KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI

UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY

Analiza danych

SKRYPT

Jan Kurzyk

Wydział Fizyki, Matematyki i Informatyki
Politechniki Krakowskiej

Kraków 2012
Zmiany i rozszerzenia 2013-2016

Materiały dydaktyczne zostały przygotowane w ramach Projektu „Politechnika XXI wieku - Program rozwojowy Politechniki Krakowskiej – najwyższej jakości dydaktyka dla przyszłych polskich inżynierów" współfinansowanego ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego. Umowa o dofinansowanie nr UDA-POKL.04.01.01-00-029/10-00

Spis treści

1. Wprowadzenie ... 3
2. Elementy teorii prawdopodobieństwa ... 3
   2.1. Definicje podstawowych pojęć ... 3
3. Rozkłady zmiennych losowych ... 7
   3.1. Zmienne losowe ... 7
   3.2. Rozkłady zmiennych losowych ... 8
   3.3. Funkcje zmiennej losowej ... 12
   3.4. Charakterystyki rozkładów ... 13
4. Wybrane rozkłady zmiennych losowych ... 21
   4.1. Rozkład jednostajny ciągły ... 21
   4.2. Rozkład Cauchy'ego ... 22
   4.3. Rozkład Lorentza ... 24
   4.4. Rozkład dwumianowy ... 24
   4.5. Rozkład Poissona ... 25
   4.6. Rozkład Gaussa (rozkład normalny) ... 27
   4.7. Rozkład χ^2 (chi kwadrat) ... 30
5. Funkcja charakterystyczna rozkładu ... 33
6. Rozkłady wielu zmiennych losowych ... 35
   6.1. Dystrybuanta i gęstość prawdopodobieństwa dwu zmiennych ... 35
   6.2. Wartości oczekiwane, wariancje, kowariancje dla dwu zmiennych losowych ... 36
   6.3. Opis wielu zmiennych losowych ... 38
7. Sploty rozkładów ... 42
8. Elementy teorii estymacji ... 45
   8.1. Metoda momentów ... 47
   8.2. Metoda największej wiarygodności ... 49
   8.3. Estymacja przedziałowa ... 50
      8.3.1. Estymacja przedziałowa $E(X)$, gdy znamy $\sigma^2(X)$ ... 51
      8.3.1. Estymacja przedziałowa $E(X)$, gdy nie znamy $\sigma^2(X)$ ... 52
      8.3.2. Estymacja przedziałowa $\sigma^2(X)$ ... 53
9. Weryfikacja hipotez statystycznych ... 56
   9.1. Porównanie wariancji z liczbą ... 58
   9.2. Test równości wariancji (test *F* Fischera-Snedecora) ... 59
   9.3. Porównanie wartości średniej z liczbą ... 60
   9.4. Porównanie wartości średnich dwu populacji ... 61
   9.5. Analiza wariancji (test ANOVA) ... 63

9.6. Test zgodności χ^2 Pearsona ... 65
9.7. Test zgodności λ Kołmogorowa-Smirnowa ... 67
9.8. Test znaków ... 68
10. Metoda najmniejszych kwadratów ... 71
10.1. Pomiary bezpośrednie o równej dokładności ... 72
10.2. Pomiary bezpośrednie o różnej dokładności ... 73
10.3. Regresja liniowa ... 74
10.4. Regresja liniowa w przypadku stałych wariancji zmiennych ... 74
10.5. Regresja liniowa w przypadku wariancji zmiennych zależnych od wartości zmiennych ... 77
10.6. Regresja nieliniowa ... 78
10.1. Wariancje parametrów metody najmniejszych kwadratów ... 80
11. Metoda Monte Carlo ... 83
11.1. Liczby pseudolosowe ... 83
11.2. Generowanie liczb pseudolosowych o dowolnym rozkładzie ... 84
11.3. Liczenie całek metodą Monte Carlo ... 90
11.4. Zastosowanie metody Monte Carlo do modelowania komputerowego ... 94
Literatura ... 96

1. Wprowadzenie

Każda dziedzina wiedzy starająca się opisać ilościowo zjawiska jakimi się zajmuje musi posługiwać się *pomiarami.* W wyniku pomiarów uzyskujemy informacje o mierzonej wielkości lub zjawisku w postaci wyników pomiaru nazywanych też *danymi pomiarowymi* lub po prostu *danymi*. Otrzymane w wyniku pomiarów dane musimy następnie przetworzyć w celu uzyskania pożądanych informacji o badanym zjawisku lub wielkości fizycznej i wyciągnięcia na tej podstawie użytecznych wniosków. Ten proces nazywany *analizą danych* znajduje zastosowanie w różnorodnych dziedzinach nauki i techniki. Analiza danych zajmuje się wieloma aspektami danych pomiarowych i używa do tego ogromnej palety metod i narzędzi. Niektóre z nich będą zaprezentowane w niniejszym skrypcie.

Poniżej przedstawiono kilka typowych problemów, jakimi zajmuje się analiza danych.

1. Wykonano serię pomiarów prędkości światła w próżni, co na tej podstawie można powiedzieć o rzeczywistej wartości tej prędkości? Jak możemy porównać wyniki pomiarów uzyskane przez różnych badaczy różnymi metodami?
2. W kontroli produkcji wybrano losowo pewną liczbę produktów i poddano je badaniu. Jakie wnioski możemy wyciągnąć o jakości pozostałych produktów? Jak duża powinna być próba kontrolna, żeby wnioski nt. całej partii produktów były wiarygodne?
3. W badaniach klinicznych przebadano dwa leki. Dla jednego uzyskano 45% skuteczności, a dla drugiego 40%. Czy pierwszy z tych leków jest skuteczniejszy? Czy liczba pacjentów wybrana do badań obu leków była wystarczająca, aby móc rozróżnić skuteczność obu leków?
4. W celu sprawdzenia pewnej teorii wykonano pewne doświadczenie. Czy wynik tego doświadczenia można uznać za zgodny z przewidywaniami teorii?
5. Pewne zjawisko cechuje zmienność np. w czasie. Z teorii wiemy, jaka funkcja opisuje tę zmienność, ale parametry tej funkcji muszą być znalezione eksperymentalnie (przykładem może być stała połowicznego rozpadu).
6. Wykonanie pewnego eksperymentu jest bardzo kosztowne znane są jednak pewne cechy statystyczne zjawisk, od których zależy ten eksperyment. Można zatem przed wykonaniem właściwego eksperymentu wykonać symulację komputerową stosując tzw. *metody Monte Carlo*.

Istotną cechą wspólną podanych wyżej przykładów jest niejednoznaczność danych doświadczalnych. Wyniki pomiarów podlegają fluktuacjom statystycznym w związku z tym są tzw. *zmiennymi losowymi*. Losowy charakter wyników pomiarów ma bardzo wiele przyczyn. Źródło losowości może tkwić w naturze samego obiektu badanego, ale również jest wynikiem niedoskonałości metod i przyrządów pomiarowych, a także wpływu różnych czynników zewnętrznych związanych z procesem pomiarowym. Stochastyczność wyników pomiarów sprawia, że metody analizy danych są przeważnie metodami statystycznymi.

2. Elementy teorii prawdopodobieństwa

2.1. Definicje podstawowych pojęć

Z powodów wymienionych w poprzednim rozdziale wyniki pomiarów mogą różnić się między sobą. Wynik pojedynczego pomiaru możemy w myśl teorii prawdopodobieństwa nazwać *zdarzeniem elementarnym*. Zbiór wszystkich zdarzeń elementarnych, które z definicji wzajemnie się wykluczają tworzy tzw. *przestrzeń prób* (*przestrzeń zdarzeń*) danego doświadczenia. Zbiór ten będziemy oznaczać literą $\mathcal{E}$. Wyczerpuje on wszystkie możliwości wyniku

doświadczenia (każde możliwe zdarzenie elementarne zawiera się w tym zbiorze). Zwykle, dokładne wyznaczenie przestrzeni prób jest trudne, a czasami wręcz niemożliwe, do wyznaczenia i dlatego często w analizie danych posługujemy się przestrzenią prób większą od rzeczywistej, zawierająca jednak prawdziwą przestrzeń prób jako podprzestrzeń. Przestrzeń prób może tworzyć obszar (lub zbiór obszarów) ciągły (będzie tak przypadku wielkości, które mogą przyjmować dowolną wartość z pewnego obszaru np. napięcie elektryczne, długość obiektu, temperatura obiektu) lub dyskretny (skokowy), czyli składający się z przeliczalnej, choć niekoniecznie skończonej liczby elementów (będzie tak w przypadku zmiennych tzw. dyskretnych np. energia atomu wodoru, liczba cząstek, data urodzenia).

Dowolny podzbiór zdarzeń elementarnych zbioru $\mathcal{E}$ nazywamy *zdarzeniem.* Wśród zdarzeń możemy wyróżnić: *zdarzenie pewne,* czyli zdarzenie zawierające wszystkie zdarzenia $\mathcal{E}$ oraz *zdarzenie niemożliwe* nie zawierające żadnego zdarzenia zbioru $\mathcal{E}$, czyli zbiór pusty $\emptyset$. Relacje między zdarzeniami oraz działania na nich są analogiczne do relacji i działań na zbiorach:

- *Zdarzenie A zawiera się w zdarzeniu B,* jeżeli każde zdarzenie elementarne należące do zbioru A należy do zbioru B $(A \subset B)$.
- *Zdarzenia A* i *B są równe*, gdy $A \subset B$ i $B \subset A$.
- *Zdarzenie C* jest *sumą zdarzeń A* i *B* $(C = A + B)$, jeśli zawiera te zdarzenia elementarne i tylko te zdarzenia elementarne, które należą do któregokolwiek ze zdarzeń A, B (Zbiór zdarzeń C jest sumą logiczną zbiorów zdarzeń elementarnych zbiorów A i B $C = A \cup B$).
- *Zdarzenie C* jest *różnicą zdarzeń A* i *B* $(C = A - B)$, jeśli zawiera te i tylko te zdarzenia elementarne, które należą do zbioru zdarzeń A, a nie należą do zbioru zdarzeń B. (Zbiór zdarzeń C jest różnicą logiczną zbiorów zdarzeń elementarnych zbiorów A i B $C = A/B$).
- *Zdarzenie C* jest *iloczynem zdarzeń A* i *B* $(C = A \cdot B)$, gdy zawiera te i tylko te zdarzenia elementarne, które należą zarówno do zdarzenia A, jak i do zdarzenia B. (W języku zbiorów $C = A \cap B$).
- Dwa zdarzenia A i B nazywamy *zdarzeniami rozłącznymi*, jeśli ich iloczyn jest zdarzeniem niemożliwym. (Iloczyn zbiorów zdarzeń elementarnych A i B jest zbiorem pustym $A \cap B = \emptyset$).
- *Zdarzeniem przeciwnym* do zdarzenia A nazywamy różnicę zdarzeń $\mathcal{E} - A$ i oznaczmy $\bar{A}$.

O zdarzeniu, co do którego nie możemy mieć pewności, czy w danych warunkach zajdzie, czy też nie, mówimy, że jest *zdarzeniem losowym*[1]. Ze zdarzeniami losowymi mamy do czynienia w procesie pomiarowym. Zdarzeniem losowym może być zarówno pojedynczy wynik pomiaru, jak i zbiór złożony z większej liczby elementów np. wyniki serii pomiarów.

Miarą szansy zajścia zdarzenia losowego jest *prawdopodobieństwo.* W zastosowaniach praktycznych prawdopodobieństwo możemy określić w oparciu o częstość występowania danego zdarzenia. Korzystamy przy tym z następującej procedury. Powtarzamy dane doświadczenie dużą liczbę razy N i liczymy liczbę n zajścia zdarzenia A. Prawdopodobieństwo zajścia zdarzenia A definiujemy jako

$$P(A) = \lim_{N\to\infty} \frac{n}{N}. \tag{2.1}$$

[1] Nie jest to ścisła definicja zdarzenia losowego, ale takie intuicyjne określenie jest wystarczające dla naszych rozważań.

Będziemy posługiwać się tą definicją w dalszych rozważaniach, gdyż jest ona wystarczająca w większości zastosowań praktycznych, trzeba jednak zaznaczyć, że nie jest ona zadowalająca z matematycznego punktu widzenia, m.in. dlatego, że wymaga wykonania nieskończenie wielu pomiarów. Bardziej precyzyjna definicja prawdopodobieństwa definiuje tę wielkość poprzez odpowiedni zestaw aksjomatów. Ogólnie przyjętym minimalnym zestawem aksjomatów teorii prawdopodobieństwa jest zestaw aksjomatów podany przez rosyjskiego matematyka Andrieja Kołmogorowa (1903-1987):

1. Każdemu zdarzeniu A możemy przypisać nieujemną liczbę zwaną prawdopodobieństwem tego zdarzenia
$$P(A) \geq 0.$$
2. Prawdopodobieństwo zajścia dowolnego zdarzenia ze zbioru $\mathcal{E}$ jest równe jedności
$$P(\mathcal{E}) = 1.$$
3. Jeżeli zdarzenia A i B są zdarzeniami rozłącznymi (wykluczającymi się), to prawdopodobieństwo zajścia zdarzenia A lub B wynosi
$$P(A + B) = P(A) + P(B).$$

W niektórych podejściach jako czwarty aksjomat podaje się wzór określający tzw. *prawdopodobieństwo warunkowe*

$$P(A|B) = \frac{P(A \cdot B)}{P(B)}, \tag{2.2}$$

gdzie $P(A|B)$ oznacza prawdopodobieństwo zajścia zdarzenia A pod warunkiem, że zachodzi zdarzenie B. Powyższą definicję podaje się często w postaci wzoru

$$P(A \cdot B) = P(B)P(A|B). \tag{2.3}$$

Druga z tych definicji, w przeciwieństwie do pierwszej nie wymaga, aby prawdopodobieństwo zdarzenia B było niezerowe $P(B) \neq 0$. Ponieważ iloczyn zdarzeń jest z definicji przemienny $(A \cdot B = B \cdot A)$, to równanie (2.3) możemy rozszerzyć do postaci

$$P(A \cdot B) = P(B)P(A|B) = P(A)P(B|A) = P(B \cdot A). \tag{2.4}$$

Z aksjomatów Kołmogorowa wynika natychmiast kilka własności prawdopodobieństwa. Z aksjomatów 2) i 3) otrzymujemy

$$P(A + \bar{A}) = P(A) + P(\bar{A}) = 1. \tag{2.5}$$

Z powyższego równania oraz aksjomatu 1) mamy

$$0 \leq P(A) \leq 1. \tag{2.6}$$

Aksjomat 3) można uogólnić na przypadek dowolnego ciągu wzajemnie wykluczających się zdarzeń *A, B, C,…*

$$P(A + B + C + \cdots) = P(A) + P(B) + P(C) + \cdots. \tag{2.7}$$

Zdarzeniami niezależnymi nazywamy zdarzenia, w przypadku których fakt wystąpienia jednego z nich nie zmienia prawdopodobieństwa drugiego, czyli zdarzenie A jest niezależne od zdarzenia B, gdy

$$P(A|B) = P(A). \tag{2.8}$$

Zauważmy, że jeżeli zdarzenie A nie zależy od B, to również zdarzenie B nie zależy od A. Aby to udowodnić skorzystajmy z równań (2.4) i (2.8)

$$P(A \cdot B) = P(B)P(A|B) = P(B) \cdot P(A) = P(A)P(B|A), \tag{2.9}$$

a stąd dostajemy

$$P(B|A) = P(B). \tag{2.10}$$

Z przeprowadzonego dowodu (patrz równanie (2.9)) wynika ważna własność zdarzeń niezależnych – prawdopodobieństwo iloczynu zdarzeń niezależnych jest równe iloczynowi prawdopodobieństw tych zdarzeń

$$P(A \cdot B) = P(A) \cdot P(B). \tag{2.11}$$

Powyższa równość jest warunkiem koniecznym i wystarczającym niezależności zdarzeń. Warunek ten można w naturalny sposób rozszerzyć na dowolną liczbę zdarzeń A, B, C,… z których każda para stanowi parę niezależnych zdarzeń

$$P(A \cdot B \cdot C \cdot \cdots) = P(A) \cdot P(B) \cdot P(C) \cdots. \tag{2.12}$$

Zadanie 2.1

Pokaż, że prawdopodobieństwo sumy dowolnych zdarzeń A i B wynosi

$$P(A + B) = P(A) + P(B) - P(A \cdot B).$$

Wskazówka

Zapisz zdarzenia $A + B$ oraz B w postaci sumy zdarzeń rozłącznych

$$A + B = A + (B - A \cdot B),$$

$$B = A \cdot B + (B - A \cdot B),$$

a następnie skorzystaj z aksjomatu 3) definicji prawdopodobieństwa do obu zdarzeń $A + B$ oraz B i równania na otrzymane prawdopodobieństwa odejmij stronami od siebie.

Zadanie 2.2

Rozważ rzuty dwiema symetrycznymi kostkami do gry. Niech zmienną losową będzie n oznaczającą sumę oczek $n_1 + n_2$ wyrzuconych na kostce o numerze 1 i o numerze 2. Znajdź prawdopodobieństwo $P_2(n)$ uzyskania sumy oczek n dla $n = 0,1,2,\ldots,12$. Sprawdź warunek normalizacyjny, czyli sprawdź czy suma obliczonych prawdopodobieństw jest równa jedności (drugi aksjomat Kołmogorowa).

Wskazówka

Prawdopodobieństwo wyrzucenia na *i*-tej $(i = 1,2)$ kostce n_i oczek $(n_i = 1,2,\dots,6)$ jest równe $P(n_i) = 1/6$. Liczby oczek wyrzucone na każdej z kostek są od siebie niezależne, a zatem $P(n_1, n_2) = P(n_1) \cdot P(n_2) = 1/36$. Rozważ na ile sposobów można uzyskać sumę n oczek. Na przykład $n = 3$ można uzyskać na dwa sposoby: $\{1,2\}$ i $\{2,1\}$. A zatem

$$P_2(3) = P(1,2) + P(2,1) = 2/36.$$

Zadanie 2.3

Z talii liczącej 52 karty losujemy jedną kartę. Oblicz prawdopodobieństwo, że wylosowana karta jest w kolorze pik lub jest asem.

Wskazówka

Niech zdarzenie *A* oznacza, że wylosowano asa, a zdarzenie *B*, że wylosowano pik. Zdarzenie $A \cdot B$ oznacza, że wylosowano asa pik. Znajdź prawdopodobieństwa zdarzeń $P(A), P(B)$ i $P(A \cdot B)$, a następnie skorzystaj z wyniku zadania 2.1.

Zadanie 2.4

Pokaż, że zdarzenia *A* i *B* oznaczające odpowiednio: wylosowanie asa i wylosowanie pika z talii liczącej 52 karty są zdarzeniami niezależnymi.

Wskazówka

Znajdź prawdopodobieństwa zdarzeń $P(A), P(B)$ i $P(A \cdot B)$, a następnie skorzystaj z warunku (2.11) niezależności zdarzeń.

Zadanie 2.5

Pokaż, że zdarzenia opisane w poprzednim zadaniu są zależne jeśli zwykłą talię liczącą 52 karty uzupełnimy o dżokera.

3. Rozkłady zmiennych losowych

3.1. Zmienne losowe

Pojęcie zdarzenia losowego jest ściśle związane z pojęciem *zmiennej losowej*. Pod pojęciem *zmiennej losowej* rozumiemy jednoznaczną funkcję określoną na zbiorze $\mathcal{E}$, która przyporządkowuje zdarzeniom elementarnym liczby rzeczywiste. W przypadku pomiarów wielkości fizycznych naturalnymi zmiennymi losowymi będą funkcje przyporządkowujące poszczególnym pomiarom miary mierzonych wielkości w danych jednostkach, np. pomiarom natężenia prądu przyporządkowujące tym pomiarom wartości tych prądów wyrażone w amperach. W przypadku zdarzeń losowych, którym nie można w tak naturalny sposób jak powyższy przyporządkować miary liczbowej, możemy przypisać te liczby w sposób arbitralny (ale jednoznaczny). Najprostszym przykładem może być doświadczenie polegające na rzutach monetą. Przestrzeń prób $\mathcal{E}$ składa się w tym przypadku ze zbioru zawierającemu dwa elementy $\{\text{orzeł}, \text{reszka}\}$[2]. Naszą zmienną losową może być w tym przypadku funkcja przyporządkowująca zdarzeniu $\{\text{orzeł}\}$ wartości liczbowej 0, a zdarzeniu $\{\text{reszka}\}$ wartości liczbowej 1.

W zależności od rodzaju wielkości mierzonej (zmiennej) wyniki pomiarów, czyli ich wartości mogą różnić się sposobami zakodowania (zapisu) wyniku tych danych oraz możli-

[2] Pomijamy tu bardzo mało prawdopodobne zdarzenie polegające na ustawieniu się monety na krawędzi.

wościami wykonywania na nich operacji. Pod tym względem zmienne możemy podzielić na następujące typy:

- ***zmienne nominalne*** – zmienne, których wartości nie da się uporządkować w sposób wynikający z natury tej wielkości. Wartości tych zmiennych zapisujemy zwykle używając tzw. etykiet, które nie muszą być wartościami liczbowymi. Nawet, jeśli wynikom pomiarów zmiennych nominalnych przypiszemy wartości liczbowe, to będą one miały jedynie charakter umowny i wykonywanie działań arytmetycznych na nich, a nawet porównywanie nie ma sensu. Przykładami takich zmiennych są np. nazwy miejscowości, płeć osób, rodzaje skał itp.
- ***zmienne porządkowe*** – zmienne, których wartości możemy w naturalny sposób uporządkować, a zatem stosować do nich relacje porządku $(=, <, >, \leq, \geq)$, ale wykonywanie na tych wartościach działań arytmetycznych typu suma, różnica, iloczyn lub iloraz nie mają sensu. Przykładami tego typu zmiennych mogą być np. alfabetyczna kolejność uczniów w dzienniku, wykształcenie itp.
- ***zmienne interwałowe (przedziałowe)*** – zmienne, dla których różnica dwóch wartości ma sensowną interpretację, ale iloraz takich wartości nie ma sensu. Zmienne tego typu mają zdefiniowaną jednostkę miary, lecz punkt odniesienia (zero) jest przyjęty umownie. Przykładami takich zmiennych mogą być np. temperatura obiektu w skali Celsjusza lub daty urodzenia.
- ***zmienne ilorazowe*** – zmienne, dla których sens fizyczny ma nie tylko różnica dwóch wartości, ale również ich iloraz. Zmienne ilorazowe mają przypisane jednostki miary, a wartość zerowa określona jest w sposób naturalny (bezwzględny). Większość wielkości fizycznych np. masa obiektu, napięcie elektryczne, temperatura w skali Kelwina zaliczamy do zmiennych typu ilorazowego.
- ***zmienne absolutne*** – zmienne, które mają cechy zmiennych ilorazowych jednak w przeciwieństwie do tych ostatnich, które mogą być reprezentowane przez różne liczby w zależności od arbitralnie przyjętej definicji jednostki miary, mogą być wyrażane tylko w jeden naturalny sposób. Przykładami takich zmiennych mogą być np. liczba znaków drukarskich w kolejnych wierszach tej strony tekstu, liczba nukleonów w jądrach atomowych.

Dla zmiennych typu nominalnego lub porządkowego możemy stosować tylko niektóre metody statystyczne. Głównie są to takie operacje statystyczne jak zliczanie, czy wyliczanie udziału procentowego w całości zbioru danych.

System symboli kodujących wyniki pomiaru nazywany jest *skalą pomiaru*. W zależności od typu zmiennej możemy zatem mówić o skali: nominalnej, porządkowej, interwałowej, ilorazowej i absolutnej.

Niezależnie od powyższego podziału zmienne mierzalne możemy podzielić na zmienne ciągłe i zmienne dyskretne (skokowe) o czym była już mowa w poprzednim rozdziale[3].

3.2. Rozkłady zmiennych losowych

Niech x oznacza dowolną liczbę rzeczywistą z przedziału od $-\infty$ do $+\infty$, a X pewną zmienną losową. Dla odróżnienia zmiennych losowych od zwykłych zmiennych funkcji, te

[3] Oczywiście istnieją również zmienne o charakterze mieszanym, które w pewnych przedziałach mogą przyjmować wartości dyskretne, a w pewnych ciągłe. Przykładem może tu być energia układu *elektron-proton*. Do osiągnięcia energii jonizacji atomu wodoru energia układu jest skwantowana, czyli przyjmuje wartości dyskretne, a powyżej tej energii energia układu może zmieniać się w sposób ciągły.

pierwsze będziemy oznaczać dużymi literami, a te drugie małymi. Prawdopodobieństwo zajścia zdarzenia polegającego na tym, że $X < x$ nazywamy *dystrybuantą zmiennej losowej X*

$$F(x) \equiv P(X < x). \tag{3.1}$$

Z definicji dystrybuanty oraz aksjomatów Kołmogorowa łatwo wykazać, że dystrybuanta ma następujące własności:

- $0 \leq F(x) \leq 1$,
- $\lim_{x \to -\infty} F(x) = 0$,
- $\lim_{x \to +\infty} F(x) = 1$,
- $F(x)$ jest funkcją monotoniczną niemalejącą,
- $F(x)$ jest zmienną bezwymiarową

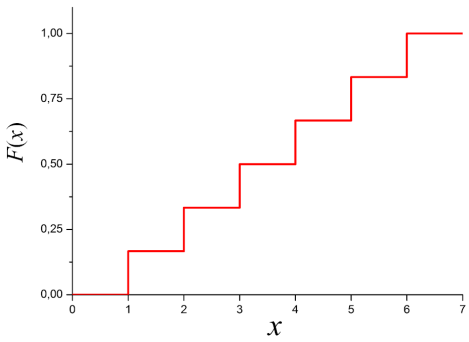
W przypadku zmiennych dyskretnych dystrybuanta będzie funkcją schodkową.

Przykład 1

Niech X będzie zmienną losową, której wartości są równe liczbie oczek wyrzuconych symetryczną kostką do gry. Dystrybuanta naszej zmiennej losowej będzie przyjmowała następujące wartości:

$$F(x) = \begin{cases} 0 \text{ dla } x < 1, \\ 1/6 \text{ dla } 1 \leq \mathrm{x} < 2, \\ 2/6 \text{ dla } 2 \leq x < 3, \\ 3/6 \text{ dla } 3 \leq x < 4, \\ 4/6 \text{ dla } 4 \leq x < 5, \\ 5/6 \text{ dla } 5 \leq x < 6, \\ 1 \text{ dla } x \geq 6. \end{cases}$$

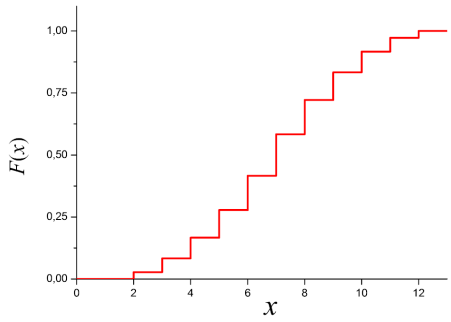
Wykresem tej dystrybuanty jest funkcja schodkowa przedstawiona na rysunku 3.1.



Rysunek 3.1. Dystrybuanta liczby oczek wyrzuconych symetryczną kostką do gry.

Przykład 2

Dla zmiennej losowej z zadania 2.2 (suma oczek, jakie wypadły w rzucie dwiema symetrycznymi kostkami do gry dystrybuanta ma postać funkcji schodkowej pokazanej rysunku 3.2.



Rysunek 3.2. Dystrybuanta sumy oczek wyrzuconych na dwóch symetrycznych kostkach do gry.

W przypadku zmiennych dyskretnych oprócz dystrybuanty własności statystyczne zmiennej losowej opisuje *rozkład prawdopodobieństwa* tej zmiennej, czyli funkcja, która każdej możliwej wartości x_i $(i = 1,2, \dots)$ przypisuje wartość prawdopodobieństwa występowania tej wartości

$$P(X = x_i) = p_i. \tag{3.2}$$

Oczywiście na mocy aksjomatu 2) Kołmogorowa zachodzi związek

$$\sum_{i=1} p_i = 1. \tag{3.3}$$

Zmiennym ciągłym nie możemy przypisać prawdopodobieństwa występowania danej wartości. Zamiast tego do opisu własności statystycznych zmiennej ciągłej używamy *gęstości prawdopodobieństwa.* Gęstością prawdopodobieństwa ciągłej zmiennej losowej nazywamy pierwszą pochodną dystrybuanty (oczywiście warunkiem jest, aby dystrybuanta była ciągła i miała pierwszą pochodną):

$$f(x) = \frac{dF(x)}{dx}. \tag{3.4}$$

Ponieważ dystrybuanta jest funkcją niemalejącą, gęstość prawdopodobieństwa jest funkcją nieujemną

$$f(x) \geq 0. \tag{3.5}$$

Wymiarem gęstości prawdopodobieństwa jest wymiar $[1/x]$ (przypomnijmy, że dystrybuanta jest bezwymiarowa).

Wielkość $f(x)dx$ interpretujemy jako prawdopodobieństwo zajścia zdarzenia $x \leq X \leq x + dx$

$$f(x)dx \equiv P(x \leq X \leq x + dx). \tag{3.6}$$

Z definicji dystrybuanty oraz gęstości prawdopodobieństwa wynika, że

$$P(X < a) = F(a) = \int_{-\infty}^{a} f(x)dx, \tag{3.7}$$

$$P(a \leq X < b) = \int_a^b f(x)dx. \tag{3.8}$$

Szczególnym przypadkiem równania (3.8) jest równanie

$$\int_{-\infty}^{+\infty} f(x)dx = 1, \tag{3.9}$$

które jest odpowiednikiem równania (3.3) dla zmiennych dyskretnych, czyli warunkiem normalizacji gęstości prawdopodobieństwa wynikającym z drugiego aksjomatu Kołmogorowa.

Funkcję opisującą gęstość prawdopodobieństwa ciągłej zmiennej losowej nazywamy *rozkładem prawdopodobieństwa* tej zmiennej. W analizie danych spotykamy wiele rozkładów prawdopodobieństwa. Najważniejsze z nich poznamy w dalszej części skryptu. Najprostszym rozkładem jest rozkład jednostajny nazywany również rozkładem równomiernym lub prostokątnym. Jest to rozkład, w którym gęstość prawdopodobieństwa przyjmuje stałą wartość na danym przedziale np. a ≤ x ≤ b oraz 0 poza tym przedziałem

$$f(x) = \begin{cases} 0, & \text{dla } x < a, \\ c, & \text{dla } a \leq x \leq b, \\ 0, & \text{dla } x > b. \end{cases} \tag{3.10}$$

Ponieważ musi zachodzić warunek (3.9), to

$$\int_{-\infty}^{+\infty} f(x)dx = c\int_a^b dx = c(b-a) = 1, \tag{3.11}$$

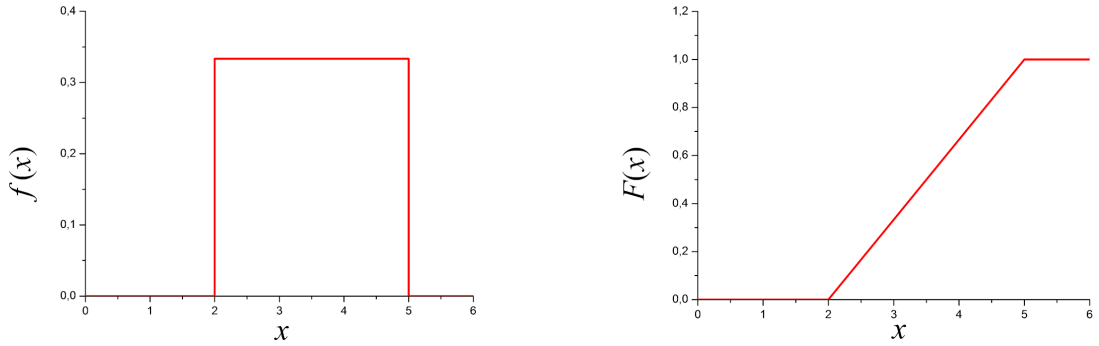
a stąd znajdujemy wartość stałej c

$$c = \frac{1}{b-a}. \tag{3.12}$$

Na przedziale $[a, b)$ dystrybuanta rozkładu jednostajnego jest funkcją liniową

$$F(x) = \int_a^x \frac{dx}{b-a} = \frac{x-a}{b-a}, \tag{3.13}$$

dla $x < a,\ F(x) = 0$, zaś dla $x > b,\ F(x) = 1$. Przykładowy wykres rozkładu jednostajnego i jego dystrybuanty przedstawia rysunek 3.3.



(a) (b)

Rysunek 3.3. Rozkład ciągły jednostajny na przedziale $[2,5)$ (a) oraz jego dystrybuanta (b).

3.3. Funkcje zmiennej losowej

Czasami interesuje nas nie tylko sama zmienna losowa X lecz jakaś funkcja tej zmiennej np.

$$H = H(X). \tag{3.14}$$

Funkcja zmiennej losowej jest również zmienną losową, czyli zmienna H ma również swoją dystrybuantę i gęstość prawdopodobieństwa. Dla nieskończenie małych przedziałów dx zmiennej X i dh zmiennej H zachodzi równość

$$f(x)dx = g(h)dh, \tag{3.15}$$

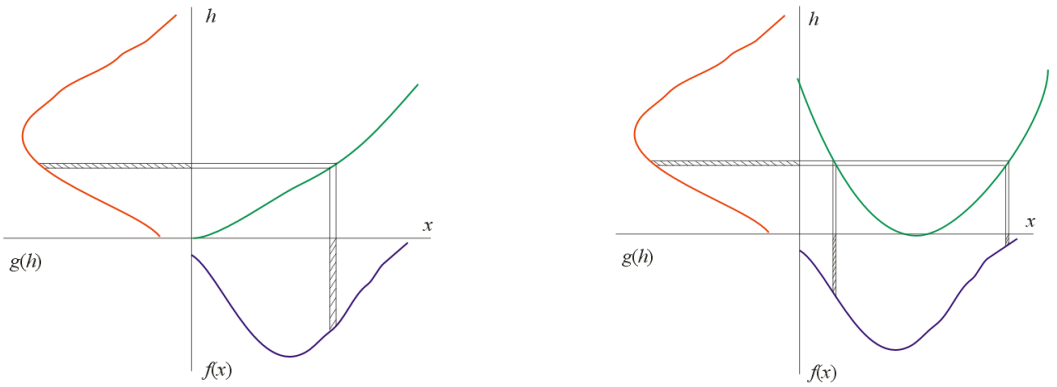
gdzie $g(h)$ jest gęstością prawdopodobieństwa zmiennej H. Z powyższego równania wynika związek pomiędzy rozkładami prawdopodobieństwa obu zmiennych

$$g(h) = \left|\frac{dx}{dh}\right| f(x(h)). \tag{3.16}$$

Wartość bezwzględna pochodnej dh/dx w powyższej relacji wynika stąd, że gęstość prawdopodobieństwa musi być nieujemna. Oczywiście wyrażenie (3.16) określa gęstość prawdopodobieństwa poprawnie jedynie wtedy, gdy funkcja $h(x)$ jest monotoniczna i różniczkowalna w całym zakresie zmienności zmiennej X. W przypadku funkcji różniczkowalnych i monotonicznych w kolejnych przedziałach zmienności zmiennej X gęstość prawdopodobieństwa zmiennej H będzie równa

$$g(h) = \sum_{i=1}^{N} \left|\frac{dx_i(h)}{dh}\right| f(x_i(h)), \tag{3.17}$$

gdzie $x_i(h),\ i = 1,2,\dots,N$ są funkcjami odwrotnymi do funkcji $h(x)$ w poszczególnych przedziałach monotoniczności funkcji $h(x)$ odpowiadającymi danemu h, dla których zmienna H przyjmuje tę samą wartość $h = h(x_1) = h(x_2) = \cdots = h(x_N)$.



