
**МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ**
**НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ**
**«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ**
**імені Ігоря Сікорського»**

# ТЕОРІЯ ЙМОВІРНОСТІ ТА МАТЕМАТИЧНА СТАТИСТИКА

Курс лекцій

Рекомендовано Методичною радою КПІ ім. Ігоря Сікорського
як навчальний посібник для здобувачів ступеня бакалавра
за освітньою програмою «Системи керування літальними апаратами та комплексами»
спеціальності 173 «Авіоніка»

**Укладач М.М.Чепілко**

Електронне мережне навчальне видання

**Київ**
**КПІ ім. Ігоря Сікорського**
**2024**

УДК 519.2(075.8)

Укладач: *Чепілко М.М., доктор фіз.-мат.наук, професор*

Рецензент: *Смірнов С.А., канд.фіз.-мат.наук, доцент, навчально - науковий фізико - технічний інститут*

Відповідальний редактор: *Пономаренко С.О., кандидат технічних наук, доцент*

Гриф надано Методичною радою КПІ ім. Ігоря Сікорського
(*протокол №8 від 20.06.2024 р.*)
за поданням Вченої ради навчально - наукового інституту аерокосмічних технологій
(*протокол №5/24 від 03.06.2024 р.*)

"Теорія ймовірності та математична статистика. Курс лекцій" [Електронний ресурс]: курс лекцій: навч. посіб. для здобувачів ступеня бакалавра за освітньою програмою «Системи керування літальними апаратами та комплексами» спеціальності 173 «Авіоніка» / КПІ ім. Ігоря Сікорського; укладач: Чепілко М.М.. – Електрон. текст. дані (1 файл). – Київ : КПІ ім. Ігоря Сікорського, 2024. – 181 с.

Метою видання навчального посібника "Теорія ймовірності та математична статистика. Курс лекцій" є допомога студентам у закріпленні та поглибленому розумінні означень, теоретичних положень та методів теорії ймовірності та математичної статистики.

У навчальному посібнику для першого розділу на основі використаної літератури відібрано традиційний матеріал теорії ймовірності, який є математичною основою для другого розділу навчального посібника, де викладаються основи математичної статистики.

Навчальний матеріал другого розділу навчального посібника викладається з використанням бібліотек (для розв'язку основних задач математичної статистики) такої мови програмування високого рівня, як Python 3.*. Такий підхід до викладання математичної статистики відповідає вимогам часу, оскільки нині Python 3.* активно використовується в аналізі великих даних, машинному навчанні, задачах штучного інтелекту.

У кожному підрозділі методичного посібника для прикладів підібрано найбільш типові задачі теорії ймовірності та математичної статистики та наведені їх роз'язки. Для частини задач наведені приклади розв'язків з використанням мови програмування Python 3.*.

Кожний підрозділ посібника також містить запитання студентам для самоконтролю їх знань.

УДК 519.2(075.8)

Реєстр. №XX/XX-XXX.. Обсяг X.X авт.арк.

Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»
Берестейський проспект, 37, Київ, 03056
https://kpi.ua
Свідоцтво про внесення до Державного реєстру видавців, виготовлювачів
і розповсюджувачів видавничої продукції ДК №5354 від 25.05.2017 р.

# Зміст

# Розділ 1

# Теорія ймовірності

## Вступ

Теорія ймовірності - математична наука, що вивчає закономірності властиві масовим випадковим явищам.

Предметом теорії імовірностей є вивчення імовірностних закономірностей масових однорідних випадкових подій. Основні поняття, методи, теореми та формули теорії імовірностей ефективно використовуються в науці, техніці, економіці, у теоріях надійності та масового обслуговування, у плануванні та організації виробництва, у страховій та податковій справах, у соціології та політології, у демографії та охороні здоров'я.

Теорія ймовірності є математичною основою для математичної статистики, методи якої нині широко застосовується у науці про великі дані, машинному навчанні, задачах штучного інтелекту і т.п.

Навчальний матеріал, викладений у розділі "Теорія ймовірності", в основному базується на використаній, при підготовці навчального посібника, літературі [1-5].

## 1. Випадкові події

### 1.1. Алгебра подій

У теорії ймовірності первинними поняттями є випробування і події [1-5]. Хай проводиться деяке випробування з випадковим результатом. Безліч всіх можливих взаємовиключних результатів даного випробування називається множиною елементарних подій. Повна сукупність всіх елементарних подій називається достовірною подією. Всяка підмножина безлічі елементарних подій називається випадковою подією.

Розглянемо деякі досліди, у результаті яких може з'явитись або не з'явитись

подія A. Прикладами таких дослідів можуть бути:

- дослід — виготовлення певного виробу, подія A — стандартність цього виробу;
- дослід — кидання монети, подія A — випав герб;
- дослід — стрільба п'ятьма пострілами у мішень, подія A — вибито 30 очок;
- дослід — введення програми у комп'ютер, подія A — безпомилковий ввід.

Загальним для усіх дослідів є те, що кожен із них може реалізуватись у певних умовах скільки завгодно разів. Такі досліди називають випробуваннями.

Події бувають достовірні, випадкові та неможливі.

- Достовірною називають таку подію, яка при розглянутих умовах обов'язково трапиться.
- Неможливою називають таку подію, яка при розглянутих умовах не може трапитись.
- Випадковою називають таку подію, яка при умовах, що розглядаються, може трапитися, а може й не трапитися.

Наприклад, якщо в урні є лише білі кулі, то добування білої кулі з урни — достовірна подія, а добування з цієї урни кулі іншого кольору — неможлива подія.

Якщо кинути монету на площину, то поява герба буде випадковою подією, тому що замість герба може з'явитися надпис.

Випадкові події позначають великими літерами, наприклад

$$A, B, C, D, X, Y, A_1, A_2, ..., A_n.$$

Кожна випадкова подія є наслідком багатьох випадкових або невідомих нам причин, які впливають на подію. Тому неможливо завбачити наслідок одночинного випробування. Але якщо розглядати випадкову подію багато разів при однакових умовах, то можна виявити певну закономірність її появи або не появи. Таку закономірність називають імовірною закономірністю масових однорідних випадкових подій.

У теорії імовірностей під масовими однорідними випадковими подіями розуміють такі події, які здійснюються багатократно при однакових умовах або багато однакових подій. Наприклад, кинути одну монету 100 разів або 100

однакових монет кинути один раз в теорії імовірностей вважають однаковими подіями.

Різновиди випадкових подій.

**Означення 1.1** Події називають несумісними, якщо поява однієї з них виключає появу інших подій в одному і тому ж випробуванні.

**Приклад 1.1** Серед однорідних деталей у ящику є стандартні та нестандартні. Навмання беруть із ящика одну деталь. Події

- A — взята стандартна деталь,
- B — взята нестандартна деталь

несумісні тому, що взята лише одна деталь, яка не може бути одночасно стандартною та нестандартною.

**Означення 1.2** Події називають сумісними, якщо поява однієї з них не виключає можливості появи інших (не обов'язково одночасно).

**Приклад 1.2** Два стрільця стріляють у мішень. Події

- $A_1$ — перший стрілок влучив у мішень,
- $A_2$ — другий стрілок влучив у мішень

будуть сумісними випадковими подіями.

**Означення 1.3** Випадкові події $A_1, A_2, \ldots, A_n$ утворюють повну групу подій, якщо внаслідок випробування хоча б одна з них з'явиться обов'язково.

**Приклад 1.3** Кидають шестигранний кубик. Позначимо події так

$A_1$ — випала грань 1; $A_2$ — випала грань 2; $A_3$ — випала грань 3; $A_4$ — випала грань 4; $A_5$ — випала грань 5; $A_6$ — випала грань 6.

Події $A_1, A_2, ..., A_6$ утворюють повну групу.

У прикладі 1.2 події $A_1$ та $A_2$ не утворюють повної групи. Але якщо позначити $A_0$ подію, що ніхто із стрільців не влучив у мішень, тоді події $A_0$, $A_1$ та $A_2$ утворюють повну групу.

**Означення 1.4** Події називають рівноможливими, якщо немає причин стверджувати, що будь-яка з них можливіша за інші.

**Приклад 1.4** Події — поява 1, 2, 3, 4, 5 або 6 очок при киданні шестигранного кубика — рівноможливі при умові, що центр його ваги не зміщений.

**Означення 1.5** Дві несумісні події, які утворюють повну групу, називають протилежними. Подія, протилежна події $A$, позначається $\overline{A}$.

**Приклад 1.5** Якщо позначити через $A$ подію, що при стрільбі по мішені

вибито 8 очок, то подія $\overline{A}$ — при стрільбі по мішені вибито будь-яке інше число очок.

Тепер розглянемо важливе поняття простору елементарних наслідків. Нехай виконується деякий експеримент, який має елементи випадковості. Кожне випробування може мати різні наслідки. Так, при киданні монети можуть бути два можливих наслідки: герб або надпис.

При киданні грального кубика можуть бути шість можливих наслідків. У випробуванні «постріл у мішень» можна розглядати такі наслідки, як влучення у мішень, або кількість вибитих очок, або координати точки влучення. Отже, що приймати за наслідок випробування, залежить від умови задачі.

**Означення 1.6** Елементарними наслідками називають такі події, які неможливо розділити на більш прості. Множину усіх можливих елементарних наслідків називають простором елементарних наслідків.

Простір елементарних наслідків може містити скінчену, злічену або незлічену множину елементів. У ролі елементарних наслідків можна розглядати точки $n$-вимірного простору, відрізок деякої лінії, точки поверхні $S$ або об'єму $V$ трьохвимірного простору, функцію однієї або багатьох змінних.

У більшості випадків що розглядаються, припускають, що елементарні наслідки рівноможливі.

- При двократному киданні монети простір елементарних наслідків містить 4 точки $\{(A,A),(A,B),(B,A),(B,B)\}$, де $A$ — означає появу герба, $B$ — появу надпису.
- Нехай по мішені стріляють одиночними пострілами до першого влучення. Можливі такі елементарні події
  — $w_1$ влучення при першому пострілі,
  — $w_2$ влучення при другому пострілі,
  — $w_3$ влучення при третьому пострілі і т.д.

  У цьому випадку простір елементарних наслідків може мати нескінченну кількість точок, які можна шляхом нумерації перелічити. Тому простір елементарних наслідків буде зліченим.
- При виробництві моніторів виникають неоднакові умови технологічного процесу, тому час роботи монітора відрізняється від його номінального значення, тобто буде випадковою подією. Простір елементарних наслідків у цьому випадку буде нескінченною незліченою множиною, елементи якої неможливо пронумерувати.

Тепер ознайомимось з алгеброю випадкових подій. Нехай $A$ та $B$ — випадкові

події.

Об'єднанням (сумою) випадкових подій $A \cup B$ (або $A + B$) називають таку випадкову подію, яка полягає у появі подій $A$ або $B$ або $A$ та $B$. Якщо $A$ та $B$ — несумісні, то $A \cup B$ означає появу події $A$ або події $B$.

- Рис.1.1а та Рис.1.1б. Подія $A$ та подія $B$.
- Рис.1.1в. Подія $B$ та протилежна їй $\overline{B}$.
- Рис.1.1г. Заштрихована площина — добуток подій $AB$.
- Рис.1.1д. Заштрихована площина — сума подій $A \cup B$.
- Рис.1.1е. Заштрихована площина — різниця подій $A - B$.

Аналогічно визначають об'єднання (суму) більшої кількості випадкових подій.

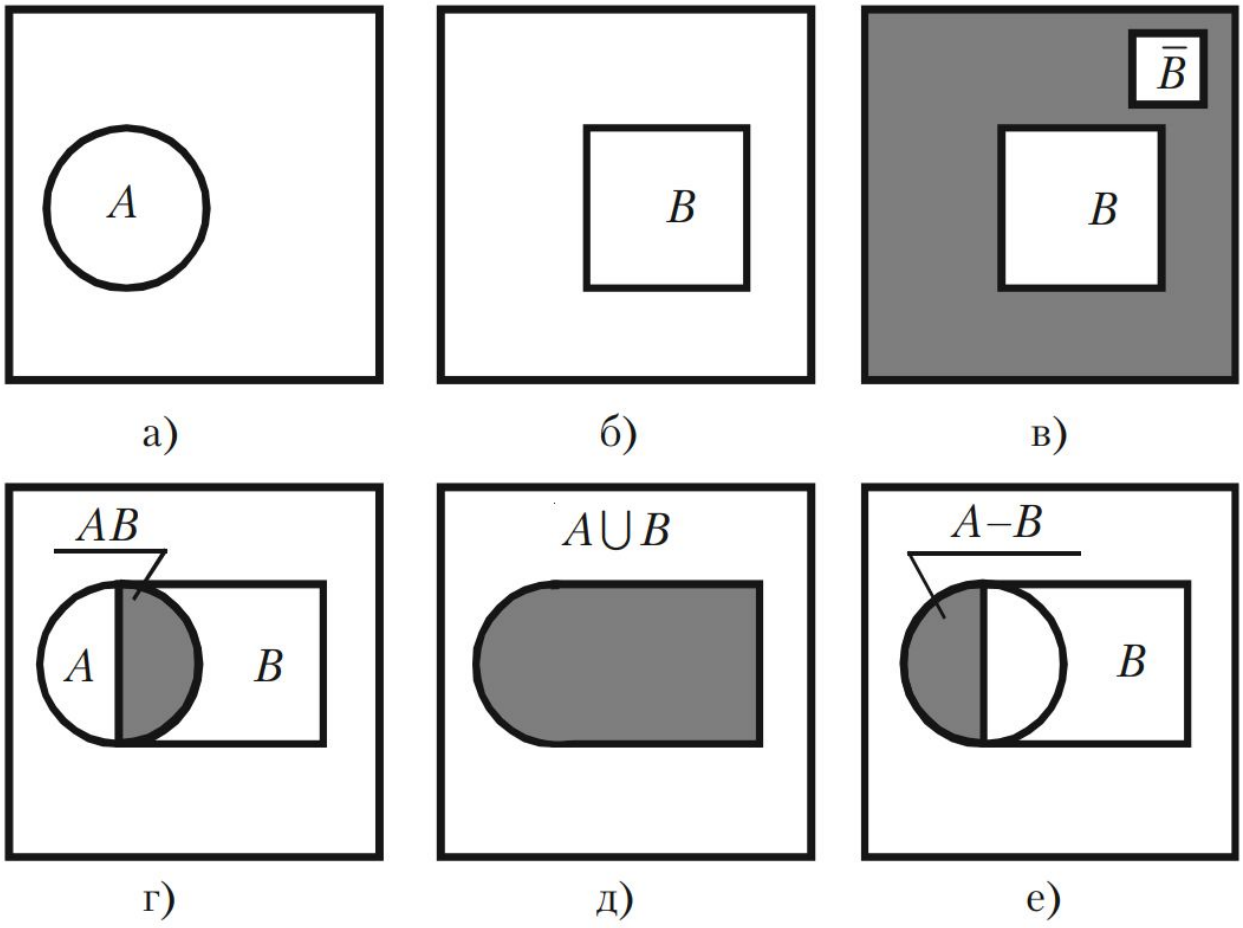


Рис. 1.1 — Математичних операцій теорії імовірностей

**Означення 1.7** Об'єднанням (сумою) випадкових подій $A_1 \cup A_2 \cup ... \cup A_n$ називають таку випадкову подію, яка полягає в появі хоча б однієї з цих подій.

Якщо події попарно несумісні, то їх сума полягає в тому, що повинна з'явитися подія $A_1$ або $A_2$ ... або $A_n$ . Нескінченну суму випадкових подій позначають

$$\bigcup_{k=1}^{\infty} A_k \tag{1.1}$$

**Означення 1.8** Стрілок робить один постріл у мішень, поділену на три області. Позначимо

- подія $A_1$ — влучення в першу область;
- подія $A_2$ — влучення у другу область;
- подія $A_3$ — влучення в третю область;
- подія $A_4$ — немає влучення у мішень;
- подія $B$ — влучення в першу або другу області;
- подія $D$ — влучення хоча б в одну область мішені.

Тоді маємо $B = A_1 \cup A_2$; $D = A_1 \cup A_2 \cup A_3$. Відмітимо, що події $A_1, A_2, A_3$ та $A_4$ — несумісні.

**Означення 1.9** Різницею $B - A$ (або $B \backslash A$) двох випадкових подій $B$, $A$ називають усі наслідки, які полягають у тому, що подія $A$ не з'являється.

**Означення 1.10** Добутком (перетином) $A \cap B$ (або $A \cdot B$) випадкових подій $A$, $B$ називають таку випадкову подію, яка полягає у появі подій А та В одночасно. Якщо $A$ та $B$ — несумісні, то добуток $A \cap B$ є множина, яка не має жодного елемента. Така множина називається пустою (порожньою) і позначається $\varnothing$.

Таким чином, у разі несумісності подій $A$, $B$ маємо

$$A \cap B = A \cdot B = \varnothing. \tag{1.2}$$

**Означення 1.11** Добутком (перетином) скінченої кількості випадкових подій $A_1, A_2, ..., A_n$, називають таку випадкову подію, яка полягає у сумісній появі усіх цих подій одночасно.

Подія $\bigcap_{k=1}^{n} A_k$ означає, що розглядаються усі події $A_k$ $(k = 1, 2, ..., n)$ одночасно.

Вказані співвідношення між подіями є звичайними співвідношеннями між множинами, які можна представити графічно (див.Рис.1.1).

**Приклад 1.6** Стрілець стріляє двічі по мішені. Описати простір елементарних наслідків. Записати подію, яка полягає в тому, що:

- стрілець влучив у мішень принаймні один раз (подія $C$);
- стрілець влучив рівно один раз (подія $D$);
- стрілець не влучив у мішень (подія $F$).

**Розв'язок задачі.** Позначимо

- подія A — влучення при першому пострілі,
- подія B — влучення при другому пострілі.

Простір елементарних наслідків складається з чотирьох подій

$$\{AB, A\overline{B}, \overline{A}B, \overline{AB}\} \tag{1.3}$$

- Якщо стілець влучив у мішень принаймні один раз, то це означає, що він влучив або при першому пострілі $A\overline{B}$ , або при другому пострілі $\overline{A}B$, або при обох . Тобто,

$$C = A\overline{B} \cup \overline{A}B \cup AB. \tag{1.4}$$

- Рівно одне влучення може бути тільки тоді, коли стрілець при першому пострілі влучив, а при другому — ні, або при першому пострілі не влучив, а при другому — влучив. Тому,

$$D = A\overline{B} \cup \overline{A}B. \tag{1.5}$$

- Якщо стілець не влучив у мішень, то це означає, що він не влучив при обох пострілах, Тобто,

$$F = \overline{AB}. \tag{1.6}$$

## 1.2. Елементи комбінаторики

Комбінаторика - розділ математи, присвячений вирішенню задач вибору і розташування $k$ елементів з деякої множини $n$ елементів відповідно до заданих правил [1-5]. Нижче розглядатимемо множини, що складаються з $n$ різних елементів.

**Означення 1.12** Перестановками називаються комбінації, що складаються з $n$ елементів і відмінні тільки порядком їх розташування. Число перестановок рівне

$$P_n = \prod_{k=1}^{n} k = 1 \cdot 2 \cdot 3 \cdot ... \cdot (n-1) \cdot n = n!, \tag{1.7}$$

де $n!$ — факторіал. Причому прийнято вважати, що $0! = 1$ (читається «нуль факторіал»).

**Означення 1.13** Поєднаннями називаються комбінації, що полягають у виборі $k$ елементів з $n$ елементів ($k \leq n$), відмінні хоча б одним елементом. Число поєднань рівно

$$C_n^k = \frac{n!}{k!(n-k)!} = n(n-1)(n-2)...(n-k+1) \tag{1.8}$$

**Означення 1.14** Розміщеннями називаються комбінації, що полягають у виборі $k$ елементів з $n$ елементів ($k \leq n$), відмінні або складом елементів, або їх порядком. Число розміщень равне

$$A_n^k = \frac{n!}{(n-k)!} = n \cdot (n-1) \cdot (n-2) \cdot ... \cdot (n-k+1). \tag{1.9}$$

**Приклад 1.7** Маємо множину, що складається з трьох елементів $\{1, 2, 3\}$. Скільки з трьох запропонованих цифр можна скласти різних чисел з цифрами, що не повторюються: а) тризначних; б) двозначних?

**Розв'язок задачі.**

- Кількість різних комбінацій з трьох цифр обчислюємо по формулі (1.1), тобто $P_3 = 3! = 6$. І дійсно рівні шість тризначних чисел можна скласти: 123, 132, 213, 231, 312, 321.
- Порядок у виборі двох елементів з трьох важливий, оскільки 23 і 32 різні числа, тому кількість способів вибрати 2 з трьох обчислюємо по формулі (1.3). Одержуємо $A_3^2 = \frac{3!}{1!} = 6$.

І дійсно рівні шість двозначних чисел можна скласти з трьох: 12, 13, 21, 23, 31, 32.

**Приклад 1.8** Скільки існує способів вибору трьох студентів з 10 на конференцію?

**Розв'язок задачі.** Порядок у виборі трьох елементів з десяти не важливий, оскільки делегація з Іванова, Петрова і Сидорова від делегації Петрова, Іванова і Сидорова не відрізняється, тому число способів вибору рівне числу поєднань з 10 студентів по 3

$$C_{10}^3 = \frac{10!}{3!7!} = 120 \tag{1.10}$$

**Означення 1.15** Біном Ньютона:

$$(a+b)^n = \sum_{k=0}^{n} C_n^k a^k b^{n-k} \tag{1.11}$$

## 1.3. Класичне визначення ймовірності

Вірогідність події - це число, що характеризує ступінь можливості появи події [1-5]. Вірогідність появи події $A$ обчислюється по формулі

$$P(A) = \frac{m}{n}, \tag{1.12}$$

де $m$ - число результатів, що сприяють настанню події $A$, $n$ — число всіх можливих результатів.

Властивості ймовірності

- Вірогідність випадкової події $A : 0 < P(A) < 1$.
- Вірогідність достовірної події $A : P(\Omega) = 1$.
- Вірогідність неможливої події $A : P(I) = 0$.

**Приклад 1.9** Монету кидають один раз. Обчислити вірогідність того, що випаде герб.

**Розв'язок задачі.** Розглянемо подію $A = \{$випав герб$\}$. При киданні монети можливі два результати $(n = 2)$ : випав герб і випала решка, і лише перший результат сприяє настанню події $A$ $(m = 1)$ Тоді по класичному визначенню ймовірності $P(A) = \frac{1}{2}$.

**Приклад 1.10** Гральну кістку підкидають один раз. Обчислити вірогідність того, що випаде просте число очок.

**Розв'язок задачі.** Розглянемо подію $A = \{$випало просте число очок$\}$. При киданні гральної кістки можливі шість результатів випадання $\{1, 2, 3, 4, 5, 6\}$ очок, тобто $n = 6$. З шести всіляких результатів тільки чотири результати $\{$випало 1 очко$\}$, $\{$випало 2 очки$\}$, $\{$випало 3 очки$\}$ і $\{$випало 5 очок$\}$ є сприяючими настанню події $A$ $(m = 4)$. Вірогідність появи простого числа очок рівна $P(A) = \frac{4}{6} = \frac{2}{3}$.

**Приклад 1.11** В урні 4 білих і 6 чорних кулі. Витягнули одночасно 3 кулі. Знайти вірогідність того, що: а) всі кулі білі; б) всі кулі чорні; в) один білий і два чорних.

**Розв'язок задачі.** Вважатимемо, що всі кулі різні і пронумеровані

$1, 2, 3, ..., 10$. Хай перші чотири номери - білі, останні - чорні. Тоді кількість всіляких способів витягнути три кулі з десяти рівно $C_{10}^3 = 240$.

- Розглянемо подію $A = \{$з урни витягнули три білі кулі$\}$. Тоді кількість способів, що сприяють появі події $A$ рівно $C_4^3 = 4$. Отримуємо $P(A) = \dfrac{4}{240} = \dfrac{1}{60}$.
- Розглянемо подію $B = \{$з урни витягнули три чорні кулі$\}$. Тоді кількість способів, що сприяють появі цієї події рівно $C_6^3 = 20$. Отримуємо $P(A) = \dfrac{20}{240} = \dfrac{1}{12}$.
- Розглянемо подію $B = \{$з урни витягнули один білий і дві чорні кулі$\}$. Тоді кількість способів, що сприяють появі цієї події рівно добутку $C_4^1 C_6^2 = 60$. Отримуємо $P(C) = \dfrac{60}{240} = \dfrac{1}{4}$.

## 1.4. Статистичне визначення ймовірності

Хай проводяться n випробувань, в результаті яких подія $A$ наступає $m$ раз. Тоді відношення $\dfrac{m}{n}$ при $n \to \infty$ називається статистичною вірогідністю.

**Приклад 1.12** Вивчали вірогідність народження хлопчика. Серед 1000 новонароджених хлопчик з'явився в 515 випадках.

**Розв'язок задачі.** Статистична вірогідність (відносна частота) народження хлопчика рівна $\dfrac{515}{1000}$.

## 1.5. Геометричне визначення ймовірності

Геометричною вірогідністю події $A$ називається відношення міри області $S$, що стосується появи події $A$, до міри всієї області $S_D$

$$P_(A) = \frac{S}{S_D}. \tag{1.13}$$

Області можуть бути одновимірними, двовимірними, тривимірними.

