicznie jak w przypadku weryfikacji hipotez dotyczących wartości przeciętnych, możemy rozpatrzeć testy do weryfikacji hipotez o wariancji – odchyleniu standardowym w zależności od przyjętych założeń, modeli.

Model 1 – przy dowolnej liczebności próby n

Badana cecha populacji ma rozkład normalny $N(\mu, \sigma)$ o nieznanych μ i σ. Weryfikujemy hipotezę $H : \sigma^2 = \sigma_0^2$ wobec hipotezy alternatywnej:

a) K: $\sigma^2 = \sigma_1^2 > \sigma_0^2$,
b) K: $\sigma^2 = \sigma_1^2 < \sigma_0^2$,
c) K: $\sigma^2 = \sigma_1^2 \neq \sigma_0^2$.

Do weryfikacji hipotezy H wykorzystuje się statystykę

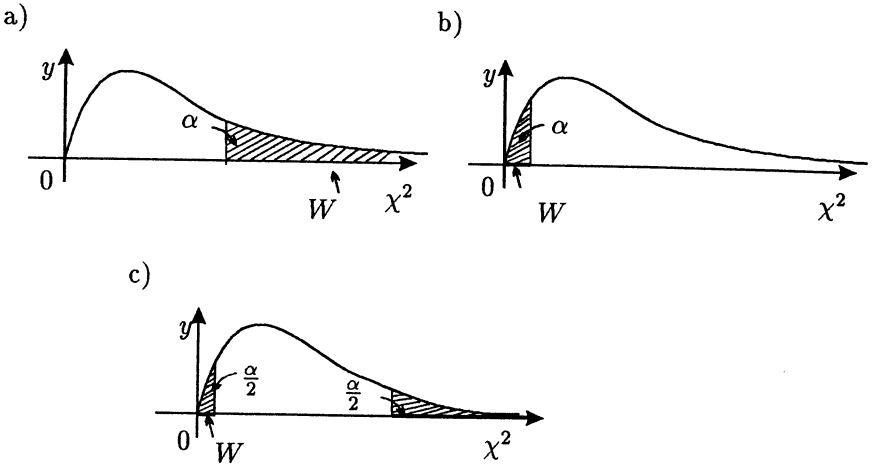
$$\chi^2 = \frac{nS^2}{\sigma_0^2},$$

która przy założeniu prawdziwości hipotezy H ma rozkład $\chi^2(n-1)$ – rozkład "chi kwadrat" o $(n-1)$ stopniach swobody.

W przypadku hipotezy alternatywnej $K : \sigma^2 > \sigma_0^2$ zbiorem krytycznym W jest przedział $[\chi^2(1-\alpha, n-1), +\infty)$, gdzie $\chi^2(1-\alpha, n-1)$ jest kwantylem rzędu $(1-\alpha)$ rozkładu χ^2 o $(n-1)$ stopniach swobody, patrz Rys. 2.5 a).

Gdy hipotezą alternatywną jest hipoteza $K : \sigma^2 < \sigma_0^2$, wówczas zbiorem krytycznym W jest przedział $(0, \chi^2(\alpha, n-1))$, patrz Rys. 2.5 b).

Jeśli hipotezą alternatywną jest $K : \sigma^2 \neq \sigma_0^2$, to zbiorem krytycznym W jest suma dwóch przedziałów $(0, \chi^2\left(\frac{\alpha}{2}, n-1\right)] \cup [\chi^2\left(1-\frac{\alpha}{2}, n-1\right), +\infty]$, patrz Rys. 2.5 c).



Rys. 2.5 Testy do weryfikacji hipotezy o odchyleniu standardowym $H : \sigma^2 = \sigma_0^2$ wobec hipotezy alternatywnej
a) $K : \sigma^2 > \sigma_0^2$, b) $K : \sigma^2 < \sigma_0^2$, c) $K : \sigma^2 \neq \sigma_0^2$

Hipotezę $H : \sigma^2 = \sigma_0^2$ odrzucamy na przyjętym poziomie ufności α, gdy obliczona z próby wartość statystyki $\chi^2 = nS^2/\sigma_0^2$ należy do zbioru krytycznego i przyjmujemy wtedy odpowiednią hipotezę alternatywną. W przeciwnym przypadku nie ma podstawy do odrzucenia hipotezy H.

Przykład 2.13 *W celu oszacowania dokładności pewnego przyrządu pomiarowego dokonano 8 pomiarów pewnej wielkości uzyskując wyniki*

$$18,17;\ 18,21;\ 18,05;\ 18,14;\ 18,19;\ 18,22;\ 18,06;\ 18,08.$$

Zweryfikować na poziomie istotności $\alpha = 0,05$ hipotezę $H : \sigma^2 = 0,06$ wobec hipotezy alternatywnej $\sigma^2 \neq 0,06$.

Rozwiązanie. Zbiorem krytycznym obustronnym jest suma przedziałów

$$\begin{aligned} W &= \left(0, \chi^2\left(\frac{\alpha}{2}, n-1\right)\right] \cup \left[\chi^2\left(1-\frac{\alpha}{2}, n-1\right), +\infty\right) = \\ &= \left(0, \chi^2(0,025, 7)\right] \cup \left[\chi^2(0,975, 7), +\infty\right) = \\ &= (0, 1,69] \cup [16,01, +\infty). \end{aligned}$$

Obliczona wartość wariancji z próby $s^2 = 0,064$ i wartość statystyki $\chi^2_{obl} = \frac{ns^2}{\sigma_0^2} = \frac{8 \cdot 0,0575}{0,06} = 8,54\ldots$ nie należy do W, nie ma więc podstaw do odrzucenia hipotezy $H : \sigma^2 = 0,06$.

Model 2 – przy liczebności próby $n \geq 50$

Badana cecha populacji ma rozkład $N(\mu, \sigma)$ o nieznanych parametrach μ, σ. Weryfikujemy hipotezę $H : \sigma^2 = \sigma_0^2$ wobec hipotezy alternatywnej:

a) K: $\sigma^2 = \sigma_1^2 > \sigma_0^2$,
b) K: $\sigma^2 = \sigma_1^2 < \sigma_0^2$,
c) K: $\sigma^2 = \sigma_1^2 \neq \sigma_0^2$.

Model ten wykorzystujemy, gdy liczność próby $n \geq 50$. Do weryfikacji hipotezy, jak poprzednio, stosujemy statystykę nS^2/σ_0^2 oraz fakt, że statystyka $\sqrt{2nS^2/\sigma_0^2}$ ma w przybliżeniu rozkład $N(\sqrt{2n-3}, 1)$, a więc że statystyka $U = \sqrt{2nS^2/\sigma_0^2} - \sqrt{2n-3}$ ma w przybliżeniu rozkład $N(0,1)$.

Zbiory krytyczne wyznacza się w zależności od hipotezy alternatywnej

a) $[u(1-\alpha), +\infty)$,
b) $(-\infty, -u(1-\alpha)]$,
c) $\left(-\infty, -u\left(1-\frac{\alpha}{2}\right)\right] \cup \left[u\left(1-\frac{\alpha}{2}\right), +\infty\right)$,

gdzie $u(p)$ oznacza, jak zwykle, kwantyl rzędu p rozkładu $N(0,1)$.

Przykład 2.14 *Do tarczy oddano* 50 *strzałów mierząc odległości od środka tarczy. Wyliczona wariancja tych odległości a* $s^2 = 107,3\ cm^2$. *Zakładając, że odległości mają rozkład normalny, zweryfikować na poziomie* $\alpha = 0,05$ *hipotezę* H, *że wariancja odchyleń trafień od środka tarczy przy tego rodzaju strzelaniu* $\sigma^2 = 100\ cm^2$, *jeśli hipotezą alternatywną jest hipoteza* $K : \sigma^2 > 100\ cm^2$.

Rozwiązanie. Zbiór krytyczny ma w tym przypadku postać

$$W = [u(1-\alpha), +\infty) = [u(0,95), +\infty) = [1,64\,, +\infty).$$

Wartość statystyki obliczeniowej

$$u_{obl} = \sqrt{\frac{2nS^2}{\sigma_0^2}} - \sqrt{2n-3} = \sqrt{\frac{2 \cdot 50 \cdot 107,3}{100}} - \sqrt{97} \approx 0,510.$$

Mamy więc $u_{obl} \notin W$, a zatem nie ma podstaw do odrzucenia hipotezy $H : \sigma^2 = 100$ przy poziomie istotności $\alpha = 0,05$.

Model 3 – przy liczebności próby $n \geq 100$

Badana cecha populacji ma rozkład dowolny o skończonej wariancji $\sigma^2 > 0$. Weryfikujemy hipotezę $H : \sigma^2 = \sigma_0^2$ wobec hipotezy alternatywnej:

a) K: $\sigma^2 = \sigma_1^2 > \sigma_0^2$,
b) K: $\sigma^2 = \sigma_1^2 < \sigma_0^2$,
c) K: $\sigma^2 = \sigma_1^2 \neq \sigma_0^2$.

Liczność próby $n \geq 100$.

W tym przypadku weryfikację hipotezy H opieramy na statystyce

$$U = \frac{S^{*2} - \sigma_0^2}{\sigma_0^2}\sqrt{\frac{n}{2}},$$

która przy założeniu prawdziwości hipotezy H ma dla dużych n w przybliżeniu rozkład $N(0,1)$. Zbiory krytyczne są więc określone identycznie jak w Modelu 2.

Przykład 2.15 *Z populacji generalnej pobrano 450-elementową próbkę i otrzymano* $s^{*2} = \frac{1}{n-1}\sum_1^n (x_i - \overline{x})^2 = 14,9$. *Na poziomie istotności* $\alpha = 0,05$ *zweryfikować hipotezę* $H : \sigma^2 = 16$, *jeśli hipotezą alternatywną jest* $K : \sigma^2 < 16$.

Rozwiązanie. Zbiór krytyczny ma postać

$$W = (-\infty, -u(1-\alpha)] = (-\infty, -1,64].$$

Wartość statystyki obliczeniowej u_{obl} wynosi

$$u_{obl} = \frac{s^{*2} - \sigma_0^2}{\sigma_0^2}\sqrt{\frac{n}{2}} = \frac{14,9 - 1,64}{16}\sqrt{225} = -1,03.$$

Wartość $u_0 = -1,03 \notin W = (-\infty\,, -1,64]$, a więc nie ma podstaw do odrzucenia hipotezy H na przyjętym poziomie istotności $\alpha = 0,05$.

2.6 Weryfikacja hipotezy o równości wartości przeciętnych badanej cechy

Model 1 – przy znanych wariancjach σ_1^2, σ_2^2

Badana cecha X ma w dwóch populacjach rozkłady $N(\mu_1, \sigma_1)$, $N(\mu_2, \sigma_2)$ odpowiednio o znanych σ_1, σ_2 i nieznanych μ_1, μ_2. Weryfikujemy hipotezę

$H : \mu_1 = \mu_2$, przy zadanym poziomie istotności α. Z obu populacji wybieramy dwie niezależne próby o licznościach równych n_1, n_2. Do weryfikacji tej hipotezy stosujemy test oparty na statystyce

$$U = (\overline{X_1} - \overline{X_2}) \Big/ \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} ,$$

gdzie $\overline{X_1}$, $\overline{X_2}$ są średnimi arytmetycznymi pobranych próbek. Statystyka ta przy założeniu prawdziwości tezy H ma rozkład $N(0,1)$. Zbiorami krytycznymi są, odpowiednio, przedziały

$$\begin{array}{lll} [u(1-\alpha), +\infty), & \text{gdy hipotezą alternatywną jest} & K : \mu_1 > \mu_2, \\ (-\infty, -u(1-\alpha)], & \text{---}\,\text{---}\,\text{---} & K : \mu_1 < \mu_2, \\ (-\infty, -u\,(1-\frac{\alpha}{2})] \cup [u\,(1-\frac{\alpha}{2})\,, +\infty)\,, & \text{---}\,\text{---}\,\text{---} & K : \mu_1 \neq \mu_2. \end{array}$$

Zatem wystarczy sprawdzić, czy u_{obl} obliczone z pobranych prób należy do odpowiedniego przedziału krytycznego W, i wyciągnąć stąd odpowiedne wnioski dotyczące odrzucenia lub nieodrzucenia badanej hipotezy H.

Przykład 2.16 *Na dwóch różnych wagach zważono po* 10 *kawałków* 100*-metrowych przędzy i uzyskano następujące rezultaty (wyniki w gramach):*

I waga: 5,25; 5,98; 5,83; 5,58; 5,35; 5,59; 5,41; 5,81; 5,95; 5,72;
II waga: 5,31; 5,13; 5,64; 5,89; 5,17; 5,18; 5,27; 5,73; 5,08; 5,24.

Wiadomo, że wariancja mas dla I wagi $\sigma_1^2 = 0,06$, *a dla drugiej* $\sigma_2^2 = 0,07$. *Zakładając, że rozpatrywana cecha wagi (waga stumetrowego kawałka) ma rozkład normalny, zweryfikować na poziomie istotności* $\alpha = 0,05$ *hipotezę* H, *że wartości przeciętne mas kawałków przędzy uzyskiwane przez te wagi są jednakowe, wobec hipotezy alternatywnej* $K : \mu_1 \neq \mu_2$.

Rozwiązanie. Zbiór krytyczny w tym przypadku jest sumą przedziałów

$$W = \left(-\infty, -u\left(1-\frac{\alpha}{2}\right)\right] \cup \left[u\left(1-\frac{\alpha}{2}\right), +\infty\right) = (-\infty\,, -1,98] \cup [1,98\,, +\infty).$$

Obliczenia na podstawie doświadczenia wskazują, że

$$u_{obl} = (5,65 - 5,36) \Big/ \sqrt{\frac{0,06}{10} + \frac{0,07}{10}} \approx 2,54,$$

a więc $u_{obl} \in W$, a zatem hipotezę o równości wartości przeciętnych na poziomie istotności $\alpha = 0,05$ należy odrzucić.

Model 2 – przy nieznanych wariancjach σ_1^2, σ_2^2

Badana cecha X ma w dwóch populacjach rozkłady

$$N(\mu_1, \sigma_1), \qquad N(\mu_2, \sigma_2)$$

odpowiednio o nieznanych μ_1, μ_2, σ_1, σ_2. Weryfikujemy hipotezę $H : \mu_1 = \mu_2$, na podstawie prób o dużych licznościach $n_1 \geq 100$, $n_2 \geq 100$. Test istotności dla sprawdzanej hipotezy H budujemy analogicznie jak dla Modelu 1, z tym że poprzednie znane wartości σ_1^2, σ_2^2 zastępujemy wartościami s_1^2, s_2^2 uzyskanymi z pobranych próbek:

$$U = (\overline{X_1} - \overline{X_2}) \Big/ \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} , \qquad u_{obl} = (\overline{x_1} - \overline{x_2}) \Big/ \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} .$$

Zmienna U ma w przybliżeniu rozkład normalny $N(0,1)$. Dalej postępujemy tak jak w Modelu 1.

Przykład 2.17 *Średnia prędkość autobusu miejskiego (w km/h) obliczona na podstawie zmierzonych, w środy, prędkości* 200 *autobusów była równa* 15,1, *natomiast średnia prędkość obliczona w niedziele była równa* 16,4. *Weryfikacja prędkości obliczona na podstawie tych wyników* $s_1^2 = 6,8$, $s_2^2 = 4,3$. *Na podstawie uzyskanych danych zweryfikować, na poziomie istotności* $\alpha = 0,05$, *hipotezę* $H : \mu_1 = \mu_2$, *że średnie prędkości autobusów w środy i w niedziele były takie same, jeśli hipotezą alternatywną jest hipoteza* $K : \mu_1 < \mu_2$, *że średnia prędkość w środy była mniejsza niż w niedziele.*

Rozwiązanie. Zbiór krytyczny w tym przypadku ma postać

$$W = (-\infty, -u(1-\alpha)] = (-\infty, -1,64].$$

Wartość obliczeniowa statystyki testowej u_{obl} wynosi

$$(\overline{x_1} - \overline{x_2}) \Big/ \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = -1,3 \Big/ \sqrt{\frac{6,8}{200} + \frac{4,3}{120}} \approx -4,69,$$

a więc należy do zbioru krytycznego W. Zatem hipotezę o równości średnich prędkości należy odrzucić na korzyść hipotezy alternatywnej K, że prędkość średnia w niedziele jest większa od średniej prędkości w środy.

Uwaga. Korzystając z poprzednich danych, możemy zweryfikować hipotezę $H : \mu_1 = \mu_2$ przy hipotezach alternatywnych: a) $K : \mu_1 > \mu_2$ oraz b) $K : \mu_1 \neq \mu_2$.

Odpowiednie zbiory krytyczne mają postać

$$W_a = [1,64, +\infty),$$
$$W_b = [-\infty, -1,96) \cup [1,96, +\infty).$$

Zatem $u_{obl} = -4,64 \notin W_a$, więc nie ma podstaw do odrzucenia hipotezy H przy hipotezie alternatywnej $K : \mu_1 > \mu_2$. Ponieważ $u_{obl} \in W_b$, więc hipotezę H należy odrzucić na rzecz hipotezy alternatywnej $K : \mu_1 \neq \mu_2$, że prędkości średnie autobusów w środy i niedziele są różne między sobą.

Uwaga. Możemy weryfikować jeszcze wiele różnych typów hipotez statystycznych, a w szczególności: hipotez o wskaźniku struktury populacji; hipotezy o równości wskaźników struktury dla dwóch populacji; hipotez o równości wielu wariancji.

2.7 Zasada najmniejszej sumy kwadratów

Zasada ta, wprowadzona w 1806 r. przez Legendre'a i w 1809 r. przez Gaussa, jest najczęściej stosowaną metodą oceny błędów – statystyczną oceną błędów. Daje ona praktycznie bardzo dobre rezultaty i jest oparta na następującym sposobie postępowania.

Przypuśćmy, że wyniki i-tego pomiaru y_i pewnej badanej wielkości możemy uważać za sumę odpowiedniej wielkości $\hat{y}_i$ przyjętego modelu matematycznego oraz błędu $e_i : y_i = \hat{y}_i + e_i$.

Przyjmijmy w najprostszym przypadku ten model jako model liniowy

$$\hat{y} = a + bx. \tag{2.1}$$

Dobieramy parametry modelu (2.1), to znaczy a, b tak, aby suma kwadratów błędów

$$S = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{2.2}$$

była możliwie najmniejsza.

Postać funkcyjną wielkości S daną wzorem (2.2) nazywa się często **kryterium "dobroci"** modelu. Należy znaleźć taki sposób doboru parametrów a, b, by sumy wartości bezwzględnych błędów e_i były małe lub sumy kwadratów błędów były małe.