Rysunek 3.4. Konstrukcja gęstości prawdopodobieństwa $g(h)$ zmiennej losowej $H = H(X)$ na podstawie znajomości gęstości prawdopodobieństwa $f(x)$ zmiennej X w przypadku, gdy $h(x)$ jest funkcją monotoniczną (a) oraz w przypadku, gdy $h(x)$ jest funkcją przedziałami monotoniczną (b).

Przykład 1

Niech $H = aX + b$, a, b są stałymi rzeczywistymi i $a \neq 0$. Funkcja $h(x) = ax + b$ jest monotoniczna, funkcja odwrotna do niej ma postać $x(h) = (h - b)/a$, a zatem

$$g(h) = \frac{1}{|a|} f\left(\frac{h-b}{a}\right). \tag{3.18}$$

Przykład 2

Niech $H = X^2$. Funkcja $h(x) = x^2$ ma dwa przedziały monotoniczności, jest malejąca w przedziale $(-\infty, 0)$ i rosnąca w przedziale $(0, +\infty)$. W pierwszym przedziale funkcją odwrotną do h jest funkcja $x = -\sqrt{h}$, a w drugim $x = +\sqrt{h}$, a zatem

$$g(h) = \left|\frac{1}{2\sqrt{h}}\right| f\left(-\sqrt{h}\right) + \left|-\frac{1}{2\sqrt{h}}\right| f\left(\sqrt{h}\right) = \frac{1}{2\sqrt{h}}\left(f\left(-\sqrt{h}\right) + f\left(\sqrt{h}\right)\right). \tag{3.19}$$

3.4. Charakterystyki rozkładów

Rozkład zmiennej losowej na ogół jest nieznany i musimy go znajdować z doświadczenia. Bardzo pomocne są różne parametry charakteryzujące rozkłady prawdopodobieństwa, gdyż często możliwości eksperymentu pozwalają jedynie na znajdowanie wartości tych parametrów. Jednym z najważniejszych parametrów każdego rozkładu jest *wartość średnia*, zwana też *wartością oczekiwaną* lub *wartością przeciętną*. W przypadku dyskretnych zmiennych losowych, wartość średnią definiujemy jako sumę wszystkich możliwych wartości x_i zmiennej losowe X pomnożonych przez ich prawdopodobieństwa

$$E(X) \equiv \hat{x} \equiv \sum_i x_i p_i. \tag{3.20}$$

Przez analogię do powyższej definicji, wartość średnią ciągłej zmiennej losowej definiujemy wzorem

$$E(X) \equiv \hat{x} = \int_{-\infty}^{+\infty} xf(x)dx. \tag{3.21}$$

Aby policzyć wartość oczekiwaną funkcji zmiennej losowej nie musimy wyznaczać jest rozkładu. Dla zmiennych dyskretnych wartość oczekiwana funkcji $H(X)$ jest równa

$$E\big(H(X)\big) = \sum_i H(x_i)p_i, \tag{3.22}$$

a dla zmiennych ciągłych

$$E\big(H(X)\big) = \int_{-\infty}^{+\infty} H(x)f(x)dx. \tag{3.23}$$

Z definicji wartości średniej łatwo wyprowadzić następujące własności:

$$\begin{gathered} E(C) = C, \quad C = const, \\ E(C \cdot X) = C \cdot E(X), \quad C = const, \\ E(X_1 + X_2) = E(X_1) + E(X_2). \end{gathered} \tag{3.24}$$

Parametrami rozkładów są również wartości oczekiwane funkcji

$$H(X) = (X - c)^l. \tag{3.25}$$

Nazywamy je *momentami rzędu l względem punktu c*

$$\alpha_l \equiv E((X - c)^l). \tag{3.26}$$

Szczególnie ważna rolę odgrywają *momenty centralne*, czyli momenty względem wartości średniej

$$\mu_l \equiv E((X - \hat{x})^l). \tag{3.27}$$

Pierwsze dwa momenty centralne μ_0 i μ_1 są mało interesujące, gdyż przyjmują one dla każdego rozkładu te same wartości

$$\mu_0 = 1 \ \text{ i } \ \mu_1 = 0. \tag{3.28}$$

Najniższym momentem centralnym wnoszącym informacje o rozkładzie jest moment rzędu 2

$$\mu_2 = \sigma^2(X) = var(X) \equiv E((X - \hat{x})^2). \tag{3.29}$$

Nazywamy go *wariancją* i definiujemy wzorem

$$\sigma^2(X) = \int_{-\infty}^{+\infty} (x - \hat{x})^2 f(x)dx. \tag{3.30}$$

Wariancja informuje nas o średnim odchyleniu zmiennej losowej X od swojej wartości średniej. Oba parametry rozkładów: wartość oczekiwana i wariancja mają bardzo duże znacznie w analizie danych. Jeśli wykonujemy pomiary wielkości fizycznych, to najczęściej wyniki tych pomiarów będą cechowały się pewnym rozrzutem statystycznym. Zakładając, że nasze pomiary nie są obarczone błędami systematycznymi, wyniki naszych pomiarów będą się grupować wokół wartości rzeczywistej mierzonej przez nas wielkości. Rozkład statystyczny będzie miał na ogół kształt tzw. krzywej dzwonowej. Interpretacją geometryczną wartości oczekiwanej jest współrzędna odciętej środka ciężkości krzywej rozkładu. Dlatego rozsądnym jest przyjęcie, że wartość oczekiwana jest najbardziej zbliżona do wartości rzeczywistej. Z kolei wariancja jest miarą szerokości rozkładu gęstości prawdopodobieństwa (patrz Rysunek 3.). Jeśli jej wartość jest duża, wówczas wyniki pomiarów są silnie rozproszone wokół wartości średniej, a jeśli jest mała, to zgrupowane są blisko wartości średniej. Dodatni pierwiastek z wariancji

$$\sigma = \sqrt{\sigma^2(X)} \tag{3.31}$$

jest nazywany *odchyleniem standardowym* lub (*dyspersją*). Odchylenie standardowe jest również miarą rozrzutu wyników pomiaru wokół wartości oczekiwanej, a ponieważ ma taki sam wymiar jak zmienna X jest stosowane jako miara niepewności pomiaru. Wyznacza ono wraz z wartością średnią przedział, o którym możemy założyć z odpowiednim prawdopodobieństwem, że znajduje się w nim wartość rzeczywista mierzonej przez nas wielkości fizycznej. Taka interpretacja odchylenia standardowego wynika z nierówności Czebyszewa

$$P\big(|X - E(X)| \geq a \cdot \sigma(X)\big) \leq 1/a^2. \tag{3.32}$$

Z nierówności wynika, że prawdopodobieństwo odchylenia wartości zmiennej losowej od jej wartości średniej o a-krotną wartość odchylenia standardowego jest nie większe niż $1/a^2$ (a jest tu dowolną dodatnią liczbą rzeczywistą). Nierówność Czebyszewa jest prawdziwa dla wszystkich rozkładów posiadających wariancję.

Zwróćmy jednak uwagę na to, że definicje wartości oczekiwanej (3.21) i wariancji (3.30) wymagają znajomości rozkładu gęstości prawdopodobieństwa, który opisuje dany pomiar, a ten na ogół nie jest znany. Najczęściej zatem jesteśmy w stanie znaleźć jedynie przybliżone wartości tych parametrów.

Wygodną operacją jest tzw. standaryzacja zmiennej losowej, czyli przyporządkowanie zmiennej losowej X zmiennej losowej U zdefiniowanej następująco

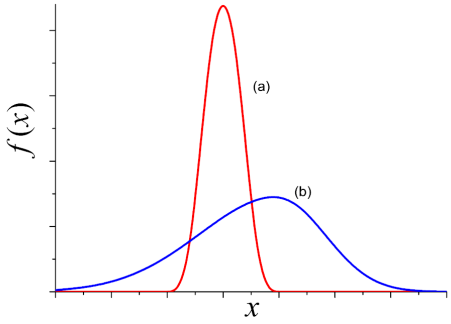
$$U = (X - \hat{x})/\sigma(X). \tag{3.33}$$

Tak zdefiniowana zmienna, ma niezależnie od rozkładu wartość oczekiwaną równą zero

$$E(U) = \frac{1}{\sigma(X)} E(X - \hat{x}) = \frac{1}{\sigma(X)} (\hat{x} - \hat{x}) = 0 \tag{3.34}$$

i wariancję równą jedności

$$\sigma^2(U) = \frac{1}{\sigma^2(X)} E((X - \hat{x})^2) = \frac{\sigma^2(X)}{\sigma^2(X)} = 1. \tag{3.35}$$



Rysunek 3.5. Dwa przykładowe rozkłady (a) o małej wariancji i (b) o dużej wariancji.

Korzystając z własności (3.24) wartości średniej możemy wyprowadzić ważną relację pomiędzy wariancją a wartością średnią

$$\sigma^2(X) \equiv E((x-\hat{x})^2) = E(x^2 - 2x\hat{x} + \hat{x}^2) = E(x^2) - \big(E(x)\big)^2 = \widehat{x^2} - \hat{x}^2. \tag{3.36}$$

Inne własności wariancji to

$$\begin{aligned} \sigma^2(C) &= 0, \\ \sigma^2(C \cdot X) &= C^2 \cdot \sigma^2(X), \\ \sigma^2(C_1 X + C_2) &= C_1^2 \sigma^2(X_1), \end{aligned} \tag{3.37}$$

gdzie C, C_1, C_2 są stałymi.

Kolejny, trzeci moment centralny

$$\mu_3(X) \equiv E((x-\hat{x})^3) \tag{3.38}$$

jest nazywany *współczynnikiem skośności*. Jest on miarą asymetrii rozkładu zmiennej losowej. Wygodniej jest używać bezwymiarowego parametru nazywanego *współczynnikiem asymetrii*

$$\gamma \equiv \frac{\mu_3}{\sigma^3}. \tag{3.39}$$

Współczynnik ten informuje nas o stopniu asymetrii rozkładu zmiennej losowej, a tym samym o możliwych różnicach pomiędzy ujemnymi i dodatnimi odchyleniami mierzonych wartości od wartości średniej. Dla rozkładów symetrycznych współczynnik asymetrii jest oczywiście zerowy, zaś dla rozkładów asymetrycznych może być zarówno ujemny, jak i dodatni zależnie od kształtu rozkładu.

Czwarty moment centralny, podobnie jak trzeci jest sprowadzany zwykle do postaci parametru bezwymiarowego nazywanego kurtozą

$$K \equiv \frac{\mu_4}{\sigma^4} - 3. \tag{3.40}$$

Kurtoza jest miarą spłaszczenia rozkładu lub mówiąc inaczej miarą skupienia wartości wokół wartości średniej. Dla jednego z najważniejszych rozkładów gęstości prawdopodobieństwa – rozkładu normalnego (poznamy go w dalszej części skryptu), kurtoza jest równa zeru, dla rozkładów bardziej wysmukłych niż rozkład normalny kurtoza jest dodania, a dla mniej wysmukłych ujemna. Wyższe momenty centralne są rzadko używane.

Wartość średnia oraz momenty centralne stanowią wystarczający zbiór parametrów określających rozkład zmiennej losowej. Mimo to często używane są dodatkowe charakterystyki opisowe.

Wartość modalna x_m (inaczej *moda, dominanta* lub *wartość najbardziej prawdopodobna*) – dla rozkładów dyskretnych jest to wartość odpowiadająca największemu prawdopodobieństwu wystąpienia, dla rozkładów ciągłych wartość odpowiadająca maksimum gęstości prawdopodobieństwa

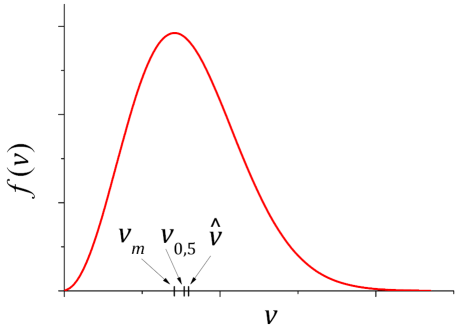
$$P(X = x_m) = max. \tag{3.41}$$

Rozkłady posiadające tylko jedno maksimum nazywamy rozkładami *jednomodalnymi*, a rozkłady o kilku modach *rozkładami wielomodalnymi*. Wartość modalna jest szczególnie użyteczna w przypadku zmiennych nominalnych lub porządkowych, dla których pojęcie wartości średniej i innych charakterystyk nie istnieje.

Mediana $x_{0,5}$ (inaczej *wartość środkowa, wartość przeciętna, drugi kwartyl*) – jest to wartość zmiennej losowej, dla której dystrybuanta przybiera wartość równą 0,5

$$F(x_{0,5}) = P(X < x_{0,5}) = 0{,}5. \tag{3.42}$$

W szeregu uporządkowanych danych poniżej i powyżej mediany mieści się taka sama liczba danych. Mediana dzieli obszar zmienności zmiennej losowej na dwa obszary o jednakowym prawdopodobieństwie (1/2). W przypadku rozkładów symetrycznych mediana jest równa wartości średniej. Jeśli dodatkowo rozkład symetryczny jest jednomodalny, to moda, mediana i wartość średnia są sobie równe. Dla rozkładów asymetrycznych wartości tych charakterystyk różnią się między sobą (patrz Rysunek 3.6).



Rysunek 3.6. Rozkład Maxwella szybkości cząstek gazu doskonałego z zaznaczoną dominantą, medianą i średnią szybkością cząstek.

Oprócz mediany (drugiego kwartyla) zdefiniowane są charakterystyki zwane pierwszym $(x_{0,25})$ i trzecim kwartylem $(x_{0,75})$

$$F(x_{0,25}) = 0{,}25, \qquad F(x_{0,75}) = 0{,}75. \tag{3.43}$$

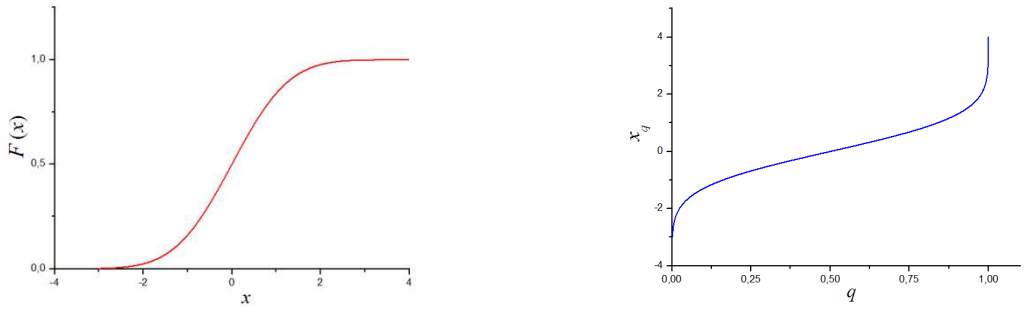
Jak widać pierwszy kwartyl dzieli obszar zmienności zmiennej losowej na dwa obszary: prawdopodobieństwo wystąpienia wartości z pierwszego z nich wynosi 0,25, a drugiego 0,75. Kwartyl trzeci dzieli ten obszar na dwa obszary z prawdopodobieństwami 0,75 i 0,25.

W podobny sposób do kwartyli zdefiniowane są *decyle*: $x_{0,1}, x_{0,2}, \ldots, x_{0,9}$.

Ogólnie parametry dzielące obszar zmienności na dwa obszary, z których prawdopodobieństwo odpowiadające pierwszemu z nich jest dowolną liczbą rzeczywistą q z przedziału $0 \leq q \leq 1$ nazywane są *kwantylami* x_q i definiowane są nastepująco

$$F(x_q) = \int_{-\infty}^{x_q} f(x)dx = q. \tag{3.44}$$

Z matematycznego punktu widzenia kwantyl $x_q(q)$ jest funkcją odwrotną do dystrybuanty.



Rysunek 3.7. Wykres dystrybuanty (a) oraz kwantyli (b) standardowego rozkładu normalnego.

Zadanie 3.1

Niech zmienną losową będzie *n* oznaczającą liczbę oczek na górnej ściance rzuconej kostki do gry. Znajdź wartość oczekiwaną, wariancję i skośność rozkładu prawdopodobieństwa tej zmiennej losowej w przypadku

a) kostki idealnie symetrycznej $\left(p_i = \frac{1}{6},\ i = 1,2,\dots,6\right)$.

b) kostki niesymetrycznej o rozkładzie: $p_1 = \frac{11}{60}, p_2 = \frac{12}{60}, p_3 = \frac{9}{60}, p_4 = \frac{9}{60}, p_5 = \frac{10}{60}, p_6 = \frac{9}{60}$.

Odp:

a) $\hat{x} = 3{,}5, \quad \sigma^2(n) = 2{,}92, \quad \mu_3 = 0, \quad \gamma = \frac{\mu_3}{\sigma^3} = 0.$

b) $\hat{x} \approx 3{,}37, \quad \sigma^2(n) \approx 2{,}97, \quad \mu_3 \approx 0{,}556, \quad \gamma = \frac{\mu_3}{\sigma^3} \approx 0{,}109.$

Zadanie 3.2

Pewna zmienna losowa *X* ma rozkład o postaci

$$f(x) = \begin{cases} 0, & x < 0 \\ -ax(x^2 - 4), & 0 \le x < 2. \\ 0, & x \ge 2 \end{cases} \tag{3.45}$$

Znajdź wartość stałej a, a następnie dystrybuantę zmiennej X.

Odp:

$a = 0{,}25, F(x) = -\frac{1}{16}x^2(x^2 - 8).$

Zadanie 3.3

Korzystając z rozkładu z zadania 3.2 znajdź: wartość średnią, wariancję, współczynnik asymetrii, dominantę i medianę rozkładu.

Odp:

$\hat{x} \approx 1{,}067,\ \sigma^2 \approx 0{,}196,\ \mu_3 \approx -0{,}0108,\ \gamma = \frac{\mu_3}{\sigma^3} \approx -0{,}125,\ x_m \approx 1{,}154,\ x_{0{,}5} \approx 1{,}082.$

Zadanie 3.4

Pewna zmienna losowa X ma rozkład trójkątny o postaci

$$f(x) = \begin{cases} 0, & x < a \\ \dfrac{2(x-a)}{a(a-b)}, & a \le x < 0 \\ \dfrac{2(x-b)}{b(a-b)}, & 0 \le x < b \\ 0, & x \ge b \end{cases}. \tag{3.46}$$

Sprawdź, czy rozkład jest unormowany, a następnie znajdź dystrybuantę zmiennej X.

Zadanie 3.5

Dla rozkładu z zadania 3.4 znajdź: wartość średnią, wariancję, dominantę i medianę rozkładu. Rozważ przypadek rozkładu symetrycznego ($a = -b$) oraz niesymetrycznego, gdy $a = -4b$.
Odp:

$$\hat{x} = \frac{a+b}{3}, \sigma^2(x) = \frac{a^2 - ab + b^2}{18},\ x_m = 0, \qquad x_{0,5} = \begin{cases} a + \sqrt{\dfrac{a(a-b)}{2}}, & a + b \le 0 \\ b - \sqrt{\dfrac{b(b-a)}{2}}, & a + b > 0 \end{cases}.$$

a) $\hat{x} = 0,\ \sigma^2 = \frac{b^2}{6}, x_m = 0,\ x_{0,5} = 0.$
b) $\hat{x} = -b,\ \sigma^2 = \frac{7}{6}b^2,\ \ x_m \approx 0,\ x_{0,5} = b\left(\sqrt{10} - 4\right) \approx -0{,}838b.$

Zadanie 3.6

Gęstość prawdopodobieństwa zmiennej określającej czas życia jąder atomowych pierwiastków promieniotwórczych opisana jest funkcją wykładniczą o postaci

$$f(t) = Ae^{-t/\tau}, \quad t \ge 0,$$

gdzie τ jest parametrem wyznaczanym doświadczalnie, charakterystycznym dla danego pierwiastka, a A stałą normalizacyjną. Znajdź

a) wartość stałej A,
b) średni czas życia atomów,
c) dystrybuantę,
d) medianę rozkładu (w fizyce mediana tego rozkładu nazywana jest czasem połowiczego zaniku),
e) wariancję rozkładu,
f) współczynnik skośności.

Zadanie 3.7

Wykonano dużą liczbę strzałów do tarczy. Gęstość prawdopodobieństwa zmiennej r określającej odległość r trafień do tarczy od jej centrum opisano funkcją

$$f(r) = Are^{-h^2r^2}, \qquad r \geq 0.$$

Znajdź

a) wartość stałej normalizacyjnej A,
b) modę rozkładu, czyli wartość r, dla której gęstość prawdopodobieństwa jest największa,
c) wartość oczekiwaną (średnią) zmiennej r,
d) wariancję zmiennej r.

Zadanie 3.8

Rozkład Maxwella opisuje rozkład szybkości (długości wektora prędkości) v atomów gazu doskonałego. Ma on postać

$$f(v) = Cv^2e^{-\frac{mv^2}{2kT}}, \qquad v \geq 0,$$

gdzie, m jest masą atomu, T temperaturą w skali bezwzględnej, a k stałą Boltzmanna. Znajdź

a) wartość stałej normalizacyjnej C,
b) średnią szybkość atomów,
c) modę rozkładu, czyli szybkość, dla której gęstość prawdopodobieństwa jest największa,
d) wariancję szybkości,
e) oszacuj szybkości otrzymane w punktach b)-d) dla azotu w temperaturze pokojowej.

Zadanie 3.9

Na podstawie rozkładu Maxwella z zadania 3.8 znajdź rozkład prawdopodobieństwa energii kinetycznej cząstek gazu doskonałego i oblicz modę tego rozkładu oraz średnią energię kinetyczną cząstek. Porównaj otrzymany wynik z prawem o ekwipartycji energii.

Wskazówka
W zadaniach od 3.6 do 3.9 skorzystaj z funkcji gamma Eulera i jej własności

$$\Gamma(t) \equiv \int_0^\infty x^{t-1}e^{-x}dx,$$

$$\Gamma(1/2) = \sqrt{\pi}, \;\; \Gamma(1) = 1, \;\; \Gamma(t+1) = t\Gamma(t).$$

lub całki, która też wykorzystuje funkcję gamma Eulera.

$$I(t) = \int_0^\infty x^t e^{-ax^2}dx = \frac{\sqrt{\pi}}{\left(2\sqrt{a}\right)^{t+1}} \frac{\Gamma(\mathrm{t}+1)}{\Gamma\left(\frac{\mathrm{t}}{2}+1\right)}$$

4. Wybrane rozkłady zmiennych losowych

4.1. Rozkład jednostajny ciągły

Najprostszym rozkładem jest rozkład jednostajny nazywany również rozkładem równomiernym lub prostokątnym. Częściowo omówiliśmy go w paragrafie 3.2. Rzadko kiedy zmienne losowe podlegają temu rozkładowi, tym nie mniej, ze względu na prostotę tego rozkładu jest on często wykorzystywany w sposób pośredni. Po pierwsze poprzez transformowanie danego rozkładu do rozkładu jednostajnego, a po drugie (znacznie częściej) do generowania liczb pseudolosowych np. w metodzie Monte Carlo. Zwykle procedury generujące liczby pseudolosowe generują liczby o rozkładzie jednostajnym z założonego przedziału, a następnie stosuje się odpowiednie transformacje zmieniające otrzymany zbiór liczb pseudolosowych na zbiór podlegający pożądanemu rozkładowi.

Jak pokazaliśmy w paragrafie 3.2 rozkład jednostajny zmiennej losowej o stałej wartości gęstości prawdopodobieństwa w przedziale $a \leq x < b$ ma postać

$$f(x) = \begin{cases} 0, & \text{dla } x < a, \\ \dfrac{1}{b-a}, & \text{dla } a \leq x \leq b, \\ 0, & \text{dla } x > b. \end{cases} \tag{4.1}$$

Dystrybuanta tego rozkładu jest dana funkcją

$$F(x) = \begin{cases} 0, & \text{dla } x < a, \\ \dfrac{x-a}{b-a}, & \text{dla } a \leq x \leq b, \\ 1, & \text{dla } x > b. \end{cases} \tag{4.2}$$

Policzmy wartość oczekiwaną zmiennej X

$$E(X) = \hat{x} = \frac{1}{b-a}\int_a^b x dx = \frac{1}{2}\frac{1}{b-a}(b^2 - a^2) = \frac{a+b}{2}. \tag{4.3}$$

Zauważmy, że wartość oczekiwana leży dokładnie w środku przedziału. Wynika to z symetrii rozkładu. Aby wyznaczyć wariancję rozkładu policzmy najpierw wartość średnią zmiennej X^2

$$\begin{aligned} E(X^2) = \frac{1}{b-a}\int_a^b x^2 dx = \frac{1}{3}\frac{1}{b-a}(b^3 - a^3) = \frac{1}{3}\frac{(b-a)(b^2+ab+a^2)}{b-a} \\ = \frac{b^2+ab+a^2}{3}. \end{aligned} \tag{4.4}$$

Teraz skorzystajmy z równania (3.36)

$$\sigma^2(X) = E(X^2) - \big(E(X)\big)^2 = \frac{b^2+ab+a^2}{3} - \left(\frac{a+b}{2}\right)^2 = \frac{(b-a)^2}{12}. \tag{4.5}$$

Ze względu na symetrię rozkładu możemy się spodziewać, że mediana będzie równa wartości oczekiwanej wyliczonej w (4.3) i rzeczywiście tak jest. Z definicji mediany mamy

$$F(x_{0,5}) = \frac{x_{0,5} - a}{b - a} = \frac{1}{2}, \tag{4.6}$$

a stąd dostajemy

$$x_{0,5} = \frac{1}{2}(b - a) + a = \frac{a + b}{2}. \tag{4.7}$$

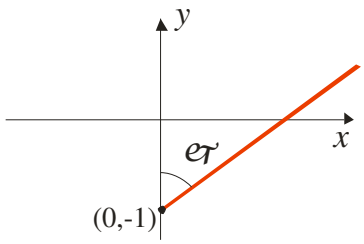
4.2. Rozkład Cauchy'ego

Wyobraźmy sobie punkt o współrzędnych $(0, -1)$ i półprostą leżącą w płaszczyźnie (x, y) wychodzącą z tego punktu. Półprosta może przyjmować losowe położenia określane kątem φ jaki tworzy ona z osią y (patrz Rys. 4.1). Kąt φ może się zmieniać w zakresie $-\frac{\pi}{2} \leq \varphi < \frac{\pi}{2}$. Niech rozkład naszej zmiennej losowej będzie rozkładem jednostajnym. Zgodnie z wynikiem z poprzedniego paragrafu, rozkład zmiennej losowej będzie miał postać

$$f(\varphi) = \frac{1}{\pi} \tag{4.8}$$

w przedziale $-\frac{\pi}{2} \leq \varphi < \frac{\pi}{2}$ i zero poza tym przedziałem.
Dokonajmy teraz transformacji $\varphi \to x$, gdzie x jest współrzędną x-ową punktu, w którym nasza półprosta przecina poziomą oś układu xy. Zgodnie z równaniem (3.16), rozkład zmiennej φ wyliczymy z równania

$$g(x) = \left|\frac{d\varphi}{dx}\right| f(\varphi). \tag{4.9}$$



Rysunek 4.1. Konstrukcja pomocna w wyprowadzeniu rozkładu Cauchy'ego.

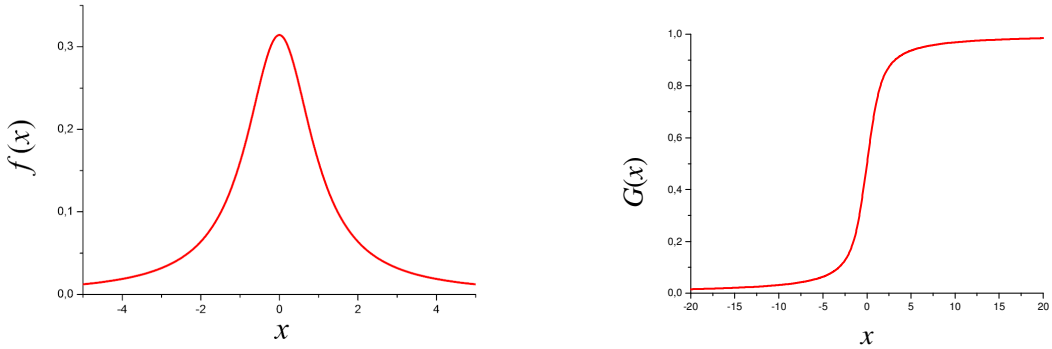
Ponieważ

$$\varphi(x) = \operatorname{arctg}(x) \quad \text{i} \quad \frac{d\varphi}{dx} = \frac{1}{1 + x^2}, \tag{4.10}$$

dostajemy

$$g(x) = \frac{1}{\pi}\frac{1}{1 + x^2}. \tag{4.11}$$

Otrzymany rozkład nazywamy *rozkładem Cauchy'ego* lub *rozkładem Cauchy'ego-Lorentza*. Rozkład Cauchy'ego przedstawia rysunek 4.2 (a), a jego dystrybuantę przedstawia rysunek 4.2.(b).



(a) (b)

Rysunek 4.2 Rozkład Cauchy'ego (a) i jego dystrybuanta (b).

Dystrybuanta rozkładu Cauchy'ego jest równa

$$G(x) = \frac{1}{\pi}\int_{-\infty}^{x}\frac{du}{1+u^2} = \frac{1}{\pi}\left(\operatorname{arctg}(x) + \frac{\pi}{2}\right). \tag{4.12}$$

Policzmy wartość oczekiwaną zmiennej x

$$E(x) = \frac{1}{\pi}\int_{-\infty}^{\infty}\frac{x dx}{1+x^2} = 0, \tag{4.13}$$

gdyż funkcja podcałkowa jest nieparzysta. Jak widać z wykresu wartość średnia pokrywa się z wartością modalną $\hat{x} = x_m = 0$. Ze wzoru na wariancję dostajemy

$$\begin{aligned}\sigma^2(x) &= \frac{1}{\pi}\int_{-\infty}^{\infty}\frac{x^2 dx}{1+x^2} = \frac{2}{\pi}\int_{0}^{\infty}\frac{x^2 dx}{1+x^2} = \frac{2}{\pi}\int_{0}^{\infty} dx - \frac{2}{\pi}\int_{0}^{\infty}\frac{dx}{1+x^2} \\ &= \frac{2}{\pi}\left(x - \operatorname{arctg}(x)\right)\Big|_0^{\infty} = \frac{2}{\pi}\lim_{x\to\infty}\left(x - \operatorname{arctg}(x)\right).\end{aligned} \tag{4.14}$$

Wartość policzonej powyżej całki jest nieskończona, co oznacza, że nie istnieje wariancja rozkładu Cauchy'ego. W takich przypadkach za miarę szerokości rozkładu przyjmuje się tzw. *szerokość połówkową*, czyli szerokość rozkładu odpowiadająca połowie wartości maksymalnej. Wartość maksymalna rozkładu Cauchy'ego wynosi

$$g(x_m) = g(0) = \frac{1}{\pi}\frac{1}{1+0^2} = \frac{1}{\pi}. \tag{4.15}$$

Połowę wartości maksymalnej, czyli $1/2\pi$ gęstość prawdopodobieństwa osiąga w punktach $x_1 = -1$ i $x_2 = 1$, a zatem szerokość połówkowa rozkładu Cauchy'ego wynosi dwa

$$\Gamma = 2. \tag{4.16}$$

4.3. Rozkład Lorentza

Uogólnieniem rozkładu Cauchy'ego jest *rozkład Lorentza* nazywany też rozkładem Breita-Wignera. Rozkład ten ma postać

$$g(x) = \frac{2}{\pi\Gamma}\frac{\Gamma^2}{4(x-a)^2+\Gamma^2}, \tag{4.17}$$

gdzie a i Γ są parametrami rozkładu, przy czym a jest dowolną liczbą rzeczywistą, a $\Gamma > 0$. Jak widać rozkład Cauchy'ego jest rozkładem Lorentza z parametrami $a = 0$ i $\Gamma = 2$. Ciekawą własnością rozkładu Lorentza jest to, że jest on unormowany niezależnie od wartości parametrów a i Γ. Dodatkowo wartość oczekiwana jest równa parametrowi a $(E(x) = a)$, a szerokość połówkowa rozkładu jest równa parametrowi Γ.

Rozkład Lorentza jest ważnym rozkładem dla fizyków. Jest on używany w różnych dziedzinach fizyki głównie w optyce i fizyce jądrowej do opisu zjawisk rezonansowych.

4.4.Rozkład dwumianowy

Rozkład dwumianowy opisuje liczbę k sukcesów w ciągu n prób, w których prawdopodobieństwo ma stałą wartość równą p. Pojedynczą próbę nazywamy próbą Bernoulliego. Rozkład doświadczenia stanowiącego próbę Bernoulliego jest rozkładem zero-jedynkowym. W literaturze anglojęzycznej jest on nazywany rozkładem Bernoulliego, podczas, gdy w literaturze polskiej nazwa rozkład Bernoulliego jest często używana wymiennie z nazwą rozkład dwumienny. W rozkładzie zero-jedynkowym, opisującym pojedynczą próbę Bernoulliego zmiennej losowej przypisujemy zwykle dwie wartości: 1 w przypadku zajścia zdarzenia A nazywanego sukcesem i 0 w przeciwnym wypadku (zajście zdarzenia $\bar{A}$). Prawdopodobieństwo sukcesu (zdarzenia A) wynosi $P(A) = p$, a zdarzenia przeciwnego $P(\bar{A}) = 1 - p = q$. Typowym przykładem próby Bernoulliego jest rzut monetą: wyrzuceniu orła możemy przypisać wartość 1, a wyrzuceniu reszki wartość 0. W przypadku symetrycznej monety $p = q = 0{,}5$.