**Приклад 1.13** Хай в квадрат, із стороною 3 см вписаний круг. Знайти вірогідність того, що точка, випадковим чином кинута в квадрат, потрапить в круг.

**Розв'язок задачі.** Позначимо подію $A = \{$точка потрапила в круг$\}$. Площа квадрата рівна 9 см$^2$ а площа круга радіусом 1.5 см рівна $2.25\pi$ см$^2$ Тоді $P(A) = \dfrac{2.25\pi}{9} = 0.25\pi$.

### 1.6. Питання для самоперевірки

- Які події називають достовірними, неможливими, випадковими?
- Що є предметом теорії імовірностей?
- Як визначають та позначають суму, добуток випадкових подій, протилежну подію, повну групу подій?
- Які події називають сумісними, несумісними, рівноможливими?
- Як визначають та в яких випадках використовують класичне та геометричне означення імовірності?
- Як визначають та позначають частость випадкової події A?
- Які основні властивості імовірності та частості?
- Що є предметом комбінаторики?
- Які комбінації називають переставленням, розміщенням, сполученням? Як позначають та обчислюють кількість цих сполук?
- Як формулюють основні принципи комбінаторики?

## 2. Головні теореми теорії ймовірності [1-5]

### 2.1. Додавання імовірностей несумісних подій

**Теорема 2.1** *Імовірність об'єднання двох випадкових несумісних подій дорівнює сумі їх імовірностей*

$$P(A \cup B) = P(A) + P(B). \tag{2.1}$$

**Доведення.** Нехай число усіх можливих елементарних наслідків появи подій $A$ та $B$ дорівнює $n$; $m1$ та $m2$ — числа наслідків, що сприяють подіям $A$ та $B$ відповідно. Тоді події $A \cup B$ будуть сприяти $m1 + m2$ наслідків. Отже, за класичним означенням імовірності, маємо

$$P(A \cup B) = \frac{m1 + m2}{n} = \frac{m1}{n} + \frac{m2}{n} = P(A) + P(B). \tag{2.2}$$

тобто твердження теореми доведено.

Зовсім аналогічно можна довести наступне твердження.

**Теорема 2.2** *Якщо випадкові події $A_1, A_2, ..., A_n$ попарно несумісні, то імовірність появи хоча б однієї з цих подій дорівнює сумі їх імовірностей*

$$P(A_1 \cup A_1 \cup ... \cup A_n) = P(A_1) + P(A_2) + ... + P(A_n). \tag{2.3}$$

**Приклад 2.1** Імовірність влучення стрілком у першу область мішені дорівнює 0.45, у другу область — 0.35, у третю — 0.15. Знайти імовірність того, що при одному пострілі стрілок влучить у першу або другу області мішені.

**Розв'язок задачі.** Позначимо за подію $A_1$ — влучення у першу область мішені; за подію $A_2$ — влучення у другу область мішені. При одному пострілі події $A_1$ та $A_2$ несумісні. Тому імовірність влучення в першу або другу області мішені буде

$$P(A_1 \cup A_2) = P(A_1) + P(A_2) = 0.45 + 0.35 = 0.8. \tag{2.4}$$

**Теорема 2.3** *Сума імовірностей повної групи випадкових подій дорівнює одиниці*

$$P(A_1 \cup A_2 + ... + A_n) = 1. \tag{2.5}$$

**Доведення.** Якщо випадкові події $A_1, A_2, ..., A_n$ утворюють повну групу, то вони попарно несумісні, а їх об'єднання буде достовірною подією. За Теоремою 2.1.2 маємо

$$P(\bigcup_{k=1}^{n} A_k) = P(A_1) + P(A_2) + ... + P(A_n). \tag{2.6}$$

Імовірність достовірної події дорівнює одиниці, тому

$$P(A_1 \cup A_2 \cup ... \cup A_n) = 1 \tag{2.7}$$

Ліві частини рівностей (2.6) та (2.7) однакові, тому праві частини будуть рівними, тобто має місце рівність (2.7). Теорема доведена.

**Наслідок.** Дві протилежні події $A_1$ та $\overline{A}_1$ утворюють повну групу, тому має місце рівність

$$P(A + \overline{A}) = 1 \tag{2.8}$$

з якої одержуємо формулу

$$P(\overline{A}) = 1 - P(A) \tag{2.9}$$

знаходження імовірності протилежної події.

**Приклад 2.2** Імовірність одержати повідомлення від певної особи на протязі доби дорівнює 0.25. Знайти імовірність того, що повідомлення на протязі доби від цієї особи не буде одержано.

**Розв'язок задачі.** Позначимо за подію $A$ — повідомлення від цієї особи на протязі доби поступить. За умовою задачі має місце співвідношення P(A) = 0.25. Протилежна подія $\overline{A}$ означає, що на протязі доби від цієї особи повідомлення не поступить. За формулою (2.9) одержимо

$$P(\overline{A}) = 1 - 0.25 = 0.75. \tag{2.10}$$

У страховій справі треба вираховувати, наприклад, таку задачу.

**Приклад 2.3** За статистичними показниками держави можна зробити висновок, що 68% чоловіків, які досягли 60-ліття, досягають також і 70-ліття. Яка імовірність того, що 60-річний чоловік не досягне свого 70-річчя?

**Розв'язок задачі.** Якщо подія $A$ — 60-річний чоловік досягає свого 70-ліття, то протилежна подія $\overline{A}$ — 60-річний чоловік не досягає свого 70-ліття. За умовою задачі $P(A) = 0.68$, тому за формулою (2.9) одержимо

$$P(\overline{A}) = 1 - 0.68 = 0.32. \tag{2.11}$$

Отже, використовуючи статистичні дані держави, можна обчислити імовірність того, що 32% 60-річних чоловіків помре на протязі 10 років.

## 2.2. Залежні та незалежні події, умовні імовірності

**Означення 2.1** Випадкові події $A$ та $B$ називають залежними, якщо імовірність появи однієї з них залежить від появи або непояви другої події. Якщо імовірність появи однієї події не залежить від появи або непояви другої, то такі події називають незалежними.

**Означення 2.2** Імовірність події $B$, обчислена при умові появи події $A$, називають умовною імовірністю події $B$ і позначають $P(B|A)$ або $P_A(B)$

**Приклад 2.4** В урні 10 куль: 3 білих і 7 чорних. Навмання беруть дві кулі. Нехай подія $A$ — взята біла куля; подія $B$ — взята чорна, куля.

Якщо кулю, яку взяли першою, повертають до урни, то імовірність появи другої кулі не залежить від того, яка взята перша куля. Якщо перша куля не повертається до урни, то імовірність другої події залежить від результату першого випробування. Якщо першою взяли білу кулю, то в урні залишилося 2 білих кулі та 7 чорних, тому

$$P_A(B) = \frac{7}{9}. \tag{2.12}$$

Якщо першою взяли чорну кулю, то в урні залишилося 3 білих кулі та 6 чорних куль, тому

$$P_B(B) = \frac{6}{9} = \frac{1}{3}. \tag{2.13}$$

Отже імовірність події $B$ залежить від появи або непояви події $A$.

**Зауваження.** Якщо події $A$ та $B$ незалежні, то умовна імовірність дорівнює безумовній імовірності, тобто

$$P_A(B) = P(B). \tag{2.14}$$

## 2.3. Множення імовірностей

**Теорема 2.4** *Імовірність сумісної появи двох випадкових подій $A$ та $B$ дорівнює добутку імовірностей однієї з цих подій та умовної імовірності другої події при умові, що перша подія з'явилась*

$$P(A \cdot B) = P(A) \cdot P_A(B) = P(B) \cdot P_B(A). \tag{2.15}$$

**Доведення.** Усі елементарні наслідки зобразимо у вигляді точок (див.рис.2.1). Нехай появі події $A$ сприяють $m_1$ наслідків, а появі події $B$

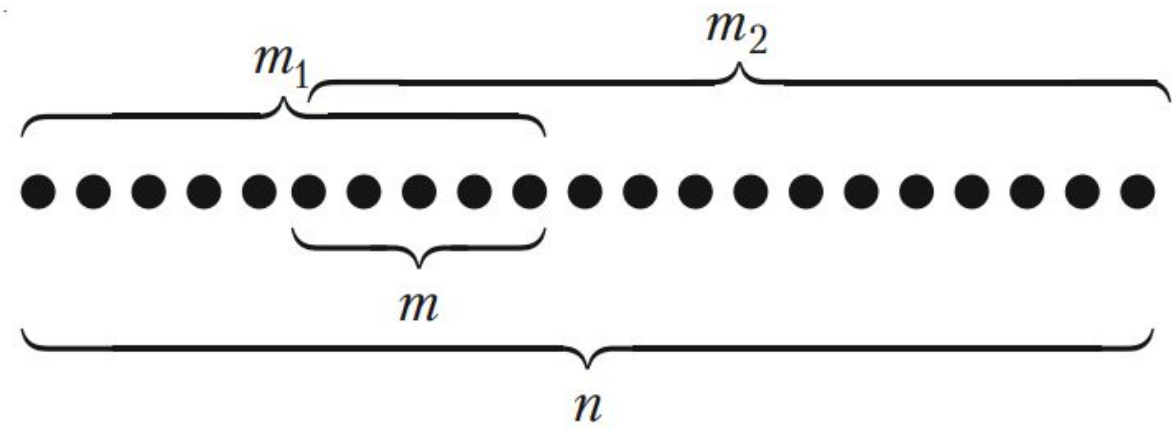


Рис. 2.1 — Множення імовірностей

— $m_2$ наслідків. Усіх можливих наслідків n, а події $A \cdot B$ будуть сприяти $m$ наслідків.

Так як

$$P(A \cdot B) = \frac{m}{n}, \quad P(A) = \frac{m_1}{n}, \quad P_A(B) = \frac{m}{m_1}, \tag{2.16}$$

то

$$P(A \cdot B) = \frac{m_1}{n}\frac{m}{m_1} = P(A) \cdot P_A(B). \tag{2.17}$$

що і треба було довести.

Співвідношення (2.15) називають формулою множення імовірностей залежних випадкових подій.

**Наслідок.** У випадку незалежних випадкових подій A та B формула (2.15) приймає вигляд

$$P(A \cdot B) = P(A) \cdot P(B) \tag{2.18}$$

і називається формулою множення імовірностей незалежних подій. У випадку скінченої кількості незалежних випадкових подій формула (2.15) приймає вигляд

$$P(A_1 \cdot A_2 \cdot ... \cdot A_n) = P(A_1) \cdot P(A_2) \cdot ... \cdot P(A_n). \tag{2.19}$$

**Приклад 2.5** У деякому людському суспільстві 70% палять, 40% хворіють на рак легенів та 25% палять та мають рак легенів. Знайти імовірність того, що навдачу взята особа з цього суспільства:

- не палить, але має рак легенів;
- палить, але не має раку легенів;
- ніколи не палить і не має раку легенів;
- палить і має рак легенів;
- або палить або має рак легенів.

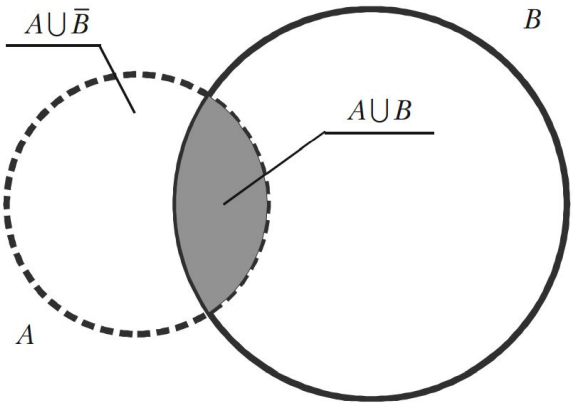


Рис. 2.2 — $P(A \cap B) + P(A \cap \overline{B}) = P(A)$

**Розв'язок задачі.** Позначимо події: $A$ — особа палить; $B$ — особа хворіє на рак легенів. Тоді за умовою задачі маємо

$$P(A) = 0.7, \quad P(B) = 0.4, \quad P(A \cdot B) = 0.25, \quad P(\overline{A}) = 0.3, \quad P(\overline{B}) = 0.6;$$

$$P(\overline{A} \cdot B) = 0.3 \cdot 0.4 = 0.12;$$

$$P(A \cdot \overline{B}) = 0.7 \cdot 0.6 = 0.42;$$

$$P(\overline{A} \cdot \overline{B}) = 0.3 \cdot 0.6 = 0.18;$$

$$P(A \cup B) = P(A) + P(B) - P(\overline{A} \cup \overline{B}) = 0.7 + 0.4 - 0.18 = 0.92;$$

$$P(A \cdot \overline{B} \cup \overline{A} \cdot B) = P(A \cdot \overline{B}) + P(\overline{A} \cdot B) = 0.42 + 0.12 = 0.54; \tag{2.20}$$

**Приклад 2.6** Навести ілюстративну діаграму властивості

$$P(A \cap B) + P(A \cap \overline{B}) = P(A). \tag{2.21}$$

**Відповідь.** Див.рис.2.2.

## 2.4. Імовірність появи хоча б однієї випадкової події

Нехай є $n$ сумісних випадкових подій $A_1, A_2, ..., A_n$. Позначимо через $A$ подію, яка полягає в тому, що з'явиться хоча б одна з цих подій. Тоді подія $\overline{A}$ полягає в тому, що події $\overline{A} = \overline{A}_1 \cdot \overline{A}_2 \cdot ... \cdot \overline{A}_n$. Події $A$ та $\overline{A}$ утворюють повну групу подій, тому

$$P(A) + P(\overline{A}) = 1, \quad P(A) = 1 - P(\overline{A}). \tag{2.22}$$

Звідси одержуємо

$$P(A) = 1 - P(\overline{A}_1 \cdot \overline{A}_2 \cdot ... \cdot \overline{A}_n). \tag{2.23}$$

За цією формулою треба обчислювати імовірність появи хоча б однієї випадкової події з $n$ сумісних подій.

**Приклад 2.7** Імовірність влучення у мішень першого стрілка дорівнює 0,7, другого стрілка — 0,8, а третього стрілка — 0,9. Знайти імовірність влучення у мішень хоча б одного стрілка.

**Розв'язок задачі.** Позначимо події

- $A_1$ — у мішень влучив перший стрілок;
- $A_2$ — у мішень влучив другий стрілок;
- $A_3$ — у мішень влучив третій стрілок;
- $A$ — у мішень влучив хоча б один стрілок.

За умовою задачі події $A_1$, $A_2$ та $A_3$ незалежні, тому події $\overline{A}_1$, $\overline{A}_2$ та $\overline{A}_3$ також незалежні.

Згідно формули (2.23) та формули множення імовірностей незалежних подій маємо

$$P(A) = 1 - P(\overline{A}_1 \cdot \overline{A}_2 \cdot \overline{A}_3) = 1 - P(\overline{A}_1) \cdot P(\overline{A}_2) \cdot P(\overline{A}_3). \tag{2.24}$$

Так як

$$\begin{aligned} &P(A_1) = 1 - 0.7 = 0.3; \quad P(A_2) = 1 - 0.8 = 0.2; \\ &P(A_3) = 1 - 0.9 = 0.1; \end{aligned} \tag{2.25}$$

то з формули (2.24) одержимо

$$P(A) = 1 - 0.3 \cdot 0.2 \cdot 0.1 = 1 - 0.006 = 0.994. \tag{2.26}$$

## 2.5. Теорема додавання імовірностей сумісних подій

**Теорема 2.5** *Якщо випадкові події $A$ та $B$ сумісні, то імовірність їх об'єднання дорівнює сумі їх імовірностей без імовірності їх сумісної появи, тобто*

$$P(A \cup B) = P(A) + P(B) - P(A \cdot B). \tag{2.27}$$

**Доведення.** Згідно з умовою теореми події $A$ та $B$ сумісні, тому $A \cup B$ з'явиться, якщо з'явиться одна з трьох несумісних подій $A \cdot \overline{B}$, $\overline{A} \cdot B$ або $A \cdot B$.

Згідно з теоремою додавання імовірностей несумісних подій одержимо

$$P(A \cup B) = P(A \cdot \overline{B}) + P(\overline{A} \cdot B) + P(A \cdot B). \tag{2.28}$$

Подія $A$ з'явиться, якщо з'явиться одна з двох несумісних подій $A \cdot \overline{B}$ або $A \cdot B$

Згідно з теоремою додавання імовірностей несумісних подій

$$P(A) = P(A \cdot \overline{B}) + P(A \cdot B), \quad P(A \cdot \overline{B}) = P(A) - P(A \cdot B). \tag{2.29}$$

Аналогічно одержимо

$$P(B) = P(\overline{A} \cdot B) + P(A \cdot B), \quad P(\overline{A} \cdot B) = P(B) - P(A \cdot B). \tag{2.30}$$

Підставимо (2.29) та (2.30) у формулу (2.28), тоді одержимо рівність (2.27), яку треба було довести.

**Зауваження.** Якщо події $A$ та $B$ незалежні, то формула (2.28) приймає вигляд

$$P(A \cup B) = P(A) + P(B) - P(A) \cdot P(B). \tag{2.31}$$

Для залежних випадкових подій маємо

$$P(A \cup B) = P(A) + P(B) - P(A) \cdot P_A(B). \tag{2.32}$$

**Приклад 2.8** У залежності від наявності сировини підприємство може виробити та відправити замовникам щодобово кількість певної продукції від 1 до 100. Яка імовірність того, що одержану кількість продукції можна розподілити без залишку?

- трьом замовникам;
- чотирьом замовникам;

- дванадцяти замовникам;
- трьом або чотирьом замовникам?

**Розв'язок задачі.** Позначимо події $A$ — одержана кількість виробів ділиться на 3 без залишку; $B$ — одержана кількість виробів ділиться на 4 без залишку. Використовуючи класичне означення імовірності, знаходимо

- $P(A) = \dfrac{33}{100}$
- $P(B) = \dfrac{25}{100}$
- $P(A \cdot B) = \dfrac{8}{100}$

Події $A$ та $B$ – сумісні, тому за формулою (2.28) одержимо

- $P(A \cup B) = P(A) + P(B) - P(A \cdot B) = \dfrac{33}{100} + \dfrac{25}{100} - \dfrac{8}{100} = \dfrac{1}{2}.$

## 2.6. Надійність системи

**Означення 2.3** Надійністю системи називають імовірність її безвідмовної роботи в певний час $t$ (гарантійний термін).

Системи складаються з елементів, поєднаних послідовно або паралельно При


Рис. 2.3 — Елементи системи поєднані послідовно

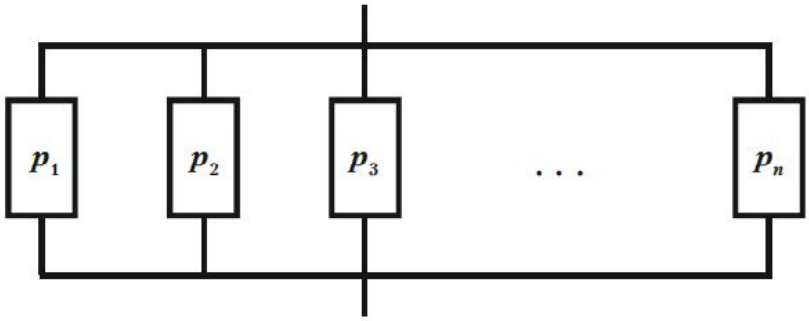


Рис. 2.4 — Елементи системи поєднані паралельно

обчисленні надійності систем необхідно виразити надійність системи через надійність елементів та блоків. Надійність елементів вважається відомою, бо вона пов'язана з технологією їх виготовлення.

Позначимо за $p_k$ надійність $k$ — того елемента, $q_k$ — імовірність виходу з строю за час $t$ $k$ — того елемента, $P$ — надійність блоку.

Розглянемо блок, усі елементи якого незалежні і з'єднані послідовно (див.рис.2.3). Такий блок буде працювати безвідмовно лише в той час, коли усі елементи працюють безвідмовно. Згідно теореми множення імовірностей незалежних подій імовірність $P$ безвідмовної роботи такого блоку буде

$$P = p_1 \cdot p_2 \cdot ... \cdot p_n. \tag{2.33}$$

Тепер розглянемо блок, елементи якого з'єднані паралельно (див.рис.2.4). Такий блок буде працювати безвідмовно, якщо хоч один елемент не вийде зі строю. Тому імовірність $P$ безвідмовної роботи буде

$$P = 1 - q_1 \cdot q_2 \cdot ... \cdot q_n. \tag{2.34}$$

Будь-яку складну систему можна розглядати як послідовне або паралельне з'єднання блоків, надійність яких обчислюють за формулами (2.33) та (2.34).

**Приклад 2.9** Прилад складено з двох блоків, з'єднаних послідовно і незалежно працюючих. Імовірність відмови блоків дорівнює 0.05 та 0.08. Знайти імовірність відмови приладу.

**Розв'язок задачі.** Відмова приладу є подія протилежна його безвідмовної роботи. Імовірності безвідмовної роботи блоків будуть

$$p_1 = 1 - 0.05; \quad p_2 = 1 - 0.08. \tag{2.35}$$

Імовірність безвідмовної роботи приладу буде згідно формули (2.33)

$$P_1 = 0.95 \cdot 0.92 = 0.874. \tag{2.36}$$

Тому імовірність відмови приладу буде

$$P = 1 - 0.874 = 0.126. \tag{2.37}$$

## 2.7. Формули повної імовірності та Байєса

**Теорема 2.6** *Якщо випадкова подія $A$ може з'явитись лише сумісно з однією із несумісних між собою подій $B_1, B_2, ..., B_n$, що утворюють повну групу, тоді імовірність події $A$ обчислюється за формулою*

$$P(A) = \sum_{k=1}^{n} P(B_k) P_{B_k}(A). \tag{2.38}$$

Доведення. За умовою теореми поява події $A$ означає появу однієї з подій

$AB_1, AB_2, ..., AB_n$, тобто

$$A = AB_1 \cup AB_2 \cup ... \cup AB_n. \tag{2.39}$$

Події $B_1, B_2, ..., B_n$ несумісні, тому й події $AB_1, AB_2, ..., AB_n$ також несумісні. Згідно з теоремою додавання імовірностей несумісних подій маємо

$$P(A) = P(AB_1 \cup AB_2 \cup ... \cup AB_n) = \sum_{k=1}^{n} P(AB_k). \tag{2.40}$$

Події $A$ та $B_k$ — залежні, тому для обчислення $P(AB_k)$ можна використати теорему множення імовірностей залежних подій, тобто

$$P(AB_k) = P(B_k)P_{B_k}(A) \tag{2.41}$$

Підставимо (2.41) у формулу (2.40) і одержимо рівність (2.38), яку треба було довести.

Формулу (2.38) називають формулою повної імовірності.

**Приклад 2.10** У першому ящику 20 деталей, з яких 15 стандартних. У другому ящику 10 деталей, з яких 9 стандартних. З другого ящика беруть навмання одну деталь і перекладають її до першого ящика. Знайти імовірність того, що взята після цього навмання деталь з першого ящика стандартна.

**Розв'язок задачі.** Позначимо такі події: $A$ – з першого ящика взято стандартну деталь; $B_1$ — з другого ящика переклали до першого стандартну деталь; $B_2$ — з другого ящика переклали до першого нестандартну деталь. Згідно з умовою задачі, з першого ящика можна взяти деталь лише після того, як здійсниться подія $B_1$ або подія $B_2$.

Події $B_1$ та $B_2$ несумісні, а подія $A$ може з'явитись лише сумісно з однією із них. Тому для знаходження імовірності події $A$ можна використати формулу повної імовірності (2.38), яка у даному випадку прийме вигляд

$$P(A) = P(B_1)P_{B_1}(A) + P(B_2)P_{B_2}(A). \tag{2.42}$$

Знайдемо потрібні імовірності

$$P(B_1) = \frac{9}{10}, \quad P(B_2) = \frac{1}{10}, \quad P_{B_1}(A) = \frac{16}{21}, \quad P_{B_2}(A) = \frac{15}{21}, \quad . \tag{2.43}$$

Підставимо ці значення у формулу (2.42) і одержимо

$$P(A) = \frac{9}{10} \cdot \frac{16}{21} + \frac{1}{10} \cdot \frac{15}{21} = \frac{144 + 15}{210} = \frac{53}{70}. \tag{2.44}$$

Тепер ознайомимось з формулами Байєса.

В умовах Теореми 2.1 невідомо, з якою подією із несумісних $B_1, B_2, ..., B_n$, подій з'явиться подія $A$. Тому кожну з подій $B_1, B_2, ..., B_n$, можна вважати гіпотезою. Тоді $P_A(B_k)$ імовірність $k$ — тої гіпотези.

Якщо випробування проведено і в результаті його подія $A$ з'явилась, то умовна імовірність $P_A(B_k)$ може не дорівнювати $P(B_k)$.

Порівняння імовірностей $P(B_k)$ та $P_A(B_k)$ дозволяє переоцінити імовірність гіпотези при умові, що подія $A$ з'явилася.