Mamy

$$y_i = \hat{y}_i + e_i, \quad i = 1, 2, \ldots n$$

lub

$$(y_i - \hat{y}_i)^2 = (y_i - a - bx_i)^2.$$

Traktując y_i jako zadane i x_i jako znane, szukamy wielkości a, b takich, żeby funkcja

$$S = S(a,b) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

miała wartości możliwie małe, tj. osiągała minimum.

W celu wyznaczenia tego minimum traktujemy funkcję S jako funkcję dwóch zmiennych i stosujemy normalną procedurę wyznaczania minimum. Uznajemy, że x_i, y_i są ustalone i znane, więc różniczkując $S(a,b)$ względem a i b, otrzymujemy **układ równań normalnych** dający warunek konieczny na ekstremum funkcji:

$$\begin{cases} \dfrac{\partial S}{\partial a} = -2\sum_{i=1}^{n}(y_i - a - bx_i) = 0, \\ \dfrac{\partial S}{\partial b} = -2\sum_{i=1}^{n}(y_i - a - bx_i)x_i = 0 \end{cases}$$

lub równoważnie

$$\begin{cases} na + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i, \\ a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i. \end{cases}$$

Wyznaczając stąd a i b, mamy:

$$\begin{cases} b = \dfrac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2}, \\ a = \dfrac{1}{n}\sum_{i=1}^{n} y_i - b \cdot \dfrac{1}{n}\sum_{i=1}^{n} x_i. \end{cases}$$

Wprowadzając oznaczenia $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, możemy parametry a i b przedstawić w postaci

$$\begin{cases} b = \dfrac{\sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2} = \dfrac{\sum_{i=1}^{n}(x_i y_i - \overline{x}\,\overline{y})}{\sum_{i=1}^{n}(x_i^2 - \overline{x}^2)}, \\ a = \overline{y} - \overline{x}b = \overline{y} - \overline{x}\dfrac{\sum_{i=1}^{n}(x_i y_i - \overline{x}\,\overline{y})}{\sum_{i=1}^{n}(x_i^2 - \overline{x}^2)} \end{cases}$$

lub

$$\begin{cases} b = \dfrac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}, \\ a = \overline{y} - b\overline{x} = \overline{y} - \overline{x}\dfrac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}. \end{cases} \tag{2.3}$$

Zatem ten optymalny model ma postać

$$\hat{y} = a + bx, \text{ gdzie } a, b \text{ są dane w (2.3)}$$

lub

$$\hat{y} = (\overline{y} - b\overline{x}) + bx, \text{ gdzie } b \text{ jest dane w (2.3).}$$

Zauważmy, że w przypadku tego optymalnego modelu suma względnych wartości błędów wynosi

$$\sum_{i=1}^{n} e_i^2 = S(a,b) = \min_{c,d} S(c,d),$$

gdzie a, b są dane w (2.3). Wstawiając wyliczone a, b do $S(a,b)$, otrzymujemy sumę względną błędów

$$\begin{aligned}\sum_{i=1}^{n} e_i &= \sum_{i=1}^{n}(\hat{y}_i - y_i) = \sum_{i=1}^{n}(a + bx_i - y_i) = \\ &= na + nb\overline{x} - n\overline{y} = n(\overline{y} - a - b\overline{x}) = 0,\end{aligned}$$

to znaczy równą zeru, a suma kwadratów błędów wynosi

$$\begin{aligned}S(a,b) &= \sum_{i=1}^{n}(y_i - bx_i - \overline{y} + b\overline{x})^2 = \sum_{i=1}^{n}[y_i - \overline{y} - b(x_i - \overline{x})]^2 = \\ &= \sum_{i=1}^{n}(y_i - \overline{y})^2 + b^2\sum_{i=1}^{n}(x_i - \overline{x})^2 - 2b\sum_{i=1}^{n}(y_i - \overline{y})(x_i - \overline{x}) = \\ &= V(Y) + b^2V(X) - 2b\sum_{i=1}^{n} x_iy_i + 2b \cdot n \cdot \overline{x} \cdot \overline{y} = \\ &= V(Y) + 2bnE(Y)E(X) + b^2V(X) - 2b\sum_{i=1}^{n} x_iy_i.\end{aligned}$$

Przykład 2.18 *Dane z pomiarów pewnej zmiennej dwuwymiarowej można zapisać następująco:*

$$(x_1, y_1) = (0,0), \quad (x_2, y_2) = (1,1), \quad (x_3, y_3) = (1,2).$$

Przyjmijmy model:

$$\hat{y} = a + bx.$$

Wyznaczyć a, b tak, by model był optymalny.

Rozwiązanie. Z podanych pomiarów i przyjętego modelu otrzymujemy:

$$\begin{aligned}\hat{y}_1 &= a + b \cdot 0 = a, \\ \hat{y}_2 &= a + b \cdot 1 = a + b, \\ \hat{y}_3 &= a + b \cdot 1 = a + b\end{aligned}$$

oraz

$$\begin{aligned}e_1 = \hat{y}_1 - y_1 &= a - 0 = a, \\ e_2 = \hat{y}_2 - y_2 &= a + b - 1, \\ e_3 = \hat{y}_3 - y_3 &= a + b - 2.\end{aligned}$$

Zatem

$$S = S(a,b) = a^2 + (a + b - 1)^2 + (a + b - 2)^2 = 3a^2 + 2b^2 + 4ab - 6a - 6b + 3$$

i układ równań normalnych ma postać

$$\begin{cases} \dfrac{\partial S}{\partial a} = 6a + 4b - 6 = 0, \\ \dfrac{\partial S}{\partial b} = 4b + 4a - 6 = 0. \end{cases}$$

Rozwiązanie tego układu to $(a,b) = (0, 3/2)$.

Zatem optymalny model liniowy można opisać Rys. 2.6.

Suma kwadratów błędów osiąga minimum

$$\min_{a,b} S(a,b) = S\left(0, \frac{3}{2}\right) = 0^2 + \left(\frac{1}{2}\right)^2 + \left(-\frac{1}{2}\right)^2 = \frac{1}{2}.$$

Istotnie

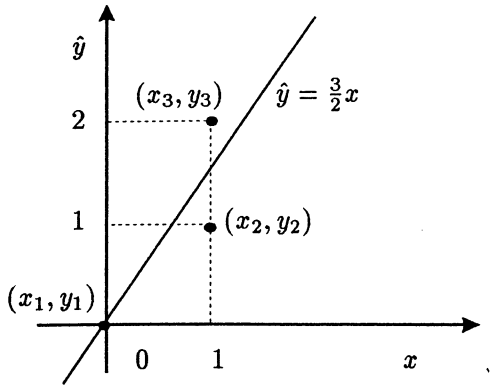
$$S(a,b) = 3a^2 + 2b^2 + 4ab - 6a - 6b + 3 \geq \frac{1}{2}.$$

Metoda najmniejszych kwadratów pozwala dla ustalonej, skończonej liczby punktów (x_k, y_k), $k = 1, 2, \ldots, n$, $x_k \neq x_l$ dla $k \neq l$, dobrać prostą

$$\hat{y} = a + bx$$

taką, żeby suma względnych wartości błędów

$$\sum_{i=1}^{n}(y_i - a - bx_i)$$



Rys. 2.6. Model liniowy dla najmniejszej sumy kwadratów błędów

była równa zeru, a suma kwadratów tych błędów

$$\sum_{i=1}^{n}(y_i - a - bx_i)^2$$

była możliwie najmniejsza, to znaczy dla każdej innej $\hat{y} = \tilde{a} + \tilde{b}x$ suma błędów jest większa. Dobrana prosta w pewnym sensie najlepiej aproksymuje liniową zależność pomiędzy x_i, y_i, $i = 1, 2, \ldots, n$.

Rozdział 3

Zadania

Zadanie 1. Iloma sposobami można ustawić 8 wież na zwykłej szachownicy, tak aby nie atakowały się wzajemnie?

Zadanie 2. Na przyjęciu gospodarz chce usadzić swoich gości przy podłużnym stole: 4 panie z jednej strony oraz 4 panów z drugiej strony.
a) Na ile sposobów może tego dokonać?
b) Jak sytuacja się zmieni, jeżeli stół będzie okrągły? (W dalszym ciągu zakładamy, że wszystkie panie siedzą kolejno oraz panowie kolejno.) Uwaga: przy okrągłym stole żadne miejsce nie jest wyróżnione.

Zadanie 3. Z talii 52 kart losujemy bez zwracania 13. Ile istnieje wyników losowania, w którym wylosujemy dwa asy?

Zadanie 4. Ile jest różnych liczb sześciocyfrowych utworzonych z cyfr $0, \ldots 5$ tak, że:
a) cyfry w liczbie się nie powtarzają, a cyfry 3 i 4 są na miejscu pierwszym lub ostatnim?
b) cyfry w liczbie mogą się powtarzać, cyfry 3 i 4 są na miejscu pierwszym lub ostatnim?

Zadanie 5. Znaleźć liczbę wszystkich podzbiorów zbioru n-elementowego.

Zadanie 6. Kości do gry w domino są znaczone dwoma liczbami. Ile różnych kości można utworzyć z liczb $0, 1, 2, \ldots, n$?

Zadanie 7. Na ile sposobów można zbudować wieżę z k białych i n czarnych ($k < n$) klocków, tak aby żaden biały klocek nie był spięty z innym białym klockiem?

Zadanie 8. W urnie znajduje się n, ponumerowanych od 1 do n, kul, $n > 3$. Losujemy dwie bez zwracania.

a) Obliczyć prawdopodobieństwo, że jedna z nich jest oznaczona liczbą mniejszą od k, a druga większą od k, przy ustalonym $k \in \mathbb{N}$, $1 < k < n$.
b) Obliczyć prawdopodobieństwo, że suma "oznaczeń" jest dokładnie równa k, $1 < k < n$.

Zadanie 9. W dwóch urnach I, II znajdują się kule odpowiednio: I – 6 czarnych i 9 białych, II – 5 czarnych i 15 białych.
a) Losujemy jedną kulę z urny I i jedną kulę z urny II. Jakie jest prawdopodobieństwo, że będą to kule tego samego koloru?
b) Losujemy jedną kulę z urny I i dwie bez zwracania z urny II. Jakie jest prawdopodobieństwo otrzymania 1 kuli białej i 2 czarnych?
c) Losujemy dwie kule z I urny i bez oglądania ich wrzucamy do urny II. Jakie jest prawdopodobieństwo wyciągnięcia kuli białej z II urny (zmienionej)?

Zadanie 10. Z urny, w której znajduje się 20 kul białych i 2 czarne, losujemy kolejno n kul. Znaleźć najmniejszą liczbę losowań n, taką że prawdopodobieństwo wylosowania chociaż raz czarnej kuli jest większe od 0,5, zakładając, że:
a) po każdym losowaniu kula wraca do urny,
b) kule po losowaniu nie wracają do urny.

Zadanie 11. W pewnej miejscowości rodzi się średnio 520 chłopców i 480 dziewczynek na 1000 niemowląt. Jakie jest prawdopodobieństwo, że w rodzinie pięciodzietnej:
a) liczba dziewcząt jest większa od liczby chłopców,
b) wszystkie dzieci są tej samej płci.

Zadanie 12. Urządzenie składa się z trzech zespołów typu A, trzech zespołów typu B oraz czterech zespołów typu C. Urządzenie przestaje działać, jeśli działają mniej niż dwa zespoły danego typu. Prawdopodobieństwo zepsucia się zespołu danego typu w pewnym określonym czasie jest równe odpowiednio $P(A) = 0,1$, $P(B) = 0,1$, $P(C) = 0,2$.
a) Obliczyć prawdopodobieństwo, że urządzenie przestanie w danym czasie działać.
b) Urządzenie przestało działać; jakie jest prawdopodobieństwo, że zdarzenie to zaszło z powodu awarii zespołu C?

Zadanie 13. Do hurtowni papieru dostarcza się towar z trzech fabryk F_1, F_2, F_3, w proporcjach 4 : 3 : 2. Liczba braków dostarczanych z tych fabryk wyraża się stosunkiem 2 : 3 : 4.
a) Jakie jest prawdopodobieństwo, że losowo wybrany towar pochodzi z fabryki F_i, $i = 1, 2, 3$?
b) Jakie jest prawdopodobieństwo, że losowo pobrany towar będzie wadliwy?
c) Pobrany losowo papier z magazynu okazał się wadliwy. Jakie jest prawdopodobieństwo, że pochodzi on z fabryki F_i, $i = 1, 2, 3$?

Zadanie 14. W lipcu przeciętnie 5 dni w ciągu tygodnia jest słonecznych. Jakie jest prawdopodobieństwo, że w czasie tygodniowej wycieczki będą co najmniej 3 dni słoneczne?

Zadanie 15. Książka ma 500 stron i zawiera 500 błędów drukarskich. Obliczyć prawdopodobieństwo, że losowo wybrana strona ma co najmniej 3 błędy.

Zadanie 16. Rzucamy n razy kostką. Wyznacz najbardziej prawdopodobną liczbę rzutów, w których wypadnie 5 oczek, przyjmując:
a) $n = 11$, **b)** $n = 20$, **c)** $n = 78$, **d)** $n = 83$.

Zadanie 17. Znaleźć prawdopodobieństwo, że 1 stycznia 3 osoby z n osobowej grupy ludzi będą obchodzić urodziny. Znaleźć najbardziej prawdopodobną liczbę osób urodzonych 1 kwietnia, dla:
a) $n = 30$, **b)** $n = 250$, **c)** $n = 1000$.

Zadanie 18. Prawdopodobieństwo pojawienia się co najmniej jednego zdarzenia A przy dwóch niezależnych próbach jest równe 0,51. Jakie jest prawdopodobieństwo pojawienia się tego zdarzenia w jednej próbie?

Zadanie 19. Odcinek długości 10 cm podzielono w sposób losowy na 3 odcinki. Obliczyć prawdopodobieństwo, że można z nich zbudować trójkąt.

Zadanie 20. Na okręgu wybrano losowo trzy punkty. Jakie jest prawdopodobieństwo:
a) że utworzą one trójkąt prostokątny,
b) że utworzą trójkąt rozwartokątny?

Zadanie 21. Płaszczyznę podzielono prostymi równoległymi o równych odległościach wynoszących $2a$. Na płaszczyznę tę rzucamy w sposób przypadkowy odcinek długości $2l < 2a$. Jakie jest prawdopodobieństwo tego, że odcinek przetnie jedną z prostych? [1]

Zadanie 22. Dwaj szachiści równej siły gry rozgrywali mecz o nagrodę 100 franków. Nagroda miała przypaść temu graczowi, który pierwszy zdobędzie 3 punkty. Umówiono się przy tym, że zwycięzca pojedynczego spotkania otrzymuje 1 punkt, przegrywający nie otrzymuje żadnego, a partie remisowe traktuje się jako niebyte. Mecz został przerwany z przyczyn niezależnych od obu graczy w chwili, gdy gracz A miał już 2 punkty, a gracz B jeden punkt. Jak należy podzielić nagrodę? [2]

Zadanie 23. Rozwiązać Zadanie 22 przyjmując, że:
a) gracz A ma 2 punkty, gracz B ma zero punktów,

[1] Zadanie Buffona.
[2] Zadanie dotyczy "problemu podziału nagrody meczowej" kawalera de Méré, którym zajmowali się wielcy matematycy XV, XVI, i XVII wieku.

b) w turnieju uczestniczy 3 szybkobiegaczy, po każdym biegu zwycięzca otrzymuje jeden punkt, pozostali zawodnicy nie otrzymują punktów, po przerwaniu z przyczyn obiektywnych turnieju zawodnik A ma 2 punkty, zawodnik B jeden punkt, zawodnik C zero punktów. Jak zawodnicy powinni podzielić 100-frankową nagrodę?

Zadanie 24. Podczas badania długości próbki zebrano wyniki:
a) 6,0; 6,1; 6,3; 5,9; 5,8; 6,0; 6,3; 6,2
b) 17, 19 15, 18, 16, 18, 16, 17.
Przedstawić zebrane wyniki w formie tabelki zakładając, że uzyskane częstości wyników są równe prawdopodobieństwom ich wystąpienia, otrzymane zaś wyniki przyjmujemy jako wartości zmiennej losowej X. Wyznaczyć wartość oczekiwaną $E(X)$ oraz wariancję $V(X)$.

Zadanie 25. Podać rozkład zmiennej losowej i dystrybuantę schematu Bernoulliego dla:
a) $n = 3, p = 0,5$, **b)** $n = 4, p = 0,5$.

Zadanie 26. Zmienna losowa X ma następujący rozkład prawdopodobieństwa:

X	-2	-1	0	1
$P(X=x)$	0,1	c	0,3	0,4

a) Znajdź stałą c.
b) Wyznacz rozkład prawdopodobieństwa zmiennej $Y = 2X + 1$.
c) Wyznacz wartości $E(X)$, $V(X)$, $E(Y)$, $V(Y)$.

Zadanie 27. Zmienna losowa Y ma rozkład prawdopodobieństwa

X	-4	-3	-2	-1	0	1	2	3	4
$P(X=x)$	1/16	1/16	1/8	1/4	1/8	1/8	c	1/16	1/16

a) Znajdź stałą c.
b) Wyznacz rozkład prawdopodobieństwa zmiennej $Y = X^2 - 4$.
c) Wyznacz wartości $E(X)$, $V(X)$, $E(Y)$, $V(Y)$.

Zadanie 28. Dla jakiej wartości parametru α funkcja $f(x)$ przedstawia rozkład prawdopodobieństwa pewnej zmiennej losowej? Wyznacz jej dystrybuantę; narysuj oba wykresy.

a) $f(x) = \begin{cases} \alpha(x+1) & \text{dla } -1 < x \le 0 \\ -\alpha(x-1) & \text{dla } 0 < x \le 1 \\ 0 & \text{dla pozostałych } x \end{cases}$. Wyznacz $P(|X| > 1/2)$.

b) $f(x) = \begin{cases} \frac{1}{x} & \text{dla } 1 < x \le \alpha \\ 0 & \text{dla pozostałych } x \end{cases}$. Wyznacz $P(X > e^2)$.

Zadanie 29. Dla jakich wartości parametru α, β funkcja $F(x)$ przedstawia dystrybuantę pewnej zmiennej losowej? Wyznacz jej funkcję gęstości, narysuj oba wykresy.

a) $F(x) = \begin{cases} 0 & \text{dla } x \le 0 \\ \sin(\alpha x) & \text{dla } 0 < x \le \pi \\ \beta & \text{dla } \pi < x \end{cases}$. Wyznacz $P(X > \pi/2)$.

b) $F(x) = \begin{cases} \frac{1}{2}e^x & \text{dla } x \le 0 \\ 1/2 & \text{dla } 0 < x \le 2 \\ \frac{\alpha x}{x+\beta} & \text{dla } 2 < x \end{cases}$. Wyznacz $P(|X| < 1)$.