Dzięki przypisaniu zmiennej X_i reprezentującej i-tą próbę Bernoulliego wartości 0 i 1 możemy zdefiniować zmienną X rozkładu dwumianowego jako sumę N zmiennych kolejnych prób Bernoulliego

$$X = \sum_{i=1}^{N} X_i. \tag{4.18}$$

Korzystając z niezależności prób Bernoulliego możemy wyliczyć prawdopodobieństwo, że k pierwszych prób Bernoulliego zakończy się sukcesem, a pozostałe n-k prób da wynik negatywny. Prawdopodobieństwo to będzie równe $p^k q^{n-k}$. Tyle samo będzie wynosiło prawdopodobieństwo dowolnej innej kombinacji k sukcesów w n próbach Bernoulliego. Zgodnie z kombinatoryką liczba takich kombinacji wynosi

$$\binom{n}{k} = \frac{n!}{k!\,(n-k)!}. \tag{4.19}$$

Szukane prawdopodobieństwo uzyskania k sukcesów w n próbach Bernoulliego wynosi zatem

$$P(k) = \frac{n!}{k!\,(n-k)!}p^k q^{n-k}. \tag{4.20}$$

Powyższy rozkład nazywamy rozkładem dwumianowym. Aby znaleźć jego niektóre charakterystyki zajmijmy się najpierw rozkładem zero-jedynkowym opisującym pojedynczą próbę Bernoulliego. Wartość oczekiwana wynosi

$$E(X_i) = 1 \cdot p + 0 \cdot q = p, \tag{4.21}$$

a wariancja

$$\begin{aligned}\sigma^2(X_i) = E((x_i - p)^2) &= (1-p)^2 p + (0-p)^2 q = 1 - 2p^2 + p^3 + p^2 q \\ &= p(1-p) - p^2(1-p) + p^2 q = pq - p^2 q + p^2 q = pq \\ &= p(1-p).\end{aligned} \tag{4.22}$$

Teraz możemy policzyć wartość oczekiwaną i wariancję rozkładu dwumianowego. Z uogólnienia trzeciej z własności (3.24) mamy

$$E(X) = E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} E(X_i) = np. \tag{4.23}$$

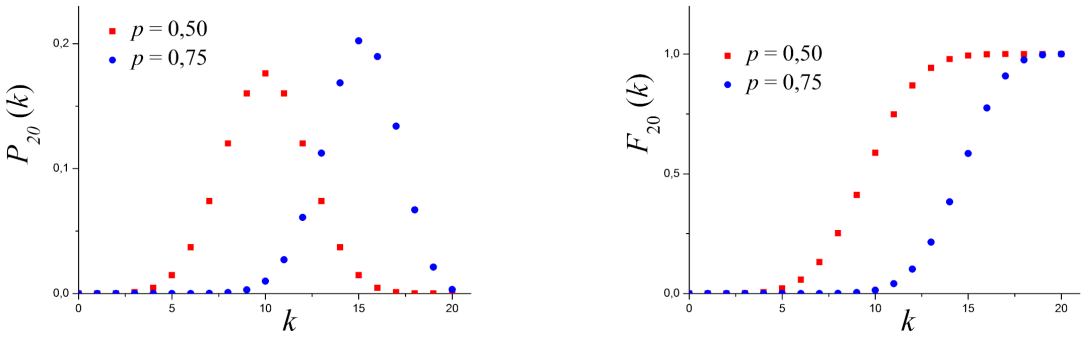
Dla zmiennych niezależnych zachodzi związek (udowodnimy go w następnym rozdziale)

$$\sigma^2\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \sigma^2(X_i). \tag{4.24}$$

Ponieważ próby Bernoulliego są zdarzeniami niezależnymi to

$$\sigma^2(X) = \sigma^2\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \sigma^2(X_i) = npq = np(1-p). \tag{4.25}$$

Przykładowy rozkłady dwumianowe i ich dystrybuanty pokazano na rysunkuRysunek 4.3.



(a) (b)

Rysunek 4.3. Przykładowy rozkład dwumianowy dla $p = 0{,}75$ *i* $n = 10$ (a) oraz jego dystrybuanta (b).

4.5. Rozkład Poissona

Załóżmy, tak jak w przypadku rozkładu dwumiennego, że mamy do czynienia ze zmienną losową określającą liczbę *k* sukcesów w ciągu *n* prób, w których prawdopodobień-

stwo ma stałą wartość równą p. Niech liczba prób n będzie bardzo duża, a p z kolei bardzo małe. Typową sytuacją pasującą do tego opisu jest zjawisko rozpadu jąder atomowych pierwiastków radioaktywnym. Prawdopodobieństwo rozpadu pojedynczego jądra w danym (małym) przedziale czasowym jest bardzo małe, zaś liczba jąder w próbce bardzo duża. Średnia liczba obserwowanych rozpadów jąder w danym czasie jest w miarę stała. Rozkład naszej zmiennej jest dobrze opisany rozkładem dwumiennym, ale jego zastosowanie byłoby w tym przypadku bardzo utrudnione ze względu na ogromne wartości n. Przypomnijmy, że liczba jader w próbce zawierającej 1 mol substancji radioaktywnej zawiera rzędu $6 \cdot 10^{23}$ jąder, a w rozkładzie dwumiennych operujemy operacją *silnia*. Przekształćmy rozkład dwumienny w opisanych wyżej warunkach do postaci wygodniejszej w użyciu. Wprowadźmy oznaczenia $\lambda \equiv np$ (przypomnijmy, że jest to równe wartości oczekiwanej zmiennej podlegającej rozkładowi dwumiennemu). Dokonajmy następujących przekształceń

$$P_n(k) = \frac{n!}{k!\,(n-k)!} p^k q^{n-k} = \frac{n!}{k!\,(n-k)!} \left(\frac{\lambda}{n}\right)^k \frac{(1-\lambda\backslash n)^n}{(1-\lambda\backslash n)^k}$$

$$= \frac{\lambda^k}{k!} \frac{n(n-1)(n-2)\ldots(n-k+1)}{n^k} \frac{(1-\lambda\backslash n)^n}{(1-\lambda\backslash n)^k} \tag{4.26}$$

$$= \frac{\lambda^k}{k!}\left(1-\frac{\lambda}{n}\right)^n \frac{\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\ldots\left(1-\frac{k-1}{n}\right)}{\left(1-\frac{\lambda}{n}\right)^k}.$$

Ponieważ zakładamy, że *n* jest bardzo duże znajdźmy postać powyższego wyrażenia w przypadku granicznym

$$\lim_{n\to\infty} \frac{\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\ldots\left(1-\frac{k-1}{n}\right)}{\left(1-\frac{\lambda}{n}\right)^k} = 1, \tag{4.27}$$

zaś

$$\lim_{n\to\infty} \left(1-\frac{\lambda}{n}\right)^n = e^{-\lambda}. \tag{4.28}$$

Ostatecznie otrzymujemy rozkład nazywany *rozkładem Poissona*

$$P_\lambda(k) = \frac{\lambda^k}{k!} e^{-\lambda}. \tag{4.29}$$

Sprawdźmy, że rozkład Poissona spełnia warunek normalizacyjny, tzn. czy suma prawdopodobieństw dla dowolnych wartości zmiennej *k* jest równa jedności.

$$\sum_{k=0}^{\infty} P_\lambda(k) = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{-\lambda} e^{\lambda} = 1. \tag{4.30}$$

Obliczmy wartość oczekiwaną dla rozkładu Poissona

$$E(K) = \sum_{k=0}^{\infty} k\frac{\lambda^k}{k!}e^{-\lambda} = e^{-\lambda}\sum_{k=1}^{\infty} k\frac{\lambda^k}{k!} = \sum_{k=1}^{\infty}\frac{\lambda\lambda^{k-1}}{(k-1)!}e^{-\lambda} = \lambda e^{-\lambda}\sum_{l=0}^{\infty}\frac{\lambda^l}{l!} = \lambda e^{-\lambda}e^{\lambda} = \lambda. \tag{4.31}$$

W podobny sposób możemy pokazać, że

$$E(K^2) = \lambda(\lambda+1). \tag{4.32}$$

Wobec tego wariancja dla rozkładu Poissona wynosi

$$\sigma^2(K) = E(K^2) - \big(E(K)\big)^2 = \lambda(\lambda+1) - \lambda^2 = \lambda. \tag{4.33}$$

Wariancja dla rozkładu Poissona jest równa wartości oczekiwanej.

4.6. Rozkład Gaussa (rozkład normalny)

Rozkład Gaussa zwany też *rozkładem normalnym* jest jednym z najważniejszych rozkładów gęstości prawdopodobieństwa. Ma on postać

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \tag{4.34}$$

Wykresem tej funkcji jest krzywa nazywana ze względu na swój kształt krzywą dzwonową (patrz Rysunek 4.4 (a)).
Większość zmiennych losowych z różnych dziedzin np.: w naukach przyrodniczych, w badaniach inżynierskich, przemysłowych, medycznych, socjologicznych, ekonomicznych podlega rozkładowi normalnemu. Jest to naturalna cecha zmiennych losowych, dla których rozrzut statystyczny wartości jest efektem wielu nakładających się czynników. Zostało to sformułowane i udowodnione w postaci tzw. *centralnego twierdzenia granicznego.* W najprostszym ujęciu stwierdza ono[4], że jeżeli zmienne losowe X_i są niezależnymi zmiennymi o wartościach oczekiwanych $E(X_i) = a$ i wariancjach $\sigma(X_i) = b$, to zmienna

$$X = \lim_{n\to\infty}\sum_{i=1}^{n} X_i \tag{4.35}$$

ma rozkład normalny o parametrach $\mu = na$, i $\sigma^2 = nb^2$. Zauważmy, że X_i mogą podlegać dowolnemu rozkładowi, twierdzenie nie narzuca postaci rozkładu dla zmiennych. Weźmy dla przykładu rozkład dwumienny. Jak pamiętamy zmienną losową tego rozkładu składaliśmy z n zmiennych losowych X_i o wartościach będących wynikami prób Bernoulliego, którym przypisywaliśmy wartości 0 albo 1. Zgodnie z centralnym twierdzeniem granicznym dla odpowiednio dużego n rozkład dwumienny można przybliżyć rozkładem Gaussa o parametrach $\mu = p$ i $\sigma^2 = np(1-p)$. Oczywiście trzeba pamiętać o tym, że rozkład dwumienny jest rozkładem dyskretnym, a rozkład Gaussa ciągłym.

Parametry rozkładu Gaussa mają następujące znaczenie: $\mu = E(X)$ i $\sigma^2 = \sigma^2(X)$, czyli parametr μ jest równy wartości średniej, a parametr σ^2 jest równy wariancji rozkładu (patrz zadanie 4.5). Ponadto można łatwo pokazać, że dla $x = \mu \pm \sigma$ rozkład Gaussa ma tzw.

[4] Istnieją również inne, ogólniejsze sformułowania tego twierdzenia.

punkty przegięcia (patrz zadanie 4.6). Dla standaryzowanej zmiennej $U = (X - \hat{x})/\sigma(X)$, dla której $E(U) = \hat{u} = 0$ i $\sigma^2(U) = 1$ rozkład Gaussa przyjmuje postać

$$f(x) = \phi_0(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}. \tag{4.36}$$

Ten rozkład nazywamy *standardowym rozkładem normalnym* oznaczany często jako $N(0,1)$.

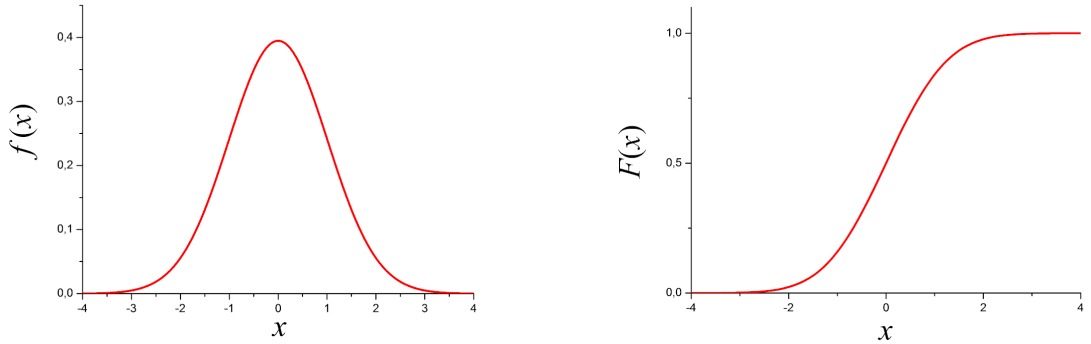
Dystrybuanty rozkładu normalnego, czyli funkcji

$$F(x) = \Psi(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt \tag{4.37}$$

lub w przypadku standardowego rozkładu Gaussa

$$F(x) = \Psi_0(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt \tag{4.38}$$

nie da się zapisać w postaci analitycznej. Kształt dystrybuanty standardowego rozkładu normalnego przedstawiono na rysunku 4.4 (b).

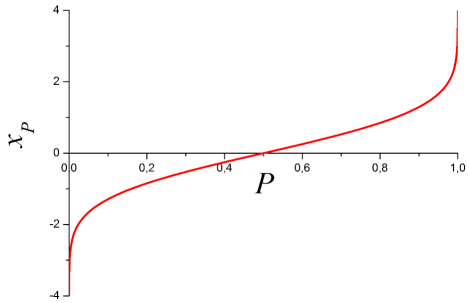


(a) (b)

Rysunek 4.4. Gęstość prawdopodobieństwa standardowego rozkładu Gaussa (a) i jego dystrybuanta (b).

Wartości tej funkcji są stablicowane, jednak dziś w dobie komputerów możemy znajdować wartości dystrybuanty rozkładu normalnego (nie koniecznie standardowego) za pomocą wielu programów statystycznych, w tym także za pomocą programu MS Excel. W programie Excel do wyznaczenia wartości gęstości rozkładu normalnego lub jego dystrybuanty korzystamy z funkcji o nazwie ROZKŁAD.NORMALNY. Funkcja przyjmuje cztery argumenty. Pierwszy – *x* jest wartością, dla której ma zostać znaleziona wartość rozkładu lub dystrybuanty, drugi jest wartością parametru μ rozkładu, trzeci jest wartością parametru σ rozkładu i ostatni, czwarty jest argumentem logicznym. Jeśli wartością ostatniego parametru jest *fałsz*, to funkcja zwraca wartość gęstości prawdopodobieństwa, a jeśli wartością ostatniego parametru jest *prawda*, to funkcja zwraca wartość dystrybuanty rozkładu normalnego.

Często potrzebne są kwantyle $x_p(P)$ rozkładu normalnego. Jak wiemy jest to funkcja odwrotna do dystrybuanty. Kwantyle standardowego rozkładu normalnego $N(0,1)$ pokazano na (4.5. Zwróćmy uwagę na własność kwantyli standardowego rozkładu normalnego: $x_P = -x_{1-p}$.



Rysunek 4.5. Kwantyle standardowego rozkładu normalnego.

Ze względu na symetrię rozkładu Gaussa względem μ

$$F(\mu - a) = F(\mu + a) - 1. \tag{4.39}$$

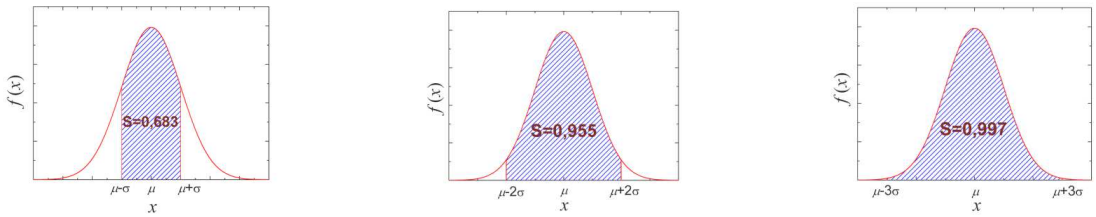
Korzystając z powyższej równości można pokazać, że prawdopodobieństwo zaobserwowania wartości zmiennej losowej wewnątrz przedziału $[\mu - a, \mu + a]$, czyli prawdopodobieństwo, że $|X - \mu| \leq a \ \ (a > 0)$, czyli wynosi

$$P(|X - \mu| \leq a) = 1 - 2F(\mu - a) = 2F(\mu + a) - 1. \tag{4.40}$$

Warto zapamiętać prawdopodobieństwa dla trzech szczególnych przedziałów symetrycznych względem μ o szerokościach $\sigma, 2\sigma, 3\sigma$:

$$\begin{aligned} P(|X - \mu| \leq \sigma) &\approx 0{,}683, \\ P(|X - \mu| \leq 2\sigma) &\approx 0{,}955, \\ P(|X - \mu| \leq 3\sigma) &\approx 0{,}997. \end{aligned} \tag{4.40}$$

Geometryczna interpretacja tych prawdopodobieństw w postaci pola powierzchni pod krzywą Gaussa w opisanych powyżej przedziałach przedstawia Rysunek 4.6.

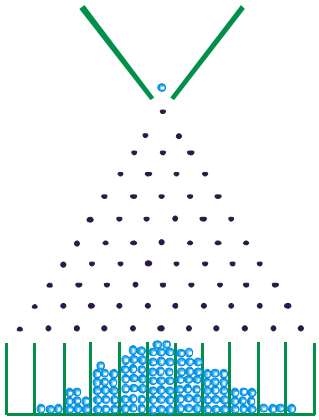


(a) (b) (c)

Rysunek 4.6. Pole powierzchni pod krzywą Gaussa dla przedziałów: $[\mu - \sigma, \mu + \sigma]$ (a), $[\mu - 2\sigma, \mu + 2\sigma]$ (b) i $[\mu - 3\sigma, \mu + 3\sigma]$ (c). Pola powierzchni odpowiadają prawdopodobieństwom wystąpienia wartości x w odpowiednim przedziale.

Rozkładowi Gaussa podlegają, w większości przypadków, błędy pomiarów fizycznych. Interpretację tego faktu podał Pierre Laplace w 1783 roku. Każdy pomiar jest zakłócany przez dużą liczbę niezależnych czynników. Laplace założył, że każdy z nich powoduje odchylenie wyniku pomiaru o pewną wartość, która z jednakowym prawdopodobieństwem może przyjmować wartości dodatnie i ujemne. Błąd pomiaru, czyli różnica między wartością rzeczywistą mierzonej wielkości, a wartością zmierzoną jest sumą wszystkich zakłóceń. Fizycznym modelem tej sytuacji jest tzw. deska Galtona. Jest to deska z nabitymi gwoździami rozmieszczonymi na powierzchni trójkąta równobocznego jak pokazano na Rysunku 4.7. Nad wierzchołkiem trójkąta umieszczony jest lejek, z którego wypadają kuleczki. Każda kuleczka zderzając się z gwoździem może z jednakowym prawdopodobieństwem odbić się w lewo lub w prawo. Po przejściu przez trójkąt z gwoździ kuleczka trafia do odpowiedniej przegródki.

Jest to dokładnie tak jak z pomiarami w interpretacji błędów pomiarowych Laplace'a, każdy gwóźdź, na który trafia kuleczka odpowiada danemu zakłóceniu. Po przejściu wielu takich kuleczek w przegródkach zobaczymy histogram rozkładu.



Rysunek 4.7. Deska Galtona obrazująca model Laplace'a błędów pomiarowych.

Rozkład prawdopodobieństwa dla modelu błędów Laplace'a wiąże się ściśle ze słynnym trójkątem Pascala, w którym każdy kolejny wiersz składa się ze współczynników Newtona, czyli współczynników rozwinięcia dwumianu $(a+b)^n$. Współczynniki te, podzielone przez 2^n $(1/(p^k q^{n-k}) = 2^n$ dla $p = q = 1/2)$ są równe prawdopodobieństwu znalezienia się kuleczki przy danym gwoździu w rzędzie o numerze n (patrz Tabela 4.1). Liczba rzędów gwoździ odpowiada liczbie zakłóceń pomiaru. Dla małej liczby zakłóceń rozkładem jest rozkład dwumianowy. Jednak zwykle zakłóceń jest dużo. W granicy $n \to \infty$ rozkład dwumianowy przechodzi w rozkład Gaussa, co wynika z centralnego twierdzenia granicznego. Przyczyny błędów pomiarowych mogą być bardzo złożone i nie istnieje jeden uniwersalny rozkład gęstości prawdopodobieństwa, który opisywałby dowolny przypadek. Rozkład Gaussa pasuje do wielu przypadków, ale nie jest rozkładem uniwersalnym, dlatego jeśli chcemy go zastosować powinniśmy przeprowadzić odpowiedni test zgodności, aby się upewnić, czy dany rozkład jest rozkładem normalnym. Niektóre z tych testów omówimy w rozdziale 9.

Tabela 4.1. Prawdopodobieństwa popełnienia błędu pomiaru w modelu Laplace'a w zależności od liczby zakłóceń. Liczniki ułamków są współczynnikami Newtona w trójkącie Pascala.

Liczba zakłóceń	różnica między wartością prawdziwą a zmierzoną						
n	-3ε	-2ε	$-\varepsilon$	0	$+\varepsilon$	$+2\varepsilon$	$+3\varepsilon$
0				1			
1			1/2		1/2		
2		1/4		2/4		1/4	
3	1/8		3/8		3/8		1/8

4.7. Rozkład χ^2 (chi kwadrat)

Rozkład χ^2 (czytaj chi kwadrat) jest bardzo ważnym rozkładem w analizie danych stosowanym w różnego rodzaju testach statystycznych. Rozkład ten zależy od liczby stopni swobody zmiennej losowej. Zmienna losowa X ma rozkład chi kwadrat o n stopniach swobody jeśli jest sumą n niezależnych zmiennych losowych X_i o standardowym rozkładzie normalnym $N(0,1)$, tzn.

$$X = X_1^2 + X_2^2 + \cdots + X_n^2 = \sum_{i=1}^{n} X_i^2. \tag{4.41}$$

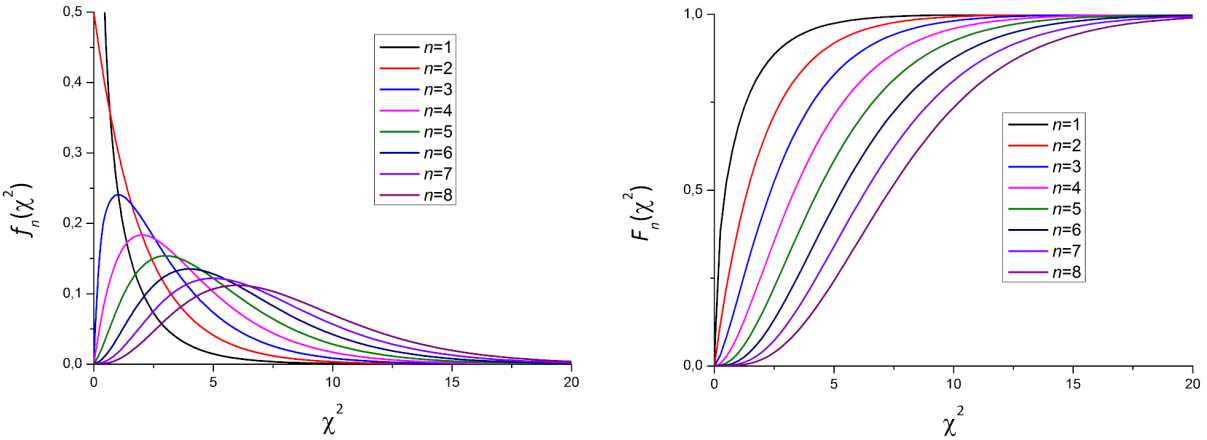
Symbol rozkładu χ^2 został wprowadzony przez Karla Pearsona. Dwójka w górnym indeksie tego symbolu podkreśla fakt, że zmienna losowa jest sumą kwadratów, ale w samym symbolu nie należy jej traktować jak symbol potęgowania. Uwaga ta dotyczy też symbolu zmiennej o rozkładzie chi kwadrat, w którym też często dodaje się w górnym indeksie dwójkę. Np. w powyższym równaniu można by było po lewej stronie użyć symbolu X^2 zamiast X i nie należy tego mylić z potęgowaniem (potęgowanie pojawia się po prawej stronie równania 4.42). Gęstość rozkładu chi kwadrat ma postać

$$f_n(x) = \frac{x^{\frac{n}{2}-1} e^{-\frac{x}{2}}}{2^{\frac{n}{2}} \Gamma\left(\frac{n}{2}\right)}, \tag{4.42}$$

gdzie Γ jest funkcją, która dla n naturalnych przyjmuje następujące wartości

$$\Gamma\left(\frac{n}{2}\right) = \begin{cases} \left(\frac{n}{2} - 1\right)!, & \text{dla } n \text{ parzystych} \\ \frac{(n-1)!}{2^{n-1}\left(\frac{n-1}{2}\right)!}\sqrt{\pi}, & \text{dla } n \text{ nieparzystych} \end{cases}. \tag{4.43}$$

Oczywiście ze względu na konstrukcję zmiennej losowej (jest sumą kwadratów zmiennych losowych) wartości zmiennej muszą być nieujemne $\left(x \in [0, \infty)\right)$. Wykresy gęstości prawdopodobieństwa rozkładów chi kwadrat o liczbach stopni swobody od $n = 1$ do $n = 8$ oraz ich dystrybuanty przedstawia rysunek 4.8.



(a) (b)

Rysunek 4.8. Gęstość prawdopodobieństwa rozkładu chi kwadrat dla liczby stopni swobody $n = 1,2,\ldots,8$ (a) oraz jej dystrybuanta (b).

Bez dowodu podajemy poniżej kilka ważnych własności rozkładu chi kwadrat.

- **Twierdzenie:** Jeśli zmienna X_1 ma rozkład chi kwadrat o n_1 stopniach swobody, a zmienna X_2 ma rozkład chi kwadrat o n_2 stopniach swobody to zmienna $X = X_1 + X_2$ ma rozkład chi kwadrat o $n = n_1 + n_2$ stopniach swobody.
- Wartość oczekiwana rozkładu chi kwadrat o n stopniach swobody jest równa liczbie stopni swobody ($E(X^2) = n$).
- Wariancja rozkładu chi kwadrat o n stopniach swobody jest równa podwojonej liczbie stopni swobody ($\sigma^2(X^2) = 2n$).

Zadanie 4.1

Pokaż, że rozkład Lorentza $g(x) = \frac{2}{\pi\Gamma}\frac{\Gamma^2}{4(x-a)^2+\Gamma^2}$ jest zawsze unormowany. Sprawdź, że szerokość połówkowa tego rozkładu jest równa parametrowi Γ, a wartość oczekiwana, wartość modalna i mediana są równe parametrowi a.

Zadanie 4.2

Transport bananów jest odrzucany przez sklep, jeśli w 10 losowo wybranych skrzynkach znajdą się co najmniej dwie z zepsutymi bananami. Załóżmy, że transporcie 2% skrzynek bananów ulega zepsuciu. Jakie jest prawdopodobieństwo, że transport zostanie odrzucony?

Wskazówka:

Zastosuj rozkład dwumienny z $n = 10$ i $p = 0{,}02$. Oblicz prawdopodobieństwo zdarzenia, że w wylosowanej próbie jest mniej niż 2 skrzynki z zepsutymi bananami $P_{10}(k < 2) = P_{10}(0) + P_{10}(1)$, a następnie prawdopodobieństwo zdarzenia przeciwnego. Do wartości wyliczenia symbolu Newtona $\binom{n}{k}$ skorzystaj z funkcji KOMBINACJE.

Zadanie 4.3

Jak duża powinna być próba losowa w problemie z zadania 4.2, aby z prawdopodobieństwem $P(K \geq 1) = 0{,}1$ znalazła się w niej co najmniej jedna skrzynka z zepsutymi bananami? Powtórz zadanie dla $P(K \geq 1) = 0{,}99$.

Wskazówka:

Jak w zadaniu 4.2 stosujemy rozkład dwumienny z $p = 0{,}02$, ale nieznanym n. Rozważ zdarzenie przeciwne w wylosowanej *n*-elementowej próbie wszystkie skrzynki zawierają niezepsute banany z prawdopodobieństwem $P_n(K = 0) = 1 - 0{,}1 = 0{,}9$ i wylicz dla jakiego n będzie to miało miejsce (wynik trzeba zaokrąglić do najmniejszej liczby naturalnej większej lub równej otrzymanej).

Zadanie 4.4

W teście jednokrotnego wyboru znajduje się 20 pytań i po cztery odpowiedzi do każdego z nich. Test jest zaliczony, jeśli student odpowie na co najmniej 12 pytań. Jakie jest prawdopodobieństwo zaliczenia testu przy wyborze odpowiedzi na chybił trafił? Jaka jest najbardziej prawdopodobna liczba poprawnych odpowiedzi?

Wskazówka:

Należy zastosować rozkład dwumienny z $n = 20$ i $p = 1/4$, a następnie policzyć prawdopodobieństwo poprawnej odpowiedzi na 12, 13, … 20 pytań $P_{20}(k \geq 12) = \sum_{k=12}^{20} P_{20}(k)$.

Zadanie 4.5

Pokaż, że parametr μ jest równy wartości średniej, a parametr σ^2 jest równy wariancji rozkładu Gaussa $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$.

Wskazówka:

Skorzystaj z funkcji gamma Eulera Γ i jej własności

$$\Gamma(t) \equiv \int_0^\infty x^{t-1} e^{-x} dx,$$
$$\Gamma(t+1) = t\Gamma(t),$$
$$\Gamma(1) = 1,$$
$$\Gamma(1/2) = \sqrt{\pi}.$$

lub całki, która też wykorzystuje funkcję gamma Eulera.

$$I(t) = \int_0^\infty x^t e^{-ax^2} dx = \frac{\sqrt{\pi}}{\left(2\sqrt{a}\right)^{t+1}} \frac{\Gamma(\mathrm{t}+1)}{\Gamma\left(\frac{\mathrm{t}}{2}+1\right)}.$$

Zadanie 4.6

Pokaż, że dla $x = \mu \pm \sigma$ rozkład Gaussa ma tzw. punkty przegięcia.

Wskazówka:

W punkcie przegięcia różniczkowalnej funkcji, pochodna tej funkcji ma lokalne ekstremum. Oblicz pochodną rozkładu Gaussa i znajdź jej punkty ekstremalne.

Zadanie 4.7

Korzystając z funkcji Excela Rozkład.Normalny sporządź wykresy gęstości prawdopodobieństwa i ich dystrybuanty dla kilku wybranych wartości parametrów μ i σ rozkładu Gaussa.

5. Funkcja charakterystyczna rozkładu

Definiując zmienne losowe w paragrafie 3.1 mówiliśmy, że przypisujemy zmiennym losowym liczby rzeczywiste. Załóżmy, że zmiennej losowej możemy przypisywać również liczby zespolone. Taką zmienną losową możemy zdefiniować jako

$$Z = X + iY \tag{5.1}$$

gdzie X, Y są rzeczywistymi zmiennymi losowymi. Wartość oczekiwaną zmiennej Z możemy zdefiniować jako

$$E(Z) = E(X) + iE(Y). \tag{5.2}$$

Zespolone zmienne losowe będziemy uważać za niezależne, jeśli odpowiednio ich części rzeczywiste i urojone są niezależne.

Niech $f(x)$ będzie gęstością prawdopodobieństwa zmiennej X. Wartość oczekiwaną wyrażenia e^{itX}, czyli

$$\varphi(t) \equiv E(e^{itx}) \tag{5.3}$$

nazywamy *funkcją charakterystyczną* rozkładu. Jeśli mamy do czynienia ze zmienną ciągłą, to funkcja charakterystyczna ma postać

$$\varphi(t) = \int_{-\infty}^{\infty} e^{itx} f(x)dx, \tag{5.4}$$

czyli jest transformatą Fouriera gęstości prawdopodobieństwa.
Zauważmy, że *n*-ta pochodna funkcji charakterystycznej w punkcie $t = 0$ wynosi

$$\frac{d^n\varphi(t)}{dt^n} \equiv \varphi^{(n)}(t) = i^n \int_{-\infty}^{\infty} x^n e^{itx} f(x)dx. \tag{5.5}$$

Jak widać jest to, z dokładnością do czynnika i^n równe *n*-temu momentowi rozkładu zmiennej względem zera:

$$\varphi^{(n)}(0) = i^n \lambda_n. \tag{5.6}$$

Dla zmiennej losowej $Y = X - \hat{X}$, funkcja charakterystyczna przyjmuje postać

$$\varphi_Y(t) = \int_{-\infty}^{\infty} e^{it(x-\hat{x})} f(x)dx = e^{-it\hat{x}}\varphi(t). \tag{5.7}$$

Jej *n*-ta pochodna w punkcie $t = 0$ wynosi

$$\varphi_Y^{(n)}(0) = i^n \int_{-\infty}^{\infty} (x - \hat{x})^n f(x)dx. \tag{5.8}$$

Wynika stąd, że *n*-ta pochodna funkcji φ_Y w zerze jest, z dokładnością do czynnika i^n, równa *n*-temu momentowi centralnemu

$$\varphi_Y(0) = i^n E((x - \hat{x})^n) = i^n \mu_n. \tag{5.9}$$

Dla przykładu wariancja rozkładu wynosi

$$\sigma^2(X) = -\varphi''(0). \tag{5.10}$$

Oprócz wyznaczania momentów rozkładów, funkcja charakterystyczna może posłużyć do wyznaczania gęstości prawdopodobieństwa. Aby otrzymać gęstość prawdopodobieństwa za pomocą funkcji charakterystycznej musimy znaleźć odwrotną transformatę Fouriera (5.4), czyli

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} \varphi(t)dt. \tag{5.11}$$

Dystrybuantę rozkładu możemy znaleźć następująco

$$F(b) - F(a) = \int_a^b f(x)dx$$
$$= \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_a^b e^{-itx} \varphi(t) dt dx = \frac{i}{2\pi} \int_{-\infty}^{\infty} \frac{e^{itb} - e^{ita}}{t} \varphi(t) dt. \tag{5.12}$$

Można udowodnić, że dystrybuanta rozkładu jest wyznaczona jednoznacznie przez jego funkcję charakterystyczną.

Zadanie 5.1

Znajdź funkcję charakterystyczną rozkładu Poissona.

Wskazówka:

W przypadku rozkładów dyskretnych jakim jest rozkład Poissona funkcję charakterystyczną wyznaczamy ze wzoru $\varphi(t) = \sum_i e^{itx_i} P(X = x_i)$.

6. Rozkłady wielu zmiennych losowych

6.1. Dystrybuanta i gęstość prawdopodobieństwa dwu zmiennych

Przez analogię do definicji dystrybuanty rozkładu jednej zmiennej, możemy zdefiniować dystrybuantę rozkładu dwu zmiennych X i Y

$$F(x, y) \equiv P(X < x, Y < y). \tag{6.1}$$

Przykład takiej dystrybuanty pokazano na rysunku 6.1.
Dla ciągłych funkcji $F(x, y)$ możemy zdefiniować *gęstość prawdopodobieństwa* dwu zmiennych losowych

$$f(x, y) = \frac{\partial}{\partial x} \frac{\partial}{\partial y} F(x, y). \tag{6.2}$$

Zgodnie z definicją (6.2) mamy

$$P(a \le x < b, c \le y < d) = \int_a^b \left[\int_c^d f(x, y) dy \right] dx. \tag{6.3}$$

W szczególnym przypadku, gdy interesuje nas prawdopodobieństwo wystąpienia wartości zmiennej $a \le x < b$ dla dowolnej wartości $-\infty < y < \infty$ dostaniemy

$$P(a \le x < b, -\infty < y < \infty) = \int_a^b \left[\int_{-\infty}^{\infty} f(x, y) dy \right] dx = \int_a^b g(x) dx, \tag{6.4}$$

gdzie

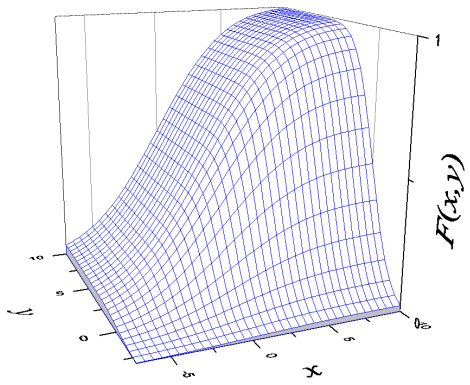
$$g(x) = \int_{-\infty}^{\infty} f(x,y)dy \tag{6.5}$$

jest tzw. *brzegową gęstością prawdopodobieństwa* zmiennej X. Analogicznie brzegową gęstość prawdopodobieństwa dla zmiennej Y definiujemy jako

$$h(y) = \int_{-\infty}^{\infty} f(x,y)dx. \tag{6.6}$$

Przez analogię do wzoru (2.11), stanowiącego warunek konieczny i wystarczający niezależności zdarzeń, możemy zdefiniować niezależność zmiennych losowych X i Y. Zmienne losowe X i Y są niezależne, gdy

$$f(x,y) = g(x)h(y). \tag{6.7}$$



Rysunek 6.1. Przykładowa dystrybuanta dwu zmiennych.