Для одержання умовної імовірності використовуємо теорему множення імовірностей залежних подій

$$\begin{aligned} P(AB_k) &= P(B_k)P_{B_k}(A) = P(A)P_A(B_k) \\ P_A(B_k) &= \frac{P(B_k)P_{B_k}(A)}{P(A)} \end{aligned} \tag{2.45}$$

Підставимо у формулу (2.45)) замість $P(A)$ її значення з формули повної імовірності. Одержимо

$$P_A(B_k) = \frac{P(B_k)P_{B_k}(A)}{\sum\limits_{k=1}^{n} P(B_k)P_{B_k}(A)}, \quad k = 1, 2, ..., n. \tag{2.46}$$

Формули (2.46) називають формулами Байєса. Вони дозволяють переоцінювати імовірності гіпотез. Це важливо при контролі або ревізіях.

**Приклад 2.11** Деталі, виготовлені цехом заводу, попадають для перевірки їх стандартності до одного з двох контролерів. Імовірність того, що деталь попаде до першого контролера, дорівнює 0.6, а до другого — 0.4. Імовірність того, що придатна деталь буде признана стандартною першим контролером, дорівнює 0.94, а другим — 0.98.

Придатна деталь при перевірці признана стандартною. Знайти імовірність того, що деталь перевіряв перший контролер.

**Розв'язок задачі.** Позначимо такі події: $A$ — придатна деталь признана стандартною; $B_1$ — деталь перевіряв перший контролер; $B_2$ — деталь перевіряв другий контролер. За умовою прикладу

$$P(B_1) = 0.6; \quad P(B_2) = 0.4; \quad P_{B_1}(A) = 0.94; \quad P_{B_2}(A) = 0.98; \tag{2.47}$$

За формулою Байєса (2.46) при $k = 1$ одержимо

$$P_A(B_1) = \frac{P(B_1)P_{B_1}(A)}{P(B_1)P_{B_1}(A) + P(B_2)P_{B_2}(A)} = \frac{0.6 \cdot 0.94}{0.6 \cdot 0.94 + 0.4 \cdot 0.98} = 0.59. \tag{2.48}$$

Відмітимо, що до появи події $A$ імовірність $P(B_1) = 0.6$ а після появи події $A$ імовірність перевірки деталі першим контролером $P_A(B_1)$ поменшала.

**Приклад 2.12** Імовірність знищення літака з одного пострілу для першої гармати дорівнює 0.2, а для другої гармати — 0.1. Кожна гармата робить по одному пострілу, причому було одне влучення у літак. Яка імовірність того, що влучила перша гармата?

**Розв'язок задачі.** Позначимо такі події: $A$ — знищення літака з одного пострілу першою гарматою; $B$ — знищення літака з одного пострілу другою гарматою; $C$ — одне влучення у літак. Маємо чотири гіпотези

$$H_1 = A \cdot B; \quad H_2 = A \cdot \overline{B}; \quad H_3 = \overline{A} \cdot B; \quad H_4 = \overline{A} \cdot \overline{B} \tag{2.49}$$

які утворюють повну групу подій. Імовірностями цих гіпотез будуть

$$\begin{aligned} P(H_1) &= 0.2 \cdot 0.1 = 0.02; \quad P(H_2) = 0.2 \cdot 0.9 = 0.18; \\ P(H_3) &= 0.8 \cdot 0.1 = 0.08; \quad P(H_4) = 0.8 \cdot 0.9 = 0.72; \end{aligned} \tag{2.50}$$

Так як сума

$$H_1 + H_2 + H_3 + H_4 \tag{2.51}$$

є достовірна подія, то

$$P(H_1) + P(H_2) + P(H_3) + P(H_4) = 1. \tag{2.52}$$

Умовні імовірності події $C$ будуть

$$P_{H_1}(C) = 0; \quad P_{H_2}(C) = 1; \quad P_{H_3}(C) = 1; \quad P_{H_4}(C) = 0. \tag{2.53}$$

Тепер за формулою Байєса знаходимо шукану імовірність

$$P_C(H_2) = \frac{0.18 \cdot 1}{0.18 \cdot 1 + 0.08 \cdot 1} = 0.7. \tag{2.54}$$

## 2.8. Питання для самоперевірки

- Як формулюють і якими формулами записують теореми додавання імовірностей сумісних та несумісних подій?
- Які випадкові події називають незалежними?
- Як визначають та позначають умовну імовірність?

- Як формулюють і якими формулами записують теореми множення імовірностей залежних та незалежних випадкових подій?
- За якою формулою можна обчислити імовірність появи хоча б однієї з $n$ сумісних подій?
- Що називають надійністю системи? Як знайти імовірність безвідмовної роботи блоку, усі елементи якого незалежні і з'єднані послідовно, паралельно? Як треба розглядати та досліджувати надійність будь-якої складної системи?
- Яким умовам повинна задовольняти подія, щоб її імовірність можна було знаходити за формулою повної імовірності? Який вигляд має ця формула?
- Коли застосовують формули Байєса та як їх записують?

# 3. Послідовності випробувань

## 3.1. Схема та формула Бернуллі

У багатьох задачах теорії імовірностей, статистики та повсякденної практики треба досліджувати послідовність (серію) $n$ випробувань. Наприклад, випробування «кинуто 1000 однакових монет» можна розглядати як послідовність 1000 більш простих випробувань — «кинута одна монета». При киданні 1000 монет імовірність появи герба або надпису на одній монеті не залежить від того, що з'явиться на інших монетах. Тому можна казати, що у цьому випадку випробування повторюються 1000 разів незалежним чином.

**Означення 3.1** Якщо усі $n$ випробувань проводити в однакових умовах і імовірність появи події $A$ в усіх випробуваннях однакова та не залежить від появи або непояви $A$ в інших випробуваннях, то таку послідовність незалежних випробувань називають схемою Бернуллі.

Нехай випадкова подія $A$ може з'явитись у кожному випробуванні з імовірністю $P(A) = p$ або не з'явитись з імовірністю $q = P(\overline{A}) = 1 - p$

Поставимо задачу: знайти імовірність того, що при $n$ випробуваннях подія $A$ з'явиться $m$ разів і не з'явиться $n - m$ разів. Шукану імовірність позначимо $P_n(m)$

Спочатку розглянемо появу події $A$ три рази в чотирьох випробуваннях. Можливі такі події

$$AAA\overline{A}, AA\overline{A}A, A\overline{A}AA, \overline{A}AAA, \tag{3.1}$$

тобто їх буде $C_4^3 = 4$

Якщо подія $A$ з'явилася 2 рази в 4 випробуваннях, то можливі такі події

$$AA\overline{A}\overline{A}, A\overline{A}A\overline{A}, A\overline{A}\overline{A}A, \overline{A}AA\overline{A}, \overline{A}A\overline{A}A, \overline{A}\overline{A}AA, \tag{3.2}$$

тобто їх буде $C_4^2 = 6$.

У загальному випадку, коли подія $A$ з'являється $m$ разів у $n$ випробуваннях, таких складних подій буде

$$C_n^m = \frac{n!}{m!(n-m)!}. \tag{3.3}$$

Обчислимо імовірність однієї складної події, наприклад,

$$\underbrace{A \cdot A \cdot \ldots \cdot A}_{m} \cdot \underbrace{\overline{A} \cdot \overline{A} \cdot \ldots \cdot \overline{A}}_{n-m} \tag{3.4}$$

Імовірність сумісної появи $n$ незалежних подій дорівнює добутку імовірностей цих подій згідно з теоремою множення імовірностей, тобто

$$\begin{aligned} P(A) = P\left( \underbrace{A \cdot A \cdot \ldots \cdot A}_{m} \cdot \underbrace{\overline{A} \cdot \overline{A} \cdot \ldots \cdot \overline{A}}_{n-m} \right) = \\ P\left( \underbrace{A \cdot A \cdot \ldots \cdot A}_{m} \right) \cdot P\left( \underbrace{\overline{A} \cdot \overline{A} \cdot \ldots \cdot \overline{A}}_{n-m} \right) = \\ P(A)^m \cdot P(\overline{A})^{n-m} = p^m \cdot q^{n-m}. \end{aligned} \tag{3.5}$$

Кількість таких складних подій $C_n^m$ і вони несумісні. Тому, згідно з теоремою додавання імовірностей несумісних подій, маємо

$$P_n(m) = C_n^m p^m q^{n-m} \tag{3.6}$$

Формулу (3.6) називають формулою Бернуллі. Вона дозволяє знаходити імовірність появи події $A$ $m$ разів при $n$ випробуваннях, які утворюють схему Бернуллі [1-5].

**Зауваження №1**. Імовірність появи події $A$ в $n$ випробуваннях схеми Бернуллі менше $m$ разів знаходять за формулою

$$P_n(k < m) = \sum_{k=0}^{m-1} P_n(k) \tag{3.7}$$

Імовірність появи події $A$ не менше $m$ разів можна знайти за формулою

$$P_n(k \geq m) = \sum_{k=m}^{n} P_n(k), \tag{3.8}$$

або за формулою

$$P_n(k < m) = 1 - \sum_{k=0}^{m-1} P_n(k). \tag{3.9}$$

Імовірність появи події $A$ хоча б один раз у $n$ випробуваннях доцільно знаходити за формулою

$$P_n(1 \le m \le n) = 1 - q^n. \tag{3.10}$$

**Зауваження №2.** У багатьох випадках треба знаходити найбільш імовірне значення числа $m_0$ появ події $A$. Це значення $m$ визначається співвідношеннями

$$np - q \le m_0 \le np + p \quad \text{або} \quad (n+1)p - 1 \le m_0 \le (n+1)p. \tag{3.11}$$

Число $m_0$ повинно бути цілим. Якщо $(n+1)p$ — ціле число, тоді найбільше значення імовірність має при двох числах

$$m_1 = (n+1)p - 1 \quad \text{та} \quad m_2 = (n+1)p. \tag{3.12}$$

**Зауваження №3.** Якщо імовірність появи події $A$ в кожному випробуванні дорівнює $p$, то кількість $n$ випробувань, які необхідно здійснити, щоб з імовірністю $P$ можна було стверджувати, що подія $A$ з'явиться хоча б один раз, знаходять за формулою

$$n > \frac{\ln(1-P)}{\ln(1-p)}. \tag{3.13}$$

**Приклад 3.1** Прилад складено з 10 блоків, надійність кожного з них 0.8. Блоки можуть виходити з ладу незалежно один від одного. Знайти імовірність того, що

- відмовлять два блоки;
- відмовить хоча б один блок;
- відмовлять не менше двох блоків.

**Розв'язок задачі.** Позначимо за подію $A$ відмову блока. Тоді імовірність події $A$ за умовою задачі буде

$$P(A) = p = 1 - 0.8 = 0.2 \quad \text{тому} \quad q = 1 - p = 1 - 0.2 = 0.8. \tag{3.14}$$

Згідно з умовою задачі $n = 10$. Використовуючи формулу Бернуллі та **Зауваження №1,** одержимо

- $P_{10}(2) = C_{10}^2 p^2 q^8 = C_{10}^2 \cdot 0.2^2 \cdot 0.8^8 = 0.202$;

- $P_{10}(1 < m \leq 10) = 1 - P_{10}(0) = 1 - C_{10}^0 \cdot 0.2^0 \cdot 0.8^{10} = 0.8926$;
- $P_{10}(2 \leq m \leq 10) = 1 - (P_{10}(0) + P_{10}(1)) =$
  $1 - (C_{10}^0 \cdot 0.2^0 \cdot 0.8^{10} + C_{10}^1 \cdot 0.2^1 \cdot 0.8^9) = 0.6244.$

**Приклад 3.2** За одну годину автомат виготовляє 20 деталей. За скільки годин імовірність виготовлення хоча б однієї бракованої деталі буде не менше 0.952, якщо імовірність браку будьякої деталі дорівнює 0.01?

**Розв'язок задачі.** Застосовуючи формулу (3.13), знайдемо спочатку таку кількість виготовлених деталей, щоб з імовірністю $P = 0.952$ можна було стверджувати про наявність хоча б однієї бракованої деталі, якщо імовірність браку за умовою $= 0.01$

$$n \geq \frac{\ln(1 - 0.952)}{\ln(1 - 0.01)} = \frac{\ln(0.048)}{\ln(0.99)} = 300. \tag{3.15}$$

Отже, за час $\Delta t = \dfrac{300}{20}$ (годин) автомат з імовірністю 0.952 виготовить хоча б одну браковану деталь.

**Приклад 3.3** При новому технологічному процесі 80% усієї виготовленої продукції має найвищу якість. Знайти найбільш імовірне число виготовлених виробів найвищої якості серед 250 виготовлених виробів.

**Розв'язок задачі.** Позначимо шукане число $m_0$. Згідно **Зауваження №2**

$$np - q \leq m_0 \leq np + q. \tag{3.16}$$

За умовою прикладу $n = 250$, $p = 0.8$, $q = 0.2$, а тому

$$199.8 \leq m_0 \leq 200.8. \tag{3.17}$$

Але $m_0$ повинно бути цілим числом, тому $m_0 = 200$.

## 3.2. Граничні теореми у схемі Бернуллі

Знаходження імовірностей $P_n(m)$ та $P_n(m_1 \leq m \leq m_2)$ за формулою Бернуллі ускладнюється при досить великих значеннях $n$ та при малих $p$ або $q$. У таких випадках часто можна використовувати замість формули Бернуллі наближені асимптотичні формули.

Вкажемо без доведення три граничні теореми, які містять наближені формули для імовірностей

$$P_n(m) \quad \text{та} \quad P_n(m_1 \leq m \leq m_2). \tag{3.18}$$

**Теорема 3.1** *(Теорема Пуассона). Якщо* $n \to \infty$ *і* $p \to 0$ *так, що* $np \to \lambda$, $0 < \lambda < \infty$, *то*

$$P_n(m) = C_n^m p^m q^{n-m} \to \frac{\lambda^m}{m!} e^{-\lambda} \tag{3.19}$$

*для будь-якого постійного* $m = 0, 1, 2, \cdots$.

**Наслідок.** Імовірність появи події $A$ $m$ разів у $n$ випробуваннях схеми Бернуллі можна знаходити за наближеною формулою Пуассона

$$P_n(m) = \frac{\lambda^m}{m!} e^{-\lambda} \tag{3.20}$$

де $\lambda = np$.

Формулу (3.20) доцільно застосовувати при великих $n$ та малих $p$.

**Приклад 3.4** Підручник надруковано тиражем 100000 екземплярів. Імовірність невірного брошурування підручника дорівнює 0.0001. Знайти імовірність того, що тираж має 5 бракованих підручників.

**Розв'язок задачі.** Брошурування кожного підручника можна розглядати як випробування. Випробування незалежні і мають однакову імовірність невірного брошурування, тому задача вкладається у схему Бернуллі. Згідно з умовою задачі $n = 100000$ досить велике; $p = 0.0001$ мала; $m = 5$. Застосовуючи формулу Пуассона (1), одержимо

$$P_{100000}(5) = \frac{10^5}{5!} e^{-10} = 0.0375. \tag{3.21}$$

Для наведення ще двох граничних теорем треба спочатку визначити локальну та інтегральну функції Лапласа та ознайомитись з їх основними властивостями.

**Означення 3.2** Локальною функцією Лапласа називають функцію вигляду

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \tag{3.22}$$

Основні властивості локальної функції Лапласа

- Функція Лапласа $\varphi(x)$ парна, тобто $\varphi(x) = \varphi(-x)$;
- Функція $\varphi(x)$ визначена для усіх $x \in (\infty, -\infty)$
- $\varphi(x) \to 0$, коли $x \to \pm\infty$
- $\varphi_{max} = \varphi(0) = \dfrac{1}{\sqrt{2\pi}}$

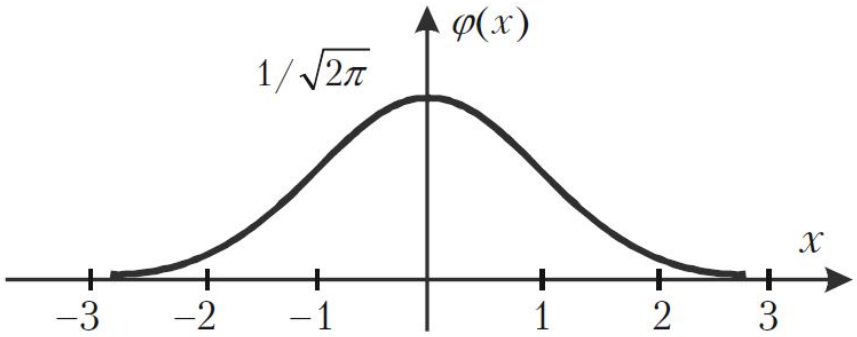


Рис. 3.1 —

Графік локальної функції Лапласа має вигляд, зображений на мал.3.1

**Означення 3.3** Інтегральною функцією Лапласа називають функцію

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{x^2}{2}} dt \tag{3.23}$$

Легко бачити, що між локальною функцією $\varphi(x)$ та інтегральною функцією

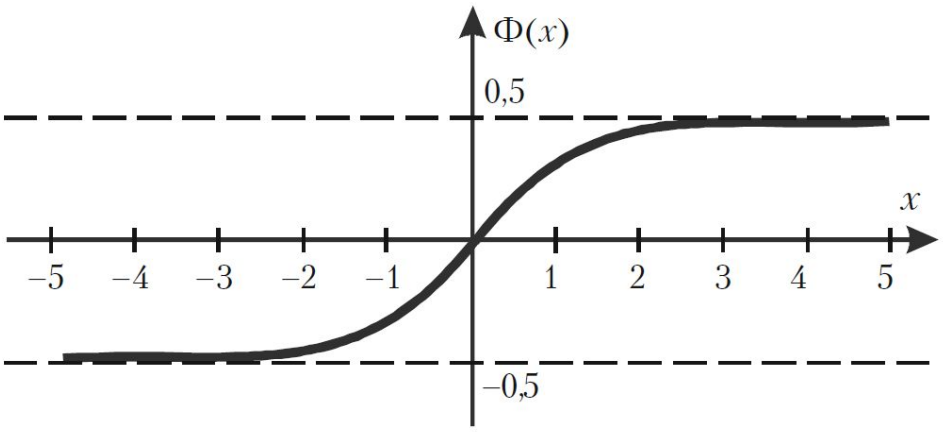


Рис. 3.2 —

$\Phi(x)$ існує простий зв'язок

$$\Phi(x) = \int_{-\infty}^{x} \varphi(t)dt. \tag{3.24}$$

Основні властивості інтегральної функції Лапласа

- Інтегральна функція Лапласа є непарною функцією $\Phi(-x) = -\Phi(x)$.
- $\Phi(0) = 0$
- $\Phi(x) = 0.5 \quad$ при $\quad x \geq 5$.

Графік інтегральної функції Лапласа зображено на мал.3.2

**Теорема 3.2** *(локальна теорема Муавра-Лапласа). Якщо у схемі Бернуллі кількість випробувань n достатньо велика, а імовірність p появи події A в усіх випробуваннях однакова, то імовірність появи події A m разів може бути знайдена за наближеною формулою*

$$P_n(m) = \frac{1}{\sqrt{npq}}\varphi(x_m), \quad x_m = \frac{m - np}{\sqrt{npq}} \tag{3.25}$$

Формулу (3.25) доцільно використовувати при $n > 100$ та $npq > 20$.

**Приклад 3.5** Гральний кубик кидають 800 разів. Яка імовірність того, що кількість очок, кратна трьом, з'явиться 267 разів.

**Розв'язок задачі.** У даному випадку $n$ та $m$ досить великі. Тому для знаходження $P_{800}(267)$ можна використати формулу (3.25). Маємо

$$P(A) = p = \frac{2}{6}, \quad q = 1 - p = \frac{2}{3}, \tag{3.26}$$

$$x_{267} = \frac{m - np}{\sqrt{npq}} = \frac{267 - 800\frac{1}{3}}{\frac{40}{3}} = 0.025. \tag{3.27}$$

Отже, за формулою (3.25) одержимо

$$P_{800}(267) = \frac{3}{40} \cdot \varphi(0.025) = \frac{3}{40} \cdot 0.3988 = 0.03. \tag{3.28}$$

**Теорема 3.3** *(Інтегральна теорема Муавра-Лапласа). Якщо у схемі Бернуллі в кожному із n незалежних випробувань подія A може з'явитися з постійною імовірністю p, тоді імовірність появи події A не менш $m_1$ та не більш $m_2$ разів може бути знайдена за формулою*

$$P_n(m_1 \le m \le m_2) = \Phi(x_2) - \Phi(x_1), \tag{3.29}$$

*де* $\Phi = (x)$ *– інтегральна функція Лапласа,*

$$x_1 = \frac{m_1 - np}{\sqrt{npq}} \quad x_2 = \frac{m_2 - np}{\sqrt{npq}}. \tag{3.30}$$

**Приклад 3.6** Гральний кубик кидають 800 разів. Яка імовірність того, що кількість очок, кратна трьом, з'явиться не менше 260 та не більше 274 разів?

**Розв'язок задачі.** Для знаходження імовірності $p_{800}(260 \le m \le 274)$

використаємо формули (3.29) та (3.30). Маємо

$$x_1 = \frac{260 - 800\frac{1}{3}}{\frac{40}{3}} = -0.5, \quad x_2 = \frac{274 - 800\frac{1}{3}}{\frac{40}{3}} = 0.55, \tag{3.31}$$

$$P_{800}(260 \le m \le 274) = \Phi(0.55) + \Phi(0.5) = 0.298840 + 0.191482 = 0.400322. \tag{3.32}$$

## 3.3. Послідовність випробувань із різними імовірностями

У схемі Бернуллі імовірність появи події $A$ в усіх випробуваннях однакова. Але у практичній діяльності іноді зустрічаються і такі випадки, коли у $n$ незалежних випробуваннях імовірності появи події $A$ різні, наприклад, вони дорівнюють $p_1, p_2, ..., p_n$. Тоді імовірності непояви події $A$ також будуть різними

$$q_1 = 1 - p_1, \quad q_2 = 1 - p_2, ..., \quad q_n = 1 - p_n. \tag{3.33}$$

У цьому випадку не можна обчислювати за формулою Бернуллі імовірність появи події $A$ $m$ разів у $n$ випробуваннях, а треба використовувати твірну функцію

$$\varphi_n(z) = \prod_{k=1}^{n}(q_k + p_k z). \tag{3.34}$$

**Правило №1.** Шукана імовірність $P_n(m)$ дорівнює коефіцієнту, що стоїть при $z^m$.

**Приклад 3.7** Імовірність відмови кожного з 4 приладів у 4 незалежних випробуваннях різні і дорівнюють

$$p_1 = 0.1, \quad p_2 = 0.2, \quad p_3 = 0.3, \quad p_4 = 0.4. \tag{3.35}$$

Знайти імовірність того, що внаслідок випробувань

1. не відмовить жоден прилад;
2. відмовлять один, два, три, чотири прилади;
3. відмовить хоча б один прилад;
4. відмовлять не менше двох приладів.

**Розв'язок задачі.** Імовірності відмови приладів у випробуваннях різні, тому застосовуємо твірну функцію 3.34, яка у даному випадку матиме вигляд

$$\varphi_n(z) = (0.9 + 0.1z) \cdot (0.8 + 0.2z) \cdot (0.7 + 0.3z) \cdot (0.6 + 0.4z). \tag{3.36}$$

Розкриємо дужки та зведемо подібні члени. Тоді матимемо

$$\varphi_n(z) = 0.302 + 0.46z + 0.205z^2 + 0.031z^3 + 0.002z^4 \tag{3.37}$$

Відповідно **Правилу №1** звідси одержуємо відповіді на питання в задачі

1. $P_4(0) = 0.302$;
2. $P_4(1) = 0.46$; $P_4(2) = 0.205$; $P_3(1) = 0.031$; $P_4(4) = 0.002$;
3. $P_4(1 \le m \le 4) = 1 - P_4(0) = 0.698$;
4. $P_4(m \ge 2) = 1 - (P_4(0) + P_4(1)) = 1 - (0.302 + 0.46) = 0.238$.

**Приклад 3.8** Працівник обслуговує три станка, що працюють незалежно один від одного. Імовірність того, що на протязі години перший станок не вимагатиме уваги працівника, дорівнює 0.9, а для другого та третього станків — 0.8 та 0.85, відповідно. Якою є імовірність того, що на протязі години

1. жоден станок не потребуватиме уваги працівника;
2. усі три станки потребують уваги працівника;
3. хоча б один станок потребує уваги працівника?