Zadanie 30. Znaleźć gęstość prawdopodobieństwa zmiennej losowej Y będącej polem koła, którego promień jest zmienną losową o rozkładzie jednostajnym, w przedziale $(0, \alpha)$.

Zadanie 31. Wyznaczyć funkcję gęstości zmiennej $Y = \ln(X+1)$, jeżeli funkcja gęstości zmiennej X wyraża się wzorem $f(x) = \ln(x+1)$ dla $x \in (0, e-1)$ oraz $f(x) = 0$ dla pozostałych x.

Zadanie 32. Wiedząc, że dystrybuanta zmiennej X wyraża się wzorem $F(x) = \ln x$ dla $x \in (1, e)$ oraz $F(x) = 0$ dla $x < 1$ i $F(x) = 1$ dla $x > e$, wyznaczyć dystrybuantę zmiennej $Y = 3X + 2$.

Zadanie 33. Funkcja gęstości zmiennej losowej X wyraża się wzorem

a) $f_X(x) = \begin{cases} 1/2 & \text{dla } 1 \le x \le 3 \\ 0 & \text{dla pozostałych } x \end{cases}$,

b) $f_X(x) = \begin{cases} 2e^{-2x} & \text{dla } x \ge 0 \\ 0 & \text{dla pozostałych } x \end{cases}$,

Wyznaczyć funkcję gęstości zmiennych losowych: $Y = 3X + 2$, $Z = (Y+1)^2$.

Zadanie 34. Wiedząc, że $E(X) = 2$, $V(X) = 1$, wyznaczyć $E(Y)$ i $V(Y)$ dla zmiennej losowej $Y = 2X + 1$.

Zadanie 35. Załóżmy, że odczytując pewien pomiar, popełniamy błąd X z przedziału $I = (-0,1; 0,1)$. Prawdopodobieństwo, że błąd ten znajdzie się w przedziale $(a, b) \subset I$, jest wprost proporcjonalne do $(b-a)$, dla dowolnych $a, b \in I$, $a \le b$.
a) Wyznaczyć współczynnik proporcjonalności.
b) Wyznaczyć $P(0 < X < 0,05)$ i $P(|X| < 0,05)$.
c) Wyznaczyć x_1, x_2, dla których $P(X < x_1) = 0,25$, $P(X > x_2) = 0,75$.
d) Wyznaczyć funkcję gęstości oraz dystrybuantę tej zmiennej losowej.
e) Wyznaczyć $E(X)$, $V(X)$.

Zadanie 36. Przyjmijmy, że czas X bezawaryjnej pracy badanego urządzenia ma rozkład $f(x) = \begin{cases} \frac{1}{10}e^{-x/10} & \text{dla } x > 0 \\ 0 & \text{dla pozostałych } x \end{cases}$.
a) Obliczyć $P(5 \le X \le 10)$, $P(X \le 100)$.

b) Wyznaczyć dystrybuantę tej zmiennej losowej.
c) Zinterpretować otrzymane wyniki na wykresach funkcji gęstości i dystrybuanty.
d) Wyznaczyć $E(X)$ i $V(X)$.
e) Po jakim czasie maszyna ulegnie awarii z prawdopodobieństwem $\alpha = 0,9$?

Zadanie 37. Korzystając z tablic funkcji wykładniczej, podać przybliżone wartości dokładnych wyników uzyskanych w Zadaniu 36 a) i e).

Zadanie 38. Wyznaczyć $E(X)$ i $V(X)$ dla zmiennych losowych opisanych w Zadaniach 28 i 29.

Zadanie 39. Prawdopodobieństwo trafienia do tarczy wynosi $p = 0,6$. Podejmujemy próbę trafienia do tarczy aż do uzyskania celu. Oznaczmy przez X liczbę prób (powtórzeń). Zakładając niezależność kolejnych doświadczeń, wyznaczyć:
a) funkcję prawdopodobieństwa zmiennej X,
b) dystrybuantę,
c) prawdopodobieństwa, że liczba prób będzie: p_1 – mniejsza od 3, p_2 – nie mniejsza niż 3, p_3 – liczbą z przedziału $[2,5)$,
d) wartość oczekiwaną i wariancję zmiennej losowej X.

Zadanie 40. Prawdopodobieństwo, że produkt poddany próbie nie wytrzyma tej próby, wynosi $p = 0,01$. Obliczyć prawdopodobieństwo, że wśród 200 takich produktów co najwyżej 2 nie wytrzymają próby.

Zadanie 41. Z partii 250 sztuk towaru zawierającej 18 sztuk wadliwych wylosowano bez zwrotu próbę 10-elementową. W procesie kontroli wyrywkowej partia zostanie odrzucona, gdy w próbce znajdą się 2 lub więcej sztuk wadliwych. Obliczyć prawdopodobieństwo przyjęcia danej partii towaru.

Zadanie 42. Niech zmienna losowa X ma rozkład $N(\mu, \sigma)$.
a) Obliczyć $P(|X - \mu| < k\sigma)$ dla: $k = 1,96$, $k = 2,55$.
b) Wyznaczyć x, dla których $P(|X - \mu| < x\sigma) = \alpha$ gdy $\alpha = 0,95$, $\alpha = 0,1$.

Zadanie 43. Niech zmienna losowa X ma rozkład $N(3,2)$.
a) Obliczyć $P(|X| < k)$ dla $k = 1,6$, $k = 5,8$,
b) wyznaczyć x, dla których $P(X < x) = \alpha$ gdy: $\alpha = 0,95$, $\alpha = 0,1$.

Zadanie 44. Pewien automat produkuje części, których długość jest zmienną losową o rozkładzie $N(2;0,2)$. Wyznaczyć prawdopodobieństwo otrzymania braku, jeśli dopuszczalne długości części powinny się zawierać w przedziale $(1,7;2,3)$.

Zadanie 45. Amplituda X kołysania bocznego statku jest zmienną losową o gęstości $f(x) = xe^{-x^2/2}$ dla $x > 0$ oraz $f(x) = 0$ dla pozostałych x.

a) Obliczyć wartość przeciętną oraz wariancję zmiennej losowej.
b) Jakie jest prawdopodobieństwo wystąpienia amplitudy większej od wartości oczekiwanej?

Zadanie 46. Pewne urządzenie składa się z dwóch elementów pracujących niezależnie od siebie, połączonych równolegle. Czas bezawaryjnej pracy każdego z nich jest zmienną losową o rozkładzie wykładniczym $f(x) = \frac{1}{\lambda}e^{\frac{-x}{\lambda}}$ dla $x > 0$ oraz $f(x) = 0$ dla pozostałych x, odpowiednio przy $\lambda_1 = 10$ oraz $\lambda_2 = 20$.
a) Obliczyć prawdopodobieństwo, że urządzenie będzie pracowało co najmniej przez 20 godzin.
b) Obliczyć prawdopodobieństwo, że urządzenie przestanie pracować przed upłynięciem dziesiątej godziny.
c) Korzystając z tablic, podać przybliżone wartości wyników uzyskanych w punktach a) i b).

Zadanie 47. Korzystając z tablic kwantyli rozkładu t Studenta,
a) wyznaczyć $P(T < 2)$, $P(T > 1,7)$, przy 25 stopniach swobody,
b) wyznaczyć x, y, dla których $P(T < x) < 0,05$, $P(T > y) < 0,01$, przy 25 stopniach swobody,
c) wyznaczyć, przy ilu stopniach swobody $P(T < 2) = 0,975$.

Zadanie 48. Korzystając z tablic kwantyli rozkładu χ^2,
a) wyznaczyć: $P(\chi^2 < 5)$, $P(\chi^2 > 29,8)$, przy 13 stopniach swobody,
b) wyznaczyć x, y, dla których $P(\chi^2 < x) < 0,005$, $P(\chi^2 > x) < 0,01$, przy 20 stopniach swobody,
c) wyznaczyć, przy ilu stopniach swobody $P(\chi^2 < 7) = 0,01$.

Zadanie 49. Korzystając z tablic rozkładu prawdopodobieństwa Poissona,
a) wyznaczyć $P(X = 5)$ oraz $P(X \leq 3)$, przy $\lambda_1 = 0,5$, $\lambda_2 = 2,5$, $\lambda_3 = 5$,
b) wyznaczyć, przy jakim λ prawdopodobieństwo uzyskania 7 sukcesów jest większe od: $p_1 = 0,0001$, $p_2 = 0,001$, $p_3 = 0,01$.

Zadanie 50. Niech zmienne losowe X, Y mają rozkład prawdopodobieństwa podany w tabelach:

x_i	0	10	20
p_i	0,3	0,4	0,3

,

y_i	0	10	20	30
p_i	0,1	0,4	0,4	0,1

.

a) Znaleźć rozkład zmiennej losowej: $Z = X + Y$, $U = XY$.
b) Zakładając, że zmienne X i Y są niezależne, wyznaczyć: $E(Z)$, $V(Z)$, $E(U)$, $V(U)$.

Zadanie 51. Dane są funkcje prawdopodobieństwa niezależnych zmiennych losowych X i Y:

a)

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

,

y_i	2	3	5
p_i	0,3	0,4	0,3

,

b)

x_i	-1	0	1
p_i	0,2	0,6	0,2

,

y_i	-2	1	2	5
p_i	0,2	0,5	0,2	0,1

.

Wyznaczyć funkcję prawdopodobieństwa zmiennych losowych: $U = X + Y$, $V = XY$.

Zadanie 52. Dwuwymiarowa zmienna losowa (X, Y) ma rozkład w postaci:

a)

Y \ X	-1	0	1
-1	1/8	1/4	1/8
1	1/8	1/8	1/4

,

b)

Y \ X	-1	0	1
0	1/12	3/12	2/12
1	1/12	4/12	1/12

.

Znaleźć rozkład prawdopodobieństwa zmiennych losowych: $Z_1 = X^2 + Y^2$, $Z_2 = 2X - 3Y$, $Z_3 = XY$. Wyznaczyć wartość oczekiwaną oraz wariancję dla zmiennych: X, Y, Z_1, Z_2, Z_3.

Zadanie 53. Dana jest gęstość dwuwymiarowej zmiennej losowej $f(x, y)$. Zbadać, czy zmienne X i Y są niezależne, jeżeli

a) $f(x, y) = 24x^2(1 - y)$ dla $0 \leq x \leq 1$ i $0 \leq y \leq 1$ oraz $f(x, y) = 0$ dla pozostałych par (x, y),

b) $f(x, y) = \frac{1}{12}(6x + 4y + 3)$ dla $-1 \leq x \leq 1$ i $-1 \leq y \leq 1$ oraz $f(x, y) = 0$ dla pozostałych par (x, y),

c) $f(x, y) = \frac{x}{2}e^{-y}$ dla $0 \leq x \leq 2$ i $0 \leq y$ oraz oraz $f(x, y) = 0$ dla pozostałych par (x, y).

Zadanie 54. Niech zmienna losowa (X, Y) wyraża się wzorem

$$f(x, y) = \begin{cases} 1/2 & \text{dla } (x, y) \in (0, 1) \times (0, 2) \\ 0 & \text{dla pozostałych par } (x, y) \end{cases}$$

a) Znaleźć dystrybuantę zmiennej $Z = (X, Y)$ oraz dystrybuanty zmiennych brzegowych X i Y.

b) Wykazać, że zmienne losowe są niezależne.

c) Wyznaczyć: $E(XY)$, $E(X^3Y^3)$.

Zadanie 55. Znaleźć dystrybuantę zmiennej losowej $Z = (X, Y)$, zbadać, czy zmienne X i Y są niezależne, jeżeli gęstość zmiennej Z wyraża się wzorem:

a) $f(x, y) = xy$ dla $0 \leq x \leq 2, 0 \leq y \leq 1$, oraz $f(x, y) = 0$ dla pozostałych par (x, y),

b) $f(x, y) = (x + y)e^{-(x+y)}$ dla $x > 0$ i $y > 0$ oraz $f(x, y) = 0$ dla pozostałych par (x, y).

Zadanie 56. Znaleźć dystrybuanty zmiennych dwuwymiarowych z Zadania 53.

Zadanie 57. Dystrybuanta dwuwymiarowej zmiennej losowej (X, Y) wyraża się funkcją

a) $$F(x, y) = \begin{cases} 0 & \text{dla } x < 0 \text{ lub } y < 0 \\ x^2y^2(3 - x - y) & \text{dla } 0 < x \leq 1 \text{ i } 0 < y \leq 1 \\ y^2(2 - y) & \text{dla } x > 1 \text{ i } 0 < y \leq 1 \\ x^2(2 - x) & \text{dla } 0 < x \leq 1 \text{ i } y > 1 \\ 1 & \text{dla } x > 1 \text{ i } y > 1 \end{cases},$$

b) $$F(x, y) = \begin{cases} 0 & \text{dla } x < 0 \text{ lub } y < 0 \\ x^2y^2/9 & \text{dla } 0 \leq x \leq 1 \text{ i } 0 \leq y \leq 3 \\ x^2 & \text{dla } 0 \leq x \leq 1 \text{ i } y > 3 \\ y^2/9 & \text{dla } x > 1 \text{ i } 0 \leq y \leq 3 \\ 1 & \text{dla } x > 1 \text{ i } y > 3 \end{cases}$$

Znaleźć gęstość prawdopodobieństwa rozkładów oraz gęstości rozkładów brzegowych i warunkowych.

Zadanie 58. Erozja T gleby przy różnym nachyleniu X, spowodowana opadami 50 mm w ciągu 164 godzin, przedstawia się następująco:

X (w stopniach)	0	2,5	5,5	8,5	11
Y (w mm)	0,03	0,9	2,8	4,7	6,3

.

a) Określić rozkład prawdopodobieństwa zmiennej (X, Y).

b) Wyznaczyć momenty zwykłe: $m_{0,1}$, $m_{1,0}$, $m_{1,1}$, $m_{0,2}$, $m_{2,0}$.

c) Napisać równania prostych regresji drugiego rodzaju.

d) Korzystając z nich, obliczyć erozję odpowiadającą nachyleniom: 2°, 6°, 10° oraz przewidywane nachylenia podczas erozji 0,5 mm, 5 mm.

Zadanie 59. Dwuwymiarowa zmienna losowa (X, Y) ma rozkład podany w tabeli:

$y_k \setminus x_i$	3	3,5	4	4,5	5
3	0,16	0	0	0	0
3,5	0,12	0,04	0	0	0
4	0,08	0,08	0,08	0	0
4,5	0	0,12	0,06	0,05	0,04
5	0	0	0,01	0,08	0,08

gdzie X jest średnią ocen w sesji egzaminacyjnej w I semestrze dla losowo wybranego studenta, Y zaś – średnią ocen w sesji tego samego studenta w VI semestrze.

a) Obliczyć współczynnik korelacji tych zmiennych losowych.

b) Napisać równanie prostej regresji drugiego rodzaju zmiennej losowej Y względem X.

Zadanie 60. Gęstość dwuwymiarowej zmiennej losowej wyraża się wzorem

$f(x,y) = x+y$ dla $(x,y) \in (0,1)\times(0,1)$, $f(x,y) = 0$ dla pozostałych par (x,y)
a) Wyznaczyć współczynnik korealcji zmiennych X i Y.
b) Czy zmienne są nieskorelowane?
c) Wyznaczyć proste regresji drugiego rodzaju.
d) Wyznaczyć rozkłady brzegowe.
e) Wyznaczyć rozkłady warunkowe.
f) Wyznaczyć linie regresji pierwszego rodzaju.

Zadanie 61. Wyznaczyć funkcje charakterystyczne rozkładów:
a) zero-jedynkowego,
b) liczby wyrzuconych orłów w rzucie trzema monetami,
c) o gęstości $f(x) = \frac{1}{\lambda}e^{-x/\lambda}$ dla $x > 0$ oraz $f(x) = 0$ dla pozostałych x (rozkład wykładniczy).

Zadanie 62. Wykazać, że niżej podane funkcje nie mogą być funkcjami charakterystycznymi zmiennych losowych:
a) $\varphi(t) = e^{-it^2}$,
b) $\varphi(t) = \frac{1}{1+i|t|}$,
c) $\varphi(t) = 2 - |t|$ dla $|t| < 1$, oraz $\varphi(t) = 0$ dla pozostałych t.

Zadanie 63. Korzystając z funkcji charakterystycznych zmiennych losowych, wyznaczyć wartość przeciętną i wariancję dla tych zmiennych, jeżeli
a) $f(x) = e^{-x}$ dla $x > 0$ oraz $f(x) = 0$ dla pozostałych x,
b) $f(x) = \frac{1}{2}e^{-|x|}$ dla $x \in \mathbb{R}$ (rozkład Laplace'a).

Zadanie 64. Podczas przeprowadzania badań statystycznych zebrano 50 wyników: 1,05; 1,20; 1,15; 1,15; 1,05; 1,20; 1,00; 1,10; 1,25; 1,25;
1,15; 1,00; 1,05; 1,25; 1,05; 1,10; 1,05; 1,20; 1,15; 1,05;
1,10; 1,10; 1,10; 1,05; 1,00; 1,15; 1,10; 1,15; 1,20; 1,15;
1,05; 1,25; 1,05; 1,10; 1,15; 1,25; 1,20; 1,00; 1,20; 1,05;
1,25; 1,20; 1,05; 1,15; 1,10; 1,20; 1,10; 1,10; 1,25; 1,00.
Sporządzić szereg rozdzielczy dla: **a)** $n = 4$, **b)** $n = 6$ klas. Wyznaczyć w obu przypadkach $\overline{x}$ i σ.

Zadanie 65. Uzupełnij tabelę, a następnie wyznacz $\overline{x}$ i σ.

Numer klasy	Przedział klasowy	Środek klasy $\overline{x}_i$	Liczność klasy n_i	Częstość klasy n_i/n	Częstość skumulowana $\sum_{l=1}^{i} x_l/n$
1	20 – ..	...	15	...	...
...	.. – ..	...	20	...	...
...	.. – ..	...	50	...	...
...	.. – ..	...	40	...	...
...	.. – 40	...	25	...	...

Zadanie 66. Wyznaczyć metodą największej wiarygodności estymator parametru p w rozkładzie geometrycznym $P(X = k) = p(1-p)^{k-1}$, $k = 1,2,\ldots$ na podstawie n-elementowej próby prostej $(X_1,\ldots,X_n)$ pochodzącej z tego rozkładu.

Zadanie 67. Cecha X ma rozkład normalny o znanej wartości średniej oraz nieznanym odchyleniu standardowym. Dla n-elementowej próby prostej ustalono estymator odchylenia standardowego $\hat{\sigma} = k\sum_{1}^{n}|X_i - \mu|$. Dobrać parametr k tak, aby był on estymatorem nieobciążonym.