6.2. Wartości oczekiwane, wariancje, kowariancje dla dwu zmiennych losowych

Wartość oczekiwaną funkcji dwu zmiennych $H(x,y)$ definiujemy następująco

$$E\big(H(x,y)\big) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} H(x,y)f(x,y)dxdy. \tag{6.8}$$

Podobnie jak dla rozkładu jednej zmiennej, możemy zdefiniować momenty rozkładu dwu zmiennych. *Momentami rzędu l, m* względem zmiennych X, Y nazywamy wartości oczekiwane funkcji o postaci $H(x,y) = x^l y^m$ (l, m liczby całkowite nieujemne)

$$\lambda_{lm} = E(x^l y^m). \tag{6.9}$$

Zauważmy, że

$$\lambda_{10} = \hat{x}$$
$$\lambda_{01} = \hat{y}. \tag{6.10}$$

Momenty względem punktów a i b są wartościami oczekiwanymi funkcji o postaci $H(x,y) = (x-a)^l(y-b)^m$

$$\alpha_{lm} = E((x-a)^l(y-b)^m). \tag{6.11}$$

Analogicznie jak w przypadku rozkładów jednej zmiennej szczególne znaczenie mają *momenty centralne*, czyli momenty względem punktów $\hat{x}$ i $\hat{y}$

$$\mu_{lm} = E((x-\hat{x})^l(y-\hat{y})^m). \tag{6.12}$$

Najniższe momenty centralne wynoszą

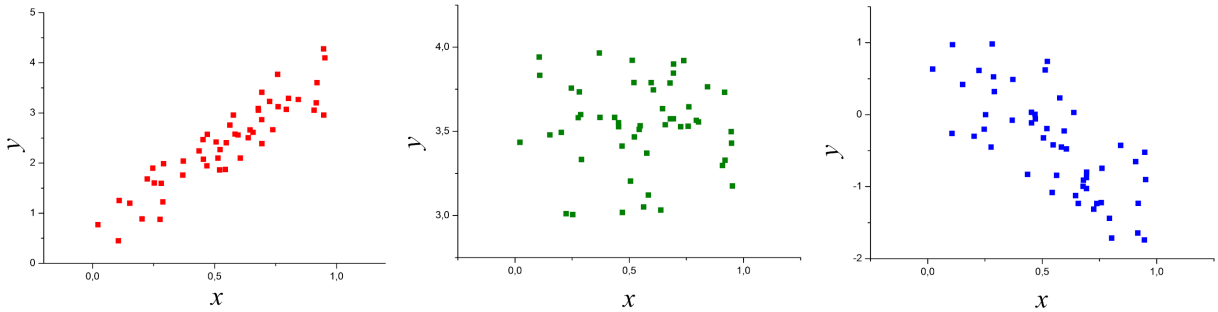
$$\mu_{00} = 1,$$
$$\mu_{10} = \mu_{01} = 0,$$
$$\mu_{11} = E\big((x-\hat{x})(y-\hat{y})\big) = \mathrm{cov}(X,Y),$$
$$\mu_{20} = \mathrm{E}((x-\hat{x})^2) = \sigma^2(X), \tag{6.13}$$
$$\mu_{02} = E((y-\hat{y})^2) = \sigma^2(Y).$$

Momenty λ_{10} i λ_{01} oraz μ_{20} i μ_{02} są podobne do wielkości zdefiniowanych dla rozkładu jednej zmiennej: wartości oczekiwanych $E(X)$ i $E(Y)$ oraz wariancji $\sigma^2(X)$ i $\sigma^2(Y)$. Natomiast moment μ_{11}, nazywany *kowariancją* $\mathrm{cov}(X,Y)$ nie ma odpowiednika wśród parametrów opisujących rozkład jednej zmiennej. Kowariancja jest miarą współzależności między zmiennymi. Dla zmiennych niezależnych kowariancja jest zerowa.

$$\mathrm{cov}(X,Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-\hat{x})(y-\hat{y})g(x)h(y)dxdy$$
$$\left(\int_{-\infty}^{\infty}(x-\hat{x})dx\right)\left(\int_{-\infty}^{\infty}(y-\hat{y})dy\right) = 0. \tag{6.14}$$

Na Rysunek 6.2 pokazano trzy przykładowe zbiory punktów, których współrzędne są wartościami pewnych zmiennych losowych X i Y. Kowariancja zmiennych w przykładzie (a) jest dodatnia, w przykładzie (b) zerowa, a w przykładzie (c) ujemna. Przykłady (a) i (c) sugerują pewną współzależność między zmiennymi, zaś w przykładzie (b) takiej zależności nie widać. Zamiast kowariancją wygodniej jest posługiwać się tzw. *współczynnikiem korelacji*

$$\rho(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sigma(X)\sigma(Y)} \tag{6.15}$$



(a) (b) (c)

Rysunek 6.2. Przykład trzech zbiorów danych o różnych kowariancjach. (a) $\mathrm{cov}(X,Y) \approx 0{,}19$, (b) $\mathrm{cov}(X,Y) \approx 0$, (c) $\mathrm{cov}(X,Y) \approx -0{,}12$.

Można pokazać, że

$$-1 \leq \rho(X,Y) \leq 1. \tag{6.16}$$

Ponadto skrajne wartości $\rho = \pm 1$ otrzymujemy w przypadku liniowej zależności między zmiennymi X i Y

$$Y = aX + b, \quad (a, b \text{ są liczbami rzeczywistymi}), \tag{6.17}$$

przy czym dla $a > 0$, $\rho = 1$, a dla $a < 0$, $\rho = -1$. Dla dwóch zmiennych losowych niezależnych kowariancja, a tym samym również współczynnik korelacji są zerowe.
Jak pokazaliśmy wcześniej kowariancja zmiennych niezależnych jest równa zeru. Jednak zerowanie się kowariancji niekoniecznie musi oznaczać niezależność zmiennych. W niektórych przypadkach pomimo zależności zmiennych kowariancja może być zerowa (zadanie 6.2)
Wyznaczmy wariancję rozkładu zmiennej $u = aX + bY$

$$\begin{gathered}\sigma^2(aX+bY) = E\left(\left((aX+bY) - E(aX+bY)\right)^2\right) = \\ E\left(\left((aX+bY) - a\hat{x} - b\hat{y}\right)^2\right) = E\left(\left(a(X-\hat{x}) + b(Y-\hat{y})\right)^2\right) = \\ E\left(a^2(X-\hat{x})^2 + b^2(Y-\hat{y})^2 + 2ab(X-\hat{x})(Y-\hat{y})\right).\end{gathered} \tag{6.18}$$

Ostatecznie dostajemy

$$\sigma^2(aX+bY) = a^2\sigma^2(X) + b^2\sigma^2(Y) + 2ab\,\mathrm{cov}(X,Y). \tag{6.19}$$

6.3. Opis wielu zmiennych losowych

Wprowadzony w poprzednich paragrafach opis dwu zmiennych losowych można przez analogie rozszerzyć na dowolną liczbę n zmiennych losowych $X_1, X_2, \ldots, X_n$.
Dystrybuantę dla n zmiennych losowych definiujemy jako

$$F(x_1, x_2, \ldots, x_n) = P(X_1 < x_1, X_2 < x_2, \ldots, X_n < x_n). \tag{6.20}$$

Jeżeli funkcja F jest ciągła i ma pochodne cząstkowe względem wszystkich zmiennych, to gęstość prawdopodobieństwa definiujemy następująco

$$f(x_1, x_2, \dots, x_n) = \frac{\partial^n}{\partial x_1 \partial x_2 \dots \partial x_n} F(x_1, x_2, \dots, x_n). \tag{6.21}$$

Gęstość rozkładu brzegowego dla *r*-tej zmiennej otrzymujemy całkując gęstość łączną po wszystkich zmiennych poza zmienną x_r

$$g_r(x_r) = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f(x_1, x_2, \dots, x_n) dx_1 \dots dx_{r-1} dx_{r+1} \dots dx_n. \tag{6.22}$$

Warunkiem koniecznym i wystarczającym niezależności zmiennych jest, aby łączna gęstość prawdopodobieństwa była iloczynem gęstości granicznych

$$f(x_1, x_2, \dots, x_n) = g_1(x_1) g_2(x_2) \dots g_n(x_n). \tag{6.23}$$

Możemy również zdefiniować łączną gęstość rozkładu brzegowego dla dowolnej liczby $l <= n$ zmiennych poprzez wycałkowanie gęstości łącznej po wszystkich pozostałych $n - l$ zmiennych. Jeśli otrzymana w ten sposób gęstość prawdopodobieństwa jest iloczynem gęstości brzegowych tych *l* zmiennych, to oznacza to, że są one niezależne.
Wartości oczekiwane funkcji

$$H(x_1, x_2, \dots, x_n) = x_1^{l_1} x_2^{l_2} \dots x_n^{l_n} \tag{6.24}$$

nazywamy momentami rzędu $l_1, l_2, \dots, l_n$ i oznaczamy przez

$$\lambda_{l_1, l_2, \dots, l_n} = E\left(x_1^{l_1} x_2^{l_2} \dots x_n^{l_n}\right). \tag{6.25}$$

Podobnie jak w przypadku dwu zmiennych

$$\begin{aligned} \lambda_{100\dots0} &= E(X_1) = \hat{x}_1, \\ \lambda_{010\dots0} &= E(X_2) = \hat{x}_2, \\ &\vdots \\ \lambda_{000\dots1} &= E(X_n) = \hat{x}_n. \end{aligned} \tag{6.26}$$

Momenty względem wartości średnich, czyli momenty centralne definiujemy jako

$$\mu_{l_1, l_2, \dots, l_n} = E((X_1 - \hat{x}_1)^{l_1} (X_2 - \hat{x}_2)^{l_2} \dots (X_n - \hat{x}_n)^{l_n}). \tag{6.27}$$

Moment centralny, dla którego wszystkie $l_i = 0$ poza jednym równym $l_r = 2$ jest wariancją zmiennej X_r

$$\begin{aligned} \mu_{200\dots0} &= E((X_1 - \hat{x}_1)^2) = \sigma^2(X_1), \\ \mu_{020\dots0} &= E((X_2 - \hat{x}_2)^2) = \sigma^2(X_2), \\ &\vdots \\ \mu_{000\dots2} &= E((X_n - \hat{x}_n)^2) = \sigma^2(X_n). \end{aligned} \tag{6.28}$$

Z kolei moment centralny, dla którego wszystkie $l_i = 0$ poza dwoma równymi $l_i = l_2 = 1$ jest kowariancją pomiędzy zmiennymi losowymi X_i, X_j

$$c_{ij} = \mathrm{cov}(X_i, X_j) = E\left((X_i - \hat{x}_i)(X_j - \hat{x}_j)\right). \tag{6.29}$$

W przypadku rozkładu wielu zmiennych wygodnie jest traktować poszczególne zmienne losowe jak współrzędne wektora w przestrzeni *n*-wymiarowej

$$\boldsymbol{X} = (X_1, X_2, \ldots, X_n). \tag{6.30}$$

Wówczas np. dystrybuantę możemy zapisać w notacji wektorowej

$$F(\boldsymbol{x}) \equiv F(x_1, x_2, \ldots, x_n). \tag{6.31}$$

Wariancje i kowariancje możemy zapisać w notacji macierzowej w postaci macierzy

$$C = \begin{pmatrix} c_{11} & c_{12} & \ldots & c_{1n} \\ c_{21} & c_{22} & & c_{2n} \\ \vdots & & & \\ c_{n1} & c_{n2} & \ldots & c_{nn} \end{pmatrix}. \tag{6.32}$$

Elementy macierzy kowariancji są zdefiniowane wzorem (6.29). Elementy diagonalne macierzy kowariancji są wariancjami $c_{ii} = \sigma^2(X_i)$, a pozadiagonalne kowariancjami. W notacji macierzowej każdy element macierzy kowariancji możemy zapisać w postaci

$$c_{ij} = E\left((X_i - \hat{x}_i)(X_j - \hat{x}_j)^T\right), \tag{6.33}$$

gdzie symbol T oznacza operację transponowania macierzy, czyli w przypadku wektora $\boldsymbol{x}$ mamy

$$\boldsymbol{x}^T = (x_1, x_2, \ldots, x_n), \qquad \boldsymbol{x} = \begin{pmatrix} x \\ x_2 \\ \vdots \\ x_n \end{pmatrix}. \tag{6.34}$$

Macierz kowariancji jest symetryczna tzn. $c_{ij} = c_{ji}$.
W zapisie macierzowym macierz kowariancji przyjmuje postać

$$C = E\left((\boldsymbol{X}_i - \widehat{\boldsymbol{x}}_i)(\boldsymbol{X}_j - \widehat{\boldsymbol{x}}_j)^T\right). \tag{6.35}$$

W przypadku tzw. pomiarów pośrednich, w których wartość mierzona jest wyliczana na podstawie związku funkcyjnego mierzonej pośrednio wielkości $\boldsymbol{Y}$, a wielkościami mierzonymi bezpośrednio $\boldsymbol{X}$ $\left(\boldsymbol{Y} = \boldsymbol{y}(\boldsymbol{X})\right)$, interesuje nas wartość oczekiwana zmiennej $\boldsymbol{Y}$ oraz jej macierz kowariancji. Często w takich wypadkach znamy wartość oczekiwaną $\widehat{\boldsymbol{X}}$ oraz jej macierz kowariancji $C(X)$. W celu oszacowania wartości oczekiwanej $\widehat{\boldsymbol{Y}}$ rozwijamy w szereg Taylora funkcję $\boldsymbol{y}$ wokół wektora $\widehat{\boldsymbol{X}}$ obcinając go na wyrazach liniowych

$$y_i \approx y_i(\widehat{\boldsymbol{X}}) + \sum_j \left(\frac{\partial y_i}{\partial x_j}\right)_{\boldsymbol{x}=\widehat{\boldsymbol{X}}} (x_j - \hat{X}_j). \tag{6.36}$$

Ponieważ wartość oczekiwana wyrażenia $(x_j - \hat{X}_j)$ znika tożsamościowo, to ostatecznie

$$\widehat{\boldsymbol{Y}} \approx \boldsymbol{y}(\hat{\boldsymbol{x}}) = \boldsymbol{y}(\hat{x}_1, \hat{x}_2, \dots, \hat{x}_n). \tag{6.37}$$

Elementy macierzy kowariancji zmiennej $\boldsymbol{Y}$ są równe

$$\mathrm{cov}(y_k, y_l) = E\big((y_k - \hat{y})(y_l - \hat{y})\big), \tag{6.38}$$

Korzystając z rozwinięcia (6.37) dostajemy

$$\mathrm{cov}(y_k, y_l) \approx \sum_{i,j} \left(\frac{\partial y_k}{\partial x_i}\right)_{\boldsymbol{x}=\widehat{\boldsymbol{X}}} \left(\frac{\partial y_l}{\partial x_j}\right)_{\boldsymbol{x}=\widehat{\boldsymbol{X}}} \mathrm{cov}\big(x_i, x_j\big). \tag{6.39}$$

W szczególnym przypadku, gdy zmienne $\boldsymbol{X}$ są niezależne niezerowe są jedynie elementy diagonalne macierzy kowariancji, czyli wariancje

$$\sigma^2(y_i) = \sum_j \left(\frac{\partial y_i}{\partial x_j}\right)_{\boldsymbol{x}=\widehat{\boldsymbol{X}}} \sigma^2\left(x_j\right). \tag{6.40}$$

Wyrażenia (6.39) i (6.40) nazywamy propagacją błędów, gdyż pozwalają one wyliczyć jak błędy pomiarów wielkości $\boldsymbol{X}$ przenoszą się na błędy pomiaru pośredniego wielkości $\boldsymbol{Y}$.

Zadania 6.1

Przedstaw graficznie dane zebrane w poniższej tabeli. Jakiego współczynnika korelacji między zmiennymi x, y spodziewasz się? Wylicz ten współczynnik.

l.p.	*x*	*y*
1	0,0	3,14
2	0,5	4,42
3	1,0	6,21
4	1,5	7,64
5	2,0	8,01
6	2,5	10,18
7	3,0	12,81
8	3,5	13,51
9	4,0	15,46
10	4,5	16,22

Zadanie 6.2

Przedstaw graficznie dane zebrane w poniższej tabeli. Jakiego współczynnika korelacji między zmiennymi x, y spodziewasz się? Wylicz ten współczynnik. Jak uzasadnisz otrzymany wynik?

l.p.	x	y
1	0,0	9,08
2	0,5	7,43
3	1,0	6,08
4	1,5	5,29
5	2,0	5,12
6	2,5	5,27
7	3,0	6,13
8	3,5	7,33
9	4,0	9,05

7. Sploty rozkładów

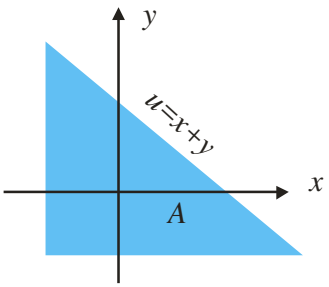
Bardzo często mierzona wielkość jest sumą dwu lub większej liczby wielkości. Każda z nich jest opisywana własnym rozkładem gęstości prawdopodobieństwa. Rozkład gęstości mierzonej wielkości będzie wówczas *splotem* obu rozkładów. Załóżmy, że wielkościami, których sumą jest wielkość U są wielkości X, Y $(U = X + Y)$. Zakładamy, że są to zmienne niezależne. w związku z tym łączną gęstość prawdopodobieństwa możemy zapisać w postaci iloczynu gęstości obu zmiennych

$$f(x, y) = f_x(x) f_y(y). \tag{7.1}$$

Dystrybuanta zmiennej U

$$F(u) = P(U < u) = P(X + Y < u) \tag{7.2}$$

jest całką z gęstości (7.1) po powierzchni określonej nierównością $x + y < u$ pokazanej na Rysunku 7.1.



Rysunek 7.1. Graficzne przedstawienie obszaru całkowania w całce

$$F(u) = \iint_A f(x,y)dxdy = \int_{-\infty}^{\infty} f_x(x)dx \int_{-\infty}^{u-x} f_y(y)dy = \int_{-\infty}^{\infty} f_y(y)dy \int_{-\infty}^{u-y} f_x(x)dx, \tag{7.3}$$

czyli

$$F(u) = \int_{-\infty}^{\infty} f_x(x)F_y(u-x)dx = \int_{-\infty}^{\infty} f_y(y)F_x(u-y)dy. \tag{7.4}$$

Różniczkując dystrybuantę F po zmiennej u dostajemy szukaną gęstość prawdopodobieństwa zmiennej u

$$f(u) = \int_{-\infty}^{\infty} f_x(x)f_y(u-x)dx = \int_{-\infty}^{\infty} f_y(y)f_x(u-y)dy. \tag{7.5}$$

W przypadku, gdy zmienne losowe X, Y mogą przyjmować wartości tylko w pewnym zwartym obszarze granice całkowania w wzorze (7.5) mogą mieć skończone wartości.

Przykład

Jako przykład rozważmy splot dwu jednakowych rozkładów równomiernych:

$$f_x(x) = \begin{cases} 1, & 0 \le x < 1 \\ 0, & \text{w przeciwnym wypadku} \end{cases} \quad \text{i} \quad f_y(y) = \begin{cases} 1, & 0 \le y < 1 \\ 0, & \text{w przeciwnym wypadku} \end{cases}.$$

Korzystając z wzoru (7.5) i postaci funkcji f_x mamy

$$f(u) = \int_0^1 f_y(u-x)dx. \tag{7.6}$$

Dokonując zamiany zmiennych $v = u - x,\ dv = -dx$ dostajemy

$$f(u) = -\int_u^{u-1} f_y(v)dv = \int_{u-1}^{u} f_y(v)dv. \tag{7.7}$$

Ponieważ $u = x + y$, to $0 \le u < 2$. W przedziale $0 \le u < 1$, zmienna v przyjmuje wartości w zakresie $-1 \le v < 1$, więc górną granicą całkowania w (7.7) może być dowolne u (z przedziału $0 \le u < 1$), a dolną granicą 0

$$f_1(u) = \int_0^u f_y(v)dv = \int_0^u dv = u. \tag{7.8}$$

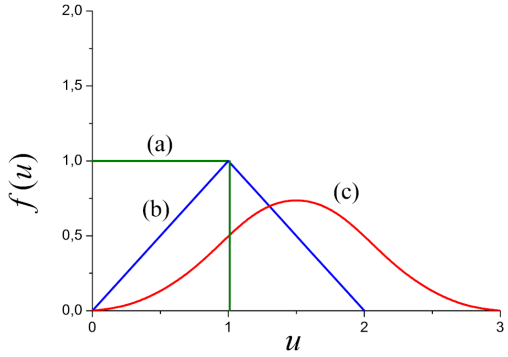
W przedziale $1 \le u < 2$, dolną granicą całkowania może być dowolne $u - 1$, zaś górną granicą jest 1

$$f_2(u) = \int_{u-1}^{1} f_y(v)dv = \int_{u-1}^{1} dv = 2 - u. \tag{7.9}$$

A zatem gęstość prawdopodobieństwa zmiennej u ma kształt trójkątny jak pokazano na Rysunku 7.2. W przypadku, gdy zmienna U będzie sumą trzech zmiennych losowych o rozkładach równomiernych takich, jakie rozważaliśmy powyżej, musimy dokonać splotu otrzymanego rozkładu trójkątnego z rozkładem równomiernym. Rozwiązaniem tego problemu jest rozkład gęstości prawdopodobieństwa złożony z trzech fragmentów parabol

$$f(u) = \begin{cases} \frac{1}{2}u^2, & 0 \le u < 1 \\ \frac{1}{2}(-2u^2 + 6u - 3), & 1 \le u < 2 \\ \frac{1}{2}(u-3)^2, & 2 \le u < 3. \end{cases} \tag{7.10}$$

Wszystkie trzy rozkłady, czyli równomierny na odcinku $[0,1]$, trójkątny otrzymany jako splot dwóch rozkładów równomiernych oraz paraboliczny (7.8) pokazano na rysunku 7.2. Warto zwrócić uwagę na to, że gęstość prawdopodobieństwa zmiennej będącej sumą trzech zmiennych o rozkładach równomiernych zaczyna przypominać swoim kształtem krzywą Gaussa. Podobieństwo byłoby coraz większe w miarę sumowania coraz większej liczby zmiennych. Jest to zgodne z centralnym twierdzeniem granicznym.



Rysunek 7.2. Gęstość prawdopodobieństwa utworzonych jako splot rozkładów jednostajnych. (a) $U = X$, (b) $U = X + X$, (c) $U = X + X + X$, gdzie X jest zmienną losową o rozkładzie równomiernym na odcinku $[0,1)$.

Zadanie 7.1

Gęstość prawdopodobieństwa dla dwu zmiennych losowych X i Y jest dana wzorem

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y} e^{-\frac{1}{2}\frac{x^2}{\sigma_x^2} - \frac{1}{2}\frac{y^2}{\sigma_y^2}}.$$

Jest to dwuwymiarowy rozkład Gaussa z $\hat{x} = 0$ i $\hat{y} = 0$.

a) Znajdź brzegowe gęstości prawdopodobieństwa $f(x)$ i $f(y)$.
b) Pokaż, że zmienne X, Y są niezależne

Wskazówka:

a) Skorzystaj ze wskazówki do zadania 4.5.
b) Sprawdź warunek (6.7).

Zadanie 7.2

Dwie zmienne losowe zdefiniowane są jako $X = \sin Z,\ Y = \cos Z$, gdzie zmienna Z jest zmienną losową o rozkładzie jednostajnym na przedziale $[0, \pi]$. Jak widać zmienne X, Y są zależne, gdyż $x^2 + y^2 = \sin^2 z + \cos^2 z = 1$. Pokaż, że pomimo tego, kowariancja jest równa zeru.

8. Elementy teorii estymacji

Na ogół nie znamy rozkładu prawdopodobieństwa, a nawet jeśli go znamy, to zwykle nie znamy bezpośrednio parametrów, od których on zależy. Zmuszeni jesteśmy do przybliżania rozkładu poprzez *rozkład częstości*, który znajdujemy doświadczalnie na podstawie skończonej liczby pomiarów, czyli tzw. *próby losowej*. Próba reprezentuje tzw. *populację generalną*, czyli zbiór wszystkich możliwych wyników obserwacji. Wyciąganie wniosków na temat rozkładu populacji generalnej na podstawie próby losowej nazywamy *estymacją*. Estymacja może mieć charakter parametryczny, gdy służy do znajdowania parametrów rozkładu lub nieparametryczny, gdy stara się znajdować postać rozkładu populacji. Estymację parametryczną możemy podzielić na *estymację punktową* i *estymację przedziałową*. Estymacja punktowa polega na znalezieniu konkretnej wartości parametrów rozkładu na podstawie pobranej próby. W estymacji przedziałowej znajdujemy przedział, w którym z określonym, przyjętym prawdopodobieństwem, tzw. *poziomem ufności* mieści się wartość szukanego parametru.

Wykonując pomiar (obserwację) powinniśmy dążyć do tego, aby elementy próby były od siebie niezależne. Nie zawsze jest to łatwe do osiągnięcia i nie sposób podać jednej ogólnej recepty na zapewnienie pełnej losowości próby. Stąd wynikają błędy zwłaszcza w badaniach socjologicznych, np. w ocenie popularności partii politycznych. Każda próba jest zmienną losową. Próbę n-elementową możemy potraktować jak n-elementowy wektor w n-wymiarowej przestrzeni prób, którego składnikami są kolejne elementy próby. Dowolna funkcja tego wektora (lub inaczej dowolna funkcja, której argumentami są elementy próby) jest też zmienną losową. Nazywamy ją *statystyką*. Bardzo często statystyka służy do szacowania parametrów rozkładu. Taką statystykę nazywamy *estymatorem*. Bardzo ważnym estymatorem jest *wartość średnia z próby* definiowana jako średnia arytmetyczna elementów próby

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i. \tag{8.1}$$

Średnia arytmetyczna z próby jest estymatorem wartości oczekiwanej rozkładu. Dobre estymatory powinny posiadać pewne cechy. Estymator powinien być *nieobciążony*. Oznacza to, że niezależnie od liczebności próby, wartość oczekiwana estymatora jest równa wartości estymowanego parametru. Intuicyjnie można to rozumieć tak: niezależnie od tego jak liczne próby pobieramy, to średnia ze średnich w poszczególnych próbach dąży do prawdziwej wartości estymowanego parametru wraz ze wzrostem liczby prób. Jeśli estymator parametru a oznaczymy przez $T_n(a)$, to dla nieobciążonego estymatora mamy

$$E\big(T_n(a)\big) = \hat{T}_n(a) = a, \quad \text{niezależnie od } n. \tag{8.2}$$

Czasami estymatory są *asymptotycznie nieobciążone,* co oznacza że obciążenie estymatora (różnica między wartością oczekiwaną estymatora, a wartością prawdziwą estymowanego parametru) znika wraz ze wzrostem liczebności próby.
Poza tym estymator powinien być *zgodny*. Intuicyjnie, przez zgodność estymatora należy rozumieć to, że dąży on do prawdziwej wartości estymowanego parametru wraz ze wzrostem liczebności próby. Jeden z warunków, który to może zapewnić jest dążenie do zera wariancji estymatora wraz ze wzrostem liczebności próby

$$\lim_{n\to\infty} \sigma^2(T_n(a)) = 0. \tag{8.3}$$

Metody znajdowania estymatorów mogą prowadzić do różnych postaci estymatorów danego parametru. Najbardziej pożądanymi są estymatory o małej wariancji. Takie estymatory nazywamy *estymatorami efektywnymi.*
Policzmy teraz wartość oczekiwaną średniej arytmetycznej próby

$$E(\bar{X}) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\sum_{i=1}^{n} \hat{X} = \hat{X}. \tag{8.4}$$

Jak widzimy wartość oczekiwana średniej arytmetycznej ze wszystkich elementów próby jest równa wartość oczekiwanej zmiennej X, a ponieważ zachodzi to dla dowolnej wartości n, wartość średniej arytmetycznej z próby jest estymatorem nieobciążonym dla wartości oczekiwanej zmiennej X. Obliczmy teraz wariancję tego estymatora

$$\sigma^2(\bar{X}) = E\left(\left(\bar{X} - \hat{X}\right)^2\right) = E\left(\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \hat{X}\right)^2\right) = \frac{1}{n^2} E\left(\left(\sum_{i=1}^{n} (X_i - \hat{X})\right)^2\right). \tag{8.5}$$

Z niezależności elementów próby wynika, że w powyższym wyrażeniu wszystkie kowariancje, czyli wyrazy typu $E\left(\left(X_i - \hat{X}\right)\left(X_j - \hat{X}\right)\right)$ z $i \neq j$ są zerowe. Wobec tego

$$\sigma^2(\bar{X}) = \frac{1}{n^2}\sum_{i=1}^{n} E\left(\left(X_i - \hat{X}\right)^2\right) = \frac{1}{n^2} n\sigma^2(X) = \frac{1}{n}\sigma^2(X). \tag{8.6}$$

Z otrzymanego wyrażenia wynika, że wariancja średniej arytmetycznej z próby dąży do zera, gdy liczebność próby rośnie do nieskończoności, czyli spełniony jest warunek (8.3). A zatem średnia arytmetyczna z próby jest estymatorem zgodnym wartości oczekiwanej (wcześniej pokazaliśmy również, że jest estymatorem nieobciążonym).
Znamy już estymator wartości oczekiwanej. Poszukajmy teraz estymatora wariancji z próby. W pierwszym przybliżeniu zdefiniujmy ten estymator jako średnią arytmetyczną odchyleń kwadratowych od wartości średniej z próby

$$s'^2 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \bar{X})^2. \tag{8.7}$$

Policzmy wartość oczekiwaną tego estymatora.

$$\frac{1}{n}E\left(\sum_{i=1}^{n}(X_i-\bar{X})^2\right)=\frac{1}{n}E\left(\sum_{i=1}^{n}\left((X_i-\hat{X})+(\hat{X}-\bar{X})\right)^2\right)=$$
$$\frac{1}{n}E\left(\sum_{i=1}^{n}(X_i-\hat{X})^2+\sum_{i=1}^{n}(\hat{X}-\bar{X})^2+2\sum_{i=1}^{n}(X_i-\hat{X})(\hat{X}-\bar{X})\right)= \tag{8.8}$$
$$\frac{1}{n}\sum_{i=1}^{n}\left(E\left((X_i-\hat{X})^2\right)-E\left((\bar{X}-\hat{X})^2\right)\right)=\frac{1}{n}\left(n\sigma^2(X)-n\frac{1}{n}\sigma^2(X)\right).$$

Czyli ostatecznie

$$E(s'^2)=\frac{n-1}{n}\sigma^2(X). \tag{8.9}$$

Znaleziony powyżej estymator wariancji z próby jest estymatorem obciążonym. Jednak jeśli we wzorze (8.7) zastąpimy czynnik $1/n$ czynnikiem $1/(n-1)$, to jak widać dostaniemy nieobciążony estymator wariancji z próby

$$s^2(X)=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2 \tag{8.10}$$

Występowanie czynnika $n-1$ zamiast n w powyższym wyrażeniu jest konsekwencją użycia we wzorach średniej arytmetycznej $\bar{X}$ w miejsce wartości oczekiwanej $\hat{X}$. Jakościowo możemy to wyjaśnić zmniejszeniem o 1 liczby niezależnych elementów, gdyż średnia arytmetyczna jest wyliczana na podstawie elementów próby, więc do jej wyliczenia wykorzystujemy część informacji zawartych w próbie.
Wykorzystując estymator wariancji z próby (8.10) oraz wzór (8.6) znajdujemy bardzo ważny estymator wariancji wartości średniej

$$s^2(\bar{X})=\frac{1}{n(n-1)}\sum_{i=1}^{n}(X_i-\bar{X})^2. \tag{8.11}$$

8.1. Metoda momentów

Jedną z metod szukania estymatorów jest metoda momentów zaproponowana przez Pearsona na przełomie XIX i XX wieku. Załóżmy, że poszukujemy estymatorów parametrów $\theta_1,\theta_2,\dots,\theta_k$, określających rozkład zmiennej losowej X. Metoda momentów podaje następujący schemat postępowania

a) Znajdujemy związki pomiędzy parametrami rozkładu a jego momentami.
b) do znalezionych związków wstawiamy estymatory momentów zdefiniowane następująco

$$M_k\equiv T_n\left(m_k(0)\right)=\frac{1}{n}\sum_{i=1}^{n}(x_i)^k.$$

c) Rozwiązujemy otrzymany w ten sposób układ równań wyrażając w ten sposób parametry $\theta_1,\theta_2,\dots,\theta_k$ przez estymatory momentów M_k

Przykład

Rozkład Gaussa ma postać

$$f(x) = \frac{1}{\sqrt{2\pi\theta_2^2}} e^{-\frac{(x-\theta_1)^2}{2\theta_2^2}}. \tag{8.12}$$

Naszym zadaniem jest znaleźć estymatory parametrów θ_1, θ_2 metodą momentów. Wiemy, że

$$\begin{aligned} \theta_1 &= \hat{x} \equiv m_1(0) \\ \theta_2^2 &= \sigma^2(X) = E(X^2) - \big(E(X)\big)^2 \equiv m_2(0) - \big(m_1(0)\big)^2. \end{aligned} \tag{8.13}$$

Estymatory momentów $m_1(0)$ i $m_2(0)$ wynoszą

$$\begin{aligned} T_n\big(m_1(0)\big) &\equiv M_1 = \frac{1}{n}\sum_{i=1}^{n} x_i, \\ T_n\big(m_2(0)\big) &\equiv M_2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2. \end{aligned} \tag{8.14}$$

Wstawiając estymator momentu M_1 do pierwszego z równań (8.13) dostajemy estymator parametru θ_1

$$T_n(\theta_1) = \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{8.15}$$

Jest to zgodne z otrzymanym wcześniej wynikiem, że estymatorem wartości oczekiwanej jest średnia arytmetyczna z próby. Po wstawieniu momentów M_1, M_2 do drugiego z równań (8.13) mamy

$$\begin{aligned} T_n(\theta_2^2) &= \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - 2\bar{X}^2 + \bar{X}^2 = \\ &\frac{1}{n}\sum_{i=1}^{n} X_i^2 - 2\bar{X}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) + \left(\frac{1}{n}\sum_{i=1}^{n} \bar{X}^2\right) = \\ &\frac{1}{n}\sum_{i=1}^{n} (X_i^2 - 2\bar{X}X_i + \bar{X}^2) = \frac{1}{n}\sum_{i=1}^{n} (X_i - \bar{X})^2. \end{aligned} \tag{8.16}$$

Otrzymany powyżej estymator parametru θ_2^2 (wariancji) rozkładu Gaussa jest *asymptotycznie nieobciążony* w przeciwieństwie do estymatora (8.10), który jest estymatorem *nieobciążonym*. Można pokazać, że estymatory znalezione metodą momentów są zgodne i asymptotycznie nieobciążone lub nieobciążone. Jednak metoda ta ma szereg wad. Estymatory znajdowane tą metodą mają na ogół większą wariancję niż estymatory znajdowane innymi metodami, czyli są mniej efektywne. Estymatory wyższych momentów wyznaczane metodą momentów są mało dokładne, co oczywiście wpływa na dokładność estymatorów parametrów. Ponadto układ równań na estymatory parametrów jest zwykle układem nieliniowym co powoduje konieczność stosowania metod numerycznych znajdowania rozwiązań.