**Розв'язок задачі.** Цей приклад можна розв'язати з використанням теорем множення та додавання імовірностей. Розв'яжемо тепер цей приклад з використанням твірної функції, яка у даному випадку прийме вигляд

$$\begin{aligned} &\varphi_n(z) = \prod_{k=1}^{3}(q_k + p_k z) = \\ &(0.1 + 0.9z) \cdot (0.2 + 0.8z) \cdot (0.15 + 0.85z) = \\ &0.003 + 0.056z + 0.0329z^2 + 0.612z^3. \end{aligned} \tag{3.38}$$

Отже, коефіцієнт при $z^k (k = 0, 1, 2, 3)$ дорівнює імовірності того, що на протязі години уваги працівника не потребують $k$ станків. Тому одержуємо відповіді на питання цього прикладу:

1. імовірність того, що усі три станка не потребують уваги працівника, дорівнює коефіцієнту при $z^3$, тобто $P_3(3) = 0.612$;
2. $P_3(0) = 0.003$;
3. $P_3(1 \le m \le 3) = 1 - P_3(3) = 1 - 0.612 = 0,388$;

## 3.4. Теорема Бернуллі

Теорема Бернуллі встановлює зв'язок теорії імовірностей з її практичними застосуваннями.

**Теорема 3.4** *Якщо у $n$ незалежних випробуваннях імовірність $p$ появи події $A$ однакова і подія $A$ з'явилася $m$ разів, то для будь-якого додатнього числа $a$ має місце рівність*

$$\lim P\left(\left|\frac{m}{n} - p\right| \geq a\right) = 0, \tag{3.39}$$

*тобто границя імовірності відхилення відносної частоти події $A$ від її імовірності на величину, що більше або дорівнює $a$, дорівнює нулеві.*

Згідно означенню границі рівність (3.39) означає, що $P\left(\left|\frac{m}{n} - p\right| \geq a\right) = \alpha$ — нескінченно мала величина. Та це означає, що подія

$$\left|\frac{m}{n} - p\right| \geq a \tag{3.40}$$

практично неможлива. Але тоді протилежна подія

$$\left|\frac{m}{n} - p\right| < a \tag{3.41}$$

практично достовірна для будь-якого додатнього числа $a$.

**Наслідок з теореми Бернуллі.** Рівність

$$\left|\frac{m}{n} - p\right| = 0 \tag{3.42}$$

може відрізнятись від практично достовірної події

$$\left|\frac{m}{n} - p\right| < a \tag{3.43}$$

$a > 0$ на нескінченно малу величину. Це означає, що $\frac{m}{n} \to p$, тобто відносна частота (частость) $W(A) = \frac{m}{n}$ події $A$ відрізняється від імовірності $p$ події $A$ на нескінченно малу величину, яку практично можна не враховувати.

**Зауваження.** Формулу (3.39) можна записати з використанням інтегральної функції Лапласа $\Phi(x)$ у вигляді

$$\lim P\left(\left|\frac{m}{n} - p\right| \leq \varepsilon\right) = 2\Phi\left(\varepsilon\sqrt{\frac{n}{pq}}\right). \tag{3.44}$$

Звідси одержуємо важливу формулу

$$P\Big(\Big|\frac{m}{n}-p\Big|\le\varepsilon\Big)\approx 2\Phi\Big(\varepsilon\sqrt{\frac{n}{pq}}\Big). \tag{3.45}$$

яка дозволяє розв'язувати багато задач.

**Приклад 3.9** Імовірність появи події в кожному із $n = 625$ незалежних випробувань дорівнює $p = 0.8$. Знайти імовірність того, що частость появи події відхиляється від імовірності за абсолютною величиною не більше ніж на 0.04.

**Розв'язок задачі.** За умовою прикладу $n = 625$, $p = 0.8$, $q = 1 - 0.8 = 0.2$, $\varepsilon = 0.004$. Треба знайти $P\Big(\Big|\frac{m}{625}-0.8\Big|\Big)\le 0.04$.

За формулою 3.45 маємо

$$P\Big(\Big|\frac{m}{625}-0.8\Big|\le 0.04\Big)\approx 2\Phi\Big(0.04\sqrt{\frac{625}{0.8\cdot 0.2}}\Big)=2\Phi(2.5). \tag{3.46}$$

Оскільки $\Phi(2.5) = 0.4938$

$$P\Big(\Big|\frac{m}{625}-0.8\Big|\le 0.04\Big)\approx 0.9876. \tag{3.47}$$

Таким чином, шукана імовірність наближено дорівнює 0.9876.

**Приклад 3.10** Імовірність появи події в кожному із незалежних випробувань дорівнює 0.5. Знайти число випробувань $n$, при якому з імовірністю 0.7698 можна чекати, що частость появи події відхиляється від її імовірності за абсолютною величиною не більше ніж на 0.02.

**Розв'язок задачі.** За умовою задачі $p = 0.5$, $q = 0.5$, $\epsilon = 0.02$.

$$P\Big(\Big|\frac{m}{n}-0.5\Big|\le 0.02\Big)=0.7698. \tag{3.48}$$

Застосуємо формулу 3.45. Тоді згідно умови одержимо

$$\begin{aligned}&2\Phi\Big(0.02\sqrt{\frac{n}{0.5\cdot 0.5}}\Big)=0.7698;\\ &\Phi(0.04\sqrt{n})=0.3849;\quad \Phi(1.2)=0.3849;\quad 0.04\sqrt{n}=1.2.\end{aligned} \tag{3.49}$$

Отже, шукана кількість випробувань $n = 900$.

**Приклад 3.11** Відділ технічного контролю перевіряє стандартність 900 виробів. Імовірність того, що виріб стандартний, дорівнює 0.9. Знайти з імовірністю 0.9544 межі інтервалу, що містить число $m$ стандартних виробів серед перевірених.

**Розв'язок задачі.** За умовою задачі $n = 900$, $p = 0.9$, $q = 0.1$.

$$\begin{aligned} &2\Phi\Big(\varepsilon\sqrt{\frac{900}{0.9\cdot 0.1}}\Big) = 0.9544; \quad \Phi(100\varepsilon) = 0.4772. \\ &\Phi(2) = 0.4772; \quad 100\varepsilon = 2; \quad \varepsilon = 0.02. \end{aligned} \tag{3.50}$$

Отже, з імовірністю 0.9544 відхилення частоти кількості стандартних виробів від імовірності 0.9 задовольняє нерівність

$$\Big|\frac{m}{900} - 0.9\Big| \le 0.02; \quad 0.88 \le \frac{m}{900} \le 0.92. \tag{3.51}$$

З останніх співвідношень випливає, що шукане число $m$ стандартних виробів серед 900 перевірених з імовірністю 0.9544 належить інтервалу $792 \le m \le 828$.

## 3.5. Проста течія подій

**Означення 3.4** Течією подій називають послідовність таких подій, які з'являються у випадкові моменти часу [1-5]. Наприклад, заява до диспетчерського пункту з викликом таксі.

**Означення 3.5** Течія подій називається пуассонівською, якщо вона:

1. **Стаціонарна,** тобто залежить від кількості $k$ появ події та часу $t$ і не залежить від моменту свого початку.
2. **Має властивість відсутності післядії,** тобто імовірність появи події не залежить від появи або не появи події раніше та не впливає на найближче майбутнє.
3. **Ординарна,** тобто імовірність появи більше однієї події в малий проміжок часу є величина нескінченно мала у порівнянні з імовірністю появи події один раз у цей проміжок часу.

**Означення 3.6** Середнє число $\lambda$ появ події $A$ в одиницю часу називають інтенсивністю течії.

**Теорема 3.5** *Якщо течія подій пуассонівська, то імовірність появи події $A$ $k$ разів за час $t$ можна знайти за формулою*

$$P_t(k) = \frac{(\lambda t)^k}{k!} e^{-\lambda t}, \tag{3.52}$$

*де $\lambda$ — інтенсивність течії.*

**Зауваження №1.** Формулу (3.52) іноді звуть математичною моделлю простої течії подій.

**Приклад 3.12** Середня кількість замовлень, що поступають до комбінату побутового обслуговування кожну годину, дорівнює 3. Знайти імовірність того, що за дві години поступлять

1. 5 замовлень;
2. менше 5 замовлень;
3. не менше 5 замовлень.

**Розв'язок задачі.** Маємо просту течію подій з інтенсивністю 3. За формулою (3.52) одержуємо

1. $P_2(5) = \dfrac{(3 \cdot 2)^5}{5!} e^{-3 \cdot 2}$;
2. $P_2(k < 5) = P(0) + P(1) + P(2) + P(3) + P(4) = 115 \cdot e^{-6}$;
3. $P_2(k \geq 5) = 1 - P_2(k < 5) = 1 - 115 \cdot e^{-6}$.

**Зауваження №2.** Прикладами простої течії подій можуть бути: поява викликів на АТС, на пункти швидкої медичної допомоги, прибуття літаків до аеропорту або клієнтів у підприємство побутового обслуговування, серія відмов елементів або блоків приладів та таке інше.

## 3.6. Питання для самоперевірки

- Яка послідовність випробувань утворює схему Бернуллі?
- Яку формулу звуть формулою Бернуллі і що вона дозволяє обчислювати?
- За якими формулами знаходять імовірність появи події A менше т або не менше за m разів у п випробуваннях схеми Бернуллі?
- За якою формулою знаходять імовірність появи події A хоча б один раз у п випробуваннях?
- Як можна знайти найбільш імовірне значення числа появ події A у схемі Бернуллі?
- Як можна знайти кількість випробувань у схемі Бернуллі, яка дозволяє з імовірністю P стверджувати, що подія A з'явиться хоча б один раз?

- У яких випадках доцільно використовувати граничні теореми у схемі Бернуллі?
- Коли доцільно застосовувати формули Пуассона, локальну або інтегральну формули Муавра
- Як визначаються і які мають властивості локальна та інтегральна функції Лапласа?
- Як знаходять Pn(m) у випадку послідовності випробувань із різними імовірностями?
- Як формулюється теорема Бернуллі і який вона має наслідок?
- Який існує зв'язок між твердженням теореми Бернуллі та інтегральною функцією Лапласа? Які задачі дозволяє розв'язувати цей зв'язок?
- За якою формулою знаходять імовірність появи у випадку простої течії?

# 4. Випадкові величини [1-5]

## 4.1. Види випадкових величин та способи їх задання

При дослідженні багатьох проблем виникають такі випадкові події, наслідком яких є поява деякого числа, заздалегідь невідомого. Тому такі числові значення — випадкові.

Прикладом такої події є: кількість очок, що випадає при киданні грального кубика; кількість студентів, які прийдуть на лекцію; кількість цукрового буряка, який чекають одержати з одного гектара.

Випадковою величиною називають таку величину, яка в наслідок випробування може прийняти лише одне числове значення, заздалегідь невідоме і обумовлене випадковими причинами.

Випадкові величини доцільно позначати великими літерами $X, Y, Z$, а їх можливі значення – відповідними малими літерами з індексами. Наприклад,

$$X : x_1, x_2, ..., x_n; \quad Z : z_1, z_2, ..., z_m.$$

Випадкові величини бувають дискретними та неперервними.

**Означення 4.1** Дискретною випадковою величиною (ДВВ) називають таку величину, яка може приймати відокремлені ізольовані одне від одного числові значення (їх можна пронумерувати) з відповідними імовірностями.

**Приклад 4.1** Кількість влучень у мішень при трьох пострілах буде $X$ : $0, 1, 2, 3$. Отже, $X$ може приймати чотири ізольовані числові значення з

різними імовірностями. Тому $X$ — дискретна випадкова величина.

Кількість викликів таксі $Y$ на диспетчерському пункті також буде дискретною випадковою величиною, але при $t \to \infty$ значення $Y$ також зростають, тобто їх кількість прямує до нескінченності $Y : 0, 1, 2, ..., n, ....$

**Означення 4.2** Неперервною випадковою величиною (НВВ) називають величину, яка може приймати будь-яке числове значення з деякого скінченого або нескінченого інтервалу $(a, b)$. Кількість можливих значень такої величини є нескінчена.

**Приклад 4.2** Величина похибки, яка може бути при вимірюванні відстані; час безвідмовної роботи приладу; зріст людини; розміри деталі, яку виготовляє станок-автомат.

**Приклад 4.3** Розглянемо випадкові величини: кількість очок, $X$ та $Y$, що можуть з'явитись при киданні правильного грального кубика та неправильного грального кубика. їх можливі значення

$$X : 1, 2, 3, 4, 5, 6; \quad Y : 1, 2, 3, 4, 5, 6$$

однакові.

Імовірність появи будь-якого значення $x_k$ дорівнює $\frac{1}{6}$, однакова для усіх можливих значень $X$, а імовірності появи можливих значень $Y$ будуть різними. Отже, випадкові величини $X$ та $Y$ не рівні тому, що при $x_k = y_k$ маємо $P(x_k) \neq P(y_k)$, $k = 1, 2, 3, 4, 5, 6$.

Таким чином, для повної характеристики випадкової величини треба вказати не тільки усі її можливі значення, але й закон, за яким знаходять імовірності кожного значення

$$p_k = P(X = x_k) = f(x_k) \quad \text{або} \quad P(X) = f(X). \tag{4.1}$$

**Означення 4.3** Законом розподілу випадкової величини називають таке співвідношення, яке встановлює зв'язок між можливими значеннями випадкової величини і відповідними їм імовірностями.

У випадку дискретної випадкової величини $X$ функціональну залежність можна задавати таблично, аналітично або графічно.

У випадку неперервної випадкової величини для її повної характеристики вводять інтегральну та диференціальну функції розподілу.

**Означення 4.4** Інтегральною функцією розподілу називають імовірність того, що випадкова величина $X$ прийме значення, менше $x$.

Функцію розподілу позначають $F(x)$. Таким чином,

$$F(x) = P(X < x). \tag{4.2}$$

Якщо НВВ $X$ може приймати будь-яке значення з $(a, b)$, то

$$P(a < X < b) = F(b) - F(a), \tag{4.3}$$

тобто імовірність прийняття величиною $X$ значень з $(a, b)$ дорівнює приросту функції розподілу.

Формулу (4.3) часто називають основною формулою теорії імовірностей.

**Зауваження.** Неперервна випадкова величина $X$, що приймає значення у проміжку $(a, b)$, має незлічену кількість можливих значень, тому набуття $X$ певних значень $X = a$ або $X = b$ буде майже неможливою подією. Це означає, що $P(X = a)$ та $P(X = b)$ будуть нескінченно малими величинами, які у практичних розрахунках можна не враховувати. Тому мають місце рівності

$$P(a < X < b) = P(a \le X < b) = P(a < X \le b) = P(a \le X \le b) \tag{4.4}$$

Означення інтегральної функції розподілу та властивості імовірності $P$ дозволяють одержати такі властивості функції розподілу:

- $0 \le F(x) \le 1$;
- $F(x)$ — зростаюча функція, тобто $F(x_2) > F(x_1)$, якщо $x_2 > x_1$;
- $F(x) = 0$ при $x \le a$ $F(x) = 1$ при $x \ge b$.

Графік функції розподілу $F(x)$ може мати вигляд, зображений, наприклад, на Рис.4.1.

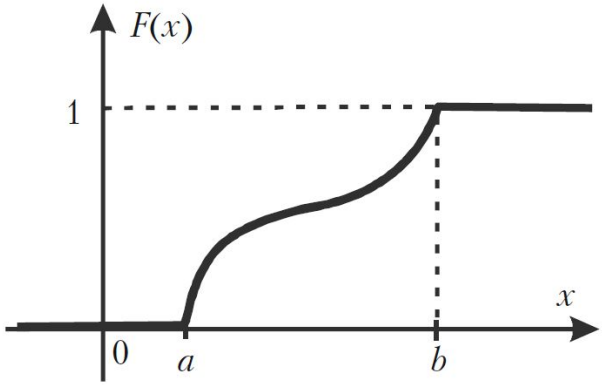


Рис. 4.1 —

**Означення 4.5** Диференціальною функцією розподілу або щільністю імовірностей неперервної випадкової величини називають похідну першого порядку від її інтегральної функції розподілу і позначають

$$f(x) = F'(x) \tag{4.5}$$

Назва "щільність імовірностей" випливає з рівності

$$f(x) = \lim_{\Delta x=0} \frac{P(X < x + \Delta x) - P(X < x)}{\Delta x}. \tag{4.6}$$

Із формули (4.5) випливає, що функція розподілу $F(x)$ є первісною для диференціальної функції розподілу $f(x)$.

**Теорема 4.1** *Імовірність того, що неперервна випадкова величина $X$ прийме значення з інтервалу $(a, b)$, можна знайти за формулою*

$$P(a < X < b) = \int_a^b f(x)dx. \tag{4.7}$$

**Доведення.** Інтегральна функція розподілу $F(x)$ — первісна для $f(x)$, тому згідно з формулою Ньютона - Лейбніца маємо

$$\int_a^b f(x)dx = F(b) - F(a). \tag{4.8}$$

Праві частини рівностей (4.3) та (4.8) рівні, тому і ліві їх частини рівні, тобто має місце рівність (4.7), яку й треба було довести.

**Наслідок.** Якщо диференціальна функція розподілу (щільність імовірності) $f(x)$ відома, то інтегральну функцію розподілу $F(x)$ можна знайти за формулою

$$F(x) = \int_{-\infty}^x f(x)dx. \tag{4.9}$$

Диференціальна функція розподілу НВВ $X(a, b)$ має такі властивості:

1. $f(x) \geq 0$ тому, що вона є похідною зростаючої функції $F(x)$;
2. $f(x) = 0$ при $x < a$ та $x > b$ тому, що є похідною $F(x) = const$;
3. $\int_{-\infty}^{+\infty} f(x)dx = 1$ тому, що подія $\{-\infty < X < +\infty\}$ — достовірна.

Графік щільності імовірності $f(x)$ називають кривою розподілу. Він може мати вигляд, зображений, наприклад, на Рис.4.2.

**Приклад 4.4** Випадкова величина має щільність імовірностей

$$f(x) = \frac{a}{1 + x^2} \tag{4.10}$$

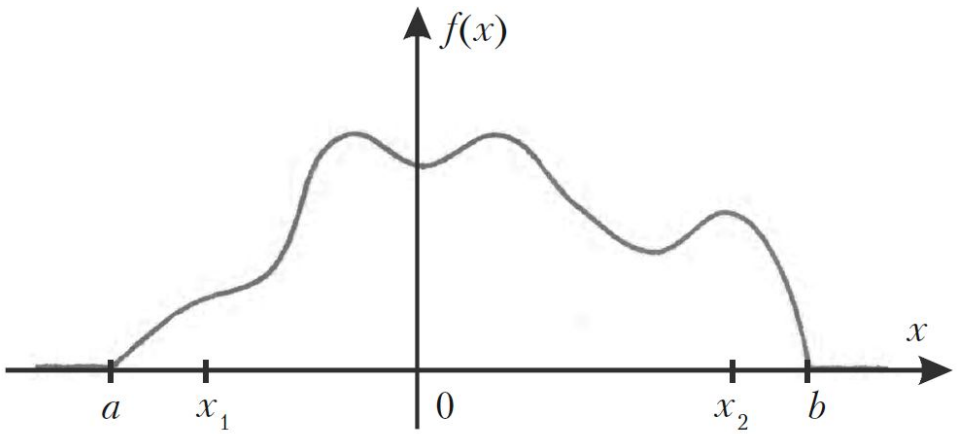


Рис. 4.2 —

Визначити параметр $a$ та функцію розподілу.

**Розв'язок задачі.** Параметр $a$ знайдено використовуючи властивість 3 диференціальної функції розподілу

$$1 = \int_{-\infty}^{+\infty} f(x)dx = \int_{-\infty}^{+\infty} \frac{a}{1+x^2}dx = a \cdot \pi. \tag{4.11}$$

Із (4.11) одержуємо

$$a = \frac{1}{\pi} \tag{4.12}$$

Функцію розподілу знайдемо за формулою (4.9)

$$F(x) = \frac{1}{\pi} \int_{-\infty}^{x} \frac{1}{1+x^2}dx = \frac{1}{\pi} \operatorname{arctg}(x) + \frac{1}{2}. \tag{4.13}$$

**Приклад 4.5** Випадкова величина $X$ задана функцією розподілу

$$F(x) = x^2 - 4x + 4. \tag{4.14}$$

Визначити область значень випадкової величини $X$ та імовірність того, що $X \geq 2.3$.

**Розв'язок задачі.** Згідно властивостям функції розподілу маємо

$$0 \leq F(x) \leq 1 \tag{4.15}$$

тому повинні виконуватись умови

$$0 \leq x^2 - 4x + 4 \leq 1 \tag{4.16}$$

Якщо область значень випадкової величини $[a, b]$, то $F(a) = 0$ та $F(b) = 1$.

Підставимо в (4.14) замість $x$ $a$ та $b$, тоді одержимо

$$\begin{aligned} &a^2 - 4a + 4 = 0, \quad (a-2)^2 = 0; \quad a = 2; \\ &b^2 - 4b + 4 = 1, \quad b^2 - 4b + 3 = 0; \quad b_1 = 3, b_2 = 1. \end{aligned} \tag{4.17}$$

Але в проміжку $[a, b]$, $b > a$, тому $b = 3$. Отже, областю значень НВВ $X$ буде $[2, 3]$.

Тепер знайдемо імовірність $P(X \geq 2.3)$. Подія $X < 2.3$ буде протилежною, тому

$$P(X \geq 2.3) = 1 - P(X < 2.3) = 1 - F(2.3). \tag{4.18}$$

З рівності (4.14) одержуємо

$$F(2.3) = 2.3^2 - 4 \cdot 2.3 + 4 = 9.29 - 9.2 = 0.09. \tag{4.19}$$

Тепер за формулою (4.19) знаходимо

$$P(X \geq 2.3) = 1 - 0.09 = 0.91. \tag{4.20}$$

## 4.2. Закони розподілу та числові характеристики дискретних випадкових величин

### Способи задання та закони розподілу

Нехай випадкова дискретна величина $X$ приймає значення $x_1, x_2, ..., x_n$ з відповідними імовірностями $p_1, p_2, ..., p_n$. Задати закон розподілу такої випадкової величини — це задати рівність $p_k = P(X = x_k)$, яку можна розглядати як функцію. Тому закон розподілу $X$ можна задати аналітично, таблично, графічно. Функція розподілу для дискретної випадкової величини має вигляд

$$F(x) = P(X < x) = \sum_{x_i < x} p(x_i). \tag{4.21}$$

**Табличний спосіб** задання ДВВ, який називають рядом розподілу і зображують у вигляді

| $X$ | 100 | 50 | 10 | 1 | 0 |
|---|---|---|---|---|---|
| $P(X)$ | 0.001 | 0.002 | 0.008 | 0.019 | 0.97 |

У першому рядку записані усі можливі значення $X$, а у другому рядку —

відповідні імовірності, які мають властивість

$$\sum_{k=1}^{n} p_k = 1. \tag{4.22}$$

**Зауваження**. Якщо випадкова дискретна величина може приймати

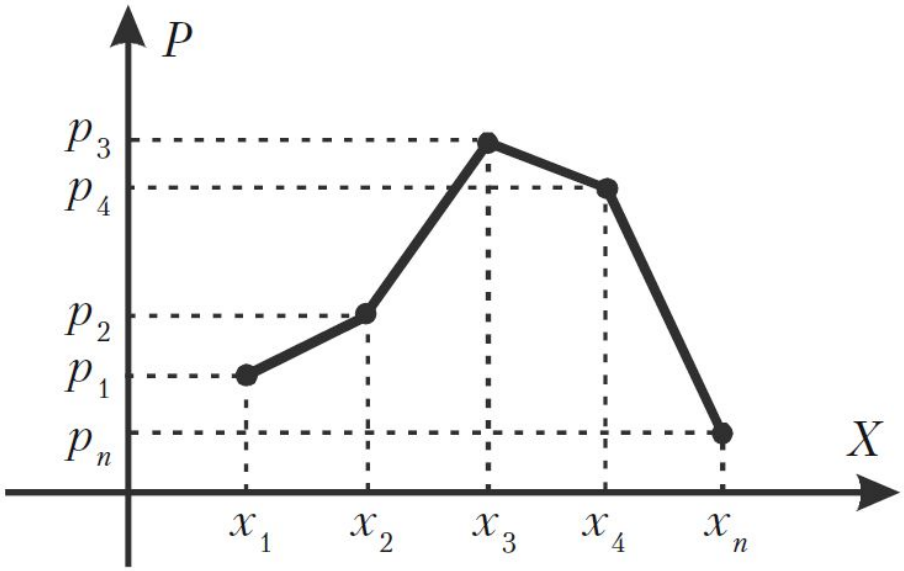


Рис. 4.3 — Графічний спосіб задання дискретних випадкових величин

нескінчену кількість значень, то її ряд розподілу (таблиця) буде мати нескінчену кількість елементів у кожному рядку, причому ряд повинен бути збіжним, а його сума повинна дорівнювати одиниці.

**Графічний спосіб**. Візьмемо прямокутну систему координат. На осі абсцис будемо відкладати можливі значення ДВВ, а на осі ординат — відповідні значення імовірності. Одержимо точки з координатами $(x_1, p_1), (x_2, p_2), \cdots, (x_n, p_n)$.

Поєднавши ці точки прямими, одержимо графік (див. Рис.4.3) у вигляді многокутника розподілу випадкової дискретної величини. Значення ДВВ, імовірність якого найбільша, називають модою. На Рис.4.3 мода — $x_3$.

**Аналітичний спосіб** задання випадкової дискретної величини базується на заданні певної функції, за якою можна знайти імовірність $p$ відповідного значення $x_k$, тобто $p_k = f(x_k)$, $k = 1, 2, \cdots, n$.