Zadanie 68. Metodą największej wiarygodności na podstawie n-elementowej próby prostej $X_1, X_2, \ldots, X_n$ wyznaczyć estymator parametru λ rozkładu wykładniczego.

Zadanie 69. Wykazać, że estymator $\hat{\lambda}^2 = \frac{1}{2n}\sum_1^n X_i^2$ wariancji rozkładu wykładniczego jest estymatorem nieobciążonym.

Zadanie 70. Przeprowadzono badania wśród studentów, zadając pytanie o czas poświęcany na naukę własną w ciągu tygodnia, i otrzymano wyniki (w godzinach):
5,5; 4,0; 7,0; 4,0; 4,0; 5,0; 6,0; 7,5; 3,0; 8,0; 2,5;
5,5; 5,0; 6,0; 5,0; 3,5; 6,0; 5,5; 8,0; 2,0; 7,0; 2,5.
Wyznaczyć przedział ufności dla wartości średniej oraz wariancji przy poziomie istotności $\alpha = 0,95$.

Zadanie 71. Z partii towaru pobrano 15-elementową próbę, na jej podstawie ustalić, na poziomie istotności $\alpha = 0,90$, przedział ufności dla wartości średniej:
4,4; 4,2; 3,7; 3,9; 4,1; 4,0; 4,1; 4,0; 4,2; 4,6; 3,7; 4,3; 4,4; 4,2; 4,0.

Zadanie 72. W celu wyznaczenia ładunku elektronu wykonano 26 pomiarów tego ładunku metodą Millikana, otrzymując wyniki

$$\overline{x} = 1,574 \cdot 10^{-19}, \qquad s = 0,043 \cdot 10^{-19}.$$

Zakładając, że błędy pomiarów mają rozkład normalny o nieznanym σ, wyznaczyć 88% przedział ufności dla prawdziwej wartości wielkości ładunku elektrycznego.

Zadanie 73. Przed dokonaniem zmian w silniku samochodu dokonano 8 pomiarów zużycia paliwa i otrzymano wyniki x_i, następnie takie same badania wykonano po dokonaniu zmian i otrzymano wyniki y_j.

x_i : 5,7; 6,5; 6,1; 5,5; 5,0; 6,1; 6,2; 5,9; y_j : 4,9; 5,0; 4,7; 5,0; 5,0; 4,7.

Załóżmy, że zużycie paliwa jest zmienną o rozkładzie normalnym i że wariancja nie uległa zmianie. Zweryfikować hipotezę o jednakowym zużyciu paliwa przed zmianą i po zmianie wobec hipotezy alternatywnej "zużycie paliwa się zmniejszyło" na poziomie istotności $\alpha = 0,1$.

Zadanie 74. W celu określenia średniej temperatury w lipcu w danej miejscowości zebrano wyniki, dotyczące średniej temperatury w lipcu, z ostatnich 8 lat:

$$23, 22, 18, 25, 20, 21, 27, 20.$$

Na poziomie istotności $\alpha = 0,05$ zweryfikować hipotezę o średniej temperaturze równej 24, przy:
a) hipotezie alternatywnej $\mu_0 \neq 24$,
b) hipotezie alternatywnej $\mu_0 < 24$.

Zadanie 75. W celu sprawdzenia poprawności maszyny ważącej zważono 10 partii towaru i otrzymano wyniki:

12,05; 12,20; 12,00; 11,95; 12,10; 12,00; 11,80; 11,95; 12,10; 12,05.

Zweryfikować na poziomie istotności $\alpha = 0,05$ hipotezę:
a) $H : \sigma^2 = 0,035$ wobec hipotezy alternatywnej $\sigma \neq 0,035$,
b) $H : \sigma^2 = 0,035$ wobec hipotezy alternatywnej $\sigma < 0,035$,
c) $H : \sigma^2 = 0,035$ wobec hipotezy alternatywnej $\sigma > 0,035$.

Zadanie 76. Otrzymano dużą partię towaru. Zweryfikować, na poziomie istotności $\alpha = 0,01$, hipotezę producenta, że średnia masa towaru wynosi 0,25 na podstawie otrzymanych wyników w 10 próbach:

x_i	0,215-0,225	0,225-0,235	0,235-0,245	0,245-0,255	0,255-0,265
n_i	2	2	3	2	1

a) wiedząc, że odchylenie standardowe w całej partii towaru wynosi $\sigma = 0,015$,
b) nie znając odchylenia standardowego.

Zadanie 77. Temperatura w chłodni ma rozkład normalny $N(-15, \sigma)$. Wykonano 60 pomiarów i na ich podstawie wyznaczono $s^2 = 5$. Zweryfikować, przy poziomie ufności $\alpha = 0,1$, hipotezę $\sigma^2 = 3$ przy hipotezie alternatywnej $\sigma^2 > 3$.

Zadanie 78. Otrzymano od producenta wagę. Chcąc zweryfikować hipotezę $\sigma = 0,1$, zważono 100-krotnie towar, otrzymano wyniki i zestawiono je w tabeli:

x_i	6,4-6,6	6,6-6,8	6,8-7,0	7,0-7,2	7,2-7,4	7,4-7,6
n_i	8	12	30	25	15	10

Czy otrzymane wyniki potwierdzają tę hipotezę, czy też świadczą o tym, że dokładność wagi jest większa, to znaczy $\sigma > 0.1$?

Zadanie 79. W celu sprawdzenia, czy natężenie ruchu, na danej trasie, w dni przedświąteczne (x_1) jest większe niż w dni powszednie (x_2), zbadano liczbę pasażerów w ciągu 15 kolejnych tygodni i otrzymano wyniki

x_1	400-600	600-800	800-1000	1000-1200	1200-1400
n_i	1	2	2	7	3

,

x_2	400-550	550-700	700-850	850-1000	1000-1150
n_i	1	2	2	7	3

.

Na poziomie istotności $\alpha = 0,002$ zweryfikować hipotezę o równych natężeniach ruchu na rzecz hipotezy alternatywnej, że średnia liczba pasażerów w dni przedświąteczne jest większa niż w dni powszednie, jeżeli:
a) nie znamy odchylenia standartowego pomiarów,
b) zakładamy, że odchylenie standardowe pomiarów wynosi $\sigma_1 = \sigma_2 = 200$.

Zadanie 80. Fabryka produkuje towar na eksport i na rynek krajowy. Przebadano 200 sztuk towaru na rynek krajowy i otrzymano wyniki $\overline{x_1} = 10,05$, $s_1^2 = 0,4$ oraz 500 sztuk przeznaczonych na eksport i otrzymano wyniki $\overline{x_2} = 9,8$, $s_2^2 = 0,2$. Na poziomie istotności $\alpha = 0,01$ zweryfikować:
a) hipotezę $H : \mu_1 = \mu_2$ przy hipotezie alternatywnej $K : \mu_1 \neq \mu_2$,
b) hipotezę $H : \sigma_1^2 = 0,2$ przy hipotezie alternatywnej $K : \sigma_1 > 0,2$,
c) hipotezę $H : \mu_1 = 10$ przy hipotezie alternatywnej $K : \mu_1 > 10$,
d) hipotezę $H : \mu_2 = 10$ przy hipotezie alternatywnej $K : \mu_2 < 10$.

Rozdział 4

Odpowiedzi i rozwiązania

1. W każdym wierszu i w każdej kolumnie może stać tylko jedna wieża. Każdej kolumnie przyporządkujemy numer wiersza, w którym stoi wieża. Liczba takich przyporządkowań jest równa liczbie permutacji zbioru 8-elementowego, $P_8 = 8!$.

2. a) Z jednej strony stołu można usadzić panie na $P_4 = 4!$ sposobów, podobnie jak panów z drugiej strony stołu. Ponieważ jednak panie mogą się zmienić z panami stroną stołu, to wszystkich możliwości jest $2(4!)^2$. **b)** Przy okrągłym stole nie rozróżniamy miejsc, zatem możliwości jest $(4!)^2$.

3. Dwa asy losujemy z czterech na $C_n^k = \binom{n}{k}$ sposobów, pozostałych 11 kart losujemy z 48 kart na $C_{48}^{11} = \binom{48}{11}$ sposobów. Wszystkich wyników losowania jest $\binom{4}{2}\binom{48}{11}$.

4. a) Najpierw uzupełniamy miejsce pierwsze i ostatnie cyframi 3 oraz 4. Dokonujemy tego na 2 sposoby. Na pozostałych 4 miejscach ustawiamy pozostałe 4 cyfry na $P_4 = 4!$ sposobów. Ostatecznie takich liczb będzie $2 \cdot 4!$. **b)** Jeżeli cyfry mogą się powtarzać, pierwsze i ostatnie miejsce ustalamy na 4 sposoby, to pozostałe 4 miejsca zapełniamy 6 cyframi (i możemy je powtarzać) na $P_6^4 = 6^4$ sposobów. Ostatecznie liczb takich będzie $4 \cdot 6^4$.

5. Zbiór pusty traktujemy jako podzbiór 0-elementowy, ogólnie k-elementowych podzbiorów zbioru n-elemetowego jest $\binom{n}{k}$, $n \leq k$. Wszystkich podzbiorów jest

$$\binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{n-1} + \binom{n}{n} = (1+1)^n = 2^n.$$

6. Liczba kości domino, na których nie powtarzają się numery, wynosi $\binom{n+1}{2}$. Ponadto jest $(n+1)$ kości o powtarzających się numerach. Ostatecznie liczba kości wynosi $\binom{n+1}{2} + (n+1) = \frac{1}{2}(n+1)(n+2)$.

7. Układamy w rzędzie na podłodze czarne klocki, tak aby pomiędzy nimi można było dołożyć po jednym białym klocku. Białe klocki (jest ich k) dokładamy do utworzonego rzędu w $(n+1)$ miejscach ($n-1$ pomiędzy klockami czarnymi, jeden przed czarnymi klockami i jeden po nich). Białe klocki dokładamy na $C_{n+1}^k = \binom{n+1}{k}$ sposobów. Na koniec budujemy wieżę; na $\binom{n+1}{k}$ sposobów.

8. $\Omega = \{\omega : \omega = \{p, q\}, p, q = 1, 2, \ldots, n,\ p \neq q\}$, $\overline{\overline{\Omega}} = C_n^2 = \binom{n}{2}$,
a) $A_k = \{\omega \in \Omega :\ p < k \text{ i } q > k,\ 0 < k < n\}$, $\overline{\overline{A_k}} = (k-1)(n-k)$,

$P(A_k) = 2^{\frac{(k-1)(n-k)}{n(n+1)}}$, **b)** $B_k = \{\omega \in \Omega :\ p+q = k,\ 0 < k < n\}$, dla nieparzystych k: $\overline{\overline{B_k}} = \frac{1}{2}(k-1)$, $P(B_k) = \frac{k-1}{n(n-1)}$, dla parzystych k, $\overline{\overline{B_k}} = \frac{k}{2} - 1)$, $P(B_k) = \frac{k-2}{n(n-1)}$.

9. a) $\Omega = \{\omega :\ \omega = \{k, n\}, k \in U_I, n \in U_{II}\}$, $\overline{\overline{\Omega}} = \binom{15}{1}\binom{20}{1}$,
$A = \{\omega \in \Omega :\ k = n = b$ lub $k = n = c\}$, $\overline{\overline{A}} = \binom{9}{1}\binom{15}{1} + \binom{6}{1}\binom{5}{1}$, $P(A) = 11/20$.
b) $\Omega = \{\omega :\ \omega = \{k, n_1, n_2\}, k \in U_I, n_1, n_2 \in U_{II}\}$, $\overline{\overline{\Omega}} = \binom{15}{1}\binom{20}{2}$,
$A = \{\omega \in \Omega :\ (k = b$ i $n_1 = c$ i $n_2 = c)$ lub $(k = c$ i $n_1 = b$ i $n_2 = c)\}$,
$\overline{\overline{A}} = \binom{9}{1}\binom{5}{2} + \binom{6}{1}\binom{15}{1}\binom{5}{1}$, $P(A) = 18/95$.
c) Prawdopodobieństwo całkowite. A_i – zdarzenie polegające na tym, że do II urny dołożono i kul białych, $i = 0, 1, 2$, $P(A_0) = 5/35$, $P(A_1) = 18/35$, $P(A_2) = 9/35$.
B – zdarzenie polegające na wylosowaniu kuli białej z II urny,
$P(B/A_i) = (15+i)/22$, $i = 0, 1, 2$. $P(B) = \sum_0^2 P(B/A_i)P(A_i) \approx 0,67\ldots$

10. a) Schemat Bernoulliego. $p = 1/11$. Szukamy najmniejszego n, takiego że $1 - P_n(0) > 0,5$, czyli $(10/11)^n < 0,5$; wyznaczamy $n = 8$. **b)** Wybieramy n kul z urny. Jeżeli zdarzenie B_n polega na tym, że nie wylosowano kuli białej, to szukamy najmniejszego n, takiego aby $1 - P(B_n) > 0,5$, gdzie $P(B_n) = \binom{20}{n} : \binom{22}{n}$, czyli $\frac{(21-n)(22-n)}{21 \cdot 22} < 0,5$. Wyznaczamy $n = 6$.

11. Schemat Bernoulliego, p – prawdopodobieństwo urodzenia dziewczynki, $p = 0,48$. **a)** $P_5(3) + P_5(4) + P_5(5) \approx 0,299 + 0,138 + 0,025 = 0,462$. **b)** $P_5(0) + P_5(5) = (0,48)^5 + (0,52)^5 \approx 0,064$.

12. a) Prawdopodobieństwo całkowite. $P(A^*)$ – proporcjonalny udział zespołów typu A w maszynie, $P(A^*) = 0,3, P(B^*) = 0,3, P(C^*) = 0,4$. Z – zdarzenie polegające na zepsuciu się maszyny. Maszyna zepsuje się z powodu awarii podzespołów typu A lub B, jeżeli popsują się co najmniej dwa z trzech, czyli $P(Z/A^*) = P(Z/B^*) = P_3(3) + P_3(2) = 0,028$, gdzie $p = P(A) = 0,1$, podobnie $P(Z/C^*) = P_4(4) + P_4(3) = 0,0272$. Korzystamy ze wzoru na prawdopodobieństwo całkowite $P(Z) = P(Z/A^*)P(A^*) + P(Z/B^*)P(B^*) + P(Z/C^*)P(C^*) = 0,02768$. **b)** Wzór Bayesa. $P(C^*/Z) = \frac{P(C^*)P(Z/C^*)}{P(Z)} = 0,393$.

13. a) $P(F_1) = 4/9, P(A_2) = 3/9, P(A_3) = 2/9$. **b)** Prawdopodobieństwo całkowite, $P(B_i)$ – prawdopodobieństwo wyprodukowania braku w fabryce F_i, $P(B_1) = 2/9, P(B_2) = 3/9, P(B_3) = 4/9$, B – pobranie z magazynu wadliwego towaru, $P(B) = \sum_1^3 P(B_i)P(F_i) = 25/81$. **c)** Wzór Bayesa, $P(F_i/B) = \frac{P(F_i)P(B/F_i)}{P(B)} = \frac{P(F_i)P(B_i)}{P(B)}$, $P(F_1/B) = 12/25$, $P(F_2/B) = 9/25$, $P(F_3/B) = 4/25$.

14. Schemat Bernoulliego, p – prawdopodobieństwo, że lipcowy dzień będzie słoneczny, $p = 5/7$, $n = 7$, $k \geq 3$, $P_7(k \geq 3) = 1 - P_7(k < 3) = 1 - (P_7(0) + P_7(1) + P_7(2)) = 1 - ((2/7)^7 + 7(5/7)(2/7)^6 + 21(5/7)^2(2/7)^5) \approx 0,976$.

15. Uogólnijmy zadanie, zakładając, że zarówno stron, jak i błędów jest n. Liczba wszystkich możliwych rozłożeń n błędów na n stronach jest $\overline{\overline{\Omega}} = n^n$, oznaczmy przez A_i – zdarzenie, że na losowo wybranej stronie (np. pierwszej) jest i błędów, $\overline{\overline{A_0}} = (n-1)^n$, $\overline{\overline{A_1}} = n(n-1)^{n-1}$, $\overline{\overline{A_1}} = \binom{n}{2}(n-1)^{n-2}$, ogólnie $\overline{\overline{A_i}} = \binom{n}{i}(n-1)^{n-i}$, $i = 0, 1, \ldots$ Jeżeli przez A oznaczymy zdarzenie sprzyjające, to

$$P(A) = 1 - \sum_0^2 \frac{\binom{n}{i}(n-1)^{n-i}}{n^n} = \left(\frac{n-1}{n}\right)^{n-1}\left(\frac{5}{2} - 1\right).$$

Kładąc $n = 500$, otrzymujemy $P(A) \approx 0,08$.

16. Schemat Bernoulliego, najbardziej prawdopodobna liczba sukcesów k_0 mieści się w przedziale $(n+1)p - 1 \leq k_0 \leq (n+1)p$, $p = 1/6$. **a)** $n = 11$, $1 \leq k_0 \leq 2$, przyjmujemy $k_0 = 1$ oraz $k_0 = 2$, **b)** $n = 20$, $k_0 = 13$, **c)** $n = 78$, $K_0 = 13$, **d)** $n = 83$, $k_0 = 13$ oraz $k_0 = 14$.

17. Schemat Bernoulliego, $p = 1/365$, najbardziej prawdopodobna liczba sukcesów k_0 mieści się w przedziale $(n+1)p-1 \leq k_0 \leq (n+1)p$. **a)** $n = 30$, $p = P_{30}(3) = 0,0^4 77$, $k_0 = 0$, **b)** $n = 250$, $p = P_{250}(3) = 0,027$, $k_0 = 0$, **c)** $n = 1000$, $p = P_{1000}(3) = 0,221$, $k_0 = 2$,

18. Schemat Bernoulliego, nieznane p, wiemy, że $P_2(1) + P_2(2) = 0,51$, zatem $2p(1-p) + p^2 = 0,51$, czyli $p = 0,3$.

19. Prawdopodobieństwo geometryczne. Każdy podział odcinka można opisać parą liczb (x, y), gdzie x – długość pierwszego odcinka, y – długość drugiego odcinka, trzeci odcinek ma wówczas długość $10 - x - y$, oczywiście $x + y < 10$. Zbiór punktów spełniających te warunki tworzy przestrzeń Ω w kształcie trójkąta o wierzchołkach $(0,0), (10,0), (0,10)$, jego pole $m(\Omega) = 50$. Zdarzeniu sprzyjającemu odpowiadają punkty z Ω spełniające dodatkowo trzy warunki: $x+y > 10-x-y$, $x+(10-x-y) > y$, $y + (10 - x - y) > x$, gdyż suma dowolnych dwóch odcinków musi być większa od trzeciego odcinka, aby można było zbudować trójkąt. Warunki te są równoważne warunkom $x + y > 5$, $y < 5$, $x < 5$, które opisują zbiór zdarzeń sprzyjających, jego miara wynosi $m(A) = 12,5$. Zatem $P(A) = 0,25$.