8.2. Metoda największej wiarygodności

Metoda największej wiarygodności powstała w 1921 roku. Jej twórcą jest R.A. Fischer. Metoda zakłada, że wyniki obserwacji są najbardziej prawdopodobne z wszystkich możliwych. W związku z tym założeniem szukamy prawdopodobieństwa tego, że próba zawiera takie wartości jakie uzyskaliśmy pod warunkiem, że parametry rozkładu są równe $\boldsymbol{\theta}$**.** W przypadku, gdy poszczególne pomiary są niezależne, prawdopodobieństwo to będzie równe iloczynowi prawdopodobieństw poszczególnych pomiarów. W przypadku zmiennej ciągłej wystarczy wymnożyć poszczególne gęstości prawdopodobieństwa (iloczyn różniczek $dx_1 dx_2 \dots dx_n$ można opuścić)

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i|\boldsymbol{\theta}). \tag{8.17}$$

Tak zdefiniowaną funkcję nazywamy *funkcją wiarygodności*. Szukanymi przez nas parametrami będą te parametry, dla których funkcja wiarygodności osiąga maksimum. W celu wyznaczenia tego maksimum musimy policzyć pierwsze pochodne funkcji wiarygodności po wszystkich parametrach i przyrównać je do zera. Funkcja wiarygodności jest iloczynem gęstości prawdopodobieństw, wiec liczenie pochodnych jest niewygodne. Aby ułatwić sobie zadanie możemy zlogarytmować funkcję wiarygodności. Nie zmieni to położenia maksimum, a za to będzie wygodniejsze w różniczkowaniu. Logarytm funkcji wiarygodności

$$l = \ln(L) = \sum_{i=1}^{n} \ln f(x_i|\,\boldsymbol{\theta}) \tag{8.18}$$

nazywany jest *logarytmiczną funkcją wiarygodności*, a często po prostu funkcją wiarygodności. Jeśli wektor parametrów $\boldsymbol{\theta}$ ma p składowych, warunek istnienia maksimum funkcji wiarygodności prowadzi do układu p równań

$$\frac{\partial l}{\partial \theta_i} = 0, \quad i = 1,2,\dots,p. \tag{8.19}$$

Przykład

Jako przykład zastosowania metody największej wiarygodności znajdźmy ponownie estymatory parametrów rozkładu Gaussa

$$f(x) = \frac{1}{\sqrt{2\pi\theta_2^2}} e^{-\frac{(x-\theta_1)^2}{2\theta_2^2}}. \tag{8.20}$$

Funkcja wiarygodności przyjmuje w tym wypadku postać

$$L(\theta_1, \theta_2) = \frac{1}{(2\pi\theta_2^2)^{n/2}} e^{-\frac{1}{2\theta_2^2}\sum_{i=1}^{n}(x_i-\theta_1)^2} \tag{8.21}$$

Po zlogarytmowaniu powyższej funkcji dostajemy logarytmiczną funkcję wiarygodności

$$l(\theta_1, \theta_2) = -n \ln\left(\sqrt{2\pi}\right) - n \ln \theta_2 - \frac{1}{2\theta_2^2} \sum_{i=1}^{n} (x_i - \theta_1)^2. \tag{8.22}$$

Zgodnie ze schematem postępowania w metodzie największej wiarygodności musimy teraz wyliczyć pochodne funkcji (8.22) i przyrównać je do zera. Otrzymamy wówczas następujący układ równań

$$\begin{cases} \dfrac{\partial l}{\partial \theta_1} = \dfrac{1}{\theta_2^2} \displaystyle\sum_{i=1}^{n} (x_i - \theta_1) = 0, \\ \dfrac{\partial l}{\partial \theta_2} = \dfrac{-n}{\theta_2} + \dfrac{1}{\theta_2^3} \displaystyle\sum_{i=1}^{n} (x_i - \theta_1)^2 = 0. \end{cases} \tag{8.23}$$

Z pierwszego równania dostajemy estymator parametru θ_1, który tak samo jak w metodzie momentów jest równy średniej arytmetycznej

$$T(\theta_1) = \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i. \tag{8.24}$$

Przekształcając drugie równanie i wykorzystując postać estymatora $T(\theta_1)$ mamy

$$n = \frac{1}{\theta_2^2} \sum_{i=1}^{n} (x_i - \bar{x})^2 , \tag{8.25}$$

czyli

$$T(\theta_2^2) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2. \tag{8.26}$$

Jest to estymator obciążony (asymptotycznie nieobciążony) wariancji zmiennej X. Jak widzimy metoda największej wiarygodności doprowadziła nas do takich samych postaci estymatorów obu parametrów co metoda momentów.

8.3. Estymacja przedziałowa

Oprócz estymacji punktowej, w której szacujemy konkretną wartość parametrów stosuje się również estymację przedziałową. Estymacja przedziałowa polega na znajdowaniu przedziału liczbowego, w którym z określonym z góry prawdopodobieństwem, tzw. *poziomem ufności* mieści się wartość szacowanego parametru. Poziom ufności oznaczany jest literą γ lub $1 - \alpha$, gdzie α jest tzw. poziomem istotności (patrz rozdział 9). Przedział, w którym z prawdopodobieństwem równym poziomowi ufności mieści się wartość parametru, czyli przedział, dla którego prawdopodobieństwo $P(T_{n1} \leq \theta \leq T_{n2}) = \gamma = 1 - \alpha$ nazywamy *przedziałem ufności*. Poziom ufności przyjmujemy zwykle dość wysoko np. 0,9. Jednak nie powinien on być za wysoki, gdyż wówczas otrzymamy szeroki przedział ufności co zmniejsza jego wartość informacyjną.

W poniższych paragrafach omówimy kilka typowych estymacji przedziałowych. Zakładamy, że zmienne losowe podlegają rozkładowi Gaussa oraz, że wyniki pomiarów nie są obarczone błędami systematycznymi.

8.3.1. Estymacja przedziałowa $E(X)$, gdy znamy $\sigma^2(X)$

W przypadku, gdy znamy wariancję rozkładu normalnego zmiennej losowej X, w celu znalezienia przedziału ufności wartości oczekiwanej $E(X) \equiv \hat{x}$ definiujemy statystykę

$$z \equiv \frac{\bar{x} - E(\bar{x})}{\sigma(\bar{x})} = \frac{(\bar{x} - \hat{x})\sqrt{n}}{\sigma(X)}. \tag{8.27}$$

Szukany przedział ufności na poziomie ufności γ musi spełniać warunek $P(z_{\min} \leq z \leq z_{\max}) = \gamma$. Oczywiście warunek ten nie określa tego przedziału jednoznacznie. Musimy go doprecyzować. Interesuje nas przedział w miejscu, gdzie gęstość prawdopodobieństwa jest największa. Ponieważ zmienna $\bar{x}$ podlega rozkładowi Gaussa (tak jak zmienna X), to zmienna z, która jest standaryzowaną średnią arytmetyczną, ma standardowy rozkład normalny $N(0,1)$. Rozkład ten jest rozkładem symetrycznym względem zera, które jest jednocześnie modą rozkładu (gęstość prawdopodobieństwa standardowego rozkładu normalnego ma w zerze maksimum). Wobec tego przedział $[z_{\min}, z_{\max}]$ powinien być symetryczny względem zera, a zatem $z_{\max} = -z_{\min}$. Granice tego przedziału możemy wyrazić poprzez kwantyle standardowego rozkładu normalnego (patrz paragraf 4.6)

$$z_{\min} = z_{\frac{1}{2}-\frac{\gamma}{2}}, \qquad z_{\max} = z_{\frac{1}{2}+\frac{\gamma}{2}}. \tag{8.28}$$

Przekształcając wyrażenie (8.27) dostajemy

$$\hat{x} = \bar{x} - z\frac{\sigma(X)}{\sqrt{n}}, \tag{8.29}$$

a ponieważ $P(z_{\min} \leq z \leq z_{\max}) = \gamma$, to również

$$P\left(\bar{x} - z_{\max}\frac{\sigma(X)}{\sqrt{n}} \leq \hat{x} \leq \bar{x} - z_{\min}\frac{\sigma(X)}{\sqrt{n}}\right) = \gamma. \tag{8.30}$$

W ten sposób otrzymaliśmy przedział ufności, o którym możemy powiedzieć, że z prawdopodobieństwem γ zawiera w sobie wartość oczekiwaną. Granice tego przedziału możemy na kilka sposobów wyrazić poprzez kwantyle standardowego rozkładu normalnego, korzystając z własności tych kwantyli

$$z_{\frac{1}{2}-\frac{\gamma}{2}} = -z_{\frac{1}{2}+\frac{\gamma}{2}}. \tag{8.31}$$

Jeden ze sposobów zapisu przedziału ufności ma postać

$$\bar{x} - z_{\frac{1}{2}+\frac{\gamma}{2}}\frac{\sigma(X)}{\sqrt{n}} \leq \hat{x} \leq \bar{x} + z_{\frac{1}{2}+\frac{\gamma}{2}}\frac{\sigma(X)}{\sqrt{n}}. \tag{8.32}$$

Warto zwrócić uwagę na fakt, że wartość oczekiwana $\hat{x}$ nie jest zmienną losową. Ma ona konkretną wartość zależna od rozkładu, zaś granice przedziału ufności są zmiennymi losowymi, ponieważ zależą od wartości średniej, czyli wyników próby.

8.3.1. Estymacja przedziałowa $E(X)$, gdy nie znamy $\sigma^2(X)$

Bardzo często chcemy znaleźć przedział ufności wartości oczekiwanej rozkładu normalnego nie znając jego wariancji. Może tak być np. w sytuacji, gdy próba jest zbyt mała, żeby móc założyć, że estymator wariancji daje wystarczająco dobre oszacowanie wariancji. W takiej sytuacji jako statystykę przyjmujemy zmienną

$$t = \frac{\bar{x} - E(\bar{x})}{S(\bar{x})} = \frac{\bar{x} - E(\bar{x})}{S(X)}\sqrt{n}, \tag{8.33}$$

gdzie

$$S(X) = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \tag{8.34}$$

jest estymatorem odchylenia standardowego z próby. Można pokazać, że zmienna losowa t podlega rozkładowi t-Studenta o $n-1$ stopniach swobody. Rozkład ten został znaleziony przez angielskiego statystyka Williama Gosseta, który publikował swoje prace pod pseudonimem Student. Rozkład t-Studenta nazywany również po prostu rozkładem Studenta (czasami rozkładem Studenta-Fishera) ma postać

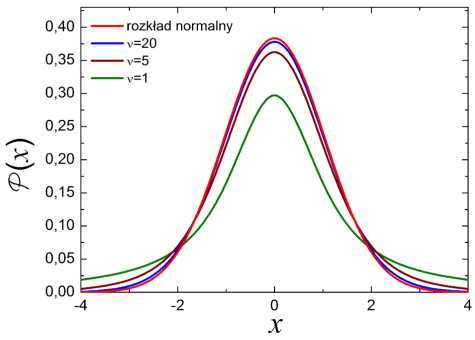
$$f(t,\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\pi\nu}}\left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}, \tag{8.35}$$

gdzie ν jest liczbą stopni swobody rozkładu (w naszym przypadku $\nu = n-1$, n jest liczbą pomiarów), a Γ jest funkcją gamma Eulera (patrz wskazówka do zadania 4.6). Postać krzywej dla rozkładu Studenta jest bardzo podobna do rozkładu normalnego i ma tę samą symetrię. W związku z tym, nasze rozumowanie z poprzedniego paragrafu jest słuszne również w tym wypadku. Jedyna różnica polega na tym, że kwantyle rozkładu normalnego musimy zastąpić kwantylami rozkładu Studenta. Przedział ufności dla poziomu ufności γ możemy zapisać w postaci

$$\bar{x} - t_{n,\frac{1}{2}+\frac{\gamma}{2}}\frac{\sigma(X)}{\sqrt{n}} \leq \hat{x} \leq \bar{x} + t_{n,\frac{1}{2}+\frac{\gamma}{2}}\frac{\sigma(X)}{\sqrt{n}}. \tag{8.36}$$

Zamiast z kwantyli rozkładu Studenta czasami korzystamy z tablic tzw. współczynników Studenta-Fishera $\tau_{\nu,\gamma}$ zależnych od liczby stopni swobody ν i poziomu ufności γ.

Dla prób składających się z dużej liczby pomiarów $n > 20$, czyli dużej liczby stopni swobody rozkład Studenta jest już bardzo bliski rozkładowi Gaussa (patrz Rysunek 8.1). Możemy wówczas znajdować granice przedziału ufności korzystając z kwantyli rozkładu normalnego zgodnie ze wzorem (8.32).



Rysunek 8.1. Rozkłady Studenta dla $\nu = 1{,}5$ i 20 stopni swobody oraz rozkład standardowy normalny.

8.3.2. Estymacja przedziałowa $\sigma^2(X)$

Czasami interesuje nas przedział ufności szacowanej wariancji lub odchylenia standardowego. W takim przypadku jako statystykę przyjmujemy

$$Y = \frac{(n-1)S^2(X)}{\sigma^2(X)}, \tag{8.37}$$

gdzie n jest liczbą pomiarów w próbie, $\sigma^2(x)$ jest wariancją, a $S^2(x)$ estymatorem wariancji

$$S^2(X) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}. \tag{8.38}$$

Można pokazać, że statystyka Y podlega rozkładowi χ^2 (czytaj chi kwadrat) o $n-1$ stopniach swobody (patrz rozdz. 4.7). Przypomnijmy, że symbolu rozkładu χ^2 nie należy interpretować jak kwadratu zmiennej losowej, symbol χ wraz z wykładnikiem 2 stanowi całość i jest zwykłą zmienną losową. Wykładnik 2 w symbolu rozkładu ma jedynie podkreślić, że rozkład χ^2 opisuje zmienne losowe będące sumą kwadratów niezależnych zmiennych losowych. Poszczególne zmienne składające się na tę sumę kwadratów podlegają rozkładowi normalnemu. Liczba tych zmiennych jest liczbą stopni swobody rozkładu. Jest to bardzo ważny rozkład w statystyce stosowany w testach statystycznych. Jeśli przez k oznaczymy liczbę stopni swobody rozkładu, to rozkład χ^2 możemy zapisać w postaci

$$f(x,k) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2}, \quad x \geq 0. \tag{8.39}$$

Ponieważ zmienna $x = \chi^2$ jest sumą kwadratów liczb rzeczywistych, to nie może być ujemna. Rozkład χ^2 jest zatem niesymetryczny, co pokazano na rysunku 4.8 (a). Dla jednego stopnia swobody $k = 1$ rozkład χ^2 dąży do nieskończoności w zerze, dla dwóch stopni swobody $k = 2$ rozkład w zerze jest równy $1/2$, a dla $k \geq 3$ rozkład w zerze jest równy 0 i przyjmuje kształt tzw. skośnej krzywej dzwonowej. Wraz ze wzrostem liczby stopni swobody kształt krzywej staje się coraz bardziej symetryczny.

Funkcja charakterystyczna rozkładu χ^2 ma postać

$$\varphi_{\chi^2}(t) = (1-2it)^{n/2}. \tag{8.40}$$

Stąd możemy wyliczyć momenty rozkładu. Wartość oczekiwana rozkładu wynosi

$$E(\chi^2) = \widehat{\chi^2} = -i\varphi'(0) = 2 \cdot \frac{n}{2} = n. \tag{8.41}$$

W celu wyliczenia wariancji policzmy najpierw wartość oczekiwaną z $(\chi^2)^2$

$$E((\chi^2)^2) = -\varphi''(0) = 4\left(\frac{n}{2}\right)^2 + 4\left(\frac{n}{2}\right) = n^2 - 2n. \tag{8.42}$$

Wariancja rozkładu wynosi zatem

$$\sigma^2(\chi^2) = \left(E(\chi^2)\right)^2 - E((\chi^2)^2) = n^2 - (n^2 - 2n) = 2n. \tag{8.43}$$

Czyli wartość oczekiwana rozkładu χ^2 jest równa jego liczbie stopni swobody, a wariancja podwojonej liczbie stopni swobody.

Wróćmy do naszego problemu. Chcemy znaleźć przedział ufności wariancji rozkładu normalnego. Podobnie jak robiliśmy to w poprzednich dwóch paragrafach, interesuje nas przedział najbardziej prawdopodobnych wartości zmiennej Y. Jednak teraz mamy do czynienia rozkładem χ^2, który nie jest symetryczny względem swojej mody. Aby jednoznacznie wyznaczyć przedział zakładamy, że prawdopodobieństwo wartości zmiennej Y mniejszych niż $Y_{\min}$ jest takie samo jak wartości powyżej $Y_{\max}$, czyli

$$P(Y < Y_{\min}) = P(Y > Y_{\max}) = \frac{1}{2} - \gamma. \tag{8.44}$$

To założenie pozwala jednoznacznie wyznaczyć granice przedziału ufności. Granice te wyrażone poprzez kwantyle rozkładu χ^2 wynoszą

$$Y_{\min} = (\chi^2_{n-1})_{\frac{1}{2}-\frac{\gamma}{2}}, \qquad Y_{\max} = (\chi^2_{n-1})_{\frac{1}{2}+\frac{\gamma}{2}} \tag{8.45}$$

Tym razem inaczej niż w poprzednich przypadkach nie ma związku pomiędzy obiema wartościami, gdyż rozkład χ^2 jest niesymetryczny. Ponieważ związek pomiędzy wariancją i statystyką Y jest monotoniczną funkcją

$$\sigma^2(X) = \frac{(n-1)S^2(X)}{Y}, \tag{8.46}$$

to prawdopodobieństwo, że statystyka Y znajduje się w przedziale $[Y_{\min}, Y_{\max}]$ jest równe prawdopodobieństwu, że szacowana wariancja leży w przedziale $\left[\frac{(n-1)S^2(X)}{Y_{\max}}, \frac{(n-1)S^2(X)}{Y_{\min}}\right]$. Ostatecznie przedział ufności wariancji $\sigma^2(X)$ na poziomie ufności γ jest określony nierównościami

$$\frac{(n-1)S^2(X)}{(\chi^2_{n-1})_{\frac{1}{2}+\frac{\gamma}{2}}} \leq \sigma^2(X) \leq \frac{(n-1)S^2(X)}{(\chi^2_{n-1})_{\frac{1}{2}-\frac{\gamma}{2}}}. \tag{8.47}$$

Pierwiastkowanie dodatnich liczb jest funkcją monotoniczną. Dlatego spierwiastkowanie powyższego wyrażenia wyznacza nam przedział ufności odchylenia standardowego na tym samym poziomie ufności co dla wariancji

$$\sqrt{\frac{(n-1)}{(\chi^2_{n-1})_{\frac{1}{2}+\frac{\gamma}{2}}}}S(X) \leq \sigma(X) \leq \sqrt{\frac{(n-1)}{(\chi^2_{n-1})_{\frac{1}{2}-\frac{\gamma}{2}}}}S(X). \tag{8.48}$$

Zadanie 8.1

Zmienna losowa x podlega rozkładowi Bernoulliego

$$P(k) = \frac{N!}{k!\,(N-k)!}p^k(1-p)^{N-k},$$

gdzie N jest liczbą prób, k liczbą prób zakończonych sukcesem, a p prawdopodobieństwem sukcesu w pojedynczej próbie. Oszacuj metodą momentów oraz metodą największej wiarygodności wartość parametru p wiedząc, że w doświadczeniu polegającym na wykonaniu N prób uzyskano n_0 prób zakończonych sukcesem.

Zadanie 8.2

Zmienna losowa ma rozkład równomierny na przedziale $[a, b]$. Znajdź metodą momentów estymatory parametrów a, b. Korzystając z tych estymatorów oszacuj wartości tych parametrów na podstawie następującej próby:

3,5; 7,6; 2,9; 4,3; 2,0; 3,8; 6,5; 6,4; 5,0; 2,8; 4,5; 7,8; 3,7; 6,4; 5,0; 4,4; 2,5; 6,7; 2,1; 7,7; 3,2; 6,2; 4,7; 6,3; 7,1.

Zadanie 8.3

Pewna zmienna losowa $x > 0$ ma rozkład wykładniczy o postaci

$$f(x) = Ce^{-Cx}.$$

W wyniku pomiarów tej zmiennej otrzymano następujące wartości:

2; 3; 3; 5; 6; 7; 7; 8; 8; 9; 10; 12; 13; 14; 15; 23; 24; 27; 29; 35.

Oszacuj metodą momentów oraz metodą największej wiarygodności wartość parametru C.

Zadanie 8.4

Pobrano czteroelementową próbę zmiennej losowej podlegającej rozkładowi normalnemu. Znajdź przedział ufności wartości oczekiwanej tej zmiennej dla poziomu ufności $\gamma = 0{,}9$. Wyniki pomiarów były następujące: 10,5; 10,2; 10,4; 10,5.

Wskazówka

Skorzystaj z wyrażenia (8.26). Odpowiednie kwantyle można obliczyć w programie MS Excel korzystając z funkcji ROZKŁAD.T.ODWR lub znaleźć w tablicach np. w S. Brandta *Analiza danych*, PWN 2002.

Zadanie 8.5

Oszacuj przedział ufności odchylenia standardowego na poziomie ufności $\gamma = 0{,}9$ dla danych z poprzedniego zadania.

Wskazówka

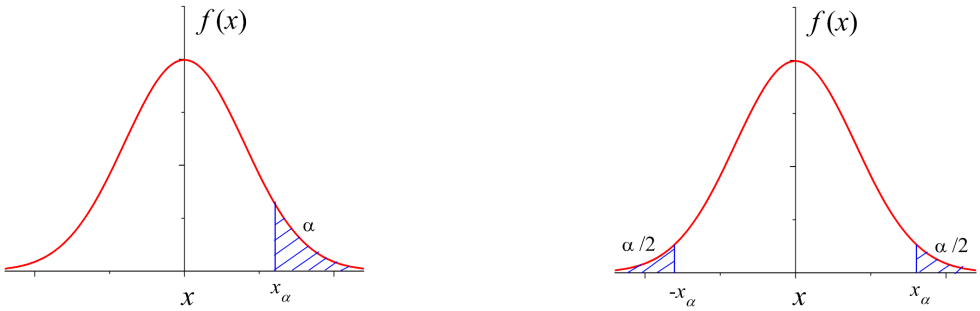
Skorzystaj z wyrażenia (8.48). Odpowiednie kwantyle można obliczyć w programie MS Excel korzystając z funkcji ROZKŁAD.CHI.ODWR lub znaleźć w tablicach np. w S. Brandta *Analiza danych*, PWN 2002.

9. Weryfikacja hipotez statystycznych

Pobrana próba statystyczna nie zawsze służy do wyznaczenia nieznanego parametru. Zdarza się, że mamy pewne przypuszczenia co do wartości parametru czy samego rozkładu wynikające np. z przewidywań modelu lub innych pomiarów. Na podstawie próby chcemy zweryfikować naszą hipotezę. Nie jest to proste, gdyż otrzymane wyniki (elementy próby) mają charakter losowy. Jeśli np. zakładamy, że wartość oczekiwana jest zerowa nie możemy liczyć na to, że otrzymana na podstawie próby wartość estymatora wartości oczekiwanej będzie równa zeru. Na ogół będzie to wartość różna od zakładanej. Musimy zatem sprawdzić, czy z założonym z góry prawdopodobieństwem wartość oczekiwana może odbiegać od zakładanej o otrzymaną wartość.

Wprowadźmy kilka nowych pojęć i definicji. Hipotezę odnosząca się do rozkładu zmiennej losowej (rozkładu prawdopodobieństwa dla zmiennej dyskretnej lub rozkładu gęstości prawdopodobieństwa dla zmiennej ciągłej) lub parametrów tego rozkładu nazywamy *hipotezą statystyczną*. Hipoteza statystyczna dotycząca parametru lub parametrów rozkładu nazywamy *hipotezą parametryczną* (np. parametr $\mu = 0$). Pozostałe hipotezy nazywamy *nieparametrycznymi* (np. rozkład zmiennej losowej X jest rozkładem normalnym). Hipoteza, którą chcemy zweryfikować jest nazywana *hipotezą zerową*. Oznaczamy ją literą H_0. Oprócz hipotezy zerowej stawiamy również hipotezę alternatywną, którą przyjmujemy wówczas, gdy w wyniku testu hipoteza zerowa zostaje odrzucona. Na ogół jest to zaprzeczenie hipotezy zerowej, ale nie musi tak być. Hipotezę alternatywną oznaczamy literą H_1. Podczas weryfikacji hipotezy możemy popełnić błąd. *Błędem pierwszego rodzaju* nazywamy sytuację, gdy odrzuciliśmy hipotezę zerową, która była prawdziwa. *Błędem drugiego rodzaju* nazywamy przyjęcie hipotezy zerowej, podczas gdy jest to hipoteza nieprawdziwa. Należy zwrócić uwagę, że przyjęcie hipotezy zerowej w wyniku testu hipotezy nie jest dowodem jej prawdziwości. Możemy jedynie powiedzieć, że na podstawie pobranej próby nie ma podstaw do odrzucenia hipotezy. Weryfikując hipotezę statystyczną przyjmujemy tzw. *poziom istotności* oznaczany zwykle literą α. Poziom istotności jest to maksymalne akceptowane przez nas prawdopodobieństwo popełnienia błędu pierwszego rodzaju. Następnie sprawdzamy, czy prawdopodobieństwo wystąpienia otrzymanych na podstawie próby wartości jest mniejsze niż przyjęty poziom istotności. Jeśli tak jest, to przyjmujemy, że jest mało prawdopodobne, żeby nasza hipoteza była prawdziwa i odrzucamy ją. Innymi słowy dzielimy przestrzeń prób na dwa rozłączne zbiory. Jeden z nich nazywamy obszarem krytycznym. Prawdopodobieństwo wystąpienia wartości parametru w obszarze krytycznym jest mniejsze lub równe założonemu poziomowi istotności. Jeśli otrzymana przez nas wartość parametru leży w obszarze krytycznym, to odrzucamy weryfikowaną hipotezę i przyjmujemy hipotezę alternatywną. Jeśli wartość parametru leży poza obszarem krytycznym, to nie mamy podstaw do odrzucenia hipotezy. Wybór poziomu istotności zależy od rozpatrywanego problemu. Zwykle przyjmuje się poziom istotności na poziomie 0,05, 0,01 lub 0,001. Wszystko zależy od tego jakie znacznie ma weryfikowana hipoteza i jakie są koszty popełnienia błędu pierwszego lub drugiego rodzaju. Czym niższy poziom istotności przyjmiemy, tym większą mamy pewność, że nie popełnimy błędu. Podczas weryfikacji hipotezy możemy się spotkać z dwiema sytuacjami: testem dwustronnym lub testem jednostronnym. Z testem jednostronnym mamy do czynienia, gdy interesuje nas wartość parametru po jednej stronie rozkładu, np. dopuszczamy tylko do-

datnie wartości parametru. Dla przykładu załóżmy, że testujemy wartość średnią jakiejś zmiennej losowej X. Wiemy, że wartość zmiennej powinna być dodatnia. Obszar krytyczny będzie wyznaczony przez nierówność $\bar{x} \geq x_\alpha$, przy czym x_α ustalamy tak, aby prawdopodobieństwo otrzymania wartości z tego obszaru było nie większe niż przyjęty poziom istotności α, $P(\bar{x} \geq x_\alpha) < \alpha$. W przypadku testu dwustronnego obszar krytyczny znajduje się po obu stronach rozkładu. Definiujemy go nierównościami $\bar{x} \leq x_{\alpha 1}$ i $\bar{x} \geq x_{\alpha 2}$, przy czym $P(\bar{x} \leq x_{\alpha 1}) < \alpha/2$ i $P(\bar{x} \geq x_{\alpha 2}) < \alpha/2$. Dla rozkładów symetrycznych obszar krytyczny definiowany jest nierównością z wartością bezwzględną $|\bar{x}| \geq x_\alpha$ z x_α definiowanym nierównością $P(|\bar{x}| \geq x_\alpha) < \alpha$. Przykładowe obszary krytyczne dla testu jednostronnego i dwustronnego pokazano na rysunku 9.1.



Rysunek 9.1. Przykładowe obszary krytyczne dla testu jednostronnego (a) i dwustronnego (b).

Podsumujmy. Idea postępowania podczas weryfikacji hipotezy statystycznej polega na tym, że odrzucamy testowaną hipotezę, gdy wyniki obserwacji przy założeniu prawdziwości tej hipotezy są dalece nieprawdopodobne. Wyniki niesprzeczne z hipotezą zerową nie są dowodem prawdziwości hipotezy, a jedynie pozwalają na stwierdzenie, że nie ma podstaw do jej odrzucenia. Oczywiście może się zdarzyć, że otrzymane wyniki leżą w obszarze krytycznym, chociaż hipoteza jest prawdziwa, gdyż małe prawdopodobieństwo wystąpienia danych wartości nie oznacza, że są one niemożliwe. Może się zatem zdarzyć, że odrzucimy hipotezę pomimo tego, że jest ona prawdziwa. Aby zminimalizować prawdopodobieństwo popełnienia błędu pierwszego rodzaju stosujemy odpowiednio małą wartość poziomu istotności. Wartość poziomu istotności zależy przede wszystkim od kosztów, jakie poniesiemy w przypadku odrzucenia prawdziwej hipotezy, czyli popełnienia błędu pierwszego rodzaju. Odrzucając hipotezę zerową akceptujemy hipotezę alternatywną, która na ogół jest zaprzeczeniem hipotezy zerowej „nie jest prawdą, że H_0".
Omówmy schemat postępowania na prostym przykładzie bez wnikania w szczegóły.

- **Formułujemy hipotezę zerową**
  H_0: zmienna losowa X mająca rozkład normalny o znanej wariancji $\sigma^2(X) = 3$ ma wartość oczekiwaną $\hat{x} = 5$.
- **Tworzymy statystykę testową**
  Statystyka testowa powinna zależeć od wielkości testowanej, a jej rozkład przy założeniu prawdziwości hipotezy H_0 powinien być znany. W naszym przykładzie najlepszą statystyką będzie średnia arytmetyczna z próby $T_n = \bar{x}$, która jest estymatorem wartości oczekiwanej. Ponadto wiemy, że rozkład średniej arytmetycznej, przy założeniu, że zmienna losowa X ma rozkład Gaussa, ma również rozkład Gaussa o wartości oczekiwanej $\hat{x} = 5$ i wariancji $\sigma^2(\bar{x}) = \sigma^2(X)/n$, gdzie n jest liczbą pomiarów w próbie. Jeśli dysponujemy jedynie tablicami standaryzowanego rozkładu Gaussa, to wygodniejszą statystyka będzie zmienna standaryzowana $T_n = (\bar{x} - \hat{x})/\sigma(\bar{x})$.
- **Ustalamy hipotezę alternatywną H_1**

Jeśli nie mamy żadnych informacji na temat wartości oczekiwanej, to jako hipotezę alternatywą przyjmujemy zaprzeczenie hipotezy H_0, czyli H_1: nie prawda, że H_0. W naszym przykładzie będzie to hipoteza $H_1: \hat{x} \neq 5$. Jeśli jednak mamy pewne informacje, np. znamy wartość średniej arytmetycznej $\bar{x} = 8{,}2$, to rozsądne jest przyjęcie hipotezy $H_1: \hat{x} > 5$.

- **Ustalamy poziom istotności α**
  Zwykle przyjmujemy poziom istotności na poziomie $\alpha = 0{,}05$ lub nawet $\alpha = 0{,}1$, jednak w sytuacji, gdy skutki odrzucenia prawdziwej hipotezy H_0 są duże przyjmujemy odpowiednio niższy poziom istotności.
- **Wyznaczamy obszar krytyczny testu**
  Obszar krytyczny testu jest to obszar wartości statystyki testowej, które przy prawdziwości hipotezy H_0 są mało prawdopodobne, a jednocześnie bardzo prawdopodobne przy prawdziwości hipotezy H_1. Jeśli hipotezą alternatywną było zaprzeczenie hipotezy H_0, to zakładając prawdziwość hipotezy H_0 najmniej prawdopodobne są duże co do modułu wartości statystyki. Dla poziomu istotności α będą to liczby określone dwiema nierównościami: $T_n < z_{\alpha/2}$ i $T_n > z_{1-\alpha/2}$, gdzie $z_{\alpha/2}$ i $z_{1-\alpha/2}$ kwantylami rozkładu Gaussa na poziomie odpowiednio $\alpha/2$ i $1-\alpha/2$. Wartości statystyki testowej z tych przedziałów są mało prawdopodobne przy prawdziwości hipotezy H_0, a jednocześnie bardzo prawdopodobne przy prawdziwości hipotezy H_1. Jeśli zaś hipotezą alternatywną było $\hat{x} > 5$, to sprzyjać tej hipotezie będą duże dodatnie liczby, które z kolei będą mało prawdopodobne przy prawdziwości hipotezy H_0. Obszarem krytycznym będzie w tym wypadku zbiór wartości statystyki testowej opisany nierównością $T_n > z_{1-\alpha}$.
- **Sprawdzamy w jakim obszarze leży wartość statystyki testowej**
  Jeśli wartość statystyki testowej leży w obszarze krytycznym, to na poziomie istotności α odrzucamy weryfikowaną hipotezę i akceptujemy hipotezę alternatywną. W przeciwnym wypadku stwierdzamy, że na poziomie istotności α nie mamy podstaw do odrzucenia hipotezy.

9.1. Porównanie wariancji z liczbą

Czasami znamy wartość wariancji i chcemy sprawdzić, czy wariancja testowana jest równa tej znanej wartości. Dla przykładu znamy dokładność jakiegoś przyrządu pomiarowego i chcemy sprawdzić, czy inny przyrząd ma taką samą dokładność. Naszą hipotezą zerową jest $H_0: \sigma^2(X) = \sigma_0^2$. Jako statystykę testową definiujemy wielkość

$$Q^2 = \frac{(n-1)S^2(X)}{\sigma_0^2}, \tag{9.1}$$

gdzie $S^2(X)$ jest znanym estymatorem wariancji, a *n* liczbą elementów próby. Jeśli hipoteza zerowa jest prawdziwa, to nasza statystyka ma rozkład χ^2_{n-1}. Do określania obszarów krytycznych korzystamy z kwantyli tego rozkładu, podobnie jak w paragrafie 7.3.2. Wybór obszarów będzie zależny od przyjętej hipotezy alternatywnej.

- $H_1: \sigma^2(X) \neq \sigma_0^2$, obszar krytyczny składa się z dwóch przedziałów $Q^2 < \chi^2_{n-1,\alpha/2}$ i $Q^2 > \chi^2_{n-1,1-\alpha/2}$.
- $H_1: \sigma^2(X) > \sigma_0^2$, $Q^2 > \chi^2_{n-1,1-\alpha}$.
- $H_1: \sigma^2(X) < \sigma_0^2$, $Q^2 < \chi^2_{n-1,\alpha}$.

9.2. Test równości wariancji (test *F* Fischera-Snedecora)

Załóżmy, że wykonaliśmy serię pomiarów pewnej wielkości fizycznej jednym przyrządem oraz drugą serię pomiarów tej samej wielkości innym przyrządem. Zakładamy, że w obu przypadkach nie popełniliśmy błędów systematycznych oraz że zmienna podlega rozkładowi normalnemu. Wartość oczekiwana mierzonej wielkości w obu przypadkach jest oczywiście taka sama. Chcemy natomiast stwierdzić, czy dokładności obu przyrządów są jednakowe. Aby to stwierdzić musimy porównać obie wariancje. Jest to typowa sytuacja, w której musimy przeprowadzić test równości wariancji.
Nasza hipoteza zerowa ma postać H_0: $\sigma^2(X_1) = \sigma^2(X_2)$. Jako statystykę testową przyjmujemy iloraz estymatorów obu wariancji

$$F = \frac{S^2(X_1)}{S^2(X_2)}. \tag{9.2}$$

Dla jednej i drugiej próby możemy zdefiniować zmienne

$$Q_1^2 = \frac{(n_1 - 1)S^2(X_1)}{\sigma^2(X_1)}, \qquad Q_2^2 = \frac{(n_2 - 1)S^2(X_2)}{\sigma^2(X_2)} \tag{9.3}$$

podlegające rozkładowi χ^2 odpowiednio dla $n_1 - 1$ i $n_2 - 1$ stopni swobody. Zakładając równość wariancji $\sigma^2(X_1) = \sigma^2(X_2)$ statystykę testową F możemy wyrazić przez

$$F = \frac{n_2 - 1}{n_1 - 1}\frac{Q_1^2}{Q_2^2}. \tag{9.4}$$

Statystyka ta przyjmuje tylko dodatnie wartości i podlega rozkładowi Fischera-Snedecora $F(n_1, n_2)$. Jako hipotezę alternatywną przyjmujemy hipotezę o braku równości wariancji: $\sigma^2(X_1) \neq \sigma^2(X_2)$ lub hipotezę, że wariancja w liczniku jest większa od wariancji w mianowniku: $\sigma^2(X_1) > \sigma^2(X_2)$. Do określania obszarów krytycznych korzystamy z kwantyli rozkładu Fischera-Snedecora.