Вкажемо деякі найважливіші закони розподілу ДВВ та задачі, в яких вони зустрічаються.

- **Біноміальний закон розподілу**. Цей закон має вигляд

$$P(X = m) = C_n^m p^m (1-p)^{n-m}, \quad m = 0, 1, 2, \cdots, n \tag{4.23}$$

і використовується у схемі Бернуллі, тобто у випадку $n$ незалежних

повторних випробувань, в кожному з яких деяка подія з'являється з імовірністю $p$.

- **Закон розподілу Пуассона**. ДВВ $X$ приймає злічену множину значень $(m = 0, 1, 2, \cdots, n)$ з імовірностями

$$P(X = m) = \frac{a^m}{m!} e^{-a}, \quad a > 0. \tag{4.24}$$

Цей розподіл використовують у задачах статистичного контролю якості, в теорії надійності, теорії масового обслуговування, для обчислення: кількості вимог на виплату страхових сум за рік, кількості дефектів однакових виробів.

Якщо у схемі незалежних повторних випробувань $n$ досить велике, а $p$ або $1-p$ прямує до нуля, то біноміальний розподіл апроксимує розподіл Пуассона, параметр якого $a = np$, причому при $p \leq 0.1$ або $p \geq 0.9$ ця апроксимація дає добрі результати незалежно від величини $n$.

**Зауваження**. Якщо у формулі Пуассона покласти $a = \lambda t$, де $\lambda$ — інтенсивність течії випадкових подій в одиницю часу, то формула прийме вигляд

$$P(X = m) = \frac{(\lambda t)^m}{m!} e^{-\lambda t} \quad (m = 0, 1, 2, \cdots). \tag{4.25}$$

- **Геометричний розподіл**. Цей розподіл має вигляд

$$P(X = m) = pq^{m-1}, \tag{4.26}$$

де $p = P(A)$ — імовірність появи події $A$ в кожному випробуванні, $q = 1 - p$, $X$ — кількість випробувань до появи події $A$ в серії незалежних повторних випробувань.

Ряд імовірностей цього розподілу буде нескінченно спадною геометричною прогресією із знаменником $q$, сума якої дорівнює одиниці.

Геометричний розподіл застосовують у різноманітних задачах статистичного контролю якості виробів, в теорії надійності та у страхових розрахунках.

- **Гіпергеометричний розподіл**. Цей розподіл має вигляд

$$P(X = m) = \frac{C_k^m \cdot C_{N-k}^{n-m}}{C_N^n}, \quad m = 0, 1, 2, \cdots, n, \quad k \geq n. \tag{4.27}$$

Він вказує імовірність появи $m$ елементів з певною властивістю серед $n$ елементів, взятих із сукупності $N$ елементів, яка містить $k$ елементів саме такої властивості. Цей розподіл використовують у багатьох задачах статистичного контролю якості.

**Зауваження**. Якщо об'єм вибірки $n$ малий у порівнянні з об'ємом $N$ сукупності, тобто $\frac{n}{N} \leq 0.1$, $\frac{n}{k} \leq 0.1$, то імовірності у гіпергеометричному розподілі будуть близькими до відповідних імовірностей біноміального розподілу з $p = \frac{k}{N}$.

У статистиці це означає, що розрахунки імовірностей для без повторної вибірки будуть мало відрізнятись від розрахунків імовірностей для повторної вибірки.

- **Поліноміальний розподіл**. Цей розподіл має вигляд

$$P_s = (X_1 = m_1; X_2 = m_2; \cdots; X_s = m_s;) = \frac{n_1!}{m_1! \cdot m_2! \cdots m_s!} \cdot p_1^{m_1} \cdot p_2^{m_2} \cdots p_s^{m_s}. \tag{4.28}$$

Він застосовується тоді, коли внаслідок кожного із здійснених повторних незалежних випробувань може з'явитися $s$ різних подій $A_i$ з імовірністю $p_i$, причому $\sum_{i=1}^{s} p_i = 1$.

## 4.3. Числові характеристики дискретних випадквих величин

Закони розподілу ДВВ повністю характеризують випадкові величини і дозволяють розв'язувати усі пов'язані з ними задачі. Але в практичній діяльності не завжди вдається одержати закон розподілу, або закон надто складний для практичних розрахунків. Тому з'явилася потреба характеризувати ДВВ за допомогою числових характеристик, які достатньо характеризують особливості випадкових величин.

Найчастіше використовують три числових характеристики: математичне сподівання, дисперсію та середнє квадратичне відхилення від математичного сподівання.

Ознайомимось із цими числовими характеристиками та їх властивостями.

### Математичне сподівання та його основні властивості

**Означення 4.6** Математичним сподіванням дискретної випадкової величини $X$ називають число, яке дорівнює сумі добутків усіх можливих значень $X$ на відповідні їм імовірності.

Математичне сподівання ДВВ $X$ позначають $M(X)$ або $m_X$, тобто

$$M(X) = \sum_{k=1}^{n} x_k \cdot p_k, \tag{4.29}$$

де $x_k$ і $p_k$ – $k$-те можливе значення дискретної випадкової величини і відповідна ймовірність, $k = 1, ..., n$.

Якщо $X$ приймає нескінчену кількість значень, то

$$M(X) = \sum_{k=1}^{\infty} x_k \cdot p_k. \tag{4.30}$$

Математичне сподівання ДВВ $X$ характеризує середнє значення випадкової величини $X$ із врахуванням імовірностей його можливих значень. У практичній діяльності під математичним сподіванням розуміють центр розподілу випадкової величини.

## Основні властивості математичного сподівання

1. Математичне сподівання постійної величини дорівнює самій постійній

$$M(C) = C. \tag{4.31}$$

2. Постійний множник можна виносити за знак математичного сподівання

$$M(CX) = C \cdot M(X). \tag{4.32}$$

Властивості 1 і 2 випливають безпосередньо з означення.

3. Математичне сподівання добутку декількох взаємно незалежних дискретних випадкових величин дорівнює добутку їх математичних сподівань, тобто

$$M(X_1 \cdot X_2 \cdot ... \cdot X_n) = M(X_1) \cdot M(X_1) \cdot ... \cdot M(X_n) \tag{4.33}$$

**Доведення.** Якщо дві величини $X$ та $Y$ розподілені за законами

| $X$ | $x_1$ | $x_2$ | $Y$ | $y_1$ | $y_2$ |
|---|---|---|---|---|---|
| $P$ | $p_1$ | $p_2$ | $G$ | $g_1$ | $g_2$ |

(для спрощення викладок взято лише по 2 можливих значення), тоді закон розподілу добутку $X \cdot Y$ буде

| $X \cdot Y$ | $x_1y_1$ | $x_2y_1$ | $x_1y_2$ | $x_2y_2$ |
|---|---|---|---|---|
| $P \cdot G$ | $p_1g_1$ | $p_2g_1$ | $p_1g_2$ | $p_2g_2$ |

За формулою (4.30) одержимо математичне сподівання

$$M(X \cdot Y) = y_1g_1(x_1p_1 + x_2p_2) + y_2g_2(x_1p_1 + x_2p_2) = \\ (x_1p_1 + x_2p_2) \cdot (y_1g_1 + y_2g_2) = M(X) \cdot M(Y). \tag{4.34}$$

У випадку трьох випадкових величин маємо

$$M(X \cdot Y \cdot Z) = M((X \cdot Y) \cdot Z) = \\ M(X \cdot Y) \cdot M(Z) = M(X) \cdot M(Y) \cdot M(Z). \tag{4.35}$$

Методом математичної індукції тепер не важко завершити доведення.

4. Математичне сподівання суми випадкових величин дорівнює сумі їх математичних сподівань, тобто

$$M(X_1 + X_2 + ... + X_n) = M(X_1) + M(X_2) + ... + M(X_n). \tag{4.36}$$

**Приклад 4.6** Незалежні випадкові величини $X$ та $Y$ розподілені так

| $X$ | 5 | 2 | 4 | $Y$ | 8 | 10 |
|---|---|---|---|---|---|---|
| $P$ | 0.6 | 0.1 | 0.3 | $P$ | 0.8 | 0.2 |

Знайти математичне сподівання випадкової величини $X \cdot Y$.

**Розв'язок задачі.** Спочатку знайдемо математичні сподівання кожної Спочатку знайдемо математичні сподівання кожної з цих величин. За формулою (1) маємо.

$$M(X) = 5 \cdot 0.6 + 2 \cdot 0.1 + 4 \cdot 0.3 = 4.4, \quad M(Y) = 8 \cdot 0.8 + 10 \cdot 0.2 = 8.4. \tag{4.37}$$

Випадкові величини $X$ і $Y$ незалежні, тому згідно властивості 3 математичного сподівання одержимо

$$M(X \cdot Y) = M(X) \cdot M(Y) = 4.4 \cdot 8.4 = 36.96. \tag{4.38}$$

**Приклад 4.7** Знайти математичне сподівання суми числа очок, які можуть з'явитися при киданні двох гральних кубиків.

**Розв'язок задачі.** Позначимо кількість очок, які можуть з'явитись на першому кубику $X$, а на другому – $Y$. Можливі значення цих величин

$1, 2, 3, 4, 5, 6$ однакові, імовірність кожного з цих значень дорівнює $\frac{1}{6}$. Тому

$$M(X) = M(Y) = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = \frac{7}{2}. \tag{4.39}$$

Згідно властивості 4 математичного сподівання, одержимо

$$M(X + Y) = M(X) + M(Y) = \frac{7}{2} + \frac{7}{2} = 7. \tag{4.40}$$

Отже, математичне сподівання суми числа очок, що можуть з'явитись при киданні двох гральних кубиків, дорівнює 7.

## Дисперсія та її властивості

Математичне сподівання характеризує центр розподілу дискретної випадкової величини. Але цієї характеристики недостатньо, бо можливе значне відхилення можливих значень від центру розподілу. Для характеристики розсіювання можливих значень X відносно центру розподілу введемо нову числову характеристику.

**Означення 4.7** Дисперсією дискретної випадкової величини $X$ називають число, яке дорівнює математичному сподіванню квадрата відхилення ДВВ $X$ від її математичного сподівання.

Дисперсію величини $X$ позначають $D(X)$ або $D_X$. Це означення математично виглядає так

$$D(X) = M\Big((X - M(X))^2\Big). \tag{4.41}$$

## Основні властивості дисперсії $D(X)$

1. Дисперсія будьякої ДВВ X невід'ємна $D(X) \geq 0$. Дійсно, $(X - M(X))^2$ невід'ємна, тому згідно означення математичного сподівання та властивостей імовірностей $p_k$, $k = 1, ..., n$, $D(X)$ також невід'ємна.

2. Дисперсія постійної величини $C$ дорівнює нулеві $D(C) = 0$. Дійсно, якщо $X = C$, то $M(C) = C$, тому $C - M(C) = 0$.

3. Постійний множник $C$ можна виносити за знак дисперсії, при цьому постійний множник треба піднести у квадрат

$$D(CX) = C^2 D(X). \tag{4.42}$$

Дійсно,

$$CX - M(CX) = C(X - M(X)), \tag{4.43}$$

тому

$$(CX - M(CX))^2 = C^2(X - M(X))^2. \tag{4.44}$$

Постійний множник $C^2$ можна виносити за знак математичного сподівання, тому з формули (4.41) випливає потрібна рівність (4.42).

4. Дисперсія ДВВ $X$ дорівнює різниці між математичним сподіванням квадрата випадкової величини $X$ та квадрата її математичного сподівання

$$D(X) = M(X^2) - (M(X))^2. \tag{4.45}$$

Дійсно,

$$\begin{aligned} &D(X) = M((X - M(X))^2) = M(X^2 - 2XM(X) + M^2(X)) = \\ &M(X^2) - 2M^2(X) + M^2(X) = M(X^2) - M^2(X). \end{aligned} \tag{4.46}$$

5. Дисперсія алгебраїчної суми ДВВ $X$ та $Y$ дорівнює сумі їх дисперсій

$$D(X \pm Y) = D(X) + D(Y). \tag{4.47}$$

Дійсно згідно з формулою(4.45) для суми двох випадкових величин маємо

$$\begin{aligned} &D(X + Y) = M((X + Y)^2) - M^2(X + Y) = \\ &M(X^2 + 2XY + Y^2) - (M(X) + M(Y))^2 = \\ &M(X^2) + 2M(X)M(Y) + M(Y^2) - \\ &M^2(X) - 2M(X)M(Y) - M^2(Y) = \\ &(M(X^2) - M^2(X)) + (M(Y^2) - M^2(Y)) = D(X) + D(Y). \end{aligned} \tag{4.48}$$

У випадку різниці $X$ та $Y$ будемо мати

$$D(X - Y) = D(X) + (-1)^2 D(Y) = D(X) + D(Y). \tag{4.49}$$

**Приклад 4.8** Знайти дисперсію випадкової величини $X$, що задана законом

| $X$ | -5 | 0 | 4 | 5 |
|---|---|---|---|---|
| $P$ | $\frac{1}{8}$ | $\frac{1}{2}$ | $\frac{1}{4}$ | $\frac{1}{8}$ |

**Розв'язок задачі.** Будемо шукати $D(X)$ з використанням формули (4.45).

Математичним сподіванням $X$ згідно з формулою (4.30) буде

$$M(X) = -5 \cdot \frac{1}{8} + 0 \cdot \frac{1}{2} + 4 \cdot \frac{1}{4} + 5 \cdot \frac{1}{8} = 1, \quad M^2(X) = 1. \tag{4.50}$$

Відмітимо, що усі значення $X^2$ отримані шляхом піднесення до квадрату відповідних значень $X$. Елементи другого рядка — імовірності цих значень не змінюються.

За формулою (4.30) знаходимо

$$M(X^2) = 26 \cdot \frac{1}{8} + 0 \cdot \frac{1}{2} + 16 \cdot \frac{1}{4} + 25 \cdot \frac{1}{8} = \frac{82}{8}. \tag{4.51}$$

Згідно з формулою (4.45) тепер одержуємо

$$D(X) = \frac{82}{8} - 1 = \frac{74}{8} = 9.25. \tag{4.52}$$

## Середнє квадратичне відхилення дискретної випадкової величини

У більшості випадків випадкова величина X має розмірність, наприклад, метр, міліметр, грам, тому її дисперсія $D(X)$ буде вимірюватись у квадратних одиницях цієї розмірності.

У практичній діяльності доцільно знати величину розсіювання випадкової величини в розмірності цієї величини. Для цього використовують середнє квадратичне відхилення, яке дорівнює квадратному кореню з дисперсії і позначається

$$\sigma(X) = \sigma_X = \sqrt{D(X)}. \tag{4.53}$$

## Моменти розподілу

**Означення 4.8 Початковим моментом порядку $k$ випадкової величини** $X$ називають математичне сподівання величини $X^k$ і позначають

$$\nu_k = M(X^k), \quad k = 1, 2, ..., n. \tag{4.54}$$

**Означення 4.9 Центральним моментом порядку $k$ випадкової величини** $X$ називають математичне сподівання величини $(X - M(X))^k$ і позначають

$$\mu_k = M((X - M(X))^k), \quad k = 1, 2, ..., n. \tag{4.55}$$

Очевидно, що

$$v_1 = M(X), \quad v_2 = M(X^2),$$

тому

$$D(X) = \nu_2 - {\nu_1}^2, \quad \mu_1 = M(X_M(X)) = 0.$$
$$\mu_2 = M((X - M(X))^2) = D(X).$$

Початкові та центральні моменти порядку $k \geq 2$ дозволяють краще враховувати вплив на математичне сподівання (центр розподілу випадкової величини $X$) тих можливих значень $X$, які великі та мають малу імовірність.

**Приклад 4.9** Дискретна випадкова величина задана законом

| $X$ | 1 | 2 | 5 | 100 |
|---|---|---|---|---|
| $P$ | 0.6 | 0.2 | 0.19 | 0.01 |

Математичним сподіванням $X$ буде

$$M(X) = 1 \cdot 0.6 + 2 \cdot 0.2 + 5 \cdot 0.19 + 100 \cdot 0.01 = 2.95.$$

Законом розподілу $X^2$ буде

| $X$ | 1 | 4 | 25 | 1000 |
|---|---|---|---|---|
| $P$ | 0.6 | 0.2 | 0.19 | 0.01 |

Тому

$$M(X^2) = 1 \cdot 0.6 + 4 \cdot 0.2 + 25 \cdot 0.19 + 10000 \cdot 0.01 = 106.15.$$

Отже, $M(X^2)$ значно більше $M(X)$, а це означає, що роль значення $X = 100$ суттєво зросла.

## 4.4. Числові характеристики законів розподілу неперервних випадкових величин

У випадку неперервних випадкових величин (НВВ) математичне сподівання, дисперсія та середнє квадратичне відхилення мають такий же смисл та властивості, як і для дискретних випадкових величин, але обчислюють їх за іншими формулами. Нехай можливі значення неперервної випадкової величини $X$ заповнюють відрізок $[a, b]$. Поділимо $[a, b]$ на $n$ частин довжиною

$$\Delta x = \frac{b - a}{n},$$

У кожній частині візьмемо точку $\xi$, $k = 1, 2, .., n$. Тоді щільність імовірності $f(x)$ в точці $\xi_k$ буде $f(\xi_k)$ — імовірність того, що $X$ прийме значення Одержимо розподіл НВВ $X$ вигляду

| $X$ | $\xi_1$ | $\xi_2$ | ... | $\xi_n$ |
|---|---|---|---|---|
| $P$ | $f(\xi_1)\Delta x$ | $f(\xi_2)\Delta x$ | ... | $f(\xi_n)\Delta x$ |

Сума

$$\sum_{k=1}^{n} \xi_k f(\xi_k)\Delta x$$

характеризує математичне сподівання $X$ тим точніше, чим менше буде $\Delta x$. Ця сума буде дорівнювати математичному сподіванню $M(X)$ неперервної величини $X$, якщо перейти до границі при $\Delta x \to 0$. Згідно з означенням визначеного інтеграла маємо

$$M(X) = \lim_{\Delta x \to 0} \sum_{k=1}^{n} \xi_k f(\xi_k)\Delta x = \int_a^b x f(x)dx. \tag{4.56}$$

Формула (4.56) є доведенням наступної теоремі.

**Теорема 4.2** *Коли неперервна випадкова величина приймає значення на відрізку* $[a, b]$ *та має щільність імовірності* $f(x)$, *то її математичне сподівання знаходиться за формулою*

$$M(X) = \int_a^b x f(x)dx. \tag{4.57}$$

**Теорема 4.3** *Якщо* $f(x)$ *є щільність імовірності* $X$, *неперервна випадкова величина* $Y$ *є функцією випадкової величини* $X$, *тобто* $Y = \varphi(X)$, *тоді математичне сподівання* $Y$ *знаходиться за формулою*

$$M(\varphi(X)) = \int_{-\infty}^{+\infty} \varphi(x) f(x)dx. \tag{4.58}$$

Якщо можливі значення $X$ належать відрізку $[a, b]$, то центр розподілу $M(X)$ величини $X$ знаходиться на цьому проміжку тому, що із нерівностей

$$\int_a^b a f(x)dx < \int_a^b x f(x)dx < \int_a^b b f(x)dx$$

та умови нормування $\int\limits_a^b f(x)dx = 1$ випливають співвідношення

$$a < M(X) = \int\limits_a^b xf(x)dx < b.$$

Якщо щільність імовірності $f(x)$ — парна функція, тобто $f(x) = f(-x)$, то центр розподілу $X$ співпадає з початком $M(X)$. Якщо графік функції $f(x)$ симетричний відносно прямої $x = a$, то $M(X) = a$.

Як і у випадку дискретних випадкових величин, дисперсію неперервних випадкових величин $X$ визначають так

$$D(X) = M((X - M(X))^2) = \int\limits_{-\infty}^{+\infty} (x - M(X))^2 f(x)dx, \tag{4.59}$$

а обчислюють за формулою

$$D(X) == \int\limits_{-\infty}^{+\infty} x^2 f(x)dx - M^2(X). \tag{4.60}$$

Якщо можливі значення $X$ належать лише скінченому проміжку $(a, b)$, то рівності (4.59) та (4.60) приймають вигляд

$$D(X) = M((X - M(X))^2) = \int\limits_a^b (x - M(X))^2 f(x)dx, \tag{4.61}$$

$$D(X) = \int\limits_a^b x^2 f(x)dx - M^2(X). \tag{4.62}$$

Середнє квадратичне відхилення неперервної випадкової величини визначають та обчислюють так

$$\sigma(X) = \sqrt{D(X)}. \tag{4.63}$$

**Приклад 4.10** Знайти числові характеристики випадкової величини $X$, яка задана функцією розподілу

$$F(x) = \begin{cases} 0 \text{ при } x \leq 0, \\ \dfrac{x^2}{25} \text{ при } 0 < x < 5, \\ 1 \text{ при } x \geq 5. \end{cases}$$

**Розв'язок задачі.** Спочатку знайдемо диференціальну функцію розподілу, тобто щільність імовірності $f(x) = F'(x)$

$$f(x) = \begin{cases} \dfrac{2x}{25} \text{ при } 0 < x < 5, \\ 0 \text{ при } x \notin [0, 5]. \end{cases}$$

Тепер за формулою (4.57) знайдемо математичне сподівання

$$M(X) = \int_0^5 x \frac{2x}{25} dx = \frac{2}{25} \frac{x^3}{3} \Big|_0^5 = \frac{10}{3}.$$

Дисперсію знайдемо за формулою (4.61)

$$D(X) = \int_0^5 x^2 \frac{2x}{25} dx - \left(\frac{10}{3}\right)^2 = \frac{2}{25} \frac{x^4}{4} \Big|_0^5 - \frac{100}{9} = \frac{25}{18}.$$

Середнє квадратичне відхилення одержимо за формулою (4.62)

$$\sigma(X) = \sqrt{\frac{25}{18}} = \frac{5\sqrt{2}}{6} = 1.17.$$

## Закони розподілу НВВ та їх числові характеристики

Основні закони розподілу неперервних випадкових величин розділяють за виглядом їх диференціальних функцій розподілу (щільності імовірностей) $f(x)$.

Найчастіше використовують наступні закони розподілу.

**Означення 4.10 Рівномірний розподіл.** Величина $X$ розподілена рівномірно на проміжку $(a, b)$, якщо усі її можливі значення належать цьому

проміжку і щільність її імовірностей на цьому проміжку постійна, тобто

$$f(x) = \begin{cases} C = \dfrac{1}{b-a} \text{ при } x \in (a,b), \\ 0 \text{ при } x \notin (a,b]. \end{cases}$$

Величина постійно $C = \dfrac{1}{b-a}$ визначається умовою нормування

$$P(a < X < b) = C(b-a) = 1.$$

Якщо $X$ рівномірно розподілена на проміжку $(a,b)$, то імовірність належності $X$ будь-якому інтервалу $(x_1, x_2) \in (a,b)$ пропорційна довжині цього інтервалу

$$P(x_1 < X < x_2) = X(x_2 - x_1) = \frac{x_2 - x_1}{b-a}.$$

Іншими словами, імовірність влучення $X$ в інтервал $(x_1, x_2)$ дорівнює відношенню довжини цього інтервалу до довжини усього проміжку $(a,b)$. Цей розподіл задовольняють, наприклад, похибки округлення різноманітних

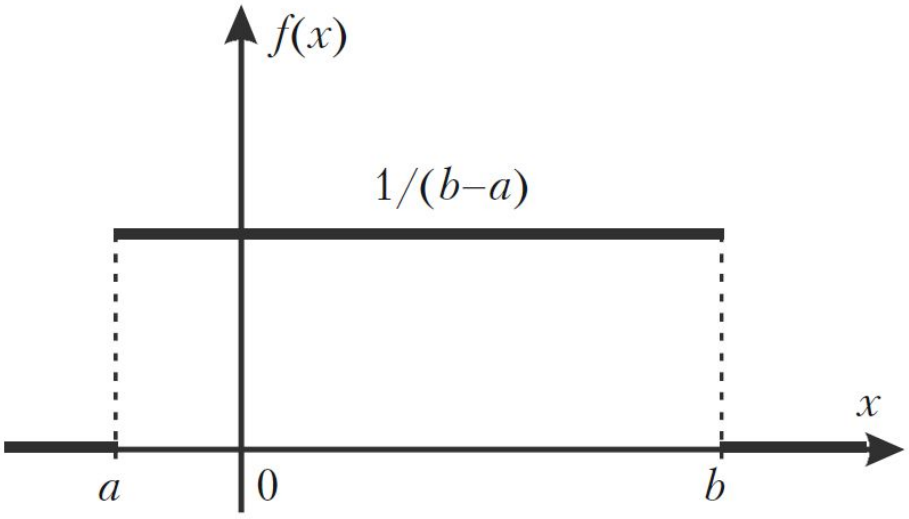


Рис. 4.4 —

розрахунків. Графік рівномірного розподілу НВВ $X$ зображено на Рис.4.4.