20. Prawdopodobieństwo geometryczne. Ustalmy, bez straty ogólności, że nasz okrąg jest jednostkowy, a jednym z punktów jest $P_1 = 1$. Każdy wybór pozostałych dwóch punktów można opisać jako wybór dwóch kątów środkowych: x – kąt środkowy oparty na jednym boku trójkąta, y – kąt środkowy oparty na drugim boku trójkąta, trzeci kąt środkowy ma miarę $2\pi - x - y$, oczywiście $x < 2\pi$, $y < 2\pi$, $x + y < 2\pi$. Wobec tego przestrzeń ω opisujemy w układzie współrzędnych jako trójkąt o wierzchołkach $(0,0), (0, 2\pi), (2\pi, 0)$, a jego miara wynosi $m(\Omega) = 2\pi^2$. **a)** Aby trójkąt był prostokątny, któryś kąt środkowy powinien być równy π, zatem $x = \pi$ lub $y = \pi$ lub $x + y = \pi$. Miara takiego zbioru wynosi 0, wobec tego $P(A) = 0$. **b)** Aby trójkąt był rozwartokątny, jeden z jego kątów musi być większy od $\pi/2$, zatem $x + y > \pi$ lub $y > \pi$ lub $x > \pi$. Część wspólna zbioru wyznaczonego z tych warunków i zbioru Ω to trzy trójkąty, każdy o polu równym $\frac{1}{2}\pi^2$, ostatecznie $P(A) = 3/4$.

21. Prawdopodobieństwo geometryczne. Załóżmy, że dane proste to proste $x = k2a$, $k \in \mathbb{Z}$. Ograniczmy się, bez straty ogólności, do przypadku, kiedy środek danego odcinka będzie leżał na osi OX, $x =\in (0, a)$, jego zaś kąt nachylenia do osi OX, $\alpha \in (\frac{1}{2}\pi, \frac{3}{2}\pi)$. Miara tak opisanego Ω wynosi $m(\Omega) = a\pi$. Zbiór zdarzeń sprzyjających jest dodatkowo opisany warunkiem: $-l\cos\alpha > x$, zatem $m(A) = -\int_{\frac{1}{2}\pi}^{\frac{3}{2}\pi} l\cos\alpha d\alpha = l$. Ostatecznie $P(A) = l/(a\pi)$.

22. Prawdopodobieństwo $P(A)$, że mecz zakończyłby się sukcesem zawodnika A, wynosi $P(A) = 0,75$, a $P(B)$, że sukces odniesie zawodnik B, wynosi $P(B) = 0,25$. I w takim stosunku należy podzielić nagrodę.

23. a) $P(A) = 7/8$, $P(B) = 1/8$ nagroda powinna zostać podzielona w stosunku $7 : 1$, **b)** $P(A) = 20/27$, gracz A może zakończyć rozgrywki po jednym meczu z

prawdopodobieństwem 1/3, po dwóch z prawdopodobieństwem 2/9, po trzech z 1/9 i po czterech z 1/27, podobnie $P(B) = 6/27$, $P(C) = 2/27$. Nagroda powinna być podzielona w stosunku 19 : 6 : 2.

24. a)

	5,8	5,9	6,0	6,1	6,2	6,3
p_i	1/8	1/8	2/8	1/8	1/8	2/8

$E(X) = 6,075$, $V(X) = 0,029$.

b)

15	16	17	18	19	
1/8	2/8	2/8	2/8	1/8	

$E(X) = 17$, $V(X) = 1,5$.

25. b) $F(x) = \begin{cases} 0 & \text{dla } x \le 0 \\ 1/8 & \text{dla } 0 < x \le 1 \\ 1/2 & \text{dla } 1 < x \le 2 \\ 7/8 & \text{dla } 2 < x \le 3 \\ 1 & \text{dla } 3 < x \end{cases}$, $F(x) = \begin{cases} 0 & \text{dla } x \le 0 \\ 1/16 & \text{dla } 0 < x \le 1 \\ 5/16 & \text{dla } 1 < x \le 2 \\ 11/16 & \text{dla } 2 < x \le 3 \\ 15/16 & \text{dla } 3 < x \le 4 \\ 1 & \text{dla } 4 < x \end{cases}$.

26. a) $c = 0,2$ b)

X	-2	-1	0	1
$2X+1$	-3	-1	1	3
$P(X = x)$	0,1	c	0,3	0,4

c) $E(X) = 0$, $V(X) = 1$, $E(Y) = 1$, $V(Y) = 4$.

27. a) $c = 1/8$,

b)

X	-4	-3	-2	-1	0	1	2	3	4
p_i	1/16	1/16	1/8	1/4	1/8	1/8	1/8	1/16	1/16
X^2-4	12	5	0	-3	-4	-3	0	5	12

dane przenosimy do tabeli:

Y	-4	-3	0	5	12
q_i	1/8	3/8	2/8	1/8	1/8

c) $E(X) = 0,75$, $V(X) = 3,94$, $E(Y) = 0,5$, $V(Y) = 20,25$.

28. a) $\alpha = 1$, $F(X) = \begin{cases} 0 & \text{dla } x \le -1 \\ (x^2+2x+1)/2 & \text{dla } -1 < x \le 0 \\ (-x^2+2x+1)/2 & \text{dla } 0 < x \le 1 \\ 1 & \text{dla } 1 < x \end{cases}$,

$P(|X| < 1/2) = 0,75$,

b) $\alpha = e$, $F(X) = \begin{cases} 0 & \text{dla } x \le 1 \\ \ln x & \text{dla } 1 < x \le e \\ 1 & \text{dla } e < x \end{cases}$, $P(X > e^2) = 0$.

29. a) $\alpha = 1/2, \beta = 1$, $f(x) = \begin{cases} \frac{1}{2}\cos(x/2) & \text{dla } 0 < x \le \pi \\ 0 & \text{dla pozostałych } x \end{cases}$, $p = 1 - \frac{\sqrt{2}}{2}$,

b) $\alpha = 1, \beta = 2$, $f(x) = \begin{cases} e^x/2 & \text{dla } x \le 0 \\ 0 & \text{dla } 0 < x \le 2 \\ 2/(x+2)^2 & \text{dla } 2 < x \end{cases}$, $p = \frac{e-1}{2e}$.

30. $f_X(x) = 1/\alpha$, dla $x \in (0, \alpha)$, oraz $f_X(x) = 0$ dla pozostałych x. Ponadto $Y = \pi X^2$, $g(x) = \pi x^2$, dla $x > 0$ g ma funkcję odwrotną $h(y) = \sqrt{y/\pi}$, $h'(y) = \frac{1}{2}\frac{1}{\sqrt{y\pi}}$, zatem $f_Y(y) = f_X(h(y))h'(y)$, dla $h(y) \in (0, \alpha)$, czyli $f_Y(y) = \dfrac{1}{2\alpha\sqrt{\pi y}}$ dla $y \in (0, \pi^2\alpha)$ oraz $f_Y(y) = 0$ dla pozostałych y.

31. Szukamy funkcji gęstości zmiennej $Y = g(X) = \ln(X+1)$ dla danego X. Podobnie jak we wcześniejszym zadaniu wyznaczamy $h(y) = e^y - 1$ i stosując odpowiedni wzór, otrzymujemy $f_{X^2}(x) = xe^x$ dla $x \in (0,1)$ oraz $f_{X^2}(x) = 0$ dla pozostałych x.

32. Wiadomo, że $f_Y(x) = f_X(h(y))|h'(y)|$, o ile $Y = g(X)$, g jest funkcją monotoniczną, o funkcji odwrotnej h, zatem, przy dodatkowym założeniu $h'(y) > 0$, mamy $F_Y(y) = \int_{-\infty}^{y} f_X(h(u))h'(u)du$. Dokonując zamiany zmiennych $h(u) = t$, otrzymujemy

$$F_Y(y) = \int_{-\infty}^{h}(y) f_X(t)\mathrm{d}t = F_X(h(y)).$$

W naszym przypadku $g(x) = 2x + 1$ jest funkcją monotoniczną, funkcją do niej odwrotną jest $h(y) = (y-1)/2$. Korzystając z wyprowadzonego wzoru, otrzymamy $F_Y(y) = \ln\frac{y-1}{2}$ dla $h(y) \in (1,e)$, tzn. $y \in (3, 2e+1)$ oraz $F_Y(y) = 0$ dla $y < 3$ i $F_Y(y) = 1$ dla $y > 2e+1$.

33. a) $f_Y(y) = \begin{cases} 1/6 & \text{dla } 5 \le y \le 11 \\ 0 & \text{dla poz. } y \end{cases}$, $f_Z(z) = \begin{cases} \frac{1}{12\sqrt{z}} & \text{dla } 36 < z < 144 \\ 0 & \text{dla poz. } z \end{cases}$.

b) $f_Y(y) = \begin{cases} \frac{2}{3}e^{-\frac{2}{3}(y-2)} & \text{dla } y \ge 2 \\ 0 & \text{dla poz. } y \end{cases}$, $f_Z(z) = \begin{cases} \frac{1}{3\sqrt{z}}e^{2-\frac{2}{3}\sqrt{z}} & \text{dla } z \ge 9 \\ 0 & \text{dla poz. } z \end{cases}$.

34. Ponieważ $V(X) = E(X^2) - (E(X))^2 = 1$, $E(X) = 2$, więc $E(X^2) = 5$, $E(Y) = E(2X+1) = 2E(X) + 1 = 5$. Podobnie $V(Y) = V(Y^2) - (V(Y))^2 = E(4X^2 + 4X + 1) - (E(Y))^2 = 4 \cdot 5 + 4 \cdot 2 + 1 - 5^2 = 4$.

35. a) Współczynnik proporcjonalności wyznaczamy z warunku $P(-0,1 < X < 0,1) = k(0,1 - (-0,1)) = 1$, zatem $k = 5$,
b) $p_1 = 5(0,05 - 0) = 0,25$, $p_2 = 5(0,05 - (-0,05)) = 0,5$, c) x_1 spełnia warunek $5(x_1 - (-0,1)) = 0,25$, $x_1 = -0,5$, podobnie x_2 spełnia warunek $5(0,1 - x_2) = 0,75$, $x_2 = -0,5$,

d) $f(x) = \begin{cases} 5 & \text{dla } x \in I \\ 0 & \text{dla poz. } x \end{cases}$, $F(x) = \begin{cases} 0 & \text{dla } x \le -0,1 \\ 5x + 0,5 & \text{dla } x \in I \\ 1 & \text{dla } 0,1 \le x \end{cases}$,

e) $E(X) = 0$, $V(X) = 1/300$.

36. a) $p_1 = e^{-1/2} - e^{-1}$, $p_2 = 1 - e^{-10}$,
b) $F(x) = \begin{cases} 0 & \text{dla } x \le 0 \\ 1 - e^{-0,1x} & \text{dla } x > 0 \end{cases}$, d) $E(X) = 10$, $V(X) = 100$, e) $x = 10\ln 10$.

37. **36a)** $e^{-1/2} = 0,6065$, $e^{-1} = 0,3679$, $e^{-}10 = 0$, $p_1 = 0,9744$, p_2=1, **36e)** $\ln 10 = y$, to $e^{-y} - 0,1$, $y = 4,6$, $x = 46$.

38. **28a)** $E(X) = 0, V(X) = 1/6$, **28b)** $E(X) = e - 1, V(X) = 0,5(e-1)(e-3)$, **29a)** $E(X) = \pi - 2, V(X) = 2\pi^2 - 8 - (\pi - 8)^2 = 4p - 12$, **29b)** $E(X) = V(X) = \infty$.

39. a) $P(X = k) = (0,4)^{k-1}(0,6)$, b) $F(x) = 0$ dla $x \le 0$, $F(x) = 1 - (0,4)^k$, dla $k - 1 < x \le k$, $k = 1,2,\ldots$, c) $p_1 = P(X < 3) = F(3) = 0,84$, $p_2 = P(X \ge 3) = 1 - F(3) = 0,16$, $p_3 = F(5) - F(2) = 0,3744$, d) $E(X) = 1,67$, $\sigma = 1,05$.

40. Rozkład Poissona, $\lambda = np = 2$, $P(A) = 0,68$.

41. Rozkład Poissona, $p = 18/250$, $\lambda = np = 0,72$, $P(A) = 0,84$.

42. Utwórzmy nową zmienną $X_n = \frac{X-\mu}{\sigma}$, $X = X_n\sigma + \mu$. Zmienna X_n ma rozkład normalny $N(0,1)$. Zatem $P(|X-\mu| < k\sigma) = P(|X_n| < k) = 2P(X_n < k) - 1$. **a)** $k = 1,96$, z tablicy rozkładu normalnego odczytujemy, że $2P(X_n < k) - 1 = 2 \cdot 0,975 - 1 = 0,95$, dla $k = 2,55$ $2P(X_n < k) - 1 = 2 \cdot 0,9963 - 1 = 0,9926$, **b)** $\alpha = 0,95$, szukamy x, dla którego $2P(X_n < x) - 1 = 0,95$, czyli $P(X_n < x) = 0,975$, odczytujemy z tablic $x = 1,96$, dla $\alpha = 0,1$, $2P(X_n < x) - 1 = 0,55$, zatem $x = 0,13$.

43. **a)** Zmienna $X_n = \frac{X-3}{2}$ ma rozkład normalny, zatem dla $k = 1,6$ mamy $P(|X| < k) = P(-2,3 < X_n < -0,7) = P(0,7 < X_n < 2,3) = 0,9893 - 0,7580 = 0,2313$, dla $k = 5,8$, $P(|X| < 5,8) = P(X_n < 1,4) - P(X < -4,4) = 0,9192$. **b)** Podobnie $P(X < x) = P(X_n < \frac{x-3}{2}) = \alpha$, stąd dla $\alpha = 0,95$ mamy $x = 6.26$, dla $\alpha = 0,1$ $x = -5,56$.

44. Zmienna $X_n = \frac{X-2}{0,2}$ ma rozkład normalny oraz $P(1,7 < X < 2,3) = 2P(X_n < 1,5) - 1 = 0,8664$, zatem prawdopodobieństwo otrzymania braku wynosi $1 - 0,8664 = 0,1336$.

45. **a)** $E(X) = \int_0^{+\infty} x^2 e^{-x^2/2}dx = \sqrt{\pi/2}$, $E(X^2) = \int_0^{+\infty} x^3 e^{-x^2/2}dx = 2$, $V(X) = E(X^2) - (E(X))^2 = 4 - \pi/2$, **b)** $P(X > E(X)) = \int_{\sqrt{\pi/2}}^{+\infty} xe^{-x^2/2}dx = e^{-\pi^2/4}$.

46. Wyznaczamy dystrybuantę $F(x) = 1 - e^{-x/\lambda}$. **a)** $P(X_1 > 20 \text{ i } X_2 > 20) = P(X_1 > 20)P(X_2 > 20) = e^{-2}e^{-1} = e^{-3}$,
b) $P(X_1 < 10 \text{ lub } X_2 < 10) = P(X_1 < 10) + P(X_2 < 20) - P(X_1 < 10 \text{ i} X_2 < 10) = e^{-1}e^{-1/2} = e^{-3/2}$.

47. **a)** $P(T < 2) = 0,975$, $P(T > 1,7) = 1 - 0,95 = 0,05$, **b)** $P(T < x_0) + P(T < -x_0) = 1$, $P(T < -x_0) = 1 - 0,005 = 0,995$ stąd $-x_0 = 2,787$, $x < -2,787$. $P(T > x_0) = 0,01$, czyli $P(T < x_0) = 0,99$ $x_0 = 2,485$, $x > 2,485$, **c)** $n = 60$.

48. **a)** $P(\chi^2 < 5) = 0,025$, $P(\chi^2 > 29,8) = 0,005$, **b)** $P(\chi^2 < x) < 0,005$, $x < 7,434$, $P(\chi^2 > x) < 0,01$, $x > 37,566$, **c)** $n = 18$.

49. **a)** $P(X = 5; 0,5) = 0,000158$, $P(X = 5; 2,5) = 0,066801$, $P(X = 5; 5) = 0,175467$, $P(X \le 3; 0,5) = 0,6065 + 0,3033 + 0,0758 = 0,0126 = 0,9982$, $P(X \le 3; 2,5) = 0,757576$, $P(X \le 3; 5) = 0,265026$, **b)** $P(X = 7; \lambda_i) > p_i$, $p_1 = 0,0001$, $\lambda_1 > 1$, $p_2 = 0,001$, $\lambda_2 \ge 1,6$ $p_3 = 0,01$, $\lambda_3 > 2,5$.

50. Poniższe tabele obrazują sposób wyznaczania dopuszczalnych wartości nowych zmiennych, są to wartości "w licznikach", oraz prawdopodobieństwa ich wystąpienia, są to wartości "w mianownikach".

$X+Y$	$\frac{0}{0,1}$	$\frac{10}{0,4}$	$\frac{20}{0,4}$	$\frac{30}{0,4}$
$\frac{0}{0,3}$	$\frac{0}{0,03}$	$\frac{10}{0,12}$	$\frac{20}{0,12}$	$\frac{30}{0,03}$
$\frac{10}{0,4}$	$\frac{10}{0,04}$	$\frac{20}{0,16}$	$\frac{30}{0,16}$	$\frac{40}{0,0,04}$
$\frac{20}{0,3}$	$\frac{20}{0,03}$	$\frac{30}{0,12}$	$\frac{40}{0,12}$	$\frac{50}{0,03}$

XY	$\frac{0}{0,1}$	$\frac{10}{0,4}$	$\frac{20}{0,4}$	$\frac{30}{0,4}$
$\frac{0}{0,3}$	$\frac{0}{0,03}$	$\frac{0}{0,12}$	$\frac{0}{0,12}$	$\frac{0}{0,03}$
$\frac{10}{0,4}$	$\frac{0}{0,04}$	$\frac{100}{0,16}$	$\frac{200}{0,16}$	$\frac{300}{0,0,04}$
$\frac{20}{0,3}$	$\frac{0}{0,03}$	$\frac{200}{0,12}$	$\frac{400}{0,12}$	$\frac{600}{0,03}$

Powyższe dane zestawiamy w tabelach

$X+Y$	0	10	20	30	40	50
p_i	0,03	0,16	0,31	0,31	0,16	0,03

XY	100	200	300	400	600
p_i	0,37	0,16	0,04	0,12	0,03

$E(Z) = 25$, $V(Z) = 125$, $E(U) = 150$, $V(U) = 23900$.