- H_1: $\sigma^2(X_1) \neq \sigma^2(X_2)$, obszar krytyczny składa się z dwóch przedziałów $F < F_{\frac{\alpha}{2}}(n_1 - 1, n_2 - 1)$ i $F > F_{1-\frac{\alpha}{2}}(n_1 - 1, n_2 - 1)$.
- H_1: $\sigma^2(X_1) > \sigma^2(X_2)$, $F > F_{1-\alpha}(n_1 - 1, n_2 - 1)$.

Czasami w tablicach kwantyli rozkładu Fischera-Snedecora podawane są tylko kwantyle dla wysokiego lub tylko niskiego poziomu ufności. Możemy wówczas skorzystać

$$F_{\frac{\alpha}{2}}(n_1, n_2) = \frac{1}{F_{1-\frac{\alpha}{2}}(n_2 n_1)}. \tag{9.5}$$

9.3.Porównanie wartości średniej z liczbą

Rozważmy zmienną losową X o rozkładzie normalnym. Wykonaliśmy n pomiarów o wartości średniej $\bar{X}$ i pytamy, czy otrzymana wartość średnia jest równa (w sensie statystycznym) znanej wartości E_0. Jest to częsta sytuacja, z jaką spotykamy się podczas porównywania wyników pomiarów wielkości takich jak np. stałe fizyczne z ich znanymi wartościami[5]. Pytamy wówczas, czy otrzymany przez nas wynik jest (w sensie statystycznym) równy wartości tablicowej. Naszą hipotezą zerową jest $H_0: \bar{X} = E_0$. Musimy tu rozróżnić dwa przypadki:

a) znamy wariancję $\sigma^2(X)$ zmiennej losowej,
b) znamy jedynie estymator wariancji $S^2(X)$.

W pierwszym przypadku statystyką testową będzie statystyka

$$z = \frac{\bar{X} - E_0}{\sigma(\bar{X})}, \tag{9.6}$$

która jak widać jest zmienną standaryzowaną podlegającą standardowemu rozkładowi Gaussa $N(0,1)$. W drugim przypadku naszą statystyka będzie statystyka

$$t = \frac{\bar{X} - E_0}{S(\bar{X})}, \tag{9.7}$$

która podlega standardowemu rozkładowi *t*-Studenta o $n-1$ stopniach swobody. Zwróćmy uwagę na to, że w statystyce z występuje odchylenie standardowe średniej arytmetycznej $\left(\sigma(\bar{X}) = \sigma(X)/\sqrt{n}\right)$, a w statystyce t, estymator odchylenia standardowego średniej arytmetycznej $\left(S(\bar{X}) = S(X)/\sqrt{n}\right)$.

W przypadku a) dla określenia obszarów granicznych wykorzystujemy kwantyle rozkładu normalnego:

- $H_1: \bar{X} \neq E_0$, $\quad |z| > z_{1-\alpha/2}$, czyli $-z_{1-\alpha/2} > z > z_{1-\alpha/2}$.
- $H_1: \bar{X} > E_0$, $\quad z > z_{1-\alpha}$.
- $H_1: \bar{X} < E_0$, $\quad z < z_{\alpha}$.

W przypadku b) dla określenia obszarów granicznych wykorzystujemy kwantyle rozkładu *t*-Studenta:

- $H_1: \bar{X} \neq E_0$, $\quad |t| > t_{n-1,1-\alpha/2}$, czyli $-t_{n-1,1-\alpha/2} > t > t_{n-1,1-\alpha/2}$.
- $H_1: \bar{X} > E_0$, $\quad t > t_{n-1,1-\alpha}$.
- $H_1: \bar{X} < E_0$, $\quad t < t_{n-1,\alpha}$.

Jak pamiętamy, zarówno standardowy rozkład Gaussa, jak i rozkład standardowy *t*-Studenta są symetryczne względem zera, więc ich kwantyle mają poniższe własności

$$z_{\alpha} = -z_{1-\alpha}, \tag{9.8}$$

[5] Najczęściej znana wartość danej wielkości jest wartością

$$t_{n-1,\alpha} = -t_{n-1,1-\alpha}.$$

Stąd dla hipotezy alternatywnej $H_1: \bar{X} \neq E_0$ dostajemy $z_{\alpha/2} > z > -z_{\alpha/2}$ oraz $t_{n-1,\alpha/2} > t > -t_{n-1,\alpha/2}$ a stąd po przekształceniach

$$\begin{aligned} \bar{X} - \sigma(\bar{X}) z_{\alpha/2} < E_0 < \bar{X} + \sigma(\bar{X}) z_{\alpha/2}, \\ \bar{X} - S(\bar{X}) t_{n-1,\alpha/2} < E_0 < \bar{X} + S(\bar{X}) t_{n-1,\alpha/2}. \end{aligned} \tag{9.9}$$

Z powyższych nierówności wynika bardzo ważny wniosek: rzeczywista wartość mierzonej wielkości, na poziomie istotności α (czyli z prawdopodobieństwem $1-\alpha$), mieści się w przedziale

$$\left(\bar{X} - \sigma(\bar{X}) z_{\alpha/2}, \bar{X} + \sigma(\bar{X}) z_{\alpha/2}\right) \tag{9.10}$$

lub

$$\left(\bar{X} - S(\bar{X}) t_{n-1,\alpha/2}, \bar{X} + S(\bar{X}) t_{n-1,\alpha/2}\right) \tag{9.11}$$

w zależności od tego, czy znamy wariancję, czy też nie. Oczywiście na ogół wariancja nie jest znana, w praktyce jednak możemy uważać, że znamy wariancję jeśli liczba pomiarów jest odpowiednio duża (estymator wariancji niewiele odbiega od wartości wariancji).

9.4. Porównanie wartości średnich dwu populacji

Często przychodzi nam porównać wartość średnią jakiejś zmiennej losowej uzyskaną na podstawie naszych pomiarów z wartością średnią uzyskaną z innej próby losowej np. przez inny zespół badawczy lub uzyskany inną metodą pomiarową. Hipotezą zerową jest hipoteza o równości wartości średnich: $H_0: \bar{X} = \bar{Y}$. Wybór odpowiedniej statystyki do przetestowania tej hipotezy będzie zależny od tego, z jakim przypadkiem, spośród poniższych, mamy do czynienia:

a) znamy obie wariancje $\sigma^2(X)$ i $\sigma^2(Y)$,
b) nie znamy obu wariancji, ale wiemy, że są one sobie równe $\sigma^2(X) = \sigma^2(Y)$,
c) nie znamy obu wariancji, ale wiemy, że są one różne $\sigma^2(X) \neq \sigma^2(Y)$.

W przypadku a) statystyką będzie

$$z = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma^2(X)}{n_x} + \frac{\sigma^2(Y)}{n_y}}}. \tag{9.12}$$

Statystyka z podlega standardowemu rozkładowi normalnemu $N(0,1)$.

W przypadku b) musimy najpierw przeprowadzić test Studenta-Snedecora, aby sprawdzić, czy wariancje obu zmiennych X, Y możemy uznać za równe, a następnie stosujemy test Studenta ze statystyką zdefiniowaną następująco:

$$t = \frac{\bar{X} - \bar{Y}}{S(X,Y)\sqrt{\frac{n_x + n_y}{n_x n_y}}}, \tag{9.13}$$

gdzie

$$S(X,Y) = \sqrt{\frac{(n_x - 1)S^2(X) + (n_y - 1)S^2(Y)}{n_x + n_y - 2}}. \tag{9.14}$$

Statystyka t podlega standardowemu rozkładowi *t*-Studenta o $\nu = n_x + n_y - 2$ stopniach swobody.

W trzecim przypadku, tzn. gdy nie znamy wariancji obu zmiennych, a test Studenta-Snedecora wykazał, że różnią się one między sobą, stosujemy statystykę t w zmodyfikowanej formie

$$t = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S^2(X)}{n_x} + \frac{S^2(Y)}{n_y}}}. \tag{9.15}$$

Taka zmienna ma rozkład, który możemy przybliżyć rozkładem *t*-Studenta o efektywnej liczbie stopni swobody n_{ef}:

$$n_{ef} = \frac{\left(\frac{S^2(X)}{n_x} + \frac{S^2(Y)}{n_y}\right)^2}{\frac{(S^2(X)/n_x)^2}{n_x + 1} + \frac{\left(S^2(Y)/n_y\right)^2}{n_y + 1}} - 2. \tag{9.16}$$

Wyliczona w powyższy sposób efektywna liczba stopni swobody jest najczęściej niecałkowita. W związku z tym należy ją zaokrąglić do liczby całkowitej, przy czym bezpieczniej stosować zaokrąglenie w dół przez co zwiększamy nieco stopień istotności.

Stosując odpowiednie kwantyle rozkładów standardowych normalnego lub Studenta znajdujemy odpowiednie obszary krytyczne (zależnie od przypadku i hipotezy alternatywnej).

Przypadek a) $\sigma(X)$, $\sigma(Y)$ są znane

- $H_1: \bar{X} \neq \bar{Y}, \qquad |z| > z_{1-\alpha/2}$
- $H_1: \bar{X} > \bar{Y}, \qquad z > z_{1-\alpha}$

Przypadek b) $\sigma(X) = \sigma(Y)$ nie są znane

- $H_1: \bar{X} \neq \bar{Y}, \qquad |t| > t_{n_x+n_y-2,1-\alpha/2}$
- $H_1: \bar{X} > \bar{Y}, \qquad t > t_{n_x+n_y-2,1-\alpha}$

Przypadek c) $\sigma(X) \neq \sigma(Y)$ nie są znane

- $H_1: \bar{X} \neq \bar{Y}, \qquad |t| > t_{n_{ef},1-\alpha/2}$
- $H_1: \bar{X} > \bar{Y}, \qquad t > t_{n_{ef},1-\alpha}$

9.5. Analiza wariancji (test ANOVA)

W latach 20. ubiegłego wielu Ronald Fischer opracował metodę statystyczną służącą do weryfikacji hipotezy o równości wartości średnich wielu prób. Istotą testu jest próba odpowiedzi na pytanie, czy różnice między wartościami średnimi różnych prób mogą być wywołane przez czynniki zewnętrzne nawet o charakterze jakościowym. Test jest popularnie nazywany ANOVA od angielskich słów Analiysis of Variance (analiza wariancji). ANOVA znalazła szerokie zastosowanie zwłaszcza w naukach medycznych i biologicznych, w których często badamy wpływ czynników jakościowych na mierzalne własności jakiegoś obiektu. Dla przykładu możemy badać wpływ jakiegoś preparatu na grzyby powodujące choroby roślin. Chcemy odpowiedzieć na pytanie, czy badany preparat ma faktycznie wpływ na rozwój grzyba. W tym celu możemy przeprowadzić hodowlę dwóch identycznych porcji grzyba, dodając do jednej z tych porcji badany preparat. Hodowlę przeprowadzamy w tych samych warunkach i po pewnym czasie mierzymy średnice otrzymanych kolonii. Próby powtarzamy kilka razy i w końcu przeprowadzamy test ANOVA. W opisanym przypadku mamy do czynienia z jednym czynnikiem zewnętrznym. Taki test nazywamy jednoczynnikowym testem ANOVA. Istnieje także wersja wieloczynnikowego testu ANOVA, służącego do badania wpływu wielu czynników na własności próby. Poniżej opiszemy tylko jednoczynnikowy test ANOVA.
Założenia metody ANOVA są następujące:

- badamy *k* populacji, które możemy scharakteryzować zmienną losową *X*.,
- zakładamy, że zmienne $X_1, X_2, \dots, X_k$ odpowiadające poszczególnym populacjom podlegają rozkładowi normalnemu i są niezależne,
- wszystkie populacje mają równe wariancje.

Jeśli spełnienie powyższych założeń nie jest oczywiste, to musimy przeprowadzić odpowiednie testy stwierdzające, czy rozkłady są normalne (np. test λ-Kołmogorowa – patrz następne rozdziały), a wariancje równe (w przypadku dwóch populacji np. test Fischera-Snedecora omówiony w rozdziale 9.2, a w przypadku większej liczby populacji np. test Barletta).

W przypadku jednoczynnikowej wersji testu ANOVA, hipoteza zerowa brzmi następująco: wartości średnie wszystkich *k* populacji są sobie równe ($H_0: \bar{X}_1 = \bar{X}_2 = \cdots \bar{X}_k$). Hipoteza alternatywna mówi, że niektóre pary wartości średnich różnią się. Aby zapisać statystykę testową musimy wprowadzić kilka oznaczeń:

- n_i – liczebność *i* tej próby,
- $N \equiv \sum_{i=1}^{k} n_i$ – łączna liczebność wszystkich prób,
- x_{ij} – *j* ty pomiar *i* tej populacji,
- $\bar{x}_i$ – średnia arytmetyczne *i* tej populacji,
- $\bar{x}$ – średnia arytmetyczna wszystkich populacji $\bar{x} \equiv \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij} = \frac{1}{N}\sum_{i=1}^{k} n_i \bar{x}_i$,
- $s_b^2 \equiv \frac{1}{k-1}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{x}_i - \bar{x})^2 = \frac{1}{k-1}\sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})^2$ – estymator tzw. wariancji międzygrupowej (stąd indeks *b* od ang. *between*). Jest to estymator wariancji liczony ze średnich poszczególnych prób.

- $s_w^2 \equiv \frac{1}{N-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(x_{ij}-\bar{x}_i\right)^2$ – estymator tzw. wariancji wewnątrzgrupowej (stąd indeks *w* od ang. *within*). Jest to estymator wariancji liczonej dla wszystkich prób uwzględniającej rozrzut wewnątrz każdej z k populacji. Ponieważ liczba stopni swobody i tej populacji wynosi $n_i - 1$, to liczba stopni swobody sumy populacji wynosi $n_1 - 1 + n_2 - 1 + \cdots + n_k - 1 = N - k$.

Jeśli wariancje wszystkich populacji są równe $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_2^2 \equiv \sigma^2$ (co jest jednym z założeniem metody ANOVA), to można pokazać, że

$$E(s_w^2) = \sigma^2 \tag{9.17}$$

oraz

$$E(s_b^2) = \sigma^2 + \frac{\sum_{i=1}^{k}\left(E(x_i) - E(x)\right)^2}{k-1} \cdot \frac{N - \frac{1}{N}\sum_{i=1}^{k} n_i^2}{k-1}. \tag{9.18}$$

Jak widzimy estymator s_w^2 jest zawsze estymatorem nieobciążonym, nawet w przypadku, gdy hipoteza zerowa jest nieprawdziwa. Natomiast estymator s_b^2 jest nieobciążony tylko wówczas, gdy hipoteza zerowa jest prawdziwa, zaś w przeciwnym wypadku jest estymatorem obciążonym dodatnio. Stosunek obu estymatorów

$$F \equiv \frac{s_b^2}{s_w^2} \tag{9.19}$$

jest statystyką testu ANOVA. Zmienna F podlega rozkładowi Fischera-Snedecora, więc do wyznaczenia obszaru krytycznego używamy kwantyli $F(k-1, N-k)$ tego rozkładu. Ze względu na to, że s_b^2 jest dodatnio obciążony stosujemy jedynie test prawostronny. Hipotezę zerową musimy odrzucić jeśli przy zadanym poziomie istotności α zachodzi nierówność

$$F > F_{1-\alpha}(k-1, N-k). \tag{9.20}$$

Wyniki analizy wariancji ANOVA zapisujemy zwykle w postaci tabeli, której przykładem jest poniższa tabela.

Tabela 9.1. Tabela jednoskładnikowej analizy wariancji ANOVA

Rodzaj wariancji	SS (sum of squares) (suma kwadratów)	DF (degrees of fredom) (liczba stopni swobody)	MS (mean square) (średnia kwadrtowa)	F statystyka testowa
Międzygrupowa	Q_b	$k-1$	$s_b^2 = \frac{Q_b}{k-1}$	$F = \frac{s_b^2}{s^2}$

Wewnątrzgrupowa	Q_w	$N-k$	$s_w^2 = \frac{Q_w}{N-k}$	
Całkowita	Q	$N-1$	$s^2 = \frac{Q}{N-1}$	

Występujące w tabeli sumy kwadratów możemy wyliczyć za pomocą wzorów

$$Q = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(x_{ij} - \bar{x}\right)^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij}^2 - N\bar{x}^2,$$

$$Q_b = \sum_{i=1}^{k}\sum_{j=1}^{n_i} n_i(\bar{x}_i - \bar{x})^2 = \sum_{i=1}^{k} n_i\bar{x}_i^2 - N\bar{x}^2, \tag{9.21}$$

$$Q_w = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(x_{ij} - \bar{x}_i\right)^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij}^2 - \sum_{i=1}^{k} n_i\bar{x}_i^2 .$$

Do obliczeń zaleca się stosowanie prawej wersji powyższych wzorów, gdyż prowadzą one do dokładniejszych wartości, a pozornie małe zaokrąglenia mogą, w przypadku tej metody, bardzo zniekształcać wyniki. Obliczona suma kwadratów Q wykorzystywana jest do sprawdzenia poprawności obliczeń, gdyż musi zachodzić związek $Q = Q_b + Q_w$.

9.6. Test zgodności χ^2 Pearsona

Większość metod statystycznych opracowanych jest dla konkretnego rozkładu statystycznego badanych zmiennych losowych. Najczęściej jest to rozkład normalny. Przystępując do analizy danych nie zawsze znamy rozkład zmiennej losowej. Musimy zatem upewnić się, czy rozkład tej zmiennej jest rozkładem zgodnym z metodą, którą chcemy zastosować. Jednym z testów służących do porównywania doświadczalnego rozkładu prawdopodobieństwa z próby z rozkładem teoretycznym jest test χ^2 Pearsona. Test ma charakter uniwersalny, możemy go stosować do porównywania rozkładu doświadczalnego z dowolnym teoretycznym rozkładem zarówno ciągłej, jak i dyskretnej zmiennej losowej.

Załóżmy, że interesująca nas zmienna losowa X ma charakter ciągły. Niech $f(x)$ oznacza gęstość prawdopodobieństwa założonego teoretycznego rozkładu, a $F(x)$ jej dystrybuantę. Podzielmy zakres zmienności zmiennej losowej na r przedziałów (niekoniecznie o stałej szerokości) $\xi_1, \xi_2, \ldots, \xi_r$. Całkując gęstość prawdopodobieństwa w poszczególnych przedziałach dostajemy prawdopodobieństwa p_i zaobserwowania zmiennej losowej X w przedziale ξ_i

$$p_i = P(X \in \xi_i) = \int_{\xi_i} f(x)dx. \tag{9.22}$$

Oczywiście $\sum_{i=1}^{r} p_i = 1$.

W przypadku zmiennej dyskretnej odpada nam problem dzielenia zakresu zmienności zmiennej na przedziały i wyliczania prawdopodobieństw dla każdego przedziału, gdyż od razu znamy prawdopodobieństwa dla poszczególnych wartości zmiennej.

Z pobranej próby o liczebności n konstruujemy szereg rozdzielczy $n_1, n_2, \dots, n_r$, gdzie n_i oznacza liczbę elementów próby, które znalazły się w przedziale ξ_i (w przypadku zmiennej dyskretnej n_i oznacza liczbę wystąpień wartości x_i w próbie). Korzystając z teoretycznej gęstości prawdopodobieństwa otrzymujemy wartości oczekiwane dla poszczególnych n_i $E(n_i) = np_i$. Dla dużych wartości liczb n_i (w praktyce dla $n_i \geq 5$) ich wariancja jest równa n_i (patrz rozdział 4.5). Wobec tego zmienne

$$u'_i = \frac{n_i - E(n_i)}{\sigma_i} = \frac{n_i - np_i}{\sqrt{n_i}} \tag{9.23}$$

w granicy dużych n_i mają standardowy rozkład Gaussa. Jest tak również wówczas, gdy w mianowniku w powyższym wyrażeniu zastąpimy wielkości doświadczalne n_i ich wartościami oczekiwanymi $E(n_i) = np_i$

$$u_i = \frac{n_i - np_i}{\sqrt{np_i}}. \tag{9.24}$$

Jeśli teraz policzymy sumę kwadratów zmiennych u_i po wszystkich przedziałach ξ_i, to otrzymamy zmienną losową

$$X^2 = \sum_{i=1}^{r} u_i^2 = \sum_{i=1}^{r} \frac{(n_i - np_i)^2}{np_i}, \tag{9.25}$$

która ma asymptotycznie (dla dużych n) rozkład chi kwadrat. Liczba stopni swobody tego rozkładu jest równa $r - 1$, czyli o jeden mniejsza niż liczba zmiennych n_i ze względu na relację wiążącą te zmienne

$$\sum_{i=1}^{r} n_i = n. \tag{9.26}$$

Często dane pomiarowe wykorzystujemy jednocześnie do estymowania parametrów rozkładu. To oczywiście zmniejsza liczbę stopni swobody rozkładu. Ostatecznie, jeśli wyznaczanych parametrów jest q, to liczba stopni swobody rozkładu chi kwadrat zmiennej X zdefiniowanej wyrażeniem 9.25 wynosi $r - q - 1$.

Do znajdowania obszaru krytycznego (tylko test prawostronny) korzystamy z kwantyli $\chi^2_{r-q-1,1-\alpha}$ rozkładu chi kwadrat o $r - q - 1$ stopni swobody. Hipotezę o zgodności rozkładu z zakładanym musimy odrzucić, na poziomie istotności α, jeśli $X^2 > \chi^2_{r-q-1,1-\alpha}$.

Rozkład chi kwadrat jest jednym z najważniejszych testów w statystyce. Powyżej przedstawiono go jako test *nieparametryczny* zastosowany do porównywania rozkładu doświadczalnego z teoretycznym, ale często stosowany jest również do testowania hipotez nt. cech mierzalnych, jeśli tylko możemy skonstruować statystykę mającą rozkład chi kwadrat, np.

$$\chi^2 = \sum_{i=1}^{n} \frac{\big(x_i - E(x_i)\big)^2}{\sigma^2(x_i)}. \tag{9.27}$$

Test chi kwadrat jest jednym z najczęściej stosowanych testów we wszystkich naukach od fizyki po nauki społeczne.

9.7. Test zgodności λ Kołmogorowa-Smirnowa

Podobnie jak test chi kwadrat Pearsona, test Kołmorowa i jego zmodyfikowana wersja test Kołmogorowa-Smirnowa jest również test stosowany do porównywania doświadczalnego rozkładu prawdopodobieństwa z założonym rozkładem teoretycznym. Może on być stosowany do dowolnych rozkładów.

Po pobraniu *n* elementowej próby $x_1, x_2, \dots, x_n$ porządkujemy dane w kierunku rosnących wartości

$$x_1^* \le x_2^* \le \cdots \le x_n^*. \tag{9.28}$$

Następnie tworzymy tzw. empiryczną dystrybuantę zmiennej losowej *X*

$$F_n(x) = \begin{cases} 0, & \text{gdy } x \le x_1^* \\ \dfrac{m}{n}, & \text{gdy } \ x_m^* < x \le x_{m+1}^*, \quad 1 \le m \le n-1 \\ 1, & \text{gdy } \ x > x_n^* \end{cases}. \tag{9.29}$$

Dystrybuanta empiryczna jest zgodnym i nieobciążonym estymatorem dystrybuanty teoretycznej (przy spełnieniu hipotezy zerowej), gdyż można pokazać, że wartość oczekiwana empirycznej dystrybuanty jest równa dystrybuancie teoretycznej

$$E\big(F_n(x)\big) = F(x), \tag{9.30}$$

a jej wariancja dąży do zera, gdy rozmiar próby *n* dąży do nieskończoności

$$\lim_{n\to\infty} \sigma^2\big(F_n(x)\big) = \lim_{n\to\infty} \frac{1}{n} F(x)\big(1 - F(x)\big) = 0. \tag{9.31}$$

Statystyka testowa, w oryginalnej wersji zaproponowanej przez Kołmogorowa przyjmuje postać

$$D_n = \sup_x |F_n(x) - F(x)|, \tag{9.32}$$

gdzie sup oznacza kres górny, czyli najmniejsze górne ograniczenie zbioru. Smirnow zaproponował dwie inne statystyki (stąd test Kołmogorowa jest często nazywany testem Kołmogorowa-Smirnowa)

$$\begin{aligned} D_n^+ &= \sup_x \big(F_n(x) - F(x)\big), \\ D_n^- &= -\inf_x \big(F_n(x) - F(x)\big). \end{aligned} \tag{9.33}$$

W praktyce stosuje się inne wzory

$$D_n^+ = \max_{1\le m\le n} \left(\frac{m}{n} - F_n(x_m^*)\right), \tag{9.34}$$

$$D_n^- = \max_{1 \le m \le n} \left(F_n(x_m^*) - \frac{m-1}{n} \right),$$

$$D_n = \max(D_n^-, D_n^+).$$

Jeśli liczba prób jest co najmniej równa 10 ($n \geq 10$), a poziom istotności $\alpha \geq 0{,}01$, to kwantyle rozkładu Kołmogorowa $D_{n,1-\alpha}$ możemy wyliczyć z dużą dokładnością z następującego wzoru

$$D_{n,1-\alpha} \approx \sqrt{\frac{1}{2n}\left(y - \frac{2y^2 - 4y - 1}{18n}\right)} - \frac{1}{6n}, \tag{9.35}$$

gdzie

$$y = -\ln\frac{\alpha}{2}. \tag{9.36}$$

Przeprowadzając test (prawostronny) Kołmogorowa, hipotezę zerową musimy odrzucić na poziomie istotności α, gdy $D_n > D_{n,1-\alpha}$.

9.8. Test znaków

Ostatnim testem jaki omówimy jest test znaków. Test znaków służy do porównywania dystrybuant dwu ciągłych zmiennych losowych X, Y. Hipotezą zerową testu jest stwierdzenie: dystrybuanty rozkładów obu zmiennych losowych X, Y są takie same $H_0: F(X) = G(Y)$. Jeśli hipoteza zerowa jest prawdziwa, to prawdopodobieństwo $P(X > Y)$, że zachodzi zdarzenie $X > Y$ jest równe prawdopodobieństwu $P(X < Y)$, że zachodzi zdarzenie $X < Y$. Ponieważ obie zmienne są zmiennymi ciągłymi, prawdopodobieństwo $P(X = Y) = 0$. Z niezależności tych trzech zdarzeń i równości $P(X > Y) = P(X < Y)$ wynika, że

$$P(X > Y) = P(X < Y) = \frac{1}{2}. \tag{9.37}$$

W teście znaków liczebność próby dla zmiennej X musi być taka sama, jak liczebność próby dla Y. Nie porządkujemy ciągu wartości obu prób i porównujemy każdą parę wartości (x_i, y_i) stawiając znak +, gdy $x_i > y_i$ lub znak – w przeciwnym wypadku (stąd nazwa testu). Zgodnie z zależnością (9.37) spodziewamy się, że znaków + będzie tyle ile znaków –. Statystyką testową jest liczba k znaków +, czyli liczba par, dla których $x_i > y_i$ spośród wszystkich n par. Przy założeniu prawdziwości hipotezy zerowej statystyka k podlega rozkładowi dwumiennemu z parametrem $p = 1/2$

$$P(k) = \binom{n}{k}\frac{1}{2^k}\frac{1}{2^{n-k}} = \binom{n}{k}\frac{1}{2^n}. \tag{9.38}$$

Przy prawostronnym teście hipotezę zerową odrzucamy, gdy $k > k_p$, dla liczby granicznej k_p spełniającej warunek

$$P(k \geq k_p) = \frac{1}{2^n}\sum_{i=k_p}^{n}\binom{n}{i} = \alpha. \tag{9.39}$$

Graniczną liczbę k_l testu lewostronnego szukamy z warunku

$$P(k \leq k_l) = \frac{1}{2^n} \sum_{i=1}^{k_l} \binom{n}{i} = \alpha \tag{9.40}$$

i hipotezę zerową odrzucamy, gdy $k < k_p$. W przypadku testu dwustronnego wartości granicznych k_l i k_p szukamy z tych samych warunków podstawiając $\alpha/2$ w miejsce α.

W opisie testu zakładaliśmy milcząco, że żadna z par nie spełnia warunku $x_i = y_i$, gdyż zgodnie z założeniem zmienne są zmiennymi ciągłymi, a dla zmiennych ciągłych prawdopodobieństwo wystąpienia takich par jest zerowe. Jednak w praktyce zarówno wyniki pomiarów jak i obliczeń mają skończoną dokładność i pojawienie się par o identycznych wartościach jest możliwe. Jeżeli liczba takich par jest niewielka, to możemy je pominąć. W przeciwnym wypadku możemy na drodze losowania wybrać wynik porównania obu wartości.

Zadanie 9.1 (Test *F*-Fishera hipotezy o równości wariancji)

Za pomocą dwóch przyrządów pomiarowych wykonano pomiary pewnej wielkości fizycznej. Wyniki pomiarów są następujące

Przyrząd 1: 18; 18; 15; 27; 26; 22; 21; 19; 21.
Przyrząd 2: 16; 23; 22; 19; 21; 22; 20.

Stosując test obustronny *F*-Fishera-Snedecora przy wartości poziomu istotności 10% określ, czy możemy uważać, że dokładności obu przyrządów (czyli wariancje obu prób) są takie same

Zadanie 9.2 (Test Studenta hipotezy o średniej równej danej liczbie)

Z populacji o rozkładzie normalnym wylosowano 30 elementów. Przyjmując poziom istotności 10% zweryfikuj hipotezę stwierdzającą, że wylosowane elementy należą do populacji o wartości średniej równej 25,5. Wartości wylosowanych elementów są następujące

25,02; 26,12; 24,78; 26,19; 25,57; 25,33; 23,45; 23,5; 25,42; 22,45
27,56; 24,67; 22,51; 28,01; 24,77; 25,66; 25,02; 27,91; 21,98; 24,55
24,51; 22,55; 25,7; 25,03; 26,72; 24,32; 26,23; 28,11; 23,79; 25,05

Zadanie 9.3 (Test Studenta hipotezy o równości wartości średnich dwóch serii pomiarów)

Wykonano pomiar stężenia kwasu neuraminowego zawartego w czerwonych ciałkach krwi u pacjentów zmarłych na skutek pewnej choroby krwi (grupa x) oraz u osób zdrowych z grupy kontrolnej (grupa y). Przyjmując poziom istotności 5%, określ czy uzyskany materiał doświadczalny jest wystarczający do wykazania związku pomiędzy stężeniem kwasu neuraminowego a śmiercią pacjenta. Wyniki obu serii pomiarów (w jednostkach umownych) są następujące:

Grupa x: 21; 24; 18; 19; 25; 17; 18; 22; 21; 23; 18; 13; 16; 23; 22; 24.
Grupa y: 16; 20; 22; 19; 18; 19; 19.

Zadanie 9.4 (Test zgodności χ^2 Pearsona dla rozkładu dyskretnego)

Komora pęcherzykowa naświetlana jest wiązką kwantów γ wykorzystywanych do badań oddziaływań fotonów z protonami. Część fotonów w wyniku zderzeń kreuje pary elektron-pozyton. Ten uboczny efekt naświetlania komory jest wykorzystywany do monitorowania

wiązki fotonów. Częstość pojawiania się zdjęć przedstawiających 0, 1, 2 itd. par elektron-pozyton powinna podlegać rozkładowi Poissona $\left(f(k) = \frac{\lambda^k}{k!}e^{-\lambda}\right)$. Analizując odchylenia rozkładu doświadczalnego od rozkładu Poissona można wnioskować o istnieniu strat, które w konsekwencji będą prowadzić do błędów systematycznych prowadzonych eksperymentów. Przeprowadź test χ^2 hipotezy: *rozkład częstości występowania par elektron-pozyton jest zgodny z rozkładem Poissona.* Przyjmij poziom istotności $\alpha = 0{,}01$. Dane dla testu znajdują się w poniższej tabelce

Uwaga: Próba składa się z $r = 8$ elementów, na jej podstawie wyznaczymy jeden parametr rozkładu (λ), czyli $p = 1$. A zatem liczba stopni swobody rozkładu χ^2 wynosi $f = r - 1 - p = 6$. Dla poziomu istotności $\alpha = 0{,}01$ i liczby stopni swobody $f = 6$ kwantyl rozkładu χ^2 jest równy $\chi^2_{1-0{,}01}(6) = \chi^2_{0{,}99}(6) = 16{,}81$.

k – liczba par e-p na zdjęciu	n_k – liczba zdjęć zawierających k par	np_k – wartość oczekiwana liczby k par e-p zgodnia z rozkładem Poissona	$\frac{(n_k - np_k)^2}{np_k}$
0	44		
1	75		
2	80		
3	73		
4	52		
5	18		
6	9		
7	1		
8	2		
	$n =$		$X^2 =$

Porównaj za pomocą histogramu rozkład doświadczalny z rozkładem teoretycznym.

Zadanie 9.5 (Test zgodności χ^2 Pearsona dla rozkładu ciągłego)

Wykonano pomiary kontrolnej próby oporników. Dane zebrano w poniższej tabeli.

193,199	195,673	195,757	196,051	196,092
196,596	196,679	196,763	196,847	197,267
197,392	197,477	198,189	198,65	198,944
199,070	199,111	199,153	199,237	199,698
199,572	199,614	199,824	199,908	200,118
200,160	200,234	200,285	200,453	200,704
200,746	200,830	200,872	200,914	200,956
200,998	200,998	201,123	201,208	201,333
201,375	201,543	201,543	201,584	201,711
201,878	201,919	202,004	202,004	202,088
202,172	202,172	202,297	202,339	202,281
202,507	202,591	202,633	202,716	202,884
203,051	203,052	203,094	203,094	203,177
203,178	203,219	203,764	203,765	203,848
203,890	203,974	204,184	204,267	204,352
204,352	204,729	205,106	205,148	205,231
205,357	205,400	205,483	206,070	206,112
206,154	206,155	206,616	206,665	206,993

207,243	207,621	208,124	208,375	208,502
208,628	208,670	208,711	210,012	211,394

a) przedstaw wyniki tych pomiarów w postaci jednowymiarowego wykresu punktowego oraz w postaci histogramu o szerokości przedziałów $2\ \Omega$,
b) korzystając z testu χ^2 zweryfikuj hipotezę zerową: zmienna losowa x podlega rozkładowi Gaussa. Przyjmij poziom istotności $\alpha = 0{,}01$. Podziel zakres zmienności zmiennej losowej x na przedziały o szerokości $2\ \Omega$.