Числовими характеристиками НВВ $X$, що розподілена за рівномірним

законом, будуть

$$M(X) = \int\limits_{-\infty}^{+\infty} xf(x)dx =$$
$$\int\limits_{-\infty}^{a} xf(x)dx + \int\limits_{a}^{b} xf(x)dx + \int\limits_{b}^{+\infty} xf(x)dx = \tag{4.64}$$
$$\int\limits_{a}^{b} \frac{x}{b-a}dx = \frac{b^2-a^2}{2(b-a)}\Big|_a^b = \frac{b+a}{2},$$

$$D(X) = \int\limits_{-\infty}^{+\infty} (x - M(X))^2 f(x)dx =$$
$$\int\limits_{a}^{b} \frac{(x-M(X))^2}{b-a} = \frac{\left(x - \dfrac{b+a}{2}\right)}{3(b-a)}\Big|_a^b = \frac{(b-a)^2}{12}, \tag{4.65}$$

$$\sigma(X) = \frac{(b-a)\sqrt{3}}{6}.$$

**Означення 4.11 Показниковий розподіл.** Випадкову величину $X$ називають розподіленою за показниковим законом, якщо щільність її імовірностей має вигляд

$$f(x) = \begin{cases} \lambda e^{-\lambda x} \text{ при } x \geq 0, \\ \\ 0 \text{ при } x < 0, \end{cases}$$

де $\lambda > 0$ — параметр.

Показниковому розподілу задовольняють: час телефонної розмови, час ремонту техніки, час безвідмовної роботи комп'ютера.

Числовими характеристиками показникового розподілу будуть

$$M(X) = \frac{1}{\lambda}, \quad D(X) = \frac{1}{\lambda^2}, \quad \sigma(X) = \frac{1}{\lambda}. \tag{4.66}$$

**Приклад 4.11** Знайти числові характеристики випадкової величини, розподіленої за законом

$$f(x) = \begin{cases} 4e^{-4x} \text{ при } x \geq 0, \\ \\ 0 \text{ при } x < 0, \end{cases}$$

**Розв'язок задачі.** У даному випадку випадкова величина $X$ розподілена

за показниковим законом із параметром $\lambda = 4$. Згідно з формулами (4.66) маємо

$$M(X) = \sigma(X) = 0.25, \quad D(X) = \frac{1}{16}.$$

Якщо випадкова величина $X$ розподілена за показниковим законом, то її функція розподілу (інтегральна функція розподілу) має вигляд $F(x) = 1 - e^{-\lambda x}$. Тому основна формула теорії імовірностей набуде вигляду

$$P(a < X < b) = e^{-a\lambda} - e^{-b\lambda}. \tag{4.67}$$

**Приклад 4.12** Величина $X$ розподілена за законом

$$f(x) = \begin{cases} 3e^{-3x} \text{ при } x \geq 0, \\ \\ 0 \text{ при } x < 0, \end{cases}$$

Знайти імовірність того, що $X$ потрапить в інтервал (0.4,1).

**Розв'язок задачі.** Випадкова величина $X$ розподілена за показниковим законом із параметром $\lambda = 3$. Використовуючи формулу (4.67), отримаємо

$$P(0.4 < X < 1) = e^{-0.4 \cdot 3} - e^{-1 \cdot 3} = e^{-1.2} - e^{-3}.$$

**Означення 4.12 Нормальний розподіл.** Випадкову величину $X$

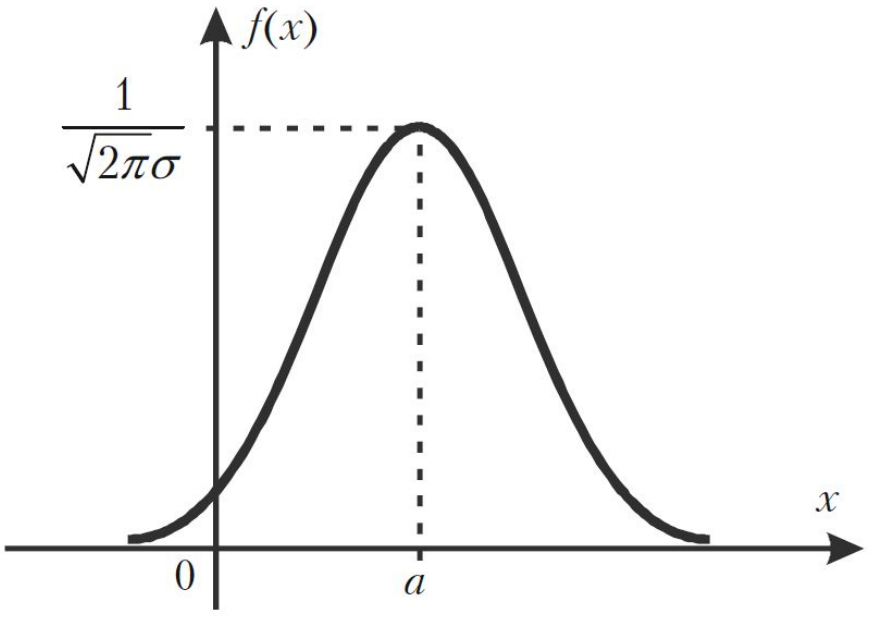


Рис. 4.5 —

називають розподіленою нормально, якщо щільність її імовірностей має вигляд

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

де $\alpha$ та $\sigma$ — параметри розподілу.

Графік цієї функції $f(x)$ називають нормальною кривою або кривою Гаусса. Повне дослідження цієї функції методами диференціального числення дозволяє побудувати графік нормальної кривої, який зображено на Рис.4.5.

При $\alpha = 0$ та $\sigma = 1$ нормальну криву називають нормованою, її рівняння буде

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

Тобто це є функція Лапласа, яка табульована.

Заміна змінної $Z = \dfrac{Z - \alpha}{\sigma}$, використання інтеграла Пуассона

$$\int_{-\infty}^{+\infty} e^{-\frac{z^2}{2}} dz = \sqrt{2\pi}$$

та формул (4.58), (4.59) та (4.62) дозволяють одержати числові характеристики нормально розподіленої НВВ $X$ у вигляді

$$M(X) = \alpha, \quad D(X) = \sigma^2, \quad \sigma(X) = \sigma.$$

Отже, математичне сподівання нормального розподілу дорівнює параметру $\alpha$ цього розподілу, а середнє квадратичне відхилення дорівнює параметру $\sigma$.

Якщо випадкова величина $X$ розподілена за нормальним законом з параметрами $\alpha$ та $\sigma$, то випадкова величина $Z = \dfrac{X - \alpha}{\sigma}$ буде розподілена за нормованим нормальним законом і $M(Z) = 0$, $\sigma = 1$.

Інтегральною функцією нормального закону розподілу буде

$$F(X) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(z-\alpha)^2}{2\sigma^2}} dz,$$

для нормованого нормального закону

$$\Phi = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{z^2}{2}} dz. \tag{4.68}$$

Імовірність влучення в інтервал $(c, d)$ нормально розподіленої випадкової

величини $X$ знаходять за формулою

$$P(c < X < d) = \Phi\Big(\frac{d-\alpha}{\sigma}\Big) - \Phi\Big(\frac{c-\alpha}{\sigma}\Big), \tag{4.69}$$

де фунція Лапласа $\Phi(x)$ має вигляд (4.68).

**Приклад 4.13** Випадкова величина $X$ розподілена за нормальним законом, її математичне сподівання дорівнює 30, а серед квадратичне відхилення — 10. Знайти імовірність того, що $X$ матиме значення з інтервалу (10,50).

**Розв'язок задачі.** Згідно умови $\alpha = 30$, $\sigma = 10$ тому за формулою (4.68) одержимо

$$P(10 < X < 50) = \Phi\Big(\frac{50-30}{10}\Big) - \Phi\Big(\frac{10-30}{10}\Big) = 2\Phi(2).$$

## Правило трьох сигм

Якщо випадкова величина $X$ розподілена нормально, то

$$P(|X-\alpha| > 3\sigma) \to 0,$$

тобто імовірність того, що абсолютна величина відхилення $X$ від її математичного сподівання більше $3\sigma$ прямує до 0, а це означає, що $|X-\alpha| < 3\sigma$.

У практиці це правило використовують так: Якщо закон розподілу випадкової величини $X$ невідомий, але $|X-\alpha| < 3\sigma$, тоді можна припустити, що $X$ розподілена нормально.

**Означення 4.13 Розподіл** $\chi^2$. Нехай $X_i (i = 1, 2, ..., n)$ — нормальні, нормовані незалежні величини, тобто їх математичне сподівання дорівнює нулю, середнє квадратичне відхилення дорівнює одиниці і кожна з них розподілена за нормальним законом. Тоді сума квадратів цих величин

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

розподілена по закону $\chi^2$ з $k = n$ степенями вільності.

Якщо величини і $X_i$ зв'язані одним лінійним співвідношенням, наприклад, $\sum_{i=1}^{n} X_i = n\overline{X}$, то число степеней вільності буде $k = n - 1$.

Диференціальна функція розподілу $\chi^2$ має вигляд

$$f(x) = \begin{cases} 0 \text{ при } x \leq 0, \\ \dfrac{1}{2^{\frac{k}{2}} \cdot \Gamma(\frac{k}{2})} \cdot e^{-\frac{x}{2}-1} \cdot x^{\frac{k}{2}-1} \text{ при } x > 0, \end{cases}$$

де $\Gamma(x) = \int\limits_0^\infty t^{x-1}e^{-t}dt$ — гама-функція, $\Gamma(n+1) = n!$.

Відмітимо, що розподіл $\chi^2$ визначається параметром — числом степеней вільності $k$. Коли $k$ зростає, розподіл $\chi^2$ прямує до нормального розподілу дуже повільно.

**Означення 4.14 Розподіл Стьюдента.** Нехай $X$ — нормальна нормована випадкова величина, а $Y$ — незалежна від $X$ величина, яка розподілена за законом $\chi^2$ з $k$ степенями вільності. Тоді величина

$$T = \frac{X}{\sqrt{\dfrac{Y}{k}}}$$

має розподіл, який називають $t$-розподілом або розподілом Стьюдента (це є псевдонім англійського статистика Вільяма Госсета) з $k$ степенями вільності.

При зростанні $k$ розподіл Стьюдента швидко наближається до нормального розподілу.

## 4.5. Закон великих чисел та центральна гранична теорема

Граничні теореми теорії імовірностей встановлюють відповідність між теоретичними та дослідними характеристиками випадкових величин або випадкових подій при великій кількості випробувань [1-5].

Граничні теореми, які встановлюють відповідність між теоретичними та дослідними характеристиками випадкових подій, об'єднують загальною назвою — закону великих чисел.

Граничні теореми, що встановлюють граничні закони розподілу випадкових величин, об'єднують загальною назвою — центральна гранична теорема.

Необхідність граничних теорем обумовлена потребою розв'язання, наприклад, таких задач:

- Коли сума багатьох випадкових величин мало відрізняється від постійної величини, тобто майже перестає бути випадковою величиною і тому її

поведінка може прогнозуватись із значною імовірністю.

- При яких умовах можна із значною імовірністю прогнозувати число появ деякої випадкової події при великій кількості незалежних випробувань.
- При яких обмеженнях сума багатьох випадкових величин буде розподілена за нормальним законом.

## Нерівності Чебишова

При доведенні різних граничних теорем, а також при розв'язанні різних задач важливу роль грає нерівність Чебишова, яка має дві форми.

**Перша форма нерівності Чебишова.** Для довільної випадкової величини $X$, яка приймає невід'ємні значення та має скінчене математичне сподівання має місце нерівність

$$P(X \geq 1) \leq M(X)$$

Якщо $X$ — дискретна випадкова величина, то

$$P(X \geq 1) = \sum_i p(x_i) \leq \sum_i x_i p(x_i) = M(X)$$

Якщо $X$ – неперервна випадкова величина, $f(x)$ — щільність її імовірностей, то

$$P(X \geq 1) = \int_1^\infty f(x)dx \leq \int_1^\infty xf(x)dx \leq \int_0^\infty xf(x)dx = M(X).$$

Якщо X приймає лише невід'ємні значення $M(X) < \infty$, $\alpha > 0$, то

$$P(X \geq \alpha) \leq \frac{M(X)}{\alpha} \tag{4.70}$$

Дійсно

$$P(X \geq \alpha) = P\Big(\frac{X}{\alpha} \geq 1\Big) \leq M\Big(\frac{X}{\alpha}\Big) = \frac{M(X)}{\alpha}.$$

Нерівність (4.70) іноді називають нерівністю Маркова.

**Друга форма нерівності Чебишова.** Якщо випадкова величина $X$ має скінчені математичне сподівання та дисперсію, то для довільного $\varepsilon > 0$ має місце нерівність

$$P(|X - M(X)| < \varepsilon) \geq 1 - \frac{D(X)}{\varepsilon^2}. \tag{4.71}$$

Доведення. Спочатку розглянемо протилежну подію $|X - M(X)| \geq \varepsilon$ Легко бачити, що ця подія еквівалентна події $(X - M(X))^2 \geq \varepsilon^2$, тому до неї можна

застосувати першу форму нерівності Чебишова

$$P(|X - M(X)| \geq \varepsilon) = P\Big(\frac{1}{\varepsilon^2}(X - M(X))^2 \geq 1\Big) \leq \frac{M((X - M(X))^2)}{\varepsilon^2} = \frac{D(X)}{\varepsilon^2} \tag{4.72}$$

Тепер імовірність протилежної події $|X - M(X)| \geq \varepsilon$ задовольняє нерівність (4.71), що і треба було довести.

**Приклад 4.14** Дисперсія випадкової величини $X$ дорівнює 0.001. Яка імовірність того, що випадкова величина $X$ відрізняється від її математичного сподівання $(X)$ більше ніж на 0.1?

**Розв'язок задачі.** За нерівністю Чебишова (4.72) маємо

$$P(|X - M(X)| > 0.1) \leq \frac{D(X)}{0.1^2} = \frac{0.001}{0.01} = 0.1.$$

## 4.6. Важливі граничні теореми

**Теорема 4.4** *(Теорема Бернуллі). Нехай імовірність появи події $A$ в кожному із $n$ незалежних повторних випробувань дорівнює $p$, $m$ — число появ події $A$ (частота події) в $n$ випробуваннях. Тоді*

$$\lim_{n=\infty} P\Big(\Big|\frac{m}{n} - p\Big| < \varepsilon\Big) = 1, \quad \varepsilon > 0. \tag{4.73}$$

**Доведення**. Частость $\dfrac{m}{n}$ можна розглядати як невід'ємну випадкову величину $X$. Знайдемо її математичне сподівання

$$M(X) = M\Big(\frac{m}{n}\Big) = \frac{1}{n}M(m) = \frac{1}{n} \cdot np = p.$$

Отже, необхідно оцінити імовірність відхилення випадкової величини $X$ від її математичного сподівання. Для цього знайдемо дисперсію цієї випадкової величини

$$D(X) = D\Big(\frac{m}{n}\Big) = \frac{1}{n^2}D(m) = \frac{1}{n^2}np(1-p) = \frac{p(p-1)}{n}.$$

За нерівністю Чебишова (4.71) одержимо

$$P\Big(\Big|\frac{m}{n} - p\Big| \leq \varepsilon\Big) = 1 - \frac{p(p-1)}{n\varepsilon^2}.$$

Звідси граничним переходом при $n \to \infty$ одержуємо (4.73), що й треба було довести.

**Теорема 4.5** *(Теорема Чебишова.) Нехай $X_1, X_2, ..., X_n$ — послідовність попарно незалежних випадкових величин, які задовольняють умовам $M(X_i) = a_i$, $D(X_i) \le c$ для усіх $i = 1, 2, ..., n$. Тоді*

$$\lim_{n\to\infty} P\left(\left|\frac{\sum_{i=1}^{n} X_i}{n} - \frac{\sum_{i=1}^{n} a_i}{n}\right|\right) = 1. \tag{4.74}$$

**Доведення**. Знайдемо математичне сподівання та дисперсію середньої випадкових величин, тобто

$$\frac{X_1 + X_2 + ... + X_n}{n}$$

$$M\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) = \frac{1}{n} M\Big(\sum_{i=1}^{n} X_i\Big) = \frac{1}{n}\sum_{i=1}^{n} M(X_i) = \frac{1}{n}\sum_{i=1}^{n} a_i;$$

$$D\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) = \frac{1}{n^2} D\Big(\sum_{i=1}^{n} X_i\Big) = \frac{1}{n^2}\sum_{i=1}^{n} D(X_i) = \frac{cn}{n^2} = \frac{c}{n}.$$

Застосуємо для випадкової величини та $\dfrac{1}{n}\sum_{i=1}^{n} X_i$ нерівність Чебишова (4.71)

$$P\left(\left|\frac{\sum_{i=1}^{n} X_i}{n} - \frac{\sum_{i=1}^{n} a_i}{n}\right| < \varepsilon\right) \ge 1 - \frac{c}{n\varepsilon^2} \tag{4.75}$$

Границя цієї імовірності при $n \to \infty$ дорівнює одиниці, тобто рівність (4.74) доведено.

**Теорема 4.6** *(Центральна гранична теорема.) Нехай задана послідовність незалежних однаково розподілених випадкових величин $X_1, X_2, ..., X_n$, $M(X_i) = 0$, $D(X_i) = b$, $i = 1, 2, ....$*

*Розглянемо випадкову величину* $Y_n = \sum_{i=1}^{n} X_i$. *Тоді*

$$M(Y_n) = \sum_{i=1}^{n} M(X_i) = 0; \quad D(Y_n) = \sum_{i=1}^{n} D(X_i) = nb.$$

*При* $n \to \infty$ *функція розподілу*

$$F_{Y_n}(x) = P(Y_n < x) \to \frac{1}{\sqrt{2\pi nb}} \int_{-\infty}^{x} e^{-\frac{z^2}{2nb}} dz,$$

*тобто сума* $Y_n$ *буде розподілена за нормальним законом з математичним сподіванням 0 та дисперсією* $\sigma = \sqrt{nb}$.

Для доведення цієї теореми треба знайти границю характеристичної функції, побудованої для нормованої випадкової величини $Z_n = \dfrac{Y_n}{\sqrt{nb}}$.

При $n \geq 30$ розподіл суми однаково розподілених випадкових величин мало відрізняється від нормального розподілу.

**Теорема 4.7** *(Теорема Ляпунова.) Нехай задана послідовність незалежних випадкових величин* $X_1, X_2, ..., X_n, ...$ *таких, що* $M(X_i) = 0$, $D(X_i) = {b_i}^2$, $i = 1, 2, ..., n, ....$

*Побудуємо суму випадкових величин* $Y_n = \sum_{i=1}^{n} X_i$. *Позначимо* ${B_n}^2 = \sum_{i=1}^{n} {b_i}^2$. *Якщо виконується умова рівномірної малості величин, що утворюють суму*

$$\frac{1}{{B_n}^3} \sum_{i=1}^{n} M(X_i)^3 \to 0 \text{ при } n \to \infty,$$

*то сума* $Y_n$ *буде розподіленою нормально з математичним сподіванням* $M(Y_n) = 0)$ *та дисперсією* $D(Y_n) = {B_n}^2$.

Доведення цієї теореми досить складне, але відмітимо, що у випадку, коли $M(X_i) = a_i$ не дорівнює нулю, можна розглядати випадкові величини ${X'}_i = X_i - a_i$, які будуть задовольняти умову теореми Ляпунова.

**Приклад 4.15** Скільки додатків треба взяти у теоремі Чебишова, щоб з надійністю 96% і точністю до 0.01 виконувалась наближена рівність

$$\frac{1}{n} \sum_{i=1}^{n} X_i = \frac{1}{n} \sum_{i=1}^{n} M(X_i).$$

**Розв'язок задачі.** В цьому прикладі $\varepsilon = 0.01$. Щоб одержати надійність 96% згідно формули (4.75) достатньо підібрати таке $n$, яке задовольняє нерівність

$$\frac{c}{\varepsilon^2 n} \leq 0.04 \to n \geq \frac{c}{0.04 \cdot 0.0001} = 250000.$$

Цей приклад показує, що навіть у випадку не дуже великих точності та надійності, треба брати значну кількість додатків ($n$ — досить велике число). Це означає, що оцінки, одержані з використанням нерівності (4.75), — завищені. Більш точні оцінки можна одержати за допомогою теореми Ляпунова.

## 4.7. Питання для самоперевірки

- Який розподіл називають біноміальним?
- Які випадкові величини розподілені за біноміальним законом?
- Навести формули для обчислення математичного сподівання, дисперсії і середнього квадратичного відхилення біноміальної величини.
- Навести формулу для обчислення ймовірності влучення значення біноміальної величини в заданий діапазон $[k_1, k_2]$.
- Який розподіл називають геометричним?
- Навести формули для обчислення математичного сподівання, дисперсії і середнього квадратичного відхилення випадкової величини, що має геометричний розподіл.
- Який розподіл називають розподілом Пуассона?
- Які випадкові величини розподілені за законом Пуассона?
- Навести формули для обчислення математичного сподівання, дисперсії і середнього квадратичного відхилення випадкової величини, що має розподіл Пуассона.
- Які випадкові величини розподілені за законом Пуассона?
- Який розподіл називають рівномірним?
- Записати в загальному вигляді інтегральну функцію рівномірно розподіленої випадкової величини. Навести формули для обчислення математичного сподівання, дисперсії і середнього квадратичного відхилення випадкової величини, що має рівномірний розподіл.

- Дати визначення випадковим величинам, розподіленим за показниковим законом.
- Записати в загальному вигляді інтегральну функцію випадкової величини, розподіленої за показниковим законом.
- Навести формули для обчислення математичного сподівання, дисперсії і середнього квадратичного відхилення випадкової величини, що має показниковий розподіл.
- Навести формулу для обчислення ймовірності влучення значення випадкової величини, розподіленої за показниковим законом, в заданий діапазон $[a, b]$, де $a$ і $b$ – невід'ємні величини.
- Який розподіл називають нормальним?
- Записати в загальному вигляді інтегральну функцію нормально розподіленої випадкової величини.
- Навести формули для обчислення математичного сподівання, дисперсії і середнього квадратичного відхилення випадкової величини, що має нормальний розподіл.
- Навести рекурентне співвідношення для визначення центральних моментів нормально розподіленої випадкової величини.
- Чому дорівнює коефіцієнт асиметрії нормально розподіленої випадкової величини?
- Чому дорівнює коефіцієнт гостровершинності нормально розподіленої випадкової величини?
- Навести формулу для обчислення ймовірності влучення значення випадкової величини, розподіленої за нормальним законом, в заданий діапазон $[c, d]$.
- В чому полягає правило трьох сигм?

# 5. Випадкові вектори і функції випадкових аргументів

## 5.1. Випадкові вектори

**Означення 5.1** Випадковим вектором називають вектор $X = (X_1, X_2, ..., X_n)$, компоненти якого являють собою випадкові величини [1-5].

Для випадкового вектора так само, як і для випадкової величини, вводяться поняття інтегральної функції розподілу, щільності розподілу, визначаються числові характеристики.

**Означення 5.2** Інтегральна функція розподілу випадкового вектора — це така функція $F(x_1, x_2, ..., x_n)$, яка при конкретних значеннях своїх аргументів чисельно дорівнює ймовірності того, що випадкові компоненти вектора виявляться менше за відповідні аргументи, тобто

$$F(x_1, x_2, ..., x_n) = P(X_1 < x_1, X_2 < x_2, ..., X_n < x_n). \tag{5.1}$$

Надалі будуть розглядатися тільки двовимірні випадкові вектори $Z = (X, Y)$, де $(X, Y)$ — компоненти вектора. Однак усі наведені положення в однаковій мірі справедливі і для багатовимірних векторів, або легко узагальнюються на випадок багатовимірних векторів.

За визначенням інтегральна функція $F(x, y)$ двовимірного випадкового вектора $Z = (X, Y)$ — це функція , яка при кожних конкретних значеннях своїх аргументів $x$ і $y$ чисельно дорівнює ймовірності того, що випадкові компоненти вектора виявляться менше за відповідні аргументи, тобто $F(x, y) = P(X < x, Y < y)$. В цьому разі геометричний зміст функції розподілу — ймовірність влучення випадкового вектора в безмежний квадрат з вершиною в точці $(x_0, y_0)$ (див. рис. 5.1). Багатовимірні випадкові величини

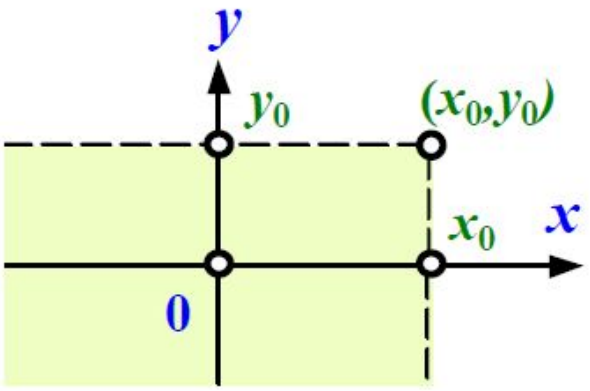


Рис. 5.1 —

бувають дискретними та неперервними (компоненти цих величин відповідно будуть дискретними та неперервними).