51. **a)**

z_i	3	4	5	6	7	8
p_i	0,09	0,27	0,26	0,17	0,15	0,06

u_i	2	3	4	5	6	9	10	15
p_i	0,09	0,12	0,15	0,09	0,26	0,08	0,15	0,06

b)

z_i	-3	-2	-1	0	1	2	3	4	5	6
p_i	0,04	0,12	0,04	0,1	0,34	0,22	0,04	0,02	0,06	0,02

u_i	-5	-2	-1	0	1	2	5
p_i	0,02	0,08	0,1	0,6	0,1	0,08	0,02

52. **a)**

$Z_1 = X^2 + Y^2$	1	2
p	3/8	5/8

$Z_3 = XY$	-1	0	1
p	2/8	3/8	3/8

$Z_2 = 2X - 3Y$	-5	-3	-1	3	5
p	1/8	1/8	3/8	2/8	1/8

$E(X) = 1/8$, $V(X) = 39/64$, $E(Y) = 0$, $V(Y) = 1$, $E(Z_1) = 13/8$, $V(Z_1) = 15/64$, $E(Z_2) = 0$, $V(Z_2) = 10$, $E(Z_3) = 1/8$, $V(Z_3) = 39/64$,

b)

$Z_1 = X^2 + Y^2$	0	1	2
p	3/12	7/12	2/12

$Z_3 = XY$	-1	0	1
p	1/12	10/12	1/12

$Z_2 = 2X - 3Y$	-5	-3	-2	-1	0	2
p	1/12	4/12	1/12	1/2	3/12	2/12

$E(X) = 1/12$, $V(X) = 59/144$, $E(Y) = 1/2$, $V(Y) = 1/4$, $E(Z_1) = 11/12$, $V(Z_1) = 59/144$, $E(Z_2) = -4/3$, $V(Z_2) = 79/18$, $E(Z_3) = 0$, $V(Z_3) = 1/6$.

53. Wyznaczamy rozkłady brzegowe **a)** $f_1(x) = \int_{-\infty}^{+\infty} f(x,u)du = 3x^2$, dla $0 \le x \le 1$, $f_2(y) = \int_{-\infty}^{+\infty} f(v,y)dv = 2(1-y)$ dla $0 \le y \le 1$. Zmienne są niezależne, gdyż $f_1(x)f_2(y) = f(x,y)$, **b)** $f_1(x) = x + 1/2$, dla $-1 \le x \le 1$, $f_2(y) = 2y/3 + 1/2$ dla $-1 \le y \le 1$. Zmienne nie są niezależne, gdyż $f_1(x)f_2(y) \ne f(x,y)$, **c)** $f_1(x) = \frac{x}{2}$, dla $0 \le x \le 2$, $f_2(y) = e^{-y}$ dla $y \ge 0$, zmienne są niezależne.

54. **a)** $F_Z(x,y) = \begin{cases} 0 & \text{dla } x \le 0 \text{ b } y \le 0 \\ xy/2 & \text{dla } 0 \le x \le 1,\ 0 < y \le 2 \\ x & \text{dla } 0 \le x \le 1,\ ,\ y > 2 \\ y/2 & \text{dla } x > 1,\ 0 < y \le 2 \\ 1 & \text{dla } x > 1,\ y > 2 \end{cases}$,

$F_1(x) = \begin{cases} 0 & \text{dla } x \le 0 \\ x & \text{dla } 0 \le x \le 1 \\ 1 & \text{dla } x > 1 \end{cases}$, $F_2(y) = \begin{cases} 0 & \text{dla } x \le 0 \\ y/2 & \text{dla } 0 \le y \le 2 \\ 1 & \text{dla } y > 2 \end{cases}$,

) $E(XY) = 1/2 = E(X^3Y^3)$.

55. **a)** $F(x,y) = \begin{cases} 0 & \text{dla } x < 0 \text{ lub } y < 0 \\ x^2y^2/4 & \text{dla } 0 \le x \le 2 \text{ i } 0 \le y \le 1 \\ x^2/4 & \text{dla } 0 \le x \le 2 \text{ i } y > 1 \\ y^2 & \text{dla } x > 2 \text{ i } 0 \le y \le 1 \\ 1 & \text{dla } x > 2 \text{ i } y > 1 \end{cases}$,

$F_1(x) = F(x,1)$, $F_2(y) = F(2,y)$,

$$F_1(x) = \begin{cases} 0 & \text{dla } x < 0 \\ x^2/4 & \text{dla } 0 \le x \le 2 \\ 1 & \text{dla } x > 2 \end{cases}, \quad F_2(y) \begin{cases} 0 & \text{dla } y < 0 \\ y^2 & \text{dla } 0 \le y \le 1 \\ 1 & \text{dla } y > 1 \end{cases}.$$

Ponieważ $F_Z(x,y) = F_1(x) \cdot F_2(y)$, zatem zmienne są niezależne.

b) $F(x,y) = \frac{1}{2}\left[2 + e^{-x}(2x+1) + e^{-y}(2y+1) + (x+y+2)e^{-(x+y)}\right]$ dla $x > 0$ i $y > 0$ oraz $F(x,y) = 0$ dla pozostałych (x,y).

$F_1(x) = \lim\limits_{y \to +\infty} F(x,y) = 1 + \frac{1}{2}e^{-x}(2x+1)$ dla $x > 0$ oraz $F_1(x) = 0$ dla pozostałych x. Podobnie $F_2(y) = 1 + \frac{1}{2}e^{-y}(2y+1)$ dla $y > 0$ oraz $F_2(y) = 0$ dla pozostałych y. Zmienne nie są niezależne.

56. a)
$$F(x,y) = \begin{cases} 0 & \text{dla } x < 0 \text{ lub } y < 0 \\ y(2-y) & \text{dla } x > 1 \text{ i } 0 \le y \le 1 \\ x^3 & \text{dla } 0 \le x \le 1 \text{ i } y > 1 \\ x^3(2-y) & \text{dla } 0 \le x \le 1 \text{ i } 0 \le y \le 1 \\ 1 & \text{dla pozostałych } x, y \end{cases},$$

b)
$$F(x,y) = \begin{cases} 0 & \text{dla } x < -1 \text{ lub } y < -1 \\ (y+1)(2y+1)/6 & \text{dla } x > 1 \text{ i } -1 \le y \le 1 \\ x(x+1)/2 & \text{dla } -1 \le x \le 1 \text{ i } y > 1 \\ (y+1)(2y+1)(3x+2y-2) & \text{dla } -1 \le x \le 1 \text{ i } -1 \le y \le 1 \\ 1 & \text{dla pozostałych } x, y \end{cases},$$

c)
$$F(x,y) = \begin{cases} 0 & \text{dla } x < 0 \text{ lub } y < 0 \\ 1 - e^{-y} & \text{dla } x > 2 \text{ i } y > 0 \\ x^2(1-e^{-y})/4 & \text{dla } 0 \le x \le 2 \text{ i } y > 0 \end{cases}.$$

57. a) Wyznaczamy gęstość prawdopodobieństwa $f(x,y) = \frac{\partial^2 F}{\partial x \partial y}(x,y) = 6xy(2-x-y)$ dla $0 \le x \le 1$ i $0 \le y \le 1$ oraz $f_Z(x,y) = 0$ dla pozostałych (x,y). Stąd wyznaczamy rozkłady brzegowe $f_1(x) = \int_0^1 f(x,y)dy = x(4-3x)$, $f_2(y) = \int_0^1 f(x,y)dx = y(4-3y)$ oraz rozkłady warunkowe $f(x|y) = \int_0^x \frac{f(u,y)}{f_2(y)}du = x^2\frac{2x+3y-6}{y(3y-4)}$, dla $0 \le x \le 1$ oraz $f(x|y) = 0$ dla poz. x, $f(y|x) = \int_0^y \frac{f(x,v)}{f_1(x)}dv = y^2\frac{3x+2y-6}{x(3x-4)}$ dla $0 \le y \le 1$ oraz $f(y|x) = 0$ dla poz. y. **b)** $f(x,y) = 4xy/9$ dla $0 \le x \le 1$ i $0 \le y \le 3$, oraz $f_1(x) = 2x$ dla $0 \le x \le 1$, $f_2(y) = 2y/9$ dla $0 \le x \le 3$. $f(x|y) = x^2$ dla dla $0 \le x \le 1$ oraz $f(x|y) = 0$ dla poz. x, $f(y|x) = y^2/9$ dla $0 \le y \le 3$ oraz $f(y|x) = 0$ dla poz. y.

58. Rozkład prawdopodobieństwa: $p_{i,j} = 0,2$ dla $i = j$ oraz $p_{i,j} = 0$ dla $i \neq j$.
a) Wobec tego momenty zwykłe wyznaczamy z zależności $m_{k,l} = \sum_1^5 x_i^k y_i$, i tak mamy: $m_{1,0} = 5,5$, $m_{0,1} = 2,946$, $m_{2,0} = 45,95$, $m_{0,2} = 14,086$, $m_{1,1} = 25,38$.
b) Momenty centralne: $\mu_{1,1} = m_{1,1} - m_{1,0}m_{0,1} = 9,177$, $\mu_{2,0} = m_{2,0} - (m_{1,0})^2 = 15,7$, $\sigma_1 = \sqrt{m_{2,0}} = 3,962$, $\mu_{0,2} = m_{0,2} - (m_{0,1})^2 = 5,422$, $\sigma_2 = \sqrt{m_{0,2}} = 2,325$,
c) Współczynnik korelacji $\rho = \frac{\mu_{1,1}}{\sigma_1\sigma_2} = 0,99$.
d) Prosta regresji (patrz (1.10)) $y - 2,946 = 0,584(x-5,5)$, $y - 2,946 = 0,589(x-5,5)$.
e) Erozję odpowiadającą kolejnym nachyleniom wyznaczamy z równania pierwszej prostej: $y(x) = 2,946 + 0,584(x-5,5)$, $y(2) = 0,9$, $y(6) = 3,26$, $y(10) = 5.58$, nachylenia zaś odpowiadające kolejnym erozjom z równania drugiej prostej: $x(y) = 1,697y + 0,5$, $x(0,5) = 1,348$, $x(5) = 8,98$.

59. a) $\rho = 0,8$, **b)** $y - 3,705 = 0,857(x - 4,06)$, $y - 3,705 = 1,306(x - 4,06)$.

60. a) Momenty zwykłe wyznaczamy ze wzoru $m_{k,l} \int\int x^k y^l f(x,y)dxdy$, $m_{0,1} = m_{1,0} = 7/12$, $m_{1,1} = 1/3$, $m_{2,0} = m_{0,2} = 5/12$, z odpowiednich wzorów wyznaczamy $\mu_{1,1} = -1/144$, $\mu_{2,0} = \mu 0,2 = 11/144$, oraz $\rho = 11/144$. **C)** proste regresji II rodzaju: $y - 7/12 = -(x - 7/12)/11$ (zmiennej Y względem X), $y - 7/12 = -11(x - 7/12)$ (zmiennej X względem Y). **d)** Rozkłady brzegowe: $f_1(x) = 1/2 + x$ dla $x \in (0,1)$, $f_1(x) = 0$ dla poz. x, $f_2(y) = 1/2 + y$ dla $y \in (0,1)$, $f_2(y) = 0$ dla poz. y.
e) Rozkłady warunkowe wyznaczamy ze wzoru $f(x/y) = \frac{f(x,y)}{f_2(y)} = 2\frac{x+y}{y+2}$ oraz $f(y/x) = 2\frac{x+y}{x+2}$. Następnie liczymy $E(X/y) = \int_0^1 x \cdot 2\frac{x+y}{x+2}dx = \frac{1}{3}\frac{2+3y}{2+y}$, $E(Y/x) = \frac{1}{3}\frac{2+3x}{2+x}$.
f) Linie regresji I rodzaju: $y = E(Y/x)$ oraz $x = E(X/y)$, czyli $y = \frac{1}{3}\frac{2+3x}{2+x}$, $x = \frac{1}{3}\frac{2+3y}{2+y}$.

61. a) $\varphi(t) = \sum_k e^{itx_k}p_k = e^{it0}q + e^{it1}p = 1 + pe^{it}$, **b)** $\varphi(t) = \frac{1}{8}(e^{it}+1)^3$,
c) $\varphi(t) = \frac{1}{1-\lambda it}$,

62. a) b) Funkcje nie spełniają warunku $\varphi(-t) = \overline{\varphi(t)}$. **c)** Funkcja nie spełnia warunku $|\varphi(t)| \le 1$.

63. a) $\varphi(t) = \frac{1}{1-it}$, $\varphi'(t) = \frac{i}{(1-it)^2}$, $\varphi''(t) = \frac{-2}{(1-it)^3}$, $E(X) = -i\varphi'(0) = 1$, $V(X) = [\varphi''(0)]^2 + \varphi''(0) = 3$, **b)** $\varphi(t) = \frac{1}{1+t^2}$, $E(X) = 0$, $V(X) = 3$.

64.

Numer klasy	Przedział klasowy	Środek klasy $\overline{x}_i$	Liczność klasy n_i	Częstość klasy n_i/n	Częstość skumulowana $\sum_{l=1}^{i} x_l/n$
1	0,975 - 1,025	1,00	6	3/25	3/25
2	1,025 - 1,075	1,05	11	11/50	17/50
3	1,075 - 1,125	1,10	10	1/5	25/50
4	1,125 - 1,175	1,15	9	9/50	17/25
5	1,175 - 1,225	1,20	7	7/50	43/50
6	2,225 - 2,275	1,25	7	7/50	1

$\overline{x} = 1,121$, $\sigma^2 = 0,0063$, $\sigma = 0,0794$.

65.

Numer klasy	Przedział klasowy	Środek klasy $\overline{x}_i$	Liczność klasy n_i	Częstość klasy n_i/n	Częstość skumulowana $\sum_{l=1}^{i} x_l/n$
1	20 - 24	22	15	1/10	1/10
2	24 - 28	26	20	2/15	7/30
3	28 - 32	30	50	1/3	17/30
4	32 - 36	34	40	4/15	5/6
5	36 - 40	38	25	1/6	1

$\overline{x} = 26,53$, $\sigma^2 = 5,582$, $\sigma = 2,362$.

66. Funkcja wiarygodności: $L = \Pi_{i=1}^n p(1-p)^{X_i - 1}$ oraz $\ln L = n \ln p + n(\overline{X} - 1)\ln(1-p)$, gdzie $\overline{X} = \frac{1}{n}\sum_1^n X_i$. Rozwiązując równanie $(\ln L)'_p = 0$, otrzymujemy $\hat{p} = \frac{1}{\overline{X}}$.

67. $E(\hat{\sigma}) = E\left(k\sum_1^n |X_i - \mu|\right) = k\sum_1^n E(|X_i - \mu|) = \frac{kn}{\sigma\sqrt{2\pi}}\int_{-\infty}^{+\infty} |x-\mu| e^{-\frac{(x-\mu)^2}{2\sigma^2}}dx = \frac{2kn}{\sigma\sqrt{2\pi}}\int_{\mu}^{+\infty}(x-\mu)e^{-\frac{(x-\mu)^2}{2\sigma^2}}dx = \frac{kn}{\sqrt{2\pi}}\int_0^{+\infty} e^{-z/2}dz = \frac{2kn\sigma}{\sqrt{2\pi}}$. (Podstawienie $((x-\mu)/\sigma = z)$. Z równości $E(\hat{\sigma}) = \sigma$ otrzymamy $k = \frac{1}{n}\sqrt{\pi/2}$.

68. Wiemy, że $f(x) = \frac{1}{\lambda}e^{-x/\lambda}$. Funkcja wiarygodności: $\Pi_1^n \frac{1}{\lambda}e^{-X_i/\lambda}$, $\ln L = \frac{1}{\lambda}n\overline{X} - \ln\lambda$. Równanie wiarygodności: $(\ln L)'_\lambda = 0$, $\frac{n\overline{X}}{\lambda^2} - \frac{1}{\lambda} = 0$, $\hat{\lambda} = n\overline{X} = \sum_1^n X_i$.

69. Wiadomo, że $f(x) = \frac{1}{\lambda}e^{-x/\lambda}$, $E(X) = \lambda$, $E(X^2) = 2\lambda^2$. Wyznaczamy $E(\hat{\lambda}^2) = \frac{1}{2n}\sum_1^n E(X_i^2) = \frac{1}{2n}\sum_1^n 2\lambda^2 = \lambda^2$. Estymator jest nieobciążony.

70. Wyznaczamy $\overline{x} = 5,11$, $s^2 = 3,02$. Przedział ufności dla wartości średniej: $W = \left(\overline{x} - \frac{t\sigma}{\sqrt{n}}; \overline{x} + \frac{t\sigma}{\sqrt{n}}\right)$, gdzie $\Phi(t) = 1 + \alpha/2 = 0,975$, $t = 1,96$, $W = (4,37; 5,86)$. Przedział ufności dla wariancji: $W = \left(\frac{nS^2}{x_2}, \frac{nS^2}{x_1}\right)$, gdzie x_2 oraz x_1 odczytujemy z tablic kwantyli rozkładu χ^2 dla $1 + \alpha/2$ i $1 - \alpha/2$ przy $n - 1$ stopniach swobody, w tym przypadku $x_2 = 35,479$, $x_1 = 10,283$, $W = (1,87; 6,461)$.

71. Wyznaczamy $\overline{x} = 4,12$, $s^2 = 0,059$. Przedział ufności dla wartości średniej i wariancji patrz poprzednie zadanie: $t = 1,64$, $W = (4,02; 4,22)$. $x_2 = 23,685$, $x_1 = 6,571$, $W = (0,037; 0,135)$.

72. Patrz poprzednie zadania, $t = 1,56$, $W = (1,561 \cdot 10^{-19}, 1,587 \cdot 10^{-19})$.

73. Patrz punkt VII° w Tabl. 5.8, $n_1 = 8$, $\overline{x_1} = 5,875$, $\sigma^2 = 0,192$, $n_1 = 6$, $\overline{x_2} = 4,88$, $\sigma^2 = 0,018$, $u_{obl} = 6,055$, $W = (1,28; +\infty)$, $u_{obl} \in W$. Hipotezę odrzucamy na rzecz hipotezy o zmniejszeniu zużycia paliwa.

74. Wyznaczamy $\overline{x} = 22$, $s^2 = 7,5$, $s = 2,74$, $t_{obl} = \frac{\overline{x} - \mu_0}{s}\sqrt{n-1} = -1,93$.
a) $W = (-\infty; -2,365) \cup (2,365; +\infty)$, $t_{obl} \notin W$. Nie ma podstaw do odrzucenia hipotezy. **b)** $W = (-\infty; -1,895)$, $t_{obl} \in W$. Hipotezę o średniej temperaturze równej 24 odrzucamy na rzecz hipotezy, że średnia temperatura jest mniejsza od 24.