W celu wyliczenia χ^2 w punkcie b) skonstruuj tabelę zawierającą

x_k	n_k	Ψ_{0+}	Ψ_{0-}	np_k	$\frac{(n_k - np_k)^2}{np_k}$
193					
195					
…					

gdzie

$$\Psi_{0\pm} \equiv \Psi_0\left(\frac{x_k \pm \frac{1}{2}\Delta x - \tilde{a}}{\tilde{\sigma}}\right),$$

Ψ_0 jest dystrybuantą standaryzowanego rozkładu normalnego, $\Delta x = 2\ \Omega$, $\tilde{a}$ jest średnią wartością próby, a $\tilde{\sigma}$ odchyleniem standardowym próby znalezione metodą największej wiarygodności.

Zadanie 9.7 (Test zgodności λ Kołmogorowa-Smirnowa)

Dla danych z zadania 9.6 wykonaj test Kołmogorowa-Smirnowa dla poziomu istotności $\alpha = 0{,}01$ i hipotezy zerowej: *zmienna losowa podlega rozkładowi Gaussa*. Narysuj wykres empirycznej dystrybuanty dla znormalizowanej zmiennej i nałóż na nią dystrybuantę standaryzowanego rozkładu normalnego.

Zadanie 9.8 (Test znaków)

Wykonano 10 niezależnych pomiarów zmiennej X i 10 niezależnych pomiarów zmiennej Y. Wyniki pomiarów w kolejności ich otrzymywania zebrano w poniższej tabeli

X	7,48	7,59	7,81	6,99	9,64	8,84	7,77	6,69	7,16	6,76
Y	5,34	8,21	7,68	8,44	5,22	7,28	6,46	8,44	6,36	6,38

Stosując test znaków na poziomie istotności $\alpha = 0{,}01$ oraz $\alpha = 0{,}1$ zweryfikuj hipotezę: *rozkłady obu zmiennych są jednakowe*.

10. Metoda najmniejszych kwadratów

Za twórców metody najmniejszych kwadratów uznaje się Legendre'a i Gaussa, którzy wprowadzili ją niezależnie od siebie. Ogólnie mówiąc jest to metoda służąca do wyrównywania danych statystycznych. Wartości w danej próbie statystycznej np. serii pomiarów fizycznych obarczone są błędami. Metoda najmniejszych kwadratów ma za zadanie, w pewnym sensie, zminimalizować wpływ tych błędów. Wynik i-tego pomiaru x_i w serii można uznać za

sumę wartości dokładnej x, której nie znamy oraz błędu pomiarowego ε_i, którego również nie znamy

$$x_i = x + \varepsilon_i. \tag{10.1}$$

Zgodnie z metodą najmniejszych kwadratów dobieramy wartości błędów ε_i tak, aby suma kwadratów błędów była najmniejsza

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (x_i - x)^2 = \min. \tag{10.2}$$

Na metodzie najmniejszych kwadratów opiera się teoria błędów pomiarowych. Początkowo była ona stosowana w astronomii i geodezji, ale szybko znalazła zastosowanie w innych dziedzinach nauki. W statystyce najczęściej używana jest do estymacji i wyznaczania tzw. linii trendu.

10.1. Pomiary bezpośrednie o równej dokładności

Załóżmy, że błędy pomiarów ε_i mają rozkład normalny z wartością oczekiwaną równą zeru

$$x_i = x + \varepsilon_i, \quad E(\varepsilon_i) = 0, \quad E(\varepsilon_i^2) = \sigma^2. \tag{10.3}$$

Błędy są na ogół sumą bardzo wielu przyczynków i dla tego, na mocy *centralnego twierdzenia granicznego* nasze założenie jest na ogół uzasadnione. Prawdopodobieństwo, że wynik pojedynczego *i*-tego pomiaru leży wewnątrz infinitezymalnie wąskiego przedziału $(x_i, x_i + dx)$ wynosi

$$f(x_i)dx = \frac{1}{\sigma\sqrt{2\pi}} e^{\left(-\frac{(x_i - x)^2}{2\sigma^2}\right)} dx. \tag{10.4}$$

Skorzystajmy teraz z estymacji metodą największej wiarygodności. Logarytmiczna funkcja wiarygodności liczona dla serii *n* pomiarów wynosi

$$l = \sum_{i=1}^{n} \left(\ln \frac{1}{\sigma\sqrt{2\pi}} - \frac{(x_i - x)}{2\sigma^2} \right) = n \ln \frac{1}{\sigma\sqrt{2\pi}} - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - x)^2. \tag{10.5}$$

Zgodnie z metodą największej wiarygodności szukamy takiej wartości x mierzonej wielkości, dla której logarytmiczna funkcja największej wiarygodności l osiągnie maksimum. Z Równania 10.5 widzimy, że warunek ten będzie spełniony wówczas, gdy drugi z wyrazów tego wyrażenia osiągnie minimum

$$\sum_{i=1}^{n} (x_i - x)^2 = \sum_{i=1}^{n} \varepsilon_i^2 = \min. \tag{10.6}$$

Jest to właśnie warunek najmniejszych kwadratów. Spełnienie tego warunku prowadzi do wniosku, że najlepszym estymatorem dla wartości x jest średnia arytmetyczna wyników pomiaru x_i

$$\tilde{x} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{10.7}$$

Wariancja tego estymatora, co pokazaliśmy już w rozdziale 8 wyraża się wzorem

$$\sigma^2(\bar{x}) = \frac{\sigma^2}{n}. \tag{10.8}$$

Jeśli przyjmiemy odchylenie standardowe średniej arytmetycznej jako miarę niepewności pomiaru, to otrzymamy

$$\Delta\bar{x} = \frac{\sigma}{\sqrt{n}}. \tag{10.9}$$

10.2. Pomiary bezpośrednie o różnej dokładności

Załóżmy ponownie, że błędy mają rozkład normalny wartości średniej równej zeru, ale teraz poszczególne pomiary mogą mieć różną dokładność

$$x_i = x + \varepsilon_i, \quad E(\varepsilon_i) = 0, \quad E(\varepsilon_i^2) = \sigma_i^2. \tag{10.10}$$

Teraz warunek największej wiarygodności wymaga, aby

$$\sum_{i=1}^{n} \frac{(x_i - x)^2}{\sigma_i^2} = \sum_{i=1}^{n} w_i \cdot (x_i - x)^2 = \sum_{i=1}^{n} w_i \varepsilon_i^2 = \min, \tag{10.11}$$

gdzie czynnik w_i równy odwrotności wariancji

$$w_i = \frac{1}{\sigma^2} \tag{10.12}$$

jest wagą poszczególnych składników sumy. Wyliczając estymator wartości x spełniający warunek 10.11 dostajemy tzw. średnią ważoną

$$\tilde{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}. \tag{10.13}$$

Jak widzimy pomiary o małej wariancji, czyli mające większą wagę mają większy wpływ na wynik pomiaru niż pomiary o dużej wariancji. Wariancja estymatora 10.13 jest równa

$$\sigma^2(\tilde{x}) = \frac{1}{\sum_{i=1}^{n}\left(\frac{1}{\sigma_i^2}\right)} = \frac{1}{\sum_{i=1}^{n} w_i}. \tag{10.14}$$

Pierwiastek tego wyrażenia możemy uznać za miarę niepewności pomiaru wielkości mierzonej x.
Estymatorem błędów ε_i będzie

$$\tilde{\varepsilon}_i = x_i - \tilde{x}. \tag{10.15}$$

Jeśli pomiary wolne są od błędów systematycznych, to możemy założyć, że wielkości $\tilde{\varepsilon}_i$ mają rozkład normalny z wartością średnią równą zeru i wariancją σ_i^2. Wobec tego wielkości $\tilde{\varepsilon}_i/\sigma_i^2$ podlegają standardowemu rozkładowi Gaussa. A jeśli tak, to wielkość

$$M = \sum_{i=1}^{n} \left(\frac{\tilde{\varepsilon}_i}{\sigma_i}\right)^2 = \sum_{i=1}^{n} \frac{(x_i - \tilde{x})^2}{\sigma_i^2} = \sum_{i=1}^{n} w_i (x_i - \tilde{x})^2 \tag{10.16}$$

ma rozkład chi kwadrat o $n-1$ stopniach swobody. Powyższy fakt może nam posłużyć do weryfikacji hipotezy o słuszności naszych założeń, w myśl których błędy pomiarowe mają rozkład Gaussa o średniej równej zeru i wariancji σ_i^2. W tym celu wyliczamy wartość statystyki M, zakładamy jakiś poziom istotności α i sprawdzamy czy M leży w obszarze krytycznym. Jeśli tak, to zachodzi podejrzenie, że oprócz czynników przypadkowych na wynik pomiaru mają wpływ również czynniki systematyczne, w wyniku czego średnia błędów nie jest zerowa.

10.3. Regresja liniowa

Regresją nazywamy metodę estymowania wartości oczekiwanej zmiennej Y dla znanej wartości zmiennej (lub zmiennych) X. Zmienna Y nazywana jest zwykle *zmienną objaśnianą* lub *zmienną zależną*, a zmienna X *zmienną objaśniającą* lub *zmienną niezależną*. W dalszych rozważaniach ograniczymy się przypadku jednej zmiennej objaśniającej. W szczególnym przypadku, jeśli zakładaną zależnością miedzy zmienną zależną, a zmienną niezależną jest zależność liniowa mówimy o *regresji liniowej*. Problem polega na ocenie wartości parametrów a, b oraz ich wariancji $\sigma^2(a), \sigma^2(b)$ prostej

$$y = ax + b, \tag{10.17}$$

która najlepiej dopasowuje się do zbioru n punktów doświadczalnych o współrzędnych $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. Nasz model statystyczny możemy zapisać w postaci

$$y_i = ax_i + b_i + \varepsilon_i, \tag{10.18}$$

gdzie ε_i jest błędem. Zakładamy, że wartość oczekiwana błędów jest zerowa $E(\varepsilon_i) = 0$ a rozkładem prawdopodobieństwa błędów jest rozkład normalny. Najstarszą i najczęściej stosowaną metodą rozwiązania tego problemu jest metoda najmniejszych kwadratów. Zgodnie z tą metodą szukamy takich wartości parametrów a, b, dla których suma kwadratów błędów jest minimalna.

10.4. Regresja liniowa w przypadku stałych wariancji zmiennych

Załóżmy że wariancje obu zmiennych nie zależą od wskaźnika i. Dodatkowo zachodzi jeden z przypadków

a) wariancje obu zmiennych są identyczne lub
b) zmienna X jest tzw. *zmienną kontrolowaną* czyli jest pozbawiona błędów lub jej błędy względne są dużo mniejsze od błędów względnych zmiennej Y: $S_x/x_i \ll S_y/y_i$.

Przypadek a) będziemy nazywać *regresją zwyczajną*, a przypadek b) *regresją klasyczną*. Przy spełnieniu powyższych założeń warunek minimalizacji sumy kwadratów błędów prowadzi do równania

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (ax_i + b - y_i)^2 = \min. \tag{10.19}$$

Warunkiem koniecznym istnienia minimum funkcji

$$f(a,b) = \sum_{i=1}^{n} (ax_i + b - y_i)^2 \tag{10.20}$$

jest zerowanie się pochodnych cząstkowych tej funkcji po argumentach a, b

$$\frac{\partial f(a,b)}{\partial a} = 2\sum_{i=1}^{n} x_i \cdot (ax_i + b - y_i) = 0 \tag{10.21}$$

oraz

$$\frac{\partial f(a,b)}{\partial b} = 2\sum_{i=1}^{n} (ax_i + b - y_i) = 0. \tag{10.22}$$

Jest to prosty układ dwóch równań liniowych na parametry a, b. Jego rozwiązanie można zapisać w różnych postaciach, przy czym do praktycznych obliczeń używa się najczęściej wzorów:

$$\begin{aligned} \bar{a} &= \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2}, \\ \bar{b} &= \frac{(\sum y_i - \bar{a}\sum x_i)}{n}, \end{aligned} \tag{10.23}$$

gdzie wszystkie sumowania przebiegają po wartościach indeksów od $i = 1$ do $i = n$. Postać estymatorów odchyleń standardowych obu parametrów będzie zależeć od, tego z którym przypadkiem mamy do czynienia. Jeśli wariancje obu zmiennych są identyczne i nie zależą od i, to estymatory te są równe

$$\begin{aligned} S_{\bar{a}} &= \sqrt{\frac{n\left(\sum y_i^2 - \bar{a}\sum x_i y_i - \bar{b}\sum y_i\right)}{(n-2)(n\sum x_i^2 - (\sum x_i)^2)}}, \\ S_{\bar{b}} &= S_{\bar{a}} \cdot \sqrt{\frac{\sum x_i^2}{n}}. \end{aligned} \tag{10.24}$$

Zaś w przypadku, gdy zmienna X jest tzw. zmienną kontrolowaną (pozbawioną błędów), a odchylenia standardowe zmiennej Y są równe S_y dla wszystkich y_i, estymatory odchyleń standardowych parametrów a, b są równe

$$S_{\bar{a}} = S_y \cdot \sqrt{\frac{n}{n\sum x_i^2 - (\sum x_i)^2}},$$
$$S_{\bar{b}} = S_{\bar{a}} \cdot \sqrt{\frac{\sum x_i^2}{n}}. \tag{10.25}$$

Jeśli mamy do czynienia z odwrotnym przypadkiem, tzn. zmienna Y jest zmienną kontrolowaną (czyli pozbawioną błędów), to należy odwrócić zależność liniową i szukać parametrów funkcji

$$x = a'y + b'. \tag{10.26}$$

Stosujemy przy tym wzory podane wyżej zastępując x_i przez y_i i odwrotnie.

Przykład

W celu wyznaczenia temperaturowego współczynnika oporu żelaza wykonano pomiary oporu elektrycznego żelaznego przewodnika w funkcji przyrostu temperatury $\Delta T = T - T_0$, gdzie $T_0 = 301$ K. Otrzymane wyniki pomiarów przedstawia tabela 10.1.
Dla większości metali zależność temperaturowa oporu elektrycznego od temperatury ma charakter liniowy

$$R(T) = R_0(1 + \alpha\Delta T), \quad \Delta T = T - T_0, \tag{10.27}$$

gdzie R_0 jest oporem w temperaturze T_0, α jest temperaturowym współczynnikiem oporu. Po wykonaniu obliczeń zgodnie ze wzorami 10.23, 10.24 otrzymujemy następujące wyniki: $\bar{a} = \bar{R}_0\bar{\alpha} = 0{,}605$ [Ω/K], $\bar{b} = R_0 = 95{,}9$ [Ω] oraz $S_{\bar{a}} = 0{,}017$ [Ω/K], $S_{\bar{b}} = 0{,}58$ [Ω]. Stąd możemy wyznaczyć szukaną wartość temperaturowego współczynnika temperatury dla żelaza:

$$\bar{\alpha} = \frac{\bar{a}}{\bar{R}_0} \approx 0{,}0060309 \ \ [\mathrm{K}^{-1}]$$

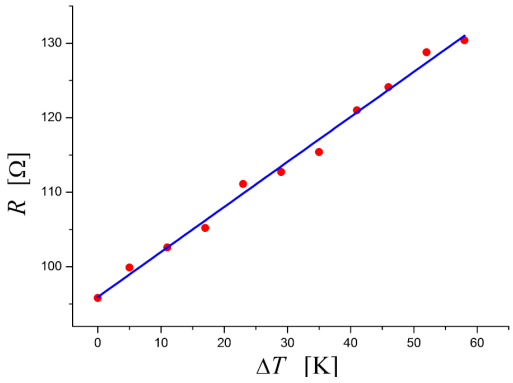
z odchyleniem standardowym równym

$$S_{\bar{\alpha}} = \sqrt{\left(\frac{\partial\alpha}{\partial a}\right)^2 S_{\bar{a}}^2 + \left(\frac{\partial\alpha}{\partial R_0}\right)^2 S_{\bar{R}_0}^2} = \bar{\alpha} \cdot \sqrt{\left(\frac{S_{\bar{a}}}{\bar{a}}\right)^2 + \left(\frac{S_{\bar{R}_0}}{\bar{R}_0}\right)^2} \approx 0{,}00168\ [\mathrm{K}^{-1}].$$

Tabela 10.1. Wyniki pomiaru oporu elektrycznego żelaznego przewodnika w funkcji temperatury.

T [K]	ΔT [K]	R [Ω]
301	0	95,8
278	5	99,9
284	11	102,6
290	17	105,2
296	23	111,1
302	29	112,7

308	35	115,4
314	41	121,0
319	46	124,1
325	52	128,8
331	58	130,4



Rysunek 10.1. Wykres przedstawia punkty pomiarowe oporu elektrycznego przewodnika metalowego w funkcji temperatury (czerwone kółeczka) oraz dopasowaną do nich metodą regresji liniowej prostą (niebieska linia).

10.5. Regresja liniowa w przypadku wariancji zmiennych zależnych od wartości zmiennych

Rozważmy teraz sytuację, gdy wariancje błędów poszczególnych punktów są różne, tzn. sytuację, gdy przy tym samym założeniu co poprzednio, że $E(\varepsilon_i) = 0$, mamy $E(\varepsilon_i^2) = \sigma_i^2$. Wówczas warunek minimalizacji sumy kwadratów błędów będzie wyglądał następująco:

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \frac{(ax_i + b - y_i)^2}{\sigma_i^2} = \min. \tag{10.28}$$

Podobnie jak poprzednio musimy rozważyć dwa przypadki. W pierwszym przypadku jedna ze zmiennych, załóżmy, że X, jest zmienną kontrolowaną, czyli $S_x/x_i \ll S_y/y_i$. Tego typu regresję będziemy nazywać *regresją ważoną*. W drugim przypadku błędy względne obu zmiennych będą porównywalne $S_x/x_i \cong S_y/y_i$. Jest to tzw. *regresja efektywna*. W obu przypadkach rozwiązanie problemu 10.27 prowadzi do następujących wzorów na parametry a, b

$$\bar{a} = \frac{\sum w_i \sum w_i x_i y_i - \sum w_i x_i \sum w_i y_i}{\sum w_i \sum w_i x_i^2 - (\sum w_i x_i)^2},$$
$$\bar{b} = \frac{\sum w_i y_i - \bar{a} \sum w_i x_i}{\sum w_i} \tag{10.29}$$

oraz na odchylenia standardowe tych parametrów

$$S_{\bar{a}} = \sqrt{\frac{\sum w_i}{n-2} \cdot \frac{\sum w_i y_i^2 - \bar{a}\sum w_i x_i y_i - \bar{b}\sum w_i y_i}{\sum w_i \sum w_i x_i^2 - (\sum w_i x_i)^2}},$$

$$S_{\bar{b}} = S_{\bar{a}} \cdot \sqrt{\frac{\sum w_i x_i^2}{\sum w_i}}. \tag{10.30}$$

W powyższych wzorach w_i są wagami statystycznymi $w_i = 1/\sigma_i^2$. W przypadku *regresji ważonej* $\sigma_i^2 = \sigma_{y_i}^2$ i wagi statystyczne są równe

$$w_i = \frac{1}{\sigma_{y_i}^2} \cong \frac{1}{S_{y_i}^2}. \tag{10.31}$$

W przypadku *regresji efektywnej* możemy wprowadzić efektywną wariancję błędów $\sigma_{ef_i} = \sigma_{y_i}^2 + a\sigma_{x_i}^2$, czyli wagi statystyczne są równe

$$w_i = \frac{1}{\sigma_{y_i}^2 + a\sigma_{x_i}^2} \cong \frac{1}{S_{y_i}^2 + aS_{x_i}^2}. \tag{10.32}$$

Zwróćmy uwagę na fakt, że przedstawione powyżej wzory zadziałają niezależnie od tego czy dane jakich użyliśmy faktycznie powiązane są zależnością liniową, gdyż metoda najmniejszych kwadratów zawsze znajdzie nam parametry funkcji, przy których suma błędów będzie minimalna. Nie oznacza to jednak, że otrzymana prosta (czy inna funkcja w przypadku regresji innego typu) będzie dobrze przybliżać układ naszych punktów pomiarowych. W skrajnych przypadkach zauważymy to wyraźnie przedstawiając nasze dane i prostą regresji graficznie. Jeśli taka ocena „na oko" nam nie wystarcza możemy przeprowadzić np. test chi kwadrat. Jako statystykę wybieramy

$$\chi^2 = \sum_{i=1}^{n} \frac{\left(y_i - f(x_i)\right)^2}{\sigma_i^2} \approx \sum_{i=1}^{n} \frac{\left(y_i - \bar{a}x_i - \bar{b}\right)^2}{S_{y_i}^2} \tag{10.33}$$

i sprawdzamy, czy z założonym poziomem istotności wartość statystyki leży poza obszarem krytycznym.

10.6. Regresja nieliniowa

W przypadku, gdy zależność między zmiennymi jest nieliniowa, warunek minimalizacji sumy kwadratów błędów prowadzi do układu równań nieliniowych, których na ogół nie da się rozwiązać analitycznie. Można to zrobić jedynie przybliżonymi metodami numerycznymi. Jest to trudna operacja dlatego wymyślono inną metodę. Polega ona na linearyzacji naszej funkcji $f(x, a_0, a_1, \dots, a_M)$ ze względu na szukane parametry i iteracyjnym dochodzeniu do ich faktycznych wartości. Załóżmy, że $a_{0,j}, a_{1,j}, \dots, a_{M,j}$ są wartościami parametrów w j-tym kroku iteracyjnym. Rozwijamy funkcję f w szereg Taylora względem parametrów urywając na wyrazach liniowych względem $\Delta a_k = a_{k,j+1} - a_{k,j}$

$$f(x_i)_{j+1} = f(x_i)_j + \sum_{k=1}^{M} \frac{\partial f(x_i)_j}{\partial a_k} \Delta a_k, \tag{10.34}$$

gdzie

$$\frac{\partial f(x_i)_j}{\partial a_k} \equiv \left. \frac{\partial f\left(x, a_{0,j}, a_{1,j}, \ldots, a_{M,j}\right)}{\partial a_k} \right|_{x=x_i}. \tag{10.35}$$

W ten sposób, w każdym kroku iteracyjnym, sprowadzamy problem do problemu liniowego ze względu na parametry, którymi są teraz Δa_k. Taki problem możemy rozwiązać metodą regresji liniowej. Po znalezieniu parametrów Δa_k wyliczamy nowe wartości interesujących nas parametrów

$$a_{k,j+1} = a_{k,j} + \Delta a_k, \tag{10.36}$$

po czym cały proces się powtarza. Obliczenia kończymy, gdy przyrosty Δa_k spadną do założonej na wstępie małej wartości.

Bardzo często zależność nieliniowa ma postać, którą, poprzez odpowiednią zamiany zmiennych można łatwo sprowadzić do postaci liniowej. W takiej sytuacji dokonujemy transformacji i przeprowadzamy regresję liniową dla nowych zmiennych, a następnie poprzez transformację odwrotną znajdujemy interesujące na parametry.

Przykład

Zakładany związek między zmiennymi x, y jest następujący

$$y = \frac{a}{x^2} + b. \tag{10.37}$$

Dokonajmy zamiany zmiennych

$$u = \frac{1}{x^2}. \tag{10.38}$$

Zależność między zmiennymi y i u jest liniowa

$$y = au + b \tag{10.39}$$

możemy zatem skorzystać z metody regresji liniowej. Zauważmy, że zarówno parametry prostej regresji 10.38, jak i ich odchylenia standardowe (poza przypadkiem regresji efektywnej) są identyczne z tymi, które nas interesują, czyli dla regresji nieliniowej (10.36). Nie musimy więc dokonywać żadnych dodatkowych obliczeń.

Przykład

Niech zależność zmiennej x od zmiennej y ma postać

$$y = Be^{ax}. \tag{10.40}$$

Problem możemy zlinearyzować jeśli zlogarytmujemy obustronnie zależność (10.39)

$$\ln y = ax + \ln B \tag{10.41}$$

i wprowadzimy nowe oznaczenia

$$u = \ln y\,, \quad b = \ln B. \tag{10.42}$$

Teraz problem regresji liniowej 10.39 sprowadza się do regresji liniowej

$$u = ax + b. \tag{10.43}$$

Po rozwiązaniu tego problemu wracamy do zależności pierwotne, przy czym z parametrem a i jego odchyleniem standardowym nie musimy już nic robić. Natomiast parametr B i jego odchylenie standardowe znajdujemy ze wzorów

$$\begin{aligned} B &= e^b, \\ S_B &= \frac{dB}{db} \cdot S_b = e^b S_b. \end{aligned} \tag{10.44}$$

10.1. Wariancje parametrów metody najmniejszych kwadratów

Omawiając regresję liniową w podrozdziałach 10.4 i 10.5 podaliśmy bez wyprowadzenia wzory na odchylenia standardowe $S_{\bar{a}}$ i $S_{\bar{b}}$. parametrów a, b funkcji liniowej. Przyjrzyjmy się teraz, w jaki można wyprowadzić wzory odchyleń standardowych parametrów funkcji w metodzie najmniejszych kwadratów (nie tylko w przypadku regresji liniowej). Załóżmy, że szukamy optymalnych, w sensie metody najmniejszych kwadratów, parametrów $a_1, a_2, \dots, a_M$ funkcji

$$y(x) = y(x, a_1, a_2, \dots, a_M). \tag{10.45}$$

Załóżmy, że dysponujemy próbą statystyczną złożoną z N punktów (x_i, y_i) $i = 1,2, \dots, N$. Dodatkowo dla uproszczenia dalszych rozważań zakładamy, że wartości zmiennej x są dokładne (ich wariancje są zerowe). Nasze zadanie sprowadza się do znalezienia parametrów $a_1, a_2, \dots, a_M$, które minimalizują funkcję

$$\chi^2(a_1, a_2, \dots, a_M) \equiv \sum_{i=1}^{N} \left(\frac{y_i - y(x_i, a_1, a_2, \dots, a_M)}{\sigma_i} \right)^2, \tag{10.46}$$

gdzie σ_i^2 są wariancjami zmiennych y_i. Jeśli dla każdego i $\sigma_i = \sigma$

$$\sigma = \sum_{i=1}^{N} \frac{\left(y_i - y(x_i, a_1, a_2, \dots, a_M)\right)^2}{N - M}, \tag{10.47}$$

to wystarczy minimalizować funkcję

$$\chi^2(a_1, a_2, \dots, a_M) \equiv \sum_{i=1}^{N} [y_i - y(x_i, a_1, a_2, \dots, a_M)]^2. \tag{10.48}$$

W tym celu obliczamy pochodne cząstkowe funkcji $\chi^2(a_1, a_2, \dots, a_M)$ po poszczególnych parametrach i przyrównujemy do zera[6]. W ten sposób dostajemy układ M równań, z których wyliczamy wartości szukanych parametrów $a_1, a_2, \dots, a_M$. Każdy z nich jest funkcją N zmiennych $x_1, x_2, \dots, x_N$ i N zmiennych $y_1, y_2, \dots, y_N$, ale ponieważ założyliśmy, że wariancje wartości zmiennej x są zerowe, to interesuje nas jedynie zależność

$$a_i(y_1, y_2, \dots, y_N), \quad i = 1,2, \dots, M. \tag{10.49}$$

W związku z tym wariancję parametrów a_i musimy policzyć następująco

$$\sigma^2(a_k) = \sum_{i=1}^{N} \left(\frac{\partial a_k}{\partial y_i}\right) \sigma_i^2, \tag{10.50}$$

gdzie σ_i^2 jest wariancją zmiennej y_i.

Przykład

Zastosujmy przedstawiony powyżej algorytm do regresji liniowej z funkcją typu

$$y(x, a) = ax. \tag{10.51}$$

Załóżmy, że zmienne x_i naszej próby statystycznej (x_i, y_i), $i = 1,2, \dots, N$ mają zerowe wariancje, zaś zmienne y_i mają jednakowe wariancje $\sigma_i^2 = \sigma^2$. Musimy zminimalizować funkcję

$$\chi^2(a) = \sum_{i=1}^{N} (y_i - ax_i)^2. \tag{10.52}$$

Z warunku koniecznego istnienia minimum mamy

$$\frac{\partial \chi^2(a)}{\partial a} = 2\sum_{i=1}^{N} x_i(y_i - ax_i) = 2\sum_{i=1}^{N} x_i y_i - 2a\sum_{i=1}^{N} x_i^2 = 0. \tag{10.53}$$

A stąd dostajemy wartość szukanego parametru

$$a = \frac{\sum_{i=1}^{N} x_i y_i}{\sum_{i=1}^{N} x_i^2}. \tag{10.54}$$

Wariancja tego parametru wynosi

$$\sigma^2(a) = \sum_{i=1}^{N} \left(\frac{\partial a}{\partial y_i}\right) \sigma^2. \tag{10.55}$$

Pochodna występująca w tym wzorze jest równa

[6] Jest to jak wiemy warunek konieczny istnienia minimum (lub maksimum), ale niewystarczający, więc należy jeszcze sprawdzić warunek wystarczający zgodnie z zasadami szukania ekstremum lokalnego funkcji wielu zmiennych.

$$\frac{\partial a}{\partial y_i} = \frac{x_i}{\sum_{i=1}^{N} x_i^2}. \tag{10.56}$$

A zatem

$$\sigma^2(a) = \frac{\sigma^2}{(\sum_{i=1}^{N} x_i^2)^2} \sum_{i=1}^{N} x_i^2 = \frac{\sigma^2}{\sum_{i=1}^{N} x_i^2}. \tag{10.57}$$

Wariancję zmiennych y_i znajdziemy ze wzoru

$$\sigma = \frac{\sum_{i=1}^{N}(y_i - ax_i)^2}{(N-1)\sum_{i=1}^{N} x_i^2}, \tag{10.58}$$

czyli

$$\begin{aligned} \sigma^2(a) &= \frac{\sum_{i=1}^{N} y_i^2 - 2a\sum_{i=1}^{N} x_i y_i + a^2 \sum_{i=1}^{N} x_i^2}{(N-1)\sum_{i=1}^{N} x_i^2} = \\ &= \frac{\sum_{i=1}^{N} y_i^2 - 2a\sum_{i=1}^{N} x_i y_i + a\frac{\sum_{i=1}^{N} x_i y_i}{\sum_{i=1}^{N} x_i^2}\sum_{i=1}^{N} x_i^2}{(N-1)\sum_{i=1}^{N} x_i^2}. \end{aligned} \tag{10.59}$$

Ostatecznie

$$\sigma^2(a) = \frac{\sum_{i=1}^{N} y_i^2 - a\sum_{i=1}^{N} x_i y_i}{(N-1)\sum_{i=1}^{N} x_i^2} \tag{10.60}$$

Zadanie 10.1

Poniżej zebrano wyniki pomiarów pewnej wielkości fizycznej wykonane przez różnych eksperymentatorów. Uśrednij te wyniki (wylicz średnią ważoną i jej odchylenie standardowe). Wykonaj test chi kwadrat na poziomie istotności $\alpha = 0{,}1$, aby sprawdzić, czy uśrednienie jest uzasadnione.

L.p.	x_i	S_{x_i}
1	450,2	1,7
2	447,7	1,1
3	449,4	0,82
4	450,1	2,4
5	450,2	0,85
6	452,6	2,1
7	449,4	0,66

Zadanie 10.2

Wykonano pomiary pewnej zmiennej y w funkcji innej zmiennej x. Teoretyczna zależność między zmiennymi ma charakter liniowy $y = ax + b$. Wyniki pomiarów zebrano w poniższej tabeli. Wylicz metodą zwyczajnej regresji liniowej parametry a, b oraz ich odchylenia stan-

dardowe. Przeprowadź test chi kwadrat na poziomie istotności $\alpha = 0{,}1$ dla sprawdzenia hipotezy zerowej H_0: zależność między zmiennymi x, y ma charakter liniowy.

l.p.	x_i	y_i
1	0,8	9,7
2	0,9	11,4
3	1,0	11,7
4	1,1	14,0
5	1,2	15,3
6	1,3	16,6
7	1,4	17,6
8	1,5	19,1
9	1,6	19,3
10	1,7	20,9

11.Metoda Monte Carlo

Twórcą metody Monte Carlo był matematyk Stanisław Ulam urodzony we Lwowie w 1909 roku. Ulam przyczynił się w bardzo istotny sposób do skonstruowania przez USA bomby wodorowej. W projekcie tym Ulam wykorzystał obliczenia wykonane wymyśloną przez siebie metodą Monte Carlo. Monte Carlo kojarzy się z kasynami gry, czyli miejscami, w których główną rolę odgrywa przypadek. Stąd właśnie wzięła się nazwa metody, gdyż istotną rolę w tej metodzie odgrywa losowanie liczb o odpowiednim rozkładzie prawdopodobieństwa odpowiadającym danym wielkościom charakteryzującym dany proces. Metoda znalazła szerokie zastosowanie w modelowaniu złożonych procesów, których rozwiązanie analityczne jest bardzo trudne lub wręcz niemożliwe. Wykorzystywane jest to zarówno do badania właściwości układów lub zjawisk, jak i w trakcie projektowania nowych eksperymentów. Metodę Monte Carlo wykorzystuje się również do wyliczania całek zwłaszcza wielowymiarowych, których nie da się wyliczyć analitycznie, a rozwiązywanie innymi metodami numerycznymi jest bardzo czasochłonne.

11.1. Liczby pseudolosowe

Podstawą metody Monte Carlo jest możliwość generowania liczb o charakterze losowym. Służą do tego tzw. generatory liczb losowych (lub pseudolosowych). Można tu wyróżnić dwa typy generatorów: generatory sprzętowe i programowe. W generatorach sprzętowych wykorzystuje się parametry fizyczne jakiegoś stochastycznego procesu np. szumu elektrycznego. Generowane w ten sposób ciągi liczb charakteryzują się nieprzewidywalnością i niepowtarzalnością. W tym sensie mają rzeczywiście charakter losowy, chociaż ich właściwości statystyczne nie zawsze są idealne. Wadą tych generatorów jest ich cena i stosunkowo wolny czas działania. W wielu przypadkach całkowicie wystarcza stosowanie generatorów programowych. Generatory programowe są znacznie szybsze od generatorów sprzętowych, a w przypadku dobrych generatorów własności statystyczne generowanych liczb są czasami lepsze niż w przypadku generatorów sprzętowych. Liczby wyliczane są w nich za pomocą mniej lub bardziej skomplikowanych procedur matematycznych. Nazywamy je generatorami liczb pseudolosowych, gdyż ciągi liczb generowanych przez nie są ściśle określone i powtarzalne. Generatory takie działają zwykle w sposób rekurencyjny. Kolejna liczba jest wyliczana na podstawie określonej liczby poprzednio wygenerowanych liczb, przy czym pierwsza liczba

znajdowana jest za pomocą tzw. ziarna generatora. Generowane liczby tworzą ciąg charakteryzujący się pewnym okresem, czyli dana sekwencja liczb powtarza się cyklicznie. Jakość generatora zależy zarówno od okresu generowanego ciągu liczb, ale również od własności statystycznych otrzymanego ciągu (następujące po sobie liczby w ramach cyklu powinny maksymalnie przypominać losowy ciąg). Dla badania generatorów liczb pseudolosowych przeprowadza się specjalne testy widmowe.

Jednym z najlepiej zbadanych algorytmów generujących liczby pseudolosowe o rozkładzie jednostajnym jest tzw. *metoda kongurencyjna*. Kolejne liczby w tej metodzie generowane są na podstawie $k+1$ poprzednich liczb według wzoru

$$x_{n+1} = (a_0 x_n + a_1 x_{n-1} + \cdots + a_k x_{n-k}) \bmod M, \tag{11.1}$$

gdzie M jest odpowiednio dużą liczbą naturalną, a_i, x_i są liczbami całkowitymi z przedziału $[0, M)$, a symbol $(a) \bmod b$ oznacza resztę z dzielenia a przez b. Szczególnymi przypadkami powyższego generatora są

- generatory Fibonacciego, stosujące relację: $x_{n+1} = x_n + x_{n-1} \bmod M$,
- generatory multiplikatywne, stosujące relację: $x_{n+1} = a_0 x_n \bmod M$,
- generatory mieszane, stosujące relację: $x_{n+1} = a_0 x_n + a_1 \bmod M$.