## 5.2. Закон розподілу імовірностей дискретної двовимірної випадкової величини

**Означення 5.3** Законом розподілу дискретної двовимірної випадкової величини називають перелік можливих значень цієї величини $(x_i, y_i)$ та їх імовірностей $p(x_i, y_i)$, $i = 1, 2, ..., n$; $k = 1, 2, ..., m$.

Найчастіше закон розподілу двовимірної дискретної випадкової величини задають у вигляді таблиці з двома входами.

У першому рядку таблиці записують усі можливі значення компоненти $X$. У першому стовпчику таблиці записують усі можливі значення компоненти $Y$. На перетині $k-$того рядка та $i-$того стовпчика записують імовірність $p(x_i, y_i)$ того, що двовимірна випадкова величина $(X, Y)$ прийме значення $p(x_i, y_i)$, $i = 1, 2, ..., n$; $k = 1, 2, ..., m$.

| $X$ \ $Y$ | $y_1$ | $y_2$ | ... | $y_k$ | ... | $y_m$ |
|---|---|---|---|---|---|---|
| $x_1$ | $p(x_1, y_1)$ | $p(x_1, y_2)$ | ... | $p(x_1, y_k)$ | ... | $p(x_1, y_m)$ |
| $x_2$ | $p(x_2, y_1)$ | $p(x_2, y_2)$ | ... | $p(x_2, y_k)$ | ... | $p(x_2, y_m)$ |
| ... | ... | ... | ... | ... | ... | ... |
| $x_i$ | $p(x_i, y_1)$ | $p(x_i, y_2)$ | ... | $p(x_i, y_k)$ | ... | $p(x_i, y_m)$ |
| ... | ... | ... | ... | ... | ... | ... |
| $x_n$ | $p(x_n, y_1)$ | $p(x_n, y_2)$ | ... | $p(x_n, y_k)$ | ... | $p(x_n, y_m)$ |

Табл. 5.1 — Таблиця розподілу

Події $(X = x_i, Y = y_k)$, $i = 1, 2, ..., n$; $k = 1, 2, ..., m$, утворюють повну групу, тому сума імовірностей таблиці дорівнює одиниці, тобто

$$\sum_{i=1}^{n} \sum_{k=1}^{m} p(x_i, y_k) = 1. \tag{5.2}$$

Закон розподілу двовимірної випадкової величини дозволяє отримати закони розподілу кожної компоненти. Дійсно, події $(x_i, y_1)$, $(x_i, y_2)$, ..., $(x_i, y_m)$ несумісні, тому імовірність $P(x_i)$ того, що $X$ прийме значення і $x_i$ за теоремою додавання імовірностей буде

$$P(x_i) = P(x_i, y_1) + P(x_i, y_2) + ... + P(x_i, y_m), \tag{5.3}$$

тобто дорівнює сумі імовірностей, що розташовані в $i-$тому стовпчику таблиці розподілу.

Аналогічно, додаванням імовірностей $k-$того рядка таблиці, одержимо імовірність

$$P(y_k) = P(x_1, y_k) + P(x_2, y_k) + ... + P(x_n, y_k). \tag{5.4}$$

**Приклад 5.1** Знайти розподіл координат випадкового вектора $Z = (X, Y)$, заданого таблицею розподілу:

| $X$ \ $Y$ | $-2$ | $-1$ | $2$ |
|---|---|---|---|
| $0$ | $0.15$ | $0.05$ | $0.25$ |
| $1$ | $0.35$ | $0.2$ | $0$ |

Табл. 5.2 — Таблиця розподілу $Z = (X, Y)$

**Розв'язок задачі.** На підставі формул (5.3), (5.4) одержимо розподіли координат $X$ і $Y$ :

| $X$ | $0$ | $1$ |
|---|---|---|
| $p$ | $0.45$ | $0.55$ |

Табл. 5.3 — Таблиця розподілу $X$

| $Y$ | -2 | -1 | 2 |
|---|---|---|---|
| $p$ | 0.5 | 0.25 | 0.25 |

Табл. 5.4 — Таблиця розподілу $Y$

**Означення 5.4 Інтегральною функцією розподілу (функцією розподілу)** двовимірної випадкової величини $(X, Y)$ називають функцію двох змінних $F(x, y)$, яка визначає для кожної пари чисел $(X, Y)$ імовірність виконання нерівностей $X < x$; $Y < y$, тобто

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

Аналогічно визначають функцію розподілу системи $n$ випадкових величин

$$F(x_1, x_2, ..., x_n) = P(X_1 < x_1, X_2 < x_2, ..., X_n < x_n).$$

Властивості імовірності P та функції розподілу дозволяють довести такі властивості функції розподілу, які у випадку двовимірної випадкової величини виглядають

1. $0 \leq F(x, y) \leq 1$
2. $F(x, y)$ не спадна функція за кожним аргументом, тобто

$$\begin{aligned} &F(x_2, y) \geq F(x_1, y) \text{ якщо } x_2 > x_1; \\ &F(x, y_2) > F(x, y_1) \text{ якщо } y_2 > y_1. \end{aligned} \tag{5.5}$$

3. Мають місце граничні співвідношення $F(-\infty, y) = 0$; $F(x, -\infty) = 0$; $F(-\infty, -\infty) = 0$; $F(+\infty, +\infty) = 1$.

4. Імовірність влучення випадкової точки до прямокутника

$$\{x_1 \le X \le x_2;\ y_1 \le Y \le y_2;\}$$

можна знайти за формулою

$$\begin{aligned} &P(x_1 \le X \le x_2,\ y_1 \le Y \le y_2) = \\ &(F(x_2, y_2) - F(x_1, y_2)) - (F(x_2, y_1) - F(x_1, y_1)). \end{aligned} \tag{5.6}$$

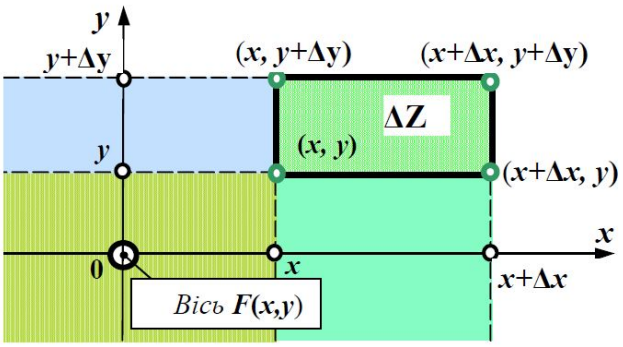


Рис. 5.2 —

Геометричний зміст функції розподілу $F(x, y)$ — це імовірність того, що випадкова точка $Z = (X, Y)$ попаде у нескінчено малий прямокутник з вершиною в точці $(x, y)$ і розміщений вижче та правіше цієї вершини (див. рис.5.2).

**Приклад 5.2** Знайти імовірність влучення випадкової точки $Z = (X, Y)$ у прямокутник, обмежений лініями

$$x = \frac{\pi}{6} \quad x = \frac{\pi}{2} \quad y = \frac{\pi}{3} \quad y = \frac{\pi}{4}$$

якщо задана функція розподілу вигляду

$$F(x, y) = \sin(x) \cdot \sin(y); \quad 0 \le x \le \frac{\pi}{2}; \quad 0 \le y \le \frac{\pi}{2}$$

**Розв'язок задачі.** У заданому випадку

$$x_1 = \frac{\pi}{6}, \quad x_2 = \frac{\pi}{2}, \quad y_1 = \frac{\pi}{4}, \quad y_2 = \frac{\pi}{2}.$$

Згідно з формулою (5.6) одержуємо

$$\begin{aligned} &P\Big(\frac{\pi}{6} \le X \le \frac{\pi}{2},\ \frac{\pi}{4} \le Y \le \frac{\pi}{3}\Big) = \\ &\Big(\sin\Big(\frac{\pi}{2}\Big)\cdot\sin\Big(\frac{\pi}{3}\Big) - \sin\Big(\frac{\pi}{6}\Big)\cdot\sin\Big(\frac{\pi}{3}\Big)\Big) - \\ &\Big(\sin\Big(\frac{\pi}{2}\Big)\cdot\sin\Big(\frac{\pi}{4}\Big) - \sin\Big(\frac{\pi}{6}\Big)\cdot\sin\Big(\frac{\pi}{4}\Big)\Big) = \\ &\Big(\frac{\sqrt{3}}{2} - \frac{1}{2}\cdot\frac{\sqrt{3}}{2}\Big) - \Big(\frac{\sqrt{2}}{2} - \frac{1}{2}\cdot\frac{\sqrt{2}}{2}\Big) = 0.08 \end{aligned} \tag{5.7}$$

## 5.3. Неперервна двовимірна випадкова величина

Двовимірну випадкову величину можна задавати функцією розподілу $F(x, y)$ або диференціальною функцією розподілу.

**Означення 5.5 Диференціальною функцією розподілу (двовимірною щільністю імовірностей)** $f(x, y)$ двовимірної випадкової величини (X, Y) називають мішану частинну похідну другого порядку від інтегральної функції розподілу

$$f(x, y) = \frac{\partial^2 F(x, y)}{\partial x \partial y} \tag{5.8}$$

Аналогічно визначають щільність імовірностей $n$-вимірної випадкової величини, тобто

$$f(x_1, x_2, ..., x_n) = \frac{\partial^2 F(x_1, x_2, ..., x_n)}{\partial x_1 \partial x_2 ... \partial x_n}. \tag{5.9}$$

Таким чином, якщо функція розподілу $F(x, y)$ двовимірної випадкової величини відома, то за формулою (5.8) можна знайти диференціальну функцію розподілу $f(x, y)$ цієї випадкової величини.

Якщо відома щільність імовірностей $f(x, y)$ двовимірної випадкової величини, то її функцію розподілу знаходять за формулою

$$F(x, y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f(x, y)dxdy, \tag{5.10}$$

тобто з використанням невласного двократного інтегралу.

Імовірність влучення випадкової точки $(X, Y)$ в довільну область $D$ знаходять за формулою

$$P((X, Y) \in D) = \iint_D f(x, y)dxdy. \tag{5.11}$$

Диференціальна функція розподілу $f(x,y)$ задовольняє властивостям:

1. $f(x,y) \geq 0$, тобто вона не від'ємна;
2. Умові нормування $\int\limits_{-\infty}^{+\infty}\int\limits_{-\infty}^{+\infty} f(x,y)dxdy = 1$.

## 5.4. Залежні та незалежні випадкові величини

Дві випадкові величини незалежні, якщо закон розподілу однієї з них не залежить від того, які можливі значення прийняла друга величина.

Випадкові величини залежні, якщо закон розподілу однієї величини залежить від того, які значення прийняла друга величина.

У теорії імовірностей доведено таке твердження.

**Теорема 5.1** *Щоб випадкові величини $X$ та $Y$ були незалежні, необхідно і достатньо, щоб інтегральна функція системи $(X,Y)$ дорівнювала добутку інтегральних функцій кожної з них*

$$F(x,y) = F_1(x) \cdot F_2(y). \tag{5.12}$$

**Наслідок.** Щоб неперервні випадкові величини $X$ та $Y$ були незалежними, необхідно і достатньо, щоб диференціальна функція системи $(X,Y)$ дорівнювала добутку диференціальних функцій складових

$$f(x,y) = f_1(x) \cdot f_2(y). \tag{5.13}$$

## 5.5. Числові характеристики двовимірної випадкової величини

Математичне сподівання двовимірної випадкової величини $(X,Y)$ характеризує координати центру розподілу випадкової величини. Ці координати у випадку неперервних величин знаходять за формулами

$$m_x = \int\limits_{-\infty}^{+\infty}\int\limits_{-\infty}^{+\infty} xf(x,y)dxdy; \quad m_y = \int\limits_{-\infty}^{+\infty}\int\limits_{-\infty}^{+\infty} yf(x,y)dxdy. \tag{5.14}$$

Дисперсії $D_X$ та $D_Y$ характеризують розсіювання випадкової точки $(x, y)$ вздовж координатних осей  та , відповідно. їх знаходять за формулами

$$\begin{aligned} D_x &= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} x^2 f(x,y)dxdy - {m_X}^2; \\ D_y &= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} y^2 f(x,y)dxdy - {m_Y}^2. \end{aligned} \tag{5.15}$$

Для опису двовимірної випадкової величини крім математичного сподівання, дисперсії та середніх квадратичних відхилень

$$\sigma_X = \sqrt{D_X}, \quad \sigma_Y = \sqrt{D_Y} \tag{5.16}$$

використовують також інші характеристики, а саме — кореляційний момент (або коваріація)

$$cov(X,Y) = K_{XY} = M((X - m_x)(Y - m_y)). \tag{5.17}$$

Для неперервних величин $X$ та $Y$

$$K_{XY} = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} (x - m_x)(y - m_y) f(x,y)dxdy. \tag{5.18}$$

## Кореляція та властивості коефіцієнта кореляції

Коефіцієнт кореляції вводиться таким чином

$$r_{XY} = \frac{K_{XY}}{\sigma_X \sigma_Y}. \tag{5.19}$$

Коефіцієнт кореляції є кількісна характеристика залежності випадкових величин $X$ та $Y$ і часто використовується в статистиці.

Якщо випадкові величини $X$ та $Y$ дискретні, то в формулах (5.14)—(5.18) знаки інтегралів замінюють знаками суми по усім можливим значенням випадкових величин.

**Означення 5.6** Випадкові величини $X$ та $Y$ звуть некорельованими, якщо їх кореляційний момент або коефіцієнт кореляції дорівнює нулеві.

Властивості коефіцієнта кореляції:

1. $|r_{XY}| \leq 1$;
2. якщо $X$ та $Y$ незалежні, то $r_{XY} = 0$;

3. якщо між $X$ та $Y$ є лінійна залежність $Y = aX + bY$, де $a$ та $b$ — постійні, то $|r_{XY}| = 1$.

Якщо момент кореляції або коефіцієнт кореляції не дорівнює нулеві, тоді випадкові величини $X$ та $Y$ — корельовані. Дві корельовані величини обов'язково залежні. Але дві залежні випадкові величини можуть бути корельованими або некорельованими, тобто їх коефіцієнт кореляції може дорівнювати нулеві, а може і не дорівнювати нулеві.

Із незалежності двох величин випливає їх некорельованість, але із некорельованості ще не випливає незалежність цих величин. У випадку нормального розподілу величин із некорельованості випадкових величин випливає їх незалежність.

## 5.6. Функції випадкової величини та їх характеристики

**Означення 5.7** Якщо вказано закон, за яким кожному можливому значенню випадкової величини $X$ відповідає певне значення випадкової величини $Y$, то $Y$ звуть функцією $X$ і позначають $Y = \varphi(X)$.

Відзначимо, що іноді різним можливим значенням випадкової величини X відповідають однакові значення Y.

Однією із можливих задач теорії імовірностей є визначення законів розподілу та числових характеристик функцій випадкового аргументу, закон розподілу якого відомий.

### Закон розподілу та числові характеристики функції дискретного випадкового аргументу

Нехай $Y = \varphi(X)$. Тут аргумент $X$ — дискретна випадкова величина з можливими значеннями $x_1, x_2, ..., x_n$ імовірності яких дорівнюють $p_1, p_2, ..., p_n$ відповідно, тобто $X$ задана законом

| $X$ | $x_1$ | $x_2$ | ... | $x_n$ |
|---|---|---|---|---|
| $P(X)$ | $p_1$ | $p_2$ | ... | $p_n$ |

Табл. 5.5 — Таблиця розподілу $X$

У цьому випадку $Y$ також дискретна випадкова величина з можливими значеннями $y_1 = \varphi(x_1)$, $y_2 = \varphi(x_2)$, ..., $y_n = \varphi(x_n)$.

Із події «величина $X$ прийняла значення $x_k$» випливає подія «величина $Y$ прийняла значення $\varphi(x_k)$, тому імовірності можливих значень $Y$ також дорівнюють $p_1, p_2, ..., p_n$. Це означає, що закон розподілу $Y$ буде мати вигляд

| $Y$ | $\varphi(x_1)$ | $\varphi(x_2)$ | ... | $\varphi(x_n)$ |
|---|---|---|---|---|
| $P(Y)$ | $p_1$ | $p_2$ | ... | $p_n$ |

Табл. 5.6 — Таблиця розподілу $Y$

Математичне сподівання, дисперсію та середнє квадратичне відхилення функції $Y$ обчислюють за формулами

$$M(Y) = \sum_{k=1}^{n} \varphi(x_k) p_k, \tag{5.20}$$

$$D(Y) = M(Y^2) - M(Y)^2 = \sum_{k=1}^{n} \varphi(x_k)^2 p_k - M(Y)^2, \tag{5.21}$$

$$\sigma(Y) = \sqrt{D(Y)}. \tag{5.22}$$

Початкові та центральні моменти розподілу знаходять за формулами

$$\nu_k = \sum_{i=1}^{n} \varphi(x_i)^k p_i, \tag{5.23}$$

$$\mu_k = \sum_{i=1}^{n} (\varphi(x_i) - M(Y))^k p_i. \tag{5.24}$$

**Приклад 5.3** Дискретна випадкова величина задана законом розподілу

| $X$ | 1 | 3 | 5 |
|---|---|---|---|
| $P$ | 0.2 | 0.5 | 0.3 |

Табл. 5.7 — Таблиця розподілу $X$

Знайти математичне сподівання функції $Y = X^2 + 1$.

**Розв'язок задачі.** Можливими значеннями $Y$ будуть

$$y_1 = 1^2 + 1 = 2; \quad y_2 = 3^2 + 1 = 10; \quad y_3 = 5^2 + 1 = 26;$$

Знаходимо математичне сподівання $Y$

$$M(Y) = M(X^2 + 1) = 2 \cdot 0.2 + 10 \cdot 0.5 + 26 \cdot 0.3 = 13.2.$$

## Закон розподілу та числові характеристики функції неперервного випадкового аргументу

Нехай $Y$ — неперервна випадкова величина, закон розподілу якої заданий диференціальною функцією розподілу (щільність ймовірностей) $f(x)$; випадкова величина $Y = \varphi(X)$.

Якщо $\varphi(X)$ — диференційовна функція, монотонна на усьому проміжку можливих значень $X$, то щільність розподілу функції $Y = \varphi(X)$ визначають так

$$g(y) = f(\varphi^{-1}(y)) \cdot \left|\frac{d\varphi^{-1}(y)}{dy}\right|, \tag{5.25}$$

де $\varphi^{-1}(y)$ — функція, обернена по відношенню до функції $\varphi(x)$.

Якщо $\varphi^{-1}$ не монотонна функція в області визначення аргументу $X$, то обернена функція неоднозначна і щільність розподілу $g(y)$ визначається як сума додатків, кількість яких дорівнює кількості значень оберненої функції, тобто

$$g(y) = \sum_{i=1}^{k} f({\varphi^{-1}}_i(y)) \cdot \left|\frac{d{\varphi^{-1}}_i(y)}{dy}\right|. \tag{5.26}$$

**Приклад 5.4** Випадкова величина $X$ розподілена за нормальним законом з математичним сподіванням, що дорівнює нулеві. Знайти закон розподілу функції $Y = X^3$.

**Розв'язок задачі.** Згідно означенню нормального розподілу неперервної випадкової величини $X$ та умови прикладу диференціальна функція розподілу $X$ має вигляд

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

Функція $Y = X^3$ диференційовна, $\dfrac{dY}{dx} = 3X^2 > 0$ тому вона зростає для усіх $x \in (-\infty, +\infty)$. Отже, можна застосувати формулу (5.25) для знаходження диференціальної функції розподілу $g(y)$ випадкової величини $Y$.

У даному випадку з рівності $X = Y^{1/3}$ формула (5.25) прийме вигляд

$$g(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{y^{2/3}}{2\sigma^2}} \cdot \left|\frac{y^{1/3}}{dy}\right| = \frac{e^{-\frac{y^{2/3}}{2\sigma^2}}}{3\sqrt{2\pi}\sigma y^{2/3}}.$$

Для знаходження математичного сподівання від $Y = \varphi(X)$ можна спочатку знайти $g(y)$ — диференціальну функцію розподілу величини Y за формулою

((5.25)) або ((5.26)), а потім використати формулу

$$M(Y) = \int\limits_{-\infty}^{+\infty} yg(y)dy.$$

Але більш доцільно знаходити математичне сподівання функції неперервного випадкового аргументу $\varphi(X)$ безпосередньо за формулою

$$M(\varphi(X)) = \int\limits_{-\infty}^{+\infty} \varphi(x)f(x)dx. \tag{5.27}$$

**Приклад 5.5** Неперервна випадкова величина $X$ задана диференціальною функцією розподілу

$$f(x) = \begin{cases} \sin(x) \text{ при } x \in \left(0, \dfrac{\pi}{2}\right); \\ 0 \text{ при } x \notin \left(0, \dfrac{\pi}{2}\right). \end{cases}$$

Знайти математичне сподівання функції $Y = X^2$.

**Розв'язок задачі.** У даному випадку тому за формулою (5.27) одержимо

$$M(Y) = M(X^2) = \int\limits_{0}^{\frac{\pi}{2}} x^2 \sin(x)dx.$$

Інтегруючи частинами два рази, одержимо потрібне математичне сподівання

$$\begin{gathered} \int\limits_{0}^{\frac{\pi}{2}} x^2 \sin(x)dx = -x^2\cos(x)\Big|_0^{\frac{\pi}{2}} + 2\int\limits_{0}^{\frac{\pi}{2}} x\cos(x)dx = \\ 2\Big(x\sin(x)\Big|_0^{\frac{\pi}{2}} - \int\limits_{0}^{\frac{\pi}{2}} \sin(x)dx\Big) = \pi + 2\cos(x)\Big|_0^{\frac{\pi}{2}} = \pi - 2. \end{gathered} \tag{5.28}$$

Отже, одержали

$$M(Y) = N(X^2) = \pi - 2.$$

Дисперсію функції $Y$ неперервного випадкового аргументу $X$ визначають звичайним чином $D(Y) = (Y^2) - M^2(Y)$, а обчислюють за формулою

$$D(Y) = D(\varphi(X)) = \int\limits_{-\infty}^{+\infty} \varphi^2(x)f(x)dx - \Big(\int\limits_{-\infty}^{+\infty} \varphi(x)f(x)dx\Big)^2. \tag{5.29}$$

У випадку, коли $X$ змінюється лише в проміжку $[a, b]$, дисперсію функції $Y = \varphi(X)$ знаходять за формулою

$$D(Y) = D(\varphi(X)) = \int_a^b \varphi^2(x) f(x) dx - \Big( \int_a^b \varphi(x) f(x) dx \Big)^2. \qquad (5.30)$$

У формулах (5.29) та (5.30) функція $f(x)$ — це щільність імовірностей неперервної випадкової величини $X$ (диференціальна функція розподілу $X$).

## 5.7. Питання для самоперевірки

- Як визначають випадкові величини, дискретні та неперервні випадкові величини?
- Якими способами можна визначити дискретну випадкову величину?
- Вказати основні закони розподілу дискретної випадкової величини та умови їх використання.
- Як визначаються і що характеризують числові характеристики дискретних випадкових величин?
- За якими формулами обчислюють числові характеристики дискретних випадкових величин?
- Як визначають функцію розподілу та щільності імовірностей неперервних випадкових величин? Які властивості мають ці функції?
- Який існує зв'язок між інтегральною та диференціальною функціями розподілу імовірностей?
- За якими формулами можна обчислити імовірність влучення випадкової величини в проміжок $(, b)$, використовуючи інтегральну або диференціальну функції розподілу?
- Які числові характеристики існують для неперервних випадкових величин та що характеризує кожна з них?
- За якими формулами обчислюють числові характеристики неперервних випадкових величин?
- Вказати основні властивості математичного сподівання та дисперсії.
- Вказати основні закони розподілу неперервних випадкових величин та їх вигляд.

- Чому дорівнюють числові характеристики основних законів розподілу дискретних та неперервних випадкових величин?
- За якими формулами треба знаходити імовірність влучення випадкової величини $X$ в проміжок $(a, b)$, якщо $X$ розподілено за рівномірним, показниковим або нормальним законами?
- Як знайти функції розподілу функції $Y = \varphi(X)$, якщо $X$ — дискретна або неперервна випадкова величина?
- За якими формулами обчислюють числові характеристики функції дискретного та неперервного випадкового аргументу?
- Як визначають початкові та центральні моменти, коефіцієнт кореляції та як пов'язані поняття кореляції, залежності та незалежності випадкових величин?
- Правило $3\sigma$ та його використання.

# Розділ 2

# Математична статистика

## 1. Основні поняття та статистичний розподіл

### 1.1. Предмет математичної статистики та її основні задачі

Математична статистика - це наука, в якій вивчаються методи реєстрації, опису і аналізу даних наглядів і експериментів з метою побудови моделей вірогідності масових випадкових явищ.

Предметом математичної статистики є вивчення випадкових подій, випадкових величин і процесів за наслідками наглядів, вимірювань, дослідів. Результати наглядів, вимірювань, дослідів називатимемо статистичними.

Методи математичної статистики носять абстрактний характер, використовуються для обробки експериментальних даних будь-якої природи, і тому вони застосовні в будь-яких областях науки, техніки, економіки, медицини, сільського господарства і т.д.