75. Patrz punkt IV° w Tabl. 5.8. $n = 10$, $\overline{x} = 12,02$, $s^2 = 0,0106$, $\chi^2_{obl} = 3,029$, **a)** $W = (0; 2,700) \cup (19,023)$, $\chi^2_{obl} \notin W$ nie ma podstaw do odrzucenia hipotezy na rzecz hipotezy $\sigma^2 \neq 0,035$, **b)** $W = (0; 3,325)$, $\chi^2_{obl} \in W$. Hipotezę odrzucamy na rzecz hipotezy $\sigma^2 < 0,035$, **c)** $W = (16,819; +\infty)$, nie ma podstaw do odrzucenia hipotezy.

76. Wyznaczamy z próby $\overline{x} = 0,238$, $s = 0,0125$. **a)** Patrz punkt II° w Tabl. 5.8 $t_{obl} = \frac{\overline{x} - \mu_0}{s}\sqrt{n-1} = -2,882$, $W = (-\infty; -2,821)$, $t_{obl} \in W$, hipotezę odrzucamy na rzecz hipotezy alternatywnej. **b)** Patrz punkt I° w Tabl. 5.8 $u_{obl} = \frac{\overline{x} - \mu_0}{\sigma}\sqrt{n} = -2,52$, $W = (-\infty, -2,33)$, $u_{obl} \in W$, hipotezę odrzucamy na rzecz hipotezy alternatywnej.

77. Patrz punkt V° w Tabl. 5.8, $u_{obl} = \sqrt{\frac{2ns^2}{\sigma_0^2}} - \sqrt{2n-3} = 3,325$, $W = (1,28, \infty)$, $u_{obl} \in W$, hipotezę odrzucamy.

78. Wyznaczamy $\overline{x} = 7,014$, $s = 0,273$, $s^* = 2,75$. Metoda I: Korzystamy z przedziału ufności dla $n \geq 50$ (punkt V° w Tabl. 5.8) i otrzymujemy $u_{obl} = 24,952$, $W = (1,28, \infty)$, $u_{obl} \in W$, hipotezę odrzucamy.
Metoda II: Korzystamy z przedziału ufności dla $n \geq 100$ (punkt VI° w Tabl. 5.8) i otrzymujemy $u_{obl} = 0,935$, $W = (1,28, \infty)$, $u_{obl} \notin W$, nie ma podstaw do odrzucenia hipotezy.

79. $\overline{x_1} = 1020$, $\overline{x_2} = 773$, $s_1 = 228$, $s_2 = 171$. **a)** Korzystamy z przedziału ufności, punkt VII° w Tabl. 5.8, $U_{obl} = 2,1224$, $W = (2,88; \infty)$, $U_{obl} \notin W$. Nie ma podstaw do odrzucenia hipotezy. **b)** Korzystamy z przedziału ufności, punkt VI° w Tabl. 5.8, $U_{obl} = 2,1007$, $W = (2,88; \infty)$, $U_{obl} \notin W$. Nie ma podstaw do odrzucenia hipotezy.

80. **a)** Patrz punkt VII° w Tabl. 5.8, $u_{obl} = 7,21$, $W = (-\infty, -2,58) \cup (2,58, +\infty)$, $u_{obl} \in W$. Hipotezę odrzucamy. **b)** Patrz punkt VI° w Tabl. 5.8, $u_{obl} = 10,1$, $W = (2,33; +\infty)$, $u_{obl} \in W$. Hipotezę odrzucamy na rzecz hipotezy $K : \sigma_1^2 > 0,2$. **c)** Patrz punkt III° w Tabl. 5.8, $u_{obl} = 0,79$, $W = (2,33; +\infty)$ $u_{obl} \notin W$. Nie ma podstaw do odrzucenia hipotezy. **d)** Podobnie jak w c), $u_{obl} = -4,47$, $W = (-\infty; -2,33)$, $u_{obl} \in W$. Hipotezę odrzucamy.

Rozdział 5

Tablice statystyczne

Tablica 5.1 Wartości funkcji wykładniczej $f(x) = e^{-x}$

x	0	1	2	3	4	5	6	7	8	9
0.00	1,0000	0,9990	0,9980	0,9970	0,9960	0,9950	0,9940	0,9930	0,9920	0,9910
0.0	1,0000	0,9900	0,9802	0,9704	0,9608	0,9512	0,9418	0,9324	0,9231	0,9139
0,1	0,9048	,8958	,8869	,8781	,8694	,8607	,8521	,8437	,8353	,8270
0,2	,8187	,8106	,8025	,7945	,7866	,7788	,7711	,7634	,7558	,7483
0,3	,7408	,7334	,7261	,7189	,7118	,7047	,6977	,6907	,6839	,6771
0,4	,6703	,6637	,6570	,6505	,6440	,6376	,6313	,6250	,6188	,6126
0,5	0,6065	0,6005	0,5945	0,5886	0,5827	0,5769	0,5712	0,5655	0,5599	0,5543
0,6	,5488	,5434	,5379	,5326	,5273	,5220	,5169	,5117	,5066	,5016
0,7	,4966	,4916	,4868	,4819	,4771	,4724	,4677	,4630	,4584	,4538
0,8	,4493	,4449	,4404	,4360	,4317	,4274	,4232	,4190	,4148	,4107
0,9	,4066	,4025	,3985	,3946	,3906	,3867	,3829	,3791	,3753	,3716
1,	0,3679	0,3329	0,3012	0,2725	0,2466	0,2231	0,2019	0,1827	0,1653	0,1496
2,	,1353	,1225	,1108	,1003	,0907	,0821	,0743	,0672	,0608	,0550
3,	,0498	,0450	,0408	,0369	,0334	,0302	,0273	,0247	,0224	,0202
4,	,0183	,0166	,0150	,0136	,0123	,0111	,0100	,0091	,0082	,0074
5,	,0067	,0061	,0055	,0050	,0045	,0041	,0037	,0034	,0030	,0027
6,	0,0025	0,0022	0,0020	0,0018	0,0017	0,0015	0,0014	0,0012	0,0011	0,0010
7,	,0009	,0008	,0007	,0007	,0006	,0006	,0005	,0005	,0004	,0004
8,	,0003	,0003	,0003	,0002	,0002	,0002	,0002	,0002	,0002	,0001
9,	,0001	,0001	,0001	,0001	,0001	,0001	,0001	,0001	,0001	,0000

Tablica 5.2 Niektóre rozkłady prawdopodobieństwa

Nazwa rozkładu	Gęstość albo funkcja prawdopodobieństwa	Parametry rozkładu	EX	D^2X
równomierny	$\frac{1}{b-a}$ dla $a \le x \le b$	$a, b \in R$	$\frac{a+b}{b}$	$\frac{(b-a)^2}{12}$
wykładniczy	$\frac{1}{\lambda}e^{-x/\lambda}$, $x > 0$	$\lambda > 0$	λ	λ^2
gamma	$\frac{x^{p-1}e^{-x/\lambda}}{\lambda^p \Gamma(p)}$, $x > 0$ [1]	$p, \lambda > 0$	λp	$\lambda^2 p$
beta	$\frac{x^{p-1}(1-x)^{q-1}}{B(p,q)}$, $0 < x < 1$ [2]	$p, q > 0$	$\frac{p}{p+q}$	$\frac{p}{(p+q)^2} \cdot \frac{q}{(p+q+1)}$
normalny	$\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$, $x \in R$	$\mu \in R$, $\sigma > 0$	μ	σ^2
Laplace'a	$\frac{1}{2\lambda} \exp\left(-\frac{\lvert x-\mu\rvert}{\lambda}\right)$, $x \in R$	$\lambda > 0$ $\mu \in R$,	μ	$2\lambda^2$
χ^2 z ν stopniami swobody	$\frac{1}{2^{\nu/2}\Gamma\left(\frac{1}{2}\nu\right)} x^{\nu/2-1}e^{-x/2}$, $x > 0$	$\nu \in N$	ν	2ν
t Studenta z ν stopniami swobody	$\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{\nu}{2}\right)\sqrt{\nu}\left(1+\frac{x^2}{\nu}\right)^{(\nu+1)/2}}$, $x \in R$	$\nu \in N$	0	$\frac{\nu}{\nu-2}$ dla $\nu > 2$
zero-jedynkowy	$P(X=1) = p$, $P(X=0) = 1-p$	$p \in (0,1)$	p	$p(1-p)$
dwumianowy	$\binom{n}{k}p^k(1-p)^{n-k}$, $k = 0, 1, \ldots, n$	$n \in N$ $p \in (0,1)$	np	$np(1-p)$
Poissona	$\frac{\lambda^k}{k!}$, $k = 0, 1, 2, \ldots$	$\lambda > 0$	λ	λ
geometryczny	$p(1-p)^{k-1}$, $k = 1, 2, \ldots$	$p \in (0,1)$	$\frac{1}{p}$	$\frac{1-p}{p^2}$

[1] $\Gamma(p) = \int_0^\infty x^{p-1}e^{-x}dx$ dla $p > 0$; dla $p_1 = n \in N$, $\Gamma(n) = (n-1)!$.

[2] $B(p,q) = \int_0^1 x^{p-1}(1-x)^{q-1}dx$, $p, q > 0$, $B(p,q) = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}$.

Tablica 5.3 Rozkład Poissona

Wartości $P(k;\lambda)$ funkcji prawdopodobieństwa rozkładu Poissona.

Dla ustalonego λ odczytujemy prawdopodobieństwo $p_k = \frac{\lambda^k}{k!}e^{-\lambda}$ uzyskania k sukcesów w rozkładzie Poissona. [1]

k	λ						
	0,001	0,002	0,003	0,004	0,005	0,006	0,007
0	0,9990	0,9980	0,9970	0,9960	0,9950	0,9940	0,9930
1	,$0^3$999	,$0^2$200	,$0^2$299	,$0^2$398	,$0^2$498	,$0^2$596	,$0^2$695
2	,$0^5$1	,$0^5$200	,$0^5$449	,$0^5$797	,$0^4$124	,$0^4$179	,$0^4$243
k	λ						
	0,008	0,009	0,010	0,020	0,030	0,040	0,050
0	0,9920	0,9910	0,9900	0,9802	0,9704	0,9608	0,9512
1	,$0^2$794	,$0^2$892	,$0^2$990	,0196	,0291	,0384	,0476
2	,$0^4$317	,$0^4$401	,$0^4$495	,$0^3$196	,$0^3$437	,$0^3$769	,$0^2$119
3	–	–	–	,$0^5$131	,$0^5$437	,$0^4$102	,$0^4$198
k	λ						
	0,06	0,07	0,08	0,09	0,10	0,20	0,30
0	0,9418	0,9324	0,9231	0,9139	0,9048	0,8187	0,7408
1	,0565	,0653	,0738	,0823	,0905	,1637	,2222
2	,$0^2$170	,$0^2$228	,$0^2$295	,$0^2$370	,$0^2$452	,0164	,0333
3	,$0^4$339	,$0^4$533	,$0^4$788	,$0^3$111	,$0^3$151	,$0^2$109	,$0^2$333
4	0,^{5}1	,$0^6$93	,$0^5$158	,$0^5$250	,$0^5$377	,$0^4$546	,$0^3$250
5	–	–	–	–	–	,$0^5$218	,$0^4$150
k	λ						
	0,4	0,5	0,6	0,7	0,8	0,9	1,0
0	0,6703	0,6065	0,5488	0,4966	0,4493	0,4066	0,3679
1	,2681	,3033	,3293	,3476	,3595	,3659	,3679
2	,0536	,0758	,0988	,1217	,1438	,1647	,1839
3	,$0^2$715	,0126	,0198	,0284	,0383	,0494	,0613
4	,$0^3$715	,$0^2$158	,$0^2$296	,$0^2$497	,$0^2$767	,0111	,0153
5	,$0^4$572	,$0^3$158	,$0^3$356	,$0^3$696	,$0^2$123	,$0^2$200	,$0^2$307
6	,$0^5$381	,$0^4$132	,$0^4$356	,$0^4$811	,$0^3$164	,$0^3$300	,$0^3$511
7	–	,$0^5$1	,$0^5$305	,$0^5$811	,$0^4$187	,$0^4$386	,$0^4$730
8	–	–	–	,$0^6$1	,$0^5$187	,$0^5$434	,$0^5$912
9	–	–	–	–	–	–	,$0^6$101

[1] Symbol 0^k234 oznacza, że 0 powinno być powtórzone k razy.

Rozkład Poissona cd.

k	λ									
	1,1	1,2	1,3	1,4	1,5	1,6	1,7	1,8	1,9	2,0
0	0,332871	0,301194	0,272532	0,246597	0,223130	0,201897	0,182684	0,165299	0,149569	0,135335
1	,366158	,361433	,354291	,345236	,334695	,323034	,310562	,297538	,284180	,270671
2	,201387	,216860	,230289	,241665	,251021	,258428	,263978	,267784	,269971	,270671
3	,073842	,086744	,099792	,112777	,125510	,137828	,149587	,160671	,170982	,180447
4	,020307	,026023	,032432	,039472	,047067	,055131	,063575	,072302	,081216	,090224
5	,004467	,006246	,008432	,011052	,014120	,017642	,021615	,026029	,030862	,036089
6	,000819	,001249	,001827	,002579	,003530	,004705	,006124	,007809	,009773	,012030
7	,000129	,000214	,000339	,000516	,000756	,001075	,001487	,002008	,002653	,003437
8	,000018	,000032	,000055	,000090	,000142	,000215	,000316	,000452	,000630	,000859
9	,000002	,000004	,000008	,000014	,000024	,000038	,000060	,000090	,000133	,000191
10	–	,000001	,000001	,000002	,000004	,000006	,000010	,000016	,000025	,000038
11	–	–	–	–	–	,000001	,000002	,000003	,000004	,000007
12	–	–	–	–	–	–	–	–	,000001	,000001

k	λ									
	2,1	2,2	2,3	2,4	2,5	2,6	2,7	2,8	2,9	3,0
0	0,122456	0,110803	0,100259	0,090718	0,082085	0,074274	0,067206	0,060810	0,055023	0,049787
1	,257159	,243767	,230595	,217723	,205212	,193111	,181455	,170268	,159567	,149361
2	,270016	,268144	,265185	,261268	,256516	,251045	,244964	,238375	,231373	,224042
3	,189012	,196639	,203308	,209014	,213763	,217572	,220468	,222484	,223660	,224042
4	,099231	,108151	,116902	,125409	,133602	,141422	,148816	,155739	,162154	,168031
5	,041677	,047587	,053775	,060196	,066801	,073539	,080360	,087214	,094049	,100819
6	,014587	,017448	,020614	,024078	,027834	,031867	,036162	,040700	,045457	,050409
7	,004376	,005484	,006773	,008255	,009941	,011836	,013948	,016280	,018832	,021604
8	,001149	,001508	,001947	,002477	,003106	,003847	,004708	,005698	,006827	,008102
9	,000268	,000369	,000498	,000660	,000863	,001111	,001412	,001773	,002200	,002701
10	,000056	,000081	,000114	,000158	,000216	,000289	,000381	,000496	,000638	,000310
11	,000011	,000016	,000024	,000035	,000049	,000068	,000094	,000126	,000168	,000221
12	,000002	,000003	,000005	,000007	,000010	,000015	,000021	,000029	,000041	,000055
13	–	,000001	,000001	,000001	,000002	,000003	,000004	,000006	,000009	,000013
14	–	–	–	–	–	,000001	,000001	,000001	,000002	,000003
15	–	–	–	–	–	–	–	–	–	,000001

Rozkład Poissona cd.

k	λ									
	3,1	3,2	3,3	3,4	3,5	3,6	3,7	3,8	3,9	4,0
0	0,045049	0,040762	0,036883	0,033373	0,030197	0,027324	0,024724	0,022371	0,020242	0,018316
1	,139653	,130439	,121714	,113469	,105691	,098365	,091477	,085009	,078943	,073263
2	,216461	,208702	,200829	,192898	,184959	,177058	,169233	,161517	,153940	,146525
3	,223677	,222616	,220912	,218617	,215785	,212469	,208720	,204588	,200122	,195367
4	,173350	,178093	,182252	,185825	,188812	,191222	,193066	,194359	,195119	,195367
5	,107477	,113979	,120286	,126361	,132169	,137680	,142869	,147713	,152193	,156293
6	,055530	,060789	,066158	,071604	,077098	,082608	,088102	0,93551	0,98925	,104196
7	,024592	,027789	,031189	,034779	,038549	,042484	,046568	,050785	,055115	,059540
8	,009529	,011116	,012865	,014781	,016865	,019118	,021538	,024123	,026869	,029770
9	,003282	,003952	,004717	,005584	,006559	,007647	,008854	,010185	,011643	,013231
10	,001018	,001265	,001557	,001899	,002296	,002753	,003276	,003870	,004541	,005292
11	,000287	,000368	,000467	,000587	,000730	,000901	,001102	,001337	,001610	,001925
12	,000074	,000098	,000128	,000166	,000213	,000270	,000340	,000423	,000523	,000642
13	,000018	,000024	,000033	,000043	,000057	,000075	,000097	,000124	,000157	,000197
14	,000004	,000006	,000008	,000011	,000014	,000019	,000026	,000034	,000044	,000056
15	,000001	,000001	,000002	,000002	,000003	,000005	,000006	,000009	,000011	,000015
16	–	–	–	,000001	,000001	,000001	,000001	,000002	,000003	,000004
17	–	–	–	–	–	–	–	–	,000001	,000001

k	λ									
	4,1	4,2	4,3	4,4	4,5	4,6	4,7	4,8	4,9	5,0
0	0,016573	0,014996	0,013569	0,012277	0,011109	0,010052	0,009095	0,008230	0,007447	0,006738
1	,067948	,062981	,058345	,054020	,049990	,046238	,042748	,039503	,036488	,033690
2	,139293	,132261	,125441	,118845	,112479	,106348	,100457	,094807	,089396	,084224
3	,190368	,185165	,179799	,174305	,168718	,163068	,157383	,151691	,146014	,140374
4	,195127	,194424	,193284	,191736	,189808	,187528	,184925	,182029	,178867	,175467
5	,160004	,163316	,166224	,168728	,170827	,172525	,173830	,174748	,175290	,175467
6	,109336	,114321	,119127	,123734	,128120	,132270	,136167	,139798	,143153	,146223
7	,064040	,068593	,073178	,077775	,082363	,086920	,091426	,095862	,100207	,104445
8	,032820	,036011	,039333	,042776	,046329	,049979	,053713	,057517	,061377	,065278
9	,014951	,016805	,018793	,020913	,023165	,025545	,028050	,030676	,033416	,036266
10	,006130	,007058	,008081	,009202	,010424	,011751	,013184	,014724	,016374	,018133
11	,002285	,002695	,003159	,003681	,004264	,004914	,005633	,006425	,007294	,008242
12	,000781	,000943	,001132	,001350	,001599	,001884	,002206	,002570	,002978	,003434
13	,000246	,000305	,000374	,000457	,000554	,000667	,000798	,000949	,001123	,001321
14	,000072	,000091	,000115	,000144	,000178	,000219	,000268	,000325	,000393	,000472
15	,000020	,000026	,000033	,000042	,000053	,000067	,000084	,000104	,000128	,000157
16	,000005	,000007	,000009	,000012	,000015	,000019	,000025	,000031	,000039	,000049
17	,000001	,000002	,000002	,000003	,000004	,000005	,000007	,000009	,000011	,000014
18	–	–	,000001	,000001	,000001	,000001	,000002	,000002	,000003	,000004
19	–	–	–	–	–	–	–	,000001	,000001	,000001

Tablica 5.4 Rozkład normalny

Dystrybuanta $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{1}{2}u^2\right) du$ rozkładu normalnego $N(0,1)$.