Jeśli liczby wygenerowane, którąś z powyższych metod podzielimy przez M to dostaniemy liczby pseudolosowe z przedziału $[0,1)$. Ważną cechą generatora jest aby liczby generowane przez niego możliwie gęsto pokrywały przedział $(0,1)$.

11.2. Generowanie liczb pseudolosowych o dowolnym rozkładzie

Generatory liczb pseudolosowych generują zwykle liczby o rozkładzie jednostajnym z przedziału $[0,1]$. Na ogół jednak potrzebne są nam liczby o innych rozkładach. Poniżej opiszemy kilka metod postępowania.

Zacznijmy od zmiennej dyskretnej. Niech nasza dyskretna zmienna losowa ma rozkład prawdopodobieństwa

$$P(x = x_i) = p_i, \qquad i = 1,2, \dots, n. \tag{11.2}$$

Jeśli chcemy wygenerować liczby o takim rozkładzie korzystając z generatora liczb losowych o rozkładzie jednostajnym z przedziału $[0,1]$ dzielimy przedział $[0,1]$ na n podprzedziałów o szerokościach $\Delta_i = p_i$. Każdemu z tych podprzedziałów i przyporządkowujemy odpowiednią liczbę x_i, której prawdopodobieństwo jest równe szerokości przedziału Po wygenerowaniu liczby γ sprawdzamy, w którym podprzedziale się mieści i zwracamy liczbę przypisaną temu przedziałowi. W sytuacji, gdy nasza zmienna dyskretna może przyjmować nieskończenie wiele wartości musimy dokonać przybliżenia – przedział $[0,1]$ dzielimy na dużą, ale skończoną liczbę n_{max} podprzedziałów tak, żeby

$$\sum_{i=1}^{n_{max}} p_i = 1 - \varepsilon, \tag{11.3}$$

gdzie ε jest z góry przyjętą odpowiednio małą liczbą.

Przejdźmy teraz do zmiennej ciągłej. Jedną z metod generowania liczb losowych o dowolnym rozkładzie bazując na generatorze liczb o rozkładzie jednostajnym jest *metoda funkcji odwrotnej*. Niech X jest zmienną o rozkładzie jednostajnym

$$f(x) = 1, \text{ gdy } 0 \le x < 1, \quad f(x) = 0, \text{ gdy } x < 0, \ x \ge 1, \tag{11.4}$$

a zmienna Y zmienną losową o rozkładzie $g(y)$. Wiemy, że $g(y)dy = f(x)dx$, a ponieważ $f(x) = 1$, to

$$g(y)dy = dx. \tag{11.5}$$

Oznaczmy dystrybuantę rozkładu $g(y)$ symbolem $G(y)$. Zgodnie z definicją gęstości prawdopodobieństwa $g(y) = dG(y)/dy$, czyli równanie 11.5 możemy przekształcić do postaci

$$dx = dG(y) = g(y)dy. \tag{11.6}$$

Po scałkowaniu powyższego równania dostajemy

$$x = G(y) = \int_{-\infty}^{y} g(u)du. \tag{11.7}$$

Jeśli istnieje funkcja odwrotna do $x = G(y)$, czyli

$$y = G^{-1}(x), \tag{11.8}$$

to na mocy równania (11.7) możemy powiedzieć, że jeżeli zmienna losowa X ma rozkład jednostajny na przedziale $[0,1]$, to zmienna $Y = G^{-1}(X)$ ma rozkład $g(y)$.

Przykład

Załóżmy, że chcemy generować liczby losowe o rozkładzie

$$f(t) = ce^{-c(t-t_0)}, \text{ gdy } t \ge 0, \quad f(t) = 0, \text{ gdy } t < 0. \tag{11.9}$$

Dystrybuanta tego rozkładu ma postać

$$G(t) = c\int_{0}^{t} ce^{-c(t'-t_0)}dt' = 1 - e^{-c(t-t_0)} \tag{11.10}$$

Funkcją odwrotną do funkcji

$$x = G(t) = 1 - e^{-c(t-t_0)} \tag{11.11}$$

jest funkcja

$$t = t_0 - \frac{1}{c}\ln(1-x). \tag{11.12}$$

A zatem w celu wygenerowania liczb losowych t o rozkładzie 11.9 generujemy liczby x o rozkładzie jednostajnym z przedziału $(0,1]$ i następnie wyliczamy liczby t zgodnie z równaniem (11.12).

Metoda funkcji odwrotnej nie zawsze jest możliwa do zastosowania. Zdarza się bowiem, że nie istnieje jednoznaczna funkcja odwrotna lub też istnieje, ale znalezienie analitycznej postaci funkcji odwrotnej jest niemożliwe zaś numeryczne metody są bardzo czaso-

chłonne. Musimy wówczas zastosować inną metodę. Jedną z nich jest *metoda eliminacji* (zwana też metodą akceptacji i odrzucania) zaproponowana przez J. Neumanna. Metodę eliminacji możemy stosować do zmiennych losowych których rozkład $f(x)$ spełnia następujące warunki:

- gęstość prawdopodobieństwa $f(x)$ jest określona na skończonym przedziale $[a, b]$, a poza tym przedziałem jest zerowa,
- $f(x)$ jest ograniczona od góry $f(x) \leq c$.

Najprostszy sposób postepowania jest następujący:

- generujemy pary liczb y_1, y_2 o rozkładzie jednostajnym z przedziału $[0,1]$
- wyliczamy liczbę $x = a + (b - a)y_1$ (w ten sposób generujemy liczby o rozkładzie jednostajnym z przedziału $[a, b]$)
- wyliczmy liczbę $y = cy_2$ (w ten sposób generujemy liczby o rozkładzie jednostajnym z przedziału $[0, c]$)
- sprawdzamy, czy $y \leq f(x)$. Jeśli tak jest to akceptujemy wartość x, a w przeciwnym wypadku odrzucamy parę x, y i powtarzamy losowanie.

W ten sposób każda z zaakceptowanych liczb jest akceptowana z prawdopodobieństwem proporcjonalnym do $f(x_i)$, a tym samym zbiór zaakceptowanych liczb ma rozkład opisany funkcją $f(x)$.

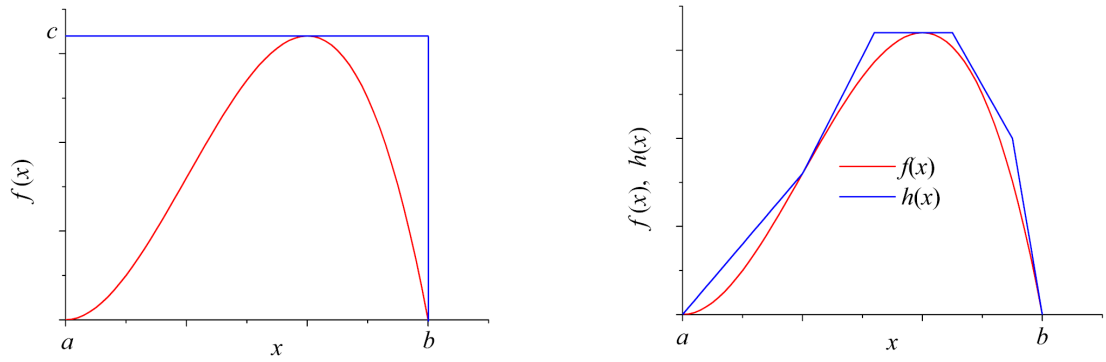
Metoda eliminacji w postaci omówionej powyżej jest mało wydajna, gdyż zwykle bardzo wiele wygenerowanych par x_i, y_i musimy odrzucić. Wydajność tej metody określa stosunek pola pod krzywą $f(x)$ w przedziale $[a, b]$ do pola całego prostokąta o bokach $b - a$ i c, czyli

$$E = \frac{\int_a^b f(x)dx}{(b-a)c} = \frac{1}{(b-a)c}. \tag{11.13}$$

Wydajność metody można znacznie poprawić, gdy zamiast prostokąta użyjemy innej figury, która od góry obejmie całą krzywą $f(x)$. W najprostszym przypadku może to być odpowiedni wielobok (patrz Rysunek 11.1 (b)), ale może to też być figura ograniczona dowolną krzywą leżącą nad krzywą $f(x)$, byle jej postać analityczna była na tyle prosta, aby można się było nią łatwo posłużyć w procesie opisanym poniżej.

Schemat postępowania w ogólnej wersji eliminacji Neumanna jest następujący:

a) wybieramy dostatecznie prostą gęstość prawdopodobieństwa $h(x)$ leżącą możliwie blisko gęstości $f(x)$ i spełniającą nierówność $h(x) \leq f(x)$,
b) generujemy liczbę losową x podlegającą rozkładowi jednostajnemu w przedziale (a, b) i drugą liczbę u o rozkładzie jednostajnym z przedziału $(0,1)$,
c) akceptujemy liczbę x, gdy $u \cdot h(x) < f(x)$, a w przeciwnym przypadku odrzucamy parę x, u,
d) powtarzamy kroki b) i c) dostatecznie wiele razy. Zbiór zaakceptowanych liczb x będzie podlegać rozkładowi $f(x)$.



(a) (b)

Rysunek 11.1. Prostokąt o bokach $(b-a),\ c$ ograniczający krzywą gęstości prawdopodobieństwa w metodzie eliminacji Neumanna (a). Wielobok ograniczający krzywą gęstości prawdopodobieństwa w metodzie eliminacji Neumanna (b). W wersji (a) metody punkty losowane będą leżeć wewnątrz prostokąta, a w wersji (b) wewnątrz wieloboku. Punkty leżące powyżej czerwonej krzywej są odrzucane, więc wersja (a) będzie mniej wydajna niż wersja (b)

Istnieją również metody wykorzystujące różne własności statystyczne funkcji niezależnych zmiennych losowych $y_1, y_2, \dots, y_n$ o rozkładzie jednostajnym. W metodach tych znajdujemy funkcję takich zmiennych spełniającą zależność $x = g(y_1, y_2, \dots, y_n)$. Następnie generujemy w jednym kroku n liczb losowych i korzystając z funkcji $x = g(y_1, y_2, \dots, y_n)$. wyliczmy szukaną liczbę losową mającą pożądany rozkład. Sposób szukania funkcji g jest zależny od własności gęstości $f(x)$ i nie istnieje ogólny przepis na jej znajdowanie. W ramach przykładu znajdźmy metodę na generowanie liczb losowych o rozkładzie normalnym.

Rozważmy wspólny rozkład dwu niezależnych zmiennych losowych X_1, X_2 o rozkładzie normalnych o parametrach μ, σ:

$$f(x_1, x_2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_1-\mu)^2}{2\sigma^2}} \cdot \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_2-\mu)^2}{2\sigma^2}} = \frac{1}{2\pi\sigma^2} e^{-\frac{(x_1-\mu)^2+(x_2-\mu)^2}{2\sigma^2}}. \tag{11.14}$$

Przejdźmy do zmiennych biegunowych r, φ zdefiniowanych następująco:

$$\begin{aligned} x_1 &= \mu + \sigma r \cos\varphi, \\ x_2 &= \mu + \sigma r \sin\varphi. \end{aligned} \tag{11.15}$$

Gęstość prawdopodobieństwa dla zmiennych r, φ ma postać

$$h(r,\varphi) = f(\mu + \sigma r\cos\varphi\,, \mu + \sigma r \sin\varphi) \begin{vmatrix} \frac{\partial x_1}{\partial r} & \frac{\partial x_2}{\partial r} \\ \frac{\partial x_1}{\partial \varphi} & \frac{\partial x_2}{\partial \varphi} \end{vmatrix} = \tag{11.16}$$

$$\frac{1}{2\pi\sigma^2} e^{-\frac{r^2}{2\sigma^2}} \cdot \sigma^2 \begin{vmatrix} \cos\varphi & \sin\varphi \\ -r\sin\varphi & r\cos\varphi \end{vmatrix} = \frac{1}{2\pi} r e^{-\frac{r^2}{2}}.$$

Jak widzimy zmienne r, φ są niezależne, gdyż rozkład $h(r, \varphi)$ jest iloczynem dwu niezależnych rozkładów

$$h_1(\varphi) = \frac{1}{2\pi}, \quad 0 \le \varphi \le 2\pi, \tag{11.17}$$

$$h_2(r) = re^{-\frac{r^2}{2}},\ \ r \geq 0.$$

Pierwszy z tych rozkładów jest rozkładem jednostajnym na przedziale $[0,2\pi]$, więc zmienną φ możemy generować zgodnie ze wzorem

$$\varphi = 2\pi y_1, \tag{11.18}$$

gdzie y_1 jest liczbą losową o rozkładzie jednostajnym z przedziału $[0,1]$. W przypadku zmiennej r możemy zastosować metodę funkcji odwrotnej. Dystrybuanta rozkładu $h_2(r)$ jest równa

$$H_2(r) = \int_0^r h_2(r')dr' = \int_0^r r'e^{-\frac{r'^2}{2}}dr' = e^{-\frac{r^2}{2}}, \tag{11.19}$$

czyli

$$y_2 = e^{-\frac{r^2}{2}}. \tag{11.20}$$

Funkcja odwrotna do funkcji 11.20 ma postać

$$r = \sqrt{-2\ln y_2}. \tag{11.21}$$

A zatem jeśli będziemy generować liczby y_2 o rozkładzie jednostajnym z przedziału $(0,1)$, to liczby (11.21) będą miały rozkład $h_2(r) = re^{-\frac{r^2}{2}}$. Możemy teraz wrócić do zmiennych x_1, x_2 zgodnie ze wzorami (11.15), czyli

$$\begin{aligned} x_1 &= g_1(y_1, y_2) = \mu + \sigma \cdot \sqrt{-2\ln y_2}\cos(2\pi y_1), \\ x_2 &= g_2(y_1, y_2) = \mu + \sigma \cdot \sqrt{-2\ln y_2}\sin(2\pi y_1). \end{aligned} \tag{11.22}$$

Opisana metoda nazywana jest metodą biegunową Boxa i Mullera. Podsumujmy jej działanie:

- generujemy dwie liczby losowe y_1, y_2 o rozkładzie jednostajnym na przedziale $(0,1]$,
- wyliczamy dwie liczby
  - $x_1 = \mu + \sigma \cdot \sqrt{-2\ln y_2}\cos(2\pi y_1)$,
  - $x_2 = \mu + \sigma \cdot \sqrt{-2\ln y_2}\sin(2\pi y_1)$, (11.23)
- obie czynności powtarzamy odpowiednią liczbę razy. Otrzymany zbiór liczb x_1, x_2 ma rozkład normalny o parametrach μ, σ.

Na zakończenie omawiania metod generowania liczb losowych o dowolnym rozkładzie przy wykorzystaniu generatora liczb o rozkładzie jednostajnym omówmy jeden przykład metody wykorzystującej przekształcenie $x = g(y_1, y_2, \ldots, y_n)$. Metoda będzie dotyczyła również generowania liczb o rozkładzie normalnym z parametrami μ i σ. Nie ma ona praktycznego zastosowania, a raczej charakter dydaktyczny. Jak pamiętamy, centralne twierdzenie graniczne mówi, że zmienna losowa, która jest sumą nieskończenie wielu zmiennych losowych ma rozkład Gaussa niezależnie od tego jakie rozkłady mają poszczególne zmienne składowe. W praktyce złożenie niewielu zmiennych daje nam zmienną o rozkładzie bardzo zbliżonym

do rozkładu Gaussa. Wprowadźmy zmienną losową Z będącą sumą n zmiennych losowych $Y_1, Y_2, \dots, Y_n$ o rozkładzie jednostajnym na przedziale $(0,1)$

$$z = g(y_1, y_2, \dots, y_n) = \sum_{i=1}^{n} y_i. \tag{11.24}$$

Wiemy, że wartość oczekiwana zmiennej o rozkładzie jednostajnym na przedziale $(0,1)$ jest równa

$$E(y_i) = \frac{1}{2(1-0)} = \frac{1}{2}, \tag{11.25}$$

a ich wariancja jest równa

$$\sigma^2(y_i) = \frac{(1-0)^2}{12} = \frac{1}{12}. \tag{11.26}$$

Wynika z tego, że wartość oczekiwana i wariancja naszej zmiennej x są równe

$$\begin{aligned} E(z) &= E\left(\sum_{i=1}^{n} y_i\right) = \sum_{i=1}^{n} E(y_i) = \frac{n}{2}, \\ \sigma^2(z) &= \sigma^2\left(\sum_{i=1}^{n} y_i\right) = \sum_{i=1}^{n} \sigma^2(y_i) = \frac{n}{12}. \end{aligned} \tag{11.27}$$

W powyższych relacjach wykorzystaliśmy dwa fakty: wartość oczekiwana sumy zmiennych jest zawsze równa sumie wartości oczekiwanych składników oraz wariancja sumy zmiennych niezależnych jest równa sumie wariancji tych zmiennych. Dla dużych wartości n zmienna z będzie miała rozkład zbliżony do rozkładu normalnego z parametrami $\mu = n/2$ i $\sigma = n/12$. Zbudujmy z niej zmienną standaryzowaną

$$u = \frac{z - E(z)}{\sigma(z)} = \frac{\left(z - \frac{n}{2}\right)}{\sqrt{\frac{n}{12}}} = \sqrt{\frac{12}{n}}\left(z - \frac{n}{2}\right). \tag{11.28}$$

A zatem chcąc generować liczby u o standardowym rozkładzie normalnym $N(0,1)$ możemy generować w każdym kroku n (n odpowiednio duża liczba) liczb losowych $y_1, y_2, \dots, y_n$ o rozkładzie jednostajnym na przedziale (0,1) i liczby u wyliczać ze wzoru

$$u = \sqrt{\frac{12}{n}}\left(\sum_{i=1}^{n} y_i - \frac{n}{2}\right). \tag{11.29}$$

Ponieważ zmienna standaryzowana jest definiowana jako

$$u = \frac{x - \hat{x}}{\sigma}, \tag{11.30}$$

to w przypadku, gdy chcemy aby generowane liczby miały rozkład normalny z parametrami μ, σ musimy wzór (11.29) zastąpić wzorem

$$x = \mu + \sigma \cdot \sqrt{\frac{12}{n}} \left(\sum_{i=1}^{n} y_i - \frac{n}{2} \right). \tag{11.31}$$

11.3. Liczenie całek metodą Monte Carlo

Liczenie całek oznaczonych metodą Monte Carlo oznacza wykorzystanie metod statystycznych do numerycznego (przybliżonego) wyliczenia wartości całki. Wymyślono wiele metod realizacji tego zadania. Jedna z tych metod opiera się koncepcyjnie na definicji Riemana całki oznaczonej, zgodnie z którą całka oznaczona jest interpretowana jako pole powierzchni po krzywą (lub uogólnienie pola powierzchni w przypadku całek wielowymiarowych). Załóżmy, że interesuje nas całka z funkcji $f(x)$ w granicach $[a, b]$, czyli

$$I = \int_a^b f(x)dx. \tag{11.32}$$

Weźmy funkcję $g(x)$ taką, że

$$g(x) \geq f(x), \quad a \leq x \leq b, \tag{11.33}$$

a ponadto znamy wartość całki

$$I_0 = \int_a^b g(x)dx. \tag{11.34}$$

W najprostszym przypadku funkcją $g(x)$ może być funkcja stała

$$g(x) = c = \max_{a \leq x \leq b} f(x). \tag{11.35}$$

Wówczas wykres funkcji będzie mieścił się w prostokącie o wysokości c podobnie jak na rysunku 11.1 (a). Wygenerujmy teraz N par liczb losowych (x_i, y_i), $i = 1,2, \dots, N$ o rozkładzie jednostajnym, takich że $a \leq x_i \leq b$ i $0 \leq y_i < c$. Policzmy liczbę n wygenerowanych par, dla których $y_i \leq f(x_i)$. Geometrycznie oznacza to wypełnienie prostokąta, w którym mieści się wykres funkcji $f(x)$ przypadkowo rozłożonymi punktami (według rozkładu jednostajnego). Spośród N takich punktów n będzie leżeć pod krzywą $f(x)$. Dla dużych wartości liczby N stosunek n/N będzie zatem bliski stosunkowi powierzchni pod krzywą $f(x)$ do pola prostokąta

$$\frac{n}{N} \approx \frac{\int_a^b f(x)dx}{c(b-a)}, \tag{11.36}$$

czyli

$$\int_a^b f(x)dx \approx \frac{n}{N} \cdot c(b-a). \tag{11.37}$$

Wydajność metody możemy poprawić, jeśli prostokąt zastąpimy inną figurą obejmującą wykres funkcji $f(x)$ o polu bardziej zbliżonym do szukanej całki np. tak, jak na rysunku 11.1 (b). Teraz jednak będziemy musieli wygenerować pary liczb losowych, które odpowiadają punktom leżącym w polu wybranej figury (pod wykresem funkcji $g(x)$), a nie w polu całego prostokąta. Możemy to zrobić metodą eliminacji Neumana. W tym przypadku losujemy, tak jak poprzednio, pary (x_i, y_i), $i = 1,2, \dots, n$ spełniające warunek $a \leq x_i \leq b$ i $0 \leq y_i < c$, ale akceptujemy tylko te, dla których $y_i \leq g(x_i)$. Jeśli przez N rozumiemy teraz liczbę zaakceptowanych par, a przez n liczbę tych par zaakceptowanych, dla których $y_i \leq f(x_i)$, to

$$\int_a^b f(x)dx = \frac{n}{N}\int_a^b g(x)dx. \tag{11.38}$$

Oczywiście, jeśli to możliwe, to zamiast metody eliminacji Neumana w powyższym schemacie możemy zastosować metodę funkcji odwrotnej.

Typowa fluktuacja liczby n jest w przybliżeniu równa $\Delta n \cong \sqrt{n}$, czyli względna dokładność wyznaczenia całki powyższą metodą jest rzędu

$$\frac{\Delta I}{I} = \frac{\Delta n}{n} = \frac{\sqrt{n}}{n} = \frac{1}{\sqrt{n}}. \tag{11.39}$$

Przyjrzyjmy się teraz innemu sposobowi liczenia całek metodami statystycznymi (czyli Monte Carlo). Weźmy funkcję $g(x)$ mającą własności gęstości prawdopodobieństwa: $g(x) > 0$ oraz $\int_a^b g(x)dx = 1$. Zapiszmy całkę (11.32) w postaci

$$\int_a^b f(x)dx = \int_a^b \frac{g(x)}{g(x)} f(x)dx = \int_a^b g(x)\frac{f(x)}{g(x)}dx. \tag{11.40}$$

Porównując powyższy wzór z definicją wartości oczekiwanej widzimy, że całką jest po prostu wartością oczekiwaną funkcji $\frac{f(x)}{g(x)}$ dla gęstości prawdopodobieństwa $g(x)$.

$$I = \int_a^b f(x)dx = \int_a^b g(x)\frac{f(x)}{g(x)}dx = E\left(\frac{f(x)}{g(x)}\right). \tag{11.41}$$

Jak pamiętamy, estymatorem wartości oczekiwanej jest średnia arytmetyczna, czyli dla dużych wartości n otrzymujemy

$$I \approx T_n(I) = \frac{1}{n}\sum_{i=1}^{n}\frac{f(x_i)}{g(x)}. \tag{11.42}$$

W najprostszym przypadku możemy jako $g(x)$ wziąć rozkład jednostajny na odcinku $[a, b]$

$$g(x) = \frac{1}{b-a}. \tag{11.43}$$

Argumenty funkcji losujemy wówczas z rozkładem jednostajnym na odcinku $[a, b]$, a całkę 11.41 przybliżamy wyrażeniem

$$\int_a^b f(x)dx \approx \frac{1}{n}\sum_{i=1}^{n}\frac{f(x_i)}{1/(b-a)} = \frac{b-a}{n}\sum_{i=1}^{n} f(x_i). \tag{11.44}$$

Jest to tzw. *podstawowa metoda liczenia całek* metodą Monte Carlo. Podobnie jak w poprzedniej metodzie, tak i w tej możemy zmniejszyć błąd całki przez zastąpienie rozkładu jednostajnego rozkładem $g(x)$ możliwie podobnym do funkcji podcałkowej . Oczywiście musimy wówczas losować argumenty funkcji według rozkładu $g(x)$ stosując metodę eliminacji Neumana lub, jeśli to możliwe, metodę funkcji odwrotnej. Jest to tzw. *metoda losowania istotnego*, gdyż częściej losowane są argumenty, dla których wartość funkcji $f(x)$ jest duża, a ich przyczynki do wartości całki są bardziej znaczące.

Jeszcze inna metoda, nazywana *losowaniem warstwowym* polega na podzieleniu przedziału całkowania na mniejsze przedziały, takie w których funkcja zmienia się możliwie mało. W każdym z tych podprzedziałów możemy zastosować *metodę podstawową*.

Estymator odchylenia standardowego (pierwiastka z wariancji) zmiennej $u = f(x)/g(x)$ wynosi

$$S(u) = S(f/g) = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(u-\bar{u})^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{f(x_i)}{g(x_i)} - T_n(I)\right)^2}, \tag{11.45}$$

zaś, jak pamiętamy, estymator odchylenia standardowego średniej arytmetycznej jest $\sqrt{n}$ razy mniejszy. Tak więc wartość estymatora odchylenia standardowego średniej arytmetycznej 11.42, którą możemy traktować jak miarę dokładności metody wynosi

$$S(I) = \frac{S(f/g)}{\sqrt{n}} = \sqrt{\frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{f(x_i)}{g(x_i)} - T_n(I)\right)^2}. \tag{11.46}$$

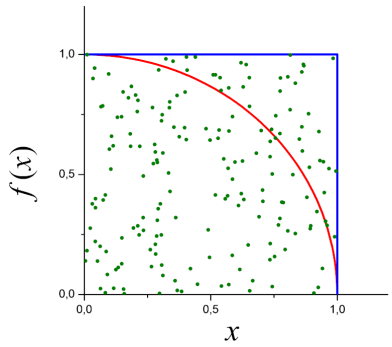
Jeśli dokładna wartość całki jest nam znana, to w powyższych wyrażeniach należy zastąpić nią estymator $T_n(I)$. Jak widać dokładność liczenia całek metodami Monte Carlo rośnie bardzo wolno z liczbą kroków n (odwrotnie proporcjonalnie do $\sqrt{n}$). W przypadku całek jednowymiarowych, przy tej samej liczbie obliczeń funkcji podcałkowej, osiągniemy znacznie lepszą dokładność stosując inne metody numerycznego całkowania, takie jak np. metoda Simpsona, Romberga, czy Gaussa. Potęga metody Monte Carlo uwidacznia się jednak, jeśli mamy do policzenia całkę w przestrzeni wielowymiarowej, gdyż błąd liczenia całki metodą Monte Carlo maleje odwrotnie proporcjonalnie do $\sqrt{n}$ niezależnie od wymiaru w przestrzeni argumentów, podczas gdy w przypadku innych metod liczba obliczeń rośnie proporcjonalnie do n^w. Dla przykładu załóżmy, że chcemy policzyć całkę w przestrzeni 10. wymiarowej przestrzeni argumentów. Jeśli w każdym z wymiarów przedział zmienności argumentu podzielilibyśmy tylko na 10 części, to funkcję podcałkową, w którejś ze zwykłych metod numerycznych musielibyśmy wyliczyć 10^{10} razy, a dokładność byłaby niewielka. Jeśli taką liczbę razy policzylibyśmy funkcję podcałkową w metodzie Monte Carlo, to uzyskalibyśmy bardzo dobrą dokładność rzędu $\frac{\sigma(f)}{10^5} = 0{,}00001 \cdot \sigma(f)$.

Często prezentowanym przykładem na zastosowanie metody Monte Carlo do liczenia całek jest znajdowanie przybliżenia liczby π poprzez obliczenie całki

$$\int_0^1 \sqrt{1-x^2}dx = \frac{\pi}{4}. \tag{11.47}$$

Wykresem funkcji podcałkowej w przedziale [0,1] jest ćwiartką okręgu o promieniu jednostkowym. Stosując metodę punktów wypełniających prostokąt, wewnątrz którego mieści się wykres funkcji, losujemy N par liczb x_i, y_i o rozkładzie jednostajnym na przedziale [0,1] ile punktów o wygenerowanych współrzędnych znajduje się poniżej wykresu (patrz rysunek 11.2), czyli ile par liczb spełnia warunek $x_i^2 + y_i^2 \leq 1$. Stosunek liczby n tych par do liczby Nwszystkich par przybliża nam wartość całki, czyli po przekształceniu dostajemy przybliżenie liczby π

$$\pi \approx \frac{4n}{N}. \tag{11.48}$$



Rysunek 11.2. Wykres funkcji $f(x) = \sqrt{1-x^2}$ w przedziale [0,1], czyli ćwiartka okręgu o promieniu 1 i środku w punkcie (0,0) wpisana w kwadrat o boku 1. Zielone kropki mają współrzędne par punktów x_i, y_i mających rozkład jednostajny na przedziale [0,1]. Stosunek liczba kropek leżących wewnątrz ćwiartki koła (na rysunku jest ich 143) do liczby wszystkich kropek (na rysunku jest ich 183) daje nam przybliżenie liczby $\pi/4$ (według sytuacji na rysunku jest to $\approx 0{,}7814$, a $\pi/4 \approx 0{,}7854$).

W metodzie liczenia całki jako wartości oczekiwanej funkcji podcałkowej podzielonej przez wybraną gęstość prawdopodobieństwa i wybierając rozkład jednostajny $g(x) = 1$ musimy wygenerować N argumentów funkcji z rozkładem jednostajnym na przedziale [0,1] i wyliczyć wartość naszej całki ze wzoru 11.44. Po przekształceniu dostajemy przybliżenie liczby π

$$\pi \approx \frac{4}{N}\sum_{i=1}^{n}\sqrt{1-x_i^2}. \tag{11.49}$$

Przykładowe wyniki obliczeń obiema metodami dla $N = 10^3, 10^4, 10^5, 10^6$ dla pięciu serii obliczeń przedstawiono w tabeli 11.1. Zwróćmy uwagę na to, że metoda wykorzystująca wzór 11.49 daje lepsze wyniki niż metoda wykorzystująca wzór (11.48).

Tabela 11.1. Przykładowe wyniki obliczeń przybliżenia liczby π metodą Monte Carlo ze wzoru 11.48 (Metoda 1) i wzoru 11.49 (Metoda 2) dla $N = 10^3, 10^4, 10^5, 10^6$

	$N = 10^3$		$N = 10^4$		$N = 10^5$		$N = 10^6$	
	$\sim\pi$	$\Delta\pi$	$\sim\pi$	$\Delta\pi$	$\sim\pi$	$\Delta\pi$	$\sim\pi$	$\Delta\pi$
Metoda 1	3,1720	-0,030	3,1940	-0,052	3,1375	0,0041	3,1405	0,0011

	3,1600	-0,018	3,1340	0,008	3,1410	0,0006	3,1387	0,0029
	3,0480	0,094	3,1664	-0,025	3,1388	0,0028	3,1430	-0,0014
	3,0200	0,122	3,1208	0,021	3,1282	0,0134	3,1408	0,0008
	3,1040	0,038	3,1536	-0,012	3,1451	-0,0035	3,1411	0,0005
Metoda 2	3,1328	0,0088	3,1452	-0,0036	3,1410	0,00062	3,1419	-0,00033
	3,1442	-0,0026	3,1314	0,0102	3,1416	0,00003	3,1422	-0,00060
	3,0724	0,0692	3,1443	-0,0027	3,1394	0,00217	3,1418	-0,00021
	3,1456	-0,0040	3,1364	0,0052	3,1423	-0,00068	3,1418	-0,00016
	3,1625	-0,0209	3,1605	-0,0189	3,1389	0,00271	3,1396	0,00204

11.4. Zastosowanie metody Monte Carlo do modelowania komputerowego

Metoda Monte Carlo wymyślona została przez Ulama i Neumana do symulowania procesu fizycznego, którego nie dało się rozwiązać innymi metodami. Chodziło o zbadanie procesu przechodzenia neutronów przez ośrodek, aby móc zaprojektować bezpieczne i niezawodne osłony reaktorów jądrowych. Metoda zdała wówczas bardzo dobrze egzamin i na stałe wpisała się na listę technik badawczych nie tylko fizyki. Większość zjawisk i procesów przyrodniczych, biologicznych, ekonomicznych, technicznych itp. ma charakter stochastyczny. Jest tak dlatego, że w procesach tych bierze udział tak wielka liczba obiektów (np. atomów), że niemożliwe staje się ścisłe opisanie tych układów[7]. W takich przypadkach naturalne staje się posługiwanie się losowymi zmiennymi opisującymi cechy układów, zjawisk, procesów itp. Mierzalne są jedynie średnie wartości tych zmiennych. Metoda Monte Carlo jest naturalną metodą pozwalającą na symulowanie tego typu zjawisk, układów itp. za pomocą komputera. Taki sposób postępowania nazywamy modelowaniem komputerowym. Modelowanie komputerowe jest bardzo użytecznym i potężnym narzędziem poznawczym. Jest to, można powiedzieć, osobna dziedzina nauki, zbyt rozległa, żeby opisywać ją szczegółowo w niniejszym skrypcie.

Zadanie 11.1 (Metoda Boxa i Mullera)

Korzystając z metody Boxa i Mullera wygeneruj ciąg 100 liczb o rozkładzie normalnym z parametrami $\mu = 3$ i $\sigma = 0{,}4$. Policz wartość średnią i odchylenie standardowe otrzymanego ciągu i porównaj z wartościami parametrów μ i σ. Zbuduj histogram dla otrzymanego ciągu i porównaj z teoretyczną krzywą Gaussa.

Zadanie 11.2 (Centralne twierdzenie graniczne)

Powtórz poprzednie zadanie generując liczby według wzoru (11.31) dla $n = 12$ czyli

$$x = \mu + \sigma \cdot \left(\sum_{i=1}^{12} y_i - 6 \right).$$

[7] Na poziomie mikroskopowym przyroda jest całkowicie niedeterministyczna i nawet dla pojedynczego obiektu, np. elektronu w atomie nie jest możliwe deterministyczne opisanie zachowania się obiektu.

Zadanie 11.3

Wylicz przybliżoną wartość liczby π metodami wykorzystującymi wzory (11.48) i (11.49). Porównaj swoje wyniki z przykładowymi wynikami zawartymi w tabeli 11.1.

Zadanie 11.4

Wylicz metodą Monte Carlo całkę potrójną

$$I = \int_{x=0}^{x=1} \int_{y=0}^{y=1} \int_{z=0}^{z=1} \sin^2(x+y+z)dxdydz.$$

Skorzystaj z metody podstawowej i wzoru

$$I \approx \frac{1}{N}\sum_{i=1}^{N} \sin^2(x_i + y_i + z_i).$$

Porównaj wynik otrzymany tą metodą dla $N = 10^k, k = 3,4, \ldots ,8$ z wynikiem dokładnym

$$I = \frac{1}{16}(8 + 3\sin 2 - 3\sin 4 + \sin 6) \approx 0{,}79493.$$

Błąd porównaj z estymatorem odchylenia standardowego otrzymanej przez nas wartości (patrz wzór (11.46)).

Literatura

1. Siegmund Brandt — Analiza Danych, Warszawa, 2002, PWN
2. M. Abramowicz, Jak analizować wyniki pomiarów, Warszawa, 1992, PWN
3. W.T.Eadie, Metody statystyczne w fizyce doświadczalnej, PWN 1989