Dla ustalonego x (pierwsza kolumna – wartość z przybliżeniem 0,01, pierwszy wiersz – setne części x) z tabeli odczytujemy $\alpha = P(N < x)$. Do korzystaniu z tablicy przydatny jest wzór $\Phi(-x) + \Phi(x) = 1$.

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	,5398	,5438	,5478	,5517	,5557	,5596	,5636	,5675	,5714	,5753
0,2	,5793	,5832	,5871	,5910	,5948	,5987	,6026	,6064	,6103	,6141
0,3	,6179	,6217	,6255	,6293	,6331	,6368	,6406	,6443	,6480	,6517
0,4	,6554	,6591	,6628	,6664	,6700	,6736	,6772	,6808	,6844	,6879
0,5	,6915	,6950	,6985	,7019	,7054	,7088	,7123	,7157	,7190	7224
0,6	,7257	,7290	,7324	,7357	,7389	,7422	,7454	,7486	,7517	,7549
0,7	,7580	,7611	,7642	,7673	,7704	,7734	,7764	,7794	,7823	,7852
0,8	,7881	,7910	,7939	,7967	,7995	,8023	,8051	,8078	,8106	,8133
0,9	,8159	,8186	,8212	,8238	,8264	,8289	,8340	,8340	,8365	,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	,8643	,8665	,8686	,8708	,8729	,8749	,8770	,8790	,8810	,8830
1,2	,8849	,8869	,8888	,8907	,8925	,8944	,8962	,8980	,8997	,9015
1,3	,9032	,9049	,9066	,9082	,9099	,9115	,9131	,9147	,9162	,9177
1,4	,9192	,9207	,9222	,9236	,9251	,9265	,9279	,9292	,9306	,9319
1,5	,9332	,9345	,9357	,9370	,9382	,9394	,9406	,9418	,9429	,9441
1,6	,9452	,9463	,9474	,9484	,9495	,9505	,9515	,9525	,9535	,9545
1,7	,9554	,9564	,9573	,9582	,9591	,9599	,9608	,9616	,9625	,9633
1,8	,9641	,9649	,9656	,9664	,9671	,9678	,9686	,9693	,9699	,9706
1,9	,9713	,9719	,9726	,9732	,9738	,9744	,9750	,9756	,9761	,9767
2,0	0,9772	0,9779	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	,9821	,9826	,9830	,9834	,9838	,9842	,9846	,9850	,9854	,9857
2,2	,9861	,9864	,9868	,9871	,9875	,9878	,9881	,9884	,9887	,9890
2,3	,9893	,9896	,9898	,9901	,9904	,9906	,9909	,9911	,9913	,9916
2,4	,9918	,9920	,9922	,9925	,9927	,9929	,9931	,9932	,9934	,9936
2,5	,9938	,9940	,9941	,9943	,9945	,9946	,9948	,9949	,9951	,9952
2,6	,9953	,9955	,9956	,9957	,9959	,9960	,9961	,9962	,9963	,9964
2,7	,9965	,9966	,9967	,9968	,9969	,9970	,9971	,9972	,9973	,9974
2,8	,9974	,9975	,9976	,9977	,9977	,9978	,9979	,9979	,9980	,9981
2,9	,9981	,9982	,9982	,9983	,9984	,9984	,9985	,9985	,9986	,9986

Tablica 5.5 Kwantyle rozkładu normalnego

Dla ustalonego p z tablicy odczytujemy wartość $u(p)$, taką że $P(N < u) = p$ lub równoważnie $P(N > u) = 1 - p$.

p	0,90	0,95	0,975	0,99	0,995
$u(p)$	1,28	1,64	1,96	2,33	2,58

Tablica 5.6 Kwantyle rozkładu t Studenta

Kwantyle $t(p,n)$ rzędu p rozkładu t Studenta o n stopniach swobody.

Przy ustalonym p i n tablica podaje wartość t_p, taką że $P(X < t_p) = p$, gdzie X jest zmienną losową rozkładu t Studenta o n stopniach swobody, to znaczy

$$\frac{1}{\sqrt{n}B(1/2, n/2)} \int_{-\infty}^{t_p} \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}} = p.$$

n	p				
	0,90	0,95	0,975	0,99	0,995
1	3,078	6,314	12,706	31,821	63,657
2	1,886	2,920	4,303	6,965	9,925
3	,638	,353	3,182	4,541	5,841
4	,533	,132	2,776	3,747	4,604
5	,476	,015	,571	,365	,032
6	1,440	1,943	2,447	3,143	3,707
7	,415	,895	,365	2,998	,499
8	,397	,859	,306	,897	,355
9	,383	,833	,262	,821	,250
10	,372	,812	,228	,764	,169
11	1,363	1,795	2,201	2,718	3,106
12	,356	,782	,179	,681	,054
13	,350	,771	,160	,650	,012
14	,345	,761	,145	,624	2,977
15	,341	,753	,131	,602	,947
16	1,337	1,746	2,120	2,583	2,921
17	,333	,740	,110	,567	,898
18	,330	,734	,101	,552	,878
19	,328	,729	,093	,539	,861
20	,325	,725	,086	,528	,845

Kwantyle rozkładu t Studenta cd.

n	p				
	0,90	0,95	0,975	0,99	0,995
21	1,323	1,721	2,080	2,518	2,831
22	,321	,717	,074	,508	,819
23	,319	,714	,069	,500	,807
24	,318	,711	,064	,492	,797
25	,316	,708	,060	,485	,787
26	1,315	1,706	2,055	2,479	2,779
27	,314	,703	,052	,473	,771
28	,312	,701	,048	,467	,763
29	,311	,699	,045	,462	,756
30	,310	,697	,042	,457	,750
31	1,309	1,695	2,039	2,453	2,744
32	,309	,694	,037	,449	,738
33	,308	,692	,034	,445	,733
34	,307	,691	,032	,441	,728
35	,306	,690	,030	,438	,724
36	1,305	1,688	2,028	2,434	2,720
37	,305	,687	,025	,431	,715
38	,304	,686	,024	,429	,712
39	,304	,685	,023	,425	,708
40	,303	,684	,021	,423	,704
41	1,303	1,683	2,019	2,421	2,701
42	,302	,682	,018	,418	,698
43	,302	,681	,017	,416	,695
44	,301	,680	,015	,414	,692
45	,301	,679	,014	,412	,690
46	1,300	1,679	2,013	2,410	2,687
47	,300	,678	,012	,408	,685
48	,299	,677	,011	,407	,682
49	,299	,677	,010	,405	,680
50	,299	,676	,009	,403	,678
55	1,297	1,673	2,004	2,396	2,668
60	,295	,671	,000	,390	,660
65	,295	,669	1,997	,385	,654
70	,294	,667	,994	,381	648
75	,293	,665	,992	,377	,643
80	1,292	1,664	1,990	2,374	2,639
90	,291	,662	,987	,369	632
100	,290	,660	,984	,364	,626
120	,289	,658	,980	,358	,617
150	,287	,655	,976	,351	,609
200	1,286	1,653	1,972	2,345	2,601
300	,284	,650	,968	,339	,592
500	,283	,648	,965	,334	,586
1000	,282	,646	,962	,330	,581
∞	,282	,645	,960	,326	,576

Tablica 5.7 Kwantyle rozkładu χ^2

Kwantyle $\chi^2(p,n)$ rzędu p rozkładu χ^2 o n stopniach swobody.

Przy ustalonym p i n tablica podaje wartości x_p, takie że $P(\chi^2 < x_p) = p$, to znaczy

$$\frac{1}{2^{n/2}\Gamma(1/2,n)} \int_{-\infty}^{x_p} x^{k/2-1} e^{-u/2} du .$$

n	p							
	0,005	0,01	0,025	0,05	0,95	0,975	0,99	0,995
1	—	—	0,001	0,004	3,841	5,024	6,635	7,879
2	0,010	0,020	0,051	0,103	5,991	7,378	9,210	10,597
3	0,072	0,115	0,216	0,352	7,815	9,348	11,345	12,838
4	0,207	0,297	0,484	0,711	9,488	11,143	13,277	14,860
5	0,412	0,554	0,831	1,145	11,071	12,833	15,086	16,750
6	0,676	0,872	1,237	1,635	12,592	14,449	16,812	18,548
7	0,989	1,239	1,690	2,167	14,067	16,013	18,475	20,278
8	1,344	1,646	2,180	2,733	15,507	17,535	20,090	21,955
9	1,735	2,088	2,700	3,325	16,919	19,023	21,666	23,589
10	2,156	2,558	3,247	3,940	18,307	20,483	23,209	25,188
11	2,603	3,053	3,816	4,575	19,675	21,920	24,725	26,757
12	3,074	3,571	4,404	5,226	21,026	23,337	26,217	28,299
13	3,565	4,107	5,009	5,892	22,362	24,736	27,688	29,819
14	4,075	4,660	5,629	6,571	23,685	26,119	29,141	31,319
15	4,601	5,229	6,262	7,261	24,996	27,488	30,578	32,801
16	5,142	5,812	6,908	7,962	26,296	28,845	32,000	34,267
17	5,697	6,408	7,564	8,672	27,587	30,191	33,409	35,718
18	6,265	7,015	8,231	9,390	28,869	31,526	34,805	37,156
19	6,844	7,633	8,907	10,117	30,144	32,852	36,191	38,582
20	7,434	8,260	9,591	10,851	31,410	34,170	37,566	39,997
21	8,034	8,897	10,283	11,591	32,671	35,479	38,932	41,401
22	8,643	9,542	10,982	12,336	33,924	36,781	40,289	42,796
23	9,260	10,196	11,689	13,091	35,172	38,076	41,638	44,181
24	9,886	10,856	12,401	13,848	36,415	39,364	42,980	45,559
25	10,520	11,524	13,120	14,611	37,652	40,646	44,314	46,928
26	11,160	12,198	13,844	15,379	38,885	41,923	45,642	48,290
27	11,808	12,879	14,573	16,151	40,113	43,194	46,963	49,645
28	12,461	13,565	15,308	16,928	41,337	44,461	48,278	50,993
29	13,121	14,257	16,047	17,708	42,557	45,722	49,588	52,336
30	13,787	14,954	16,791	18,493	43,773	46,979	50,892	53,672

Kwantyle rozkładu χ^2 cd.

n	p							
	0,005	0,01	0,025	0,05	0,95	0,975	0,99	0,995
31	14,458	15,655	17,539	19,281	44,985	48,232	52,191	55,003
32	15,134	16,362	18,291	20,072	46,194	49,480	43,486	56,328
33	15,815	17,074	19,047	20,867	47,400	50,725	54,776	57,648
34	16,501	17,789	19,806	21,664	48,602	51,966	56,061	58,964
35	17,192	18,509	20,569	22,465	49,802	53,203	57,342	60,275
36	17,887	19,233	21,336	23,269	50,998	54,437	58,619	61,581
37	18,586	19,960	22,106	24,075	52,192	55,668	59,892	62,883
38	19,289	20,691	22,878	24,884	53,384	56,896	61,162	64,181
39	19,996	21,426	23,654	25,695	54,572	58,120	62,428	65,476
40	20,707	22,164	24,433	26,509	55,758	59,342	63,691	66,766
41	21,421	22,906	25,215	27,326	56,942	60,561	64,950	68,053
42	22,138	23,650	25,999	28,144	58,124	61,777	66,206	69,336
43	22,859	24,398	26,785	28,965	59,304	62,990	67,459	70,616
44	23,584	25,148	27,575	29,787	60,481	64,201	68,710	71,893
45	24,311	25,901	28,366	30,612	61,656	65,410	69,957	73,166
46	25,041	26,657	29,160	31,439	62,830	66,617	71,201	74,437
47	25,775	27,416	29,956	32,268	64,001	67,821	72,443	75,704
48	26,511	28,177	30,755	33,098	65,171	69,023	73,683	76,969
49	27,249	28,941	31,555	33,930	66,339	70,222	74,919	78,231
50	27,991	29,707	32,357	34,764	67,505	71,420	76,154	79,490

Tablica 5.8 Podstawowe przedziały ufności

Punkt	Hipoteza	Hipoteza alternatywna	Stosowana statystyka (wartość obliczeniowa)	Stosowane tablice	Zbiór krytyczny	Uwagi
I°	$\mu = \mu_0$	$\mu = \mu_1 > \mu_0$ $\mu = \mu_1 < \mu_0$ $\mu = \mu_1 \neq \mu_0$	$\frac{\overline{X} - \mu_0}{\sigma}\sqrt{n}$	kwantyle lub dystrybuanta rozkładu normalnego	$[u(1-\alpha), \infty)$ $(-\infty, -u(1-\alpha)]$ $(-\infty, -u(1-\frac{\alpha}{2})] \cup [u(1-\frac{\alpha}{2}), \infty)$	σ znane
II°	$\mu = \mu_0$	jak w I°	$\frac{\overline{X} - \mu_0}{S}\sqrt{n-1}$	rozkład t Studenta przy $n-1$ stopniach swobody	$[t(1-\alpha, n-1), \infty)$ $(-\infty, -t(1-\alpha, n-1)]$ $(-\infty, -t(1-\frac{\alpha}{2}, n-1)] \cup [t(1-\frac{\alpha}{2}, n-1), \infty)$	σ nieznane
III°	$\mu = \mu_0$	jak w I°	$\frac{\overline{X} - \mu_0}{S}\sqrt{n}$	jak w I°	jak w I°	σ nieznane $n \geq 100$
IV°	$\sigma^2 = \sigma_0^2$	$\sigma^2 = \sigma_1^2 > \sigma_0^2$ $\sigma^2 = \sigma_1^2 < \sigma_0^2$ $\sigma^2 = \sigma_1^2 \neq \sigma_0^2$	$\frac{nS^2}{\sigma_0^2}$	rozkład χ^2 przy $n-1$ stopniach swobody	$[\chi^2(1-\alpha, n-1), \infty)$ $(0, \chi^2(\alpha, n-1)]$ $(0, \chi^2(\frac{\alpha}{2}, n-1)] \cup [\chi^2(1-\frac{\alpha}{2}, n-1), \infty)$	
V°	$\sigma^2 = \sigma_0^2$	jak w IV°	$\sqrt{\frac{2nS^2}{\sigma_0^2}} - \sqrt{2n-3}$	jak w I°	jak w I°	$n \geq 50$
VI°	$\sigma^2 = \sigma_0^2$	jak w IV°	$\frac{(S^*)^2 - \sigma_0^2}{\sigma_0^2}\sqrt{\frac{2}{n}}$	jak w I°	jak w I°	$n \geq 100$
VII°	$\mu_1 = \mu_2$	$\mu_1 > \mu_2$ $\mu_1 < \mu_2$ $\mu_1 \neq \mu_2$	$\frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$	jak w I°	jak w I°	σ_1, σ_2 znane
VIII°	$\mu_1 = \mu_2$	jak w VII°	$\frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$	jak w I°	jak w I°	σ_1, σ_2 nieznane n_1, n_2 duże

Podstawowe pojęcia kombinatoryki

Permutacją zbioru n-elementowego nazywamy każdy ciąg, jaki można utworzyć z elementów tego zbioru; permutacji takich jest

$$P_n = n!.$$

Wariacją k-elementową z powtórzeniami zbioru n-elementowego nazywamy każdy ciąg k-elementowy elementów ze zbioru n-elementowego, przy czym elementy te mogą się powtarzać; liczba takich wariacji wynosi

$$W_n^k = n^k.$$

Wariacją k-elementową bez powtórzeń zbioru n-elementowego, $k \leq$, nazywamy każdy ciąg k-elementowy elementów ze zbioru n-elementowego, przy czym elementy te nie mogą się powtarzać; liczba takich wariacji wynosi

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

Jeżeli $k = n$, pojęcie wariacji bez powtórzeń pokrywa się z pojęciem permutacji.

Kombinacją k-elementową zbioru n-elementowego nazywamy każdy k-elementowy podzbiór zbioru n-elementowego. Podzbiorów tych jest

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

Literatura

[1] Bobrowski D., *Probabilistyka w zastosowaniach technicznych*, Warszawa 1980.

[2] Fisz M., *Rachunek prawdopodobieństwa i statystyka matematyczna*, Warszawa 1976.

[3] Gajek L., Kałuszka M., *Wnioskowanie statystyczne, Modele i metody*, Warszawa 1993,1994.

[4] Greń J., *Modele i zadania statystyki matematycznej*, Warszawa 1980.

[5] Krysicki W., Bartos J., Dyczka W., Królikowska K., Wasilewski M., *Rachunek prawdopodobieństwa i statystyka matematyczna w zadaniach*, część I, Warszawa 1995.

[6] Krysicki W., Bartos J., Dyczka W., Królikowska K., Wasilewski M., *Rachunek prawdopodobieństwa i statystyka matematyczna w zadaniach*, część II, Warszawa 1997.

[7] Plucińska A., Pluciński E., *Elementy probabilistyki*, Warszawa 1979.

[8] Plucińska A., Pluciński E., *Zadania z rachunku prawdopodobieństwa i statystyki matematycznej*, Warszawa 1978.

[9] Silvey S.D., *Wnioskowanie statystyczne*, Warszawa 1978.

[10] Smirnow N.W., Dunin-Borkowski I.W., *Kurs rachunku prawdopodobieństwa i statystyki matematycznej dla zastosowań technicznych*, Warszawa 1969.

[11] Stępniak Cz., *Ćwiczenia ze statystyki matematycznej*, Lublin 1989.

[12] Zieliński R., *Tablice statystyczne*, Warszawa 1972.

[13] Zubrzycki S., *Wykłady rachunku prawdopodobieństwa i statystyki matematycznej*, Warszawa 1970.

Skorowidz terminów