Можна виділити дві основні задачі математичної статистики. Перша задача - розробка методів збору, реєстрації і угрупування статистичних даних, одержаних в результаті наглядів за випадковими процесами. Друга задача математичної статистики полягає у формуванні методів обробки і аналізу одержаних вибіркових даних.

Нині методи математичної статистики широко застосовується у науці про великі дані, машинному навчанні, задачах штучного інтелекту і т.п. У зв'язку з цим для розв'язку типових задач математичної статистики у навчальному посібнику розглядається технологія використання такої мови програмування, як Python 3.*.

Навчальний матеріал, викладений у розділі "Математична статистика", в основному базується на використаній, при підготовці навчального посібника, літературі [2,3,6,7].

## 1.2. Генеральна і вибіркова сукупності. Статистичний розподіл вибірки

Основу досліджень в математичній статистиці, як вже наголошувалося вище, складають дані наглядів.

У математичній статистиці прийнято оперувати поняттями генеральної і вибіркової совокупностей. Введемо наступні визначення.

**Означення 1.1** Сукупність всіх належних вивченню однорідних об'єктів або можливих результатів всіх мислимих наглядів, що проводяться в незмінних умовах над одним об'єктом, називається генеральною сукупністю.

Поняття генеральної сукупності в певному значенні аналогічно поняттю випадкової величини (простору вірогідності, закону розподілу вірогідності). Оскільки поняття генеральної сукупності і сукупності всіх значень випадкової величини пов'язані з випробуваннями (наглядами) в незмінних умовах, то надалі ці поняття не розрізнятимуться. Тому іноді в математичній статистиці генеральною сукупністю називають безліч можливих значень випадкової величини, що вивчається $X$. Під законом розподілу (розподілом) генеральної сукупності $X$ розумітимемо закон розподілу вірогідності випадкової величини $X$, а його числові характеристики називатимемо числовими характеристиками генеральної сукупності.

Генеральна сукупність може бути кінцевою або нескінченною, залежно від того, кінцева або нескінченна сукупність становлячих її елементів або значень випадкової величини.

**Означення 1.2** Вибірковою сукупністю або вибіркою називається сукупність об'єктів, відібраних випадковим чином з генеральної сукупності.

Більш строго вибірку визначають таким чином. Сукупність незалежних випадкових величин $X_1, X_2, ..., X_n$ кожна з яких має той же розподіл, що і випадкова величина $X$, називатимемо випадковою вибіркою з генеральної сукупності $X$ і записувати $(X_1, X_2, ..., X_n)$. Самі значення $X_1, X_2, ..., X_n$ називають також вибірковими значеннями випадкової величини $X$.

Число об'єктів (наглядів) в генеральній або вибірковій сукупності називатимемо її об'ємом і позначатимемо відповідно $N$ і $n$.

Конкретні значення вибірки, одержані в результаті наглядів (випробувань), називають реалізацією вибірки і позначають рядковими буквами $x_1, x_2, ..., x_n$.

**Приклад 1.1** Група з 15 студентів відібрана із загального потоку для тестування по теорії вірогідності. Кожний із студентів, що тестуються, може набрати від 0 до 5 балів включно. Скласти генеральну і вибіркову сукупності. Вказати реалізації вибірки.

**Розв'язок задачі.** Хай $X_i$ — кількість балів, набраних $i$-м $i = 1, 2..., 15$ студентом. Тоді $0, 1, 2, 3, 4, 5$ — всі можливі кількості балів, набраних одним студентом, утворюють генеральну сукупність.

Вибіркою буде $X_1, X_2, ..., X_n$ — результат тестування 15 студентів. Реалізаціями вибірки можуть бути наступні набори з 15 чисел:
$(3, 5, 0, 1, 4, 2, 5, 1, 3, 5, 4, 4, 3, 5, 3)$, $(2, 4, 0, 1, 0, 5, 3, 3, 5, 4, 2, 5, 5, 3, 0)$,
$(4, 4, 5, 5, 3, 2, 1, 2, 4, 3, 3, 4, 0, 2, 5)$ і т.д.

Помітимо, що вибірку можна розглядати як якийсь емпіричний (статистичний) аналог генеральної сукупності. Єство вибіркового методу в математичній статистиці полягає в тому, щоб по певній частині генеральної сукупності (вибірці) можна судити про властивості генеральної сукупності в цілому.

Для отримання якісних характеристик генеральної сукупності необхідно, щоб вибірка була репрезентативною або представницькою, тобто повинна достатньо повно представляти ознаки генеральної сукупності, що вивчаються.

З вибору починаються всі статистичні дослідження. З теорії вірогідності відомий вибір куль з урни. Він може бути встановлений в основу визначення відбору (вибору), вживаного в математичній статистиці. Замість куль вибираються числа, що становлять кінцеву генеральну сукупність.

Раніше наголошувалася найважливіша вимога до вибірки: бути репрезентативною, тобто вибірка повинна представляти всі особливості генеральної сукупності.

Іншою вимогою є вимога однорідності вибірки. Це означає, що складання вибірки повинне виконуватися в незмінних умовах. Розрізняють вибірки малі і великі, і вони відрізняються методами обробки.

У статистичній практиці прийнято рахувати вибірку з об'ємом $> 30$ великий. Ця межа для $n$ представляється умовною, оскільки в різних задачах можуть бути свої критерії.

Будь-яку функцію випадкової вибірки в математичній статистиці називають вибірковою характеристикою. Розподіл цієї випадкової величини називають вибірковим розподілом.

Вибірковий розподіл однозначно визначається сумісним розподілом випадкових величин $X_1, X_2, ..., X_n$ тобто розподілом випадкової вибірки. Статистичним розподілом вибірки називають перелік спостережуваних значень $x_i$ і відповідних їм частот $n_i$ або відносних частот $\omega_i$.

Пояснимо це поняття. Хай з генеральної сукупності витягнута вибірка, причому значення $x_1$ спостерігалося $n_1$ раз $x_2$ - $n_2$ раз ... $x_k$ - $n_k$ раз, тоді $n_1 + n_2 + ... + n_k = n$ — об'єм вибірки. В цьому випадку спостережувані значення $x_1, x_2, ..., x_n$ іноді називають варіантами, а числа наглядів $n_1, n_2, ..., n_k$, називають частотами, їх відносини до об'єму вибірки $\omega_1 = \dfrac{n_1}{n}$, $\omega_2 = \dfrac{n_2}{n}$ , ...,

$\omega_k = \dfrac{n_k}{n}$ — відносними частотами.

Відмітимо, що сума відносних частот рівна одиниці $\omega_1 + \omega_2 + ... + \omega_k = 1$.

**Приклад 1.2** Перейти від частот до відносних частот в наступному розподілі виботки з об'ємом $n = 20$

Табл. 1.1

| $x_i$ | 2 | 6 | 12 |
|---|---|---|---|
| $n_i$ | 3 | 10 | 7 |

**Розв'язок задачі.** Введемо відносні частоти

$$\omega_1 = \frac{3}{20} = 0.15; \quad \omega_1 = \frac{10}{20} = 0.50; \quad \omega_3 = \frac{7}{20} = 0.35. \tag{1.1}$$

Отримаємо наступний розподіл виборки

Табл. 1.2

| $x_i$ | 2 | 6 | 12 |
|---|---|---|---|
| $\omega_i$ | 0.15 | 0.50 | 0.35 |

Нехай $N$ — об'єм генеральної сукупності, $N_i$ — число елементів генеральної сукупності із значення ознаки $x_i$. Позначимо через $K$ і $k$ — число елементів відповідно генеральною і вибірковою сукупністями наділеними даною властивістю. Середні арифметичні розподіли ознакив генеральній і вибірковій сукупностях називаються генеральною и вибірковою середніми і позначаються $\overline{x}_{\overline{a}}$ і $\overline{x}_{\hat{a}}$ відповідно. Дисперсію цих розподілів називають генеральною і вибірковою дисперсіями і позначаються ${\sigma_{\overline{a}}}^2$ і ${\sigma_{\hat{a}}}^2$ відповідно.

Відношення числа елементів генеральної і вибіркової сукупністей, наділених певнми ознаками, до їх об'мів називаються генеральною и вибірковою частинами і позначаються і ${\omega_{\overline{a}}}^2$ і ${\omega_{\hat{a}}}^2$ відповідно.

Указані характеристики визначаються формулами, які приведені в Таблиці

Табл. 1.3

| Сукупності / Характер. | Генеральна сукупність | Вибірка |
|---|---|---|
| Середня | $\overline{X}_{\overline{a}} = \dfrac{\sum\limits_{i=1}^{K} x_i N_i}{N}$ | $\overline{X}_{\hat{a}} = \dfrac{\sum\limits_{i=1}^{k} x_i n_i}{n}$ |
| Дисперсія | ${\sigma_{\overline{a}}}^2 = \dfrac{\sum\limits_{i=1}^{K} (x_i - \overline{X}_{\overline{a}})^2 N_i}{N}$ | ${\sigma_{\hat{a}}}^2 = \dfrac{\sum\limits_{i=1}^{K} (x_i - \overline{X}_{\hat{a}})^2 n_i}{N}$ |
| Доля | $\omega_{\overline{a}} = \dfrac{K}{N}$ | $\omega_{\hat{a}} = \dfrac{k}{n}$ |

## 1.3. Варіаційні і статистичні ряди та їх графічне зображення

Перш ніж перейти до аналізу одержаних в результаті експерименту статистичних даних, звичайно проводять їх попередню обробку, яка включає впорядкування, угрупування і графічне представлення статистичних даних.

Операція розташування статистичних даних по неубуванню називається їх ранжируванням. Одержана таким чином послідовність чисел $x_{(1)}, x_{(2)}, ..., x_{(n)}$, де $x_{(1)} \leq x_{(2)} \leq ... \leq x_{(n)}$ є варіаційний ряд вибірки.

**Означення 1.3** Варіаційним рядом називається послідовність елементів вибірки, розташованих в неубуваючому поряду. Однакові елементи повторюються. Запис варіаційного ряду має такий вигляд: $x_{(1)}, x_{(2)}, ..., x_{(n)}$.

Елементи варіаційного ряду називаються упорядковими статистиками. Мінімальний і максимальний елементи вибірки называются крайніми або екстремальними елементами варіаційного ряду: $x_{min} = x_{(1)}$, $x_{max} = x_{(n)}$.

**Приклад 1.3** В результаті п'яти повторних незалежних спостережень, наприклад, вимірювання тиску в газовому балоні, одержали наступні результати:

$$x_1 = 10.4; \quad x_2 = 9.5; \quad x_3 = 10.7; \quad x_4 = 9.3; \quad x_5 = 10.1;$$

Для представленої вибірки варіаційний ряд має вигляд:

$$x_{(1)} = 9.3; \quad x_{(2)} = 9.5; \quad x_{(3)} = 10.1; \quad x_{(4)} = 10.4; \quad x_{(5)} = 10.7;$$

**Приклад 1.4** В результаті тестування група з 24 чоловік набрала наступні бали:

$$4; 0; 3; 4; 1; 0; 3; 1; 0; 4; 0; 0; 3; 1; 0; 1; 1; 3; 2; 3; 1; 2; 1; 2.$$

Побудувати варіаційний ряд.

**Рох'язок задачі.** Ранжируємо початковий ряд:

$$0; 0; 0; 0; 0; 0; 1; 1; 1; 1; 1; 1; 1; 2; 2; 2; 3; 3; 3; 3; 3; 4; 4; 4.$$

Одержали варіаційний ряд.

**Означення 1.4** Статистичним рядом вибірки або просто статистичним рядом називається послідовність різних елементів вибірки $x_1, x_2, ..., x_k$, розташованих в зростаючому порядку з вказівкою частот $n_1, n_2, ..., n_k$, з якими ці елементи містяться у вибірці.

Статистичний ряд звичайно записують в виді таблиці.

Табл. 1.4

| $x_i$ | $x_1$ | $x_2$ | ... | $x_k$ |
|---|---|---|---|---|
| $n_i$ | $n_1$ | $n_2$ | ... | $n_k$ |

Для варіаційного ряду з попереднього прикладу статистичним рядом буде таблиця:

Табл. 1.5

| $x_i$ | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| $n_i$ | 6 | 7 | 3 | 5 | 3 |

Розрізняють дискретні і інтервальні статистичні ряди. Статистичний ряд називається дискретним, якщо він представляет собою вибірку значень дискретної випадкової величини. Статистичний ряд називається інтервальним (неперервним), якщо він є вибіркою неперервної випадкової величини.

Інтервальний статистичний ряд є совокупность проміжків $\Delta_1, \Delta_2, ..., \Delta_k$ і відповідних їм частот $n_1, n_2, ..., n_k$. Під частотами тут розуміється сума чисел спостережень, що потрапили в даний напівінтервал. Вважається, що каждый проміжок (напівінтервал) містить свою ліву межу і лише останній проміжок і свою праву межу. Якщо вибіркове значення (варіанта) знаходиться на межі, то його приєднують до правого інтервалу. При такій угоді кожне вибіркове значення міститиметься в одному і лише одному проміжку $\Delta_i$.

У результаті інтервальний статистичний ряд можна представити следуючою таблицею.

Табл. 1.6

| Інтервал $\Delta_i$ | $[x_1, x_2)$ | $[x_2, x_3)$ | ... | $[x_k, x_{k+1})$ |
|---|---|---|---|---|
| Частота $n_i$ | $n_1$ | $n_2$ | ... | $n_k$ |

Іноді у верхньому рядку таблиці указують не інтервал, а його середину $x_c$, а в нижньому рядку замість частоти $n_i$ записують відносну частоту $\omega_i = \dfrac{n_i}{n}$.

Дискретний статистичний ряд в деяких випадках представляется сукупністю варіантів (групованих спостережуваних значень) і не частот $n_i$, а відносних частот $\omega_i$. Відмітимо також, що довжини $h_i = h$ інтервалів частіше за все беруть однаковими. В цьому випадку інтервальний статистичний ряд називається статистичним наряду з рівновіддаленими варіантами. Тоді

довжина інтервалу $h$ буде рівною: $h = x_2 - x_1 = x_3 - x_2 = ... = x_{k+1} - x_k$.

**Приклад 1.5** В результаті тестування група студентів набрала наступні бали: $4; 2; 2; 0; 1; 3; 5; 2; 1; 5; 4; 4; 3; 0; 1$. На основі представленої вибірки скласти: a) варіаційний ряд; b) статистичний ряд.

**Рох'язок задачі.** a) Проведемо ранжирування спостережуваних значень: $0; 0; 1; 1; 1; 2; 2; 2; 3; 3; 4; 4; 4; 5; 5$. Одержали варіаційний ряд даної вибірки.

b) Підрахуємо частоти $n_i$ і відносні частоти $\omega_i$ варіантів $x_1 = 0$, $x_2 = 1$, $x_3 = 2$, $x_4 = 3$, $x_5 = 4$, $x_6 = 6$. Вони будуть рівними таким значенням: $n_1 = 2$, $n_2 = 3$, $n_3 = 3$, $n_4 = 2$, $n_5 = 3$, $n_6 = 2$ і $\omega_1 = \frac{2}{15}$, $\omega_2 = \frac{3}{15}$, $\omega_3 = \frac{3}{15}$, $\omega_4 = \frac{2}{15}$, $\omega_5 = \frac{3}{15}$, $\omega_6 = \frac{2}{15}$. У результаті одержимо наступний статистичний розподіл вибірки або так званий дискретний статистичний ряд:

Табл. 1.7

| $x_i$ | 0 | 1 | 3 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| $n_i$ | 2 | 3 | 3 | 2 | 3 | 1 |

Табл. 1.8

| $x_i$ | 0 | 1 | 3 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| $\omega_i$ | $\frac{2}{15}$ | $\frac{3}{15}$ | $\frac{3}{15}$ | $\frac{2}{15}$ | $\frac{3}{15}$ | $\frac{2}{15}$ |

де $\sum_{i=1}^{6} n_i = 15$, $\sum_{i=1}^{6} \omega_i = 1$ — умови нормування.

Для наочності уявлення використовують графічні изображения статистичних рядів у вигляді полігону, гістограми і кумуляты.

Полігон служить, як правило, для зображення дискретних статистических рядів. Полігон можна будувати як для частот $n_i$, так і для відносних частот $\omega_i$.

Полігон частот є ламаною, сполучаючою точки площини з координатами $(x_1, n_1), (x_2, n_2), ..., (x_k, n_k)$. Полігін відносних частот є ламаною, яка з'єднує точки площини з координатами $(x_1, \omega_1), (x_2, \omega_2), ..., (x_k, \omega_k)$.

Для даних з таблиці 5.5 побудуємо полігон частот (див. перший рис. 1.1). Для побудови в даному прикладі полігону відносних частот знайдемо $\omega_i : \omega_1 = \frac{6}{24}$, $\omega_2 = \frac{7}{24}$, $\omega_3 = \frac{3}{24}$, $\omega_4 = \frac{5}{24}$, $\omega_5 = \frac{3}{24}$. Тоді полигін відносних частот матиме вигляд зображений на другому рис. 1.1. Для інтервального (безперервного)

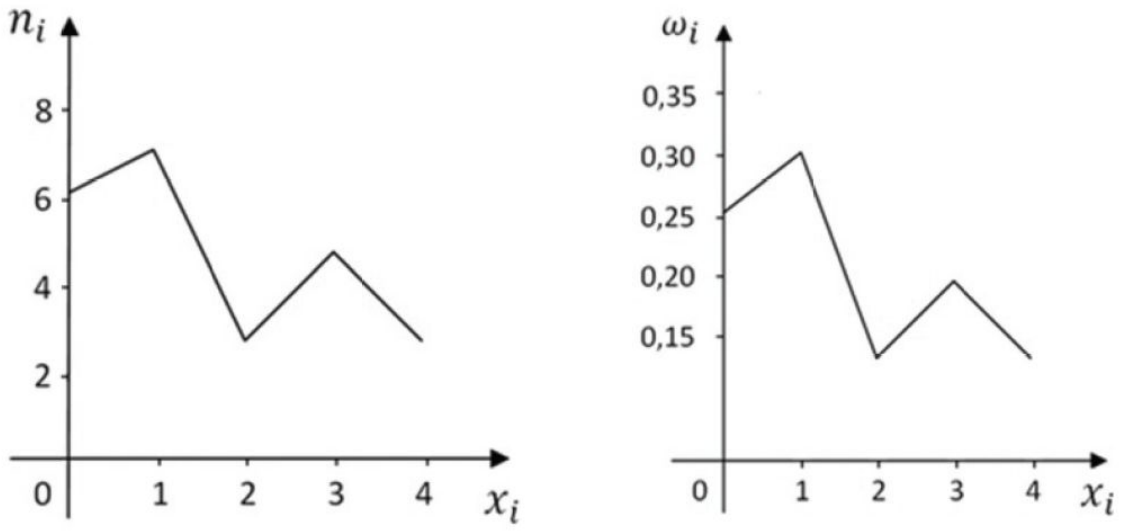


Рис. 1.1 — Полигіни частот та відносних частот

статистичного ряду можна побудувати полігон частот, узявши середини інтервалів як значень $x_1, x_2, ..., x_k$.

Інтервальні статистичні ряди частіше изображают у вигляді гістограми.

Гістограми частот служать тільки для представлення інтервальних статистичних результатів і мають вигляд ступінчатої структури, яка складається з прямокутників з основою $h$ для інтервалів, і висотами рівними відношенню $\frac{n_i}{n}$ що є щільність відносної частоти. Очевидно, що площа гістограми частот рівна об'єму вибірки, а площа гістограми відносних частот рівна одиниці.

Гістограми відносних частот є аналогом диференціальної функції розподілу або щільності розподілу $p(x)$ випадкової величини.

**Приклад 1.6** Для інтервального статистичного ряду маємо:

Табл. 1.9

| Інтервали | [150;156) | [156;162) | [162;168) | [168;174) | [174;180) | [174;180] |
|---|---|---|---|---|---|---|
| Частоти | 4 | 5 | 6 | 7 | 5 | 3 |

Треба побудувати гістограму відносних частот.

**Роз'язок задачі.** Обчислимо об'єм вибірки і відносні частоти

$$n = \sum_{i=1}^{6} n_i = 4 + 5 + 6 + 7 + 5 + 3 = 30;$$

$$\omega_1 = \frac{4}{30} = 0.13; \quad \omega_2 = \frac{4}{30} = 0.17; \quad \omega_3 = \frac{6}{30} = 0.20;$$
$$\omega_4 = \frac{7}{30} = 0.23; \quad \omega_5 = \frac{5}{30} = 0.17; \quad \omega_6 = \frac{3}{30} = 0.10.$$

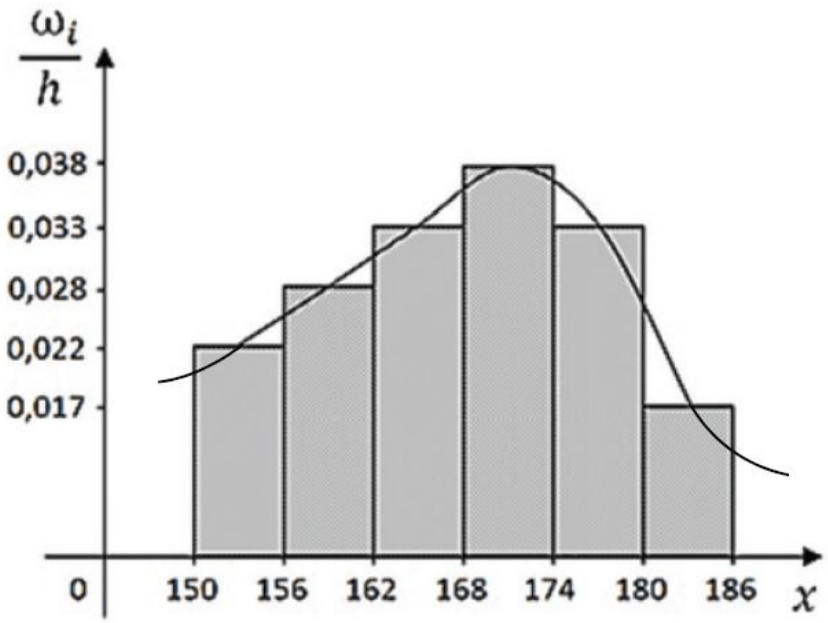


Рис. 1.2 — Гістограма відносних частот

З'єднавши середини верхніх підстав прямокутників відрізками прямої, одержимо полігон відносних частот того ж розподілу.

Як наголошувалося раніше, гістограма частот являється статистичним аналогом щільності розподілу імовірності $p(x)$ випадкової величини. Графік функції $p(x)$ приведений на рис. 1.2.

Кумулятивна крива або кумулята — це крива накопичених частот або накопичених відносних частот.

Якщо статистичний ряд дискретний, то кумулята є ламаною лінією, відрізки якої сполучають точки з координатами $x_i, {n_i}^{\text{нак}}$ або $x_i, {\omega_i}^{\text{нак}}$, $i = 1, 2, ..., n$.

Для інтервального варіаційного ряду така ламана починається з точки абсциса якої дорівнає началу першого інтервалу, а ордината — накопленій частоті або відносній частоті, рівній нулю. Інші точки цієї ламаної відповідають кінцям інтервалів.

Накопиченою частотою ${n_i}^{\text{нак}}$ називають число варіантів із значенням, менших $x_i$. Накопичена частота визначається послідовним підсумовуванням частот, тобто ${n_i}^{\text{нак}} = \sum_{k<i} = n_k$.

Аналогічно визначається накопичена відносна частота ${\omega_i}^{\text{нак}} = \dfrac{{n_i}^{\text{нак}}}{n_i}$.

**Приклад 1.7** Для статистичного ряду, заданого таблицею, необхідно побудувати кумулятивну криву:

Табл. 1.10

| $x_i$ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $n_i$ | 2 | 3 | 6 | 8 | 22 | 9 |

**Роз'язок задачі.** Задано дискретний статистичний ряд. Складемо таблицю, що містить рядки накопичених частот $n_i^{\text{нак}}$ і накопичених відносних частот $\omega_i^{\text{нак}}$, попередньо знайшовши відносні частоти. Тут $n = \sum_{i=1}^{6} n_i = 50$. Зведемо все в таблицю

Табл. 1.11

| $x_i$ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $n_i$ | 2 | 3 | 6 | 8 | 22 | 9 |
| $n_i^{\text{нак}}$ | 2 | 5 | 11 | 19 | 41 | 50 |
| $\omega_i$ | 0.04 | 0.06 | 0.12 | 0.16 | 0.44 | 0.18 |
| $\omega_i^{\text{нак}}$ | 0.04 | 0.10 | 0.22 | 0.38 | 0.82 | 1.00 |

Будуємо кумулятивні криві накопичених частот і накопичених відносних частот (див. рис. 1.3):

Як видно з малюнка кумулятивні криві $n_i^{\text{нак}}$ і $\omega_i^{\text{нак}}$ можуть відрізнятися між собою тільки масштабом, тому достатньо будувати одну кумулятивну криву.

## 1.4. Емпірична функція розподілу та її властивості

Нехай є статистичний розподіл частот деякої ознаки $X$. Позначимо через $n$ загальну кількість спостережень, тобто об'єм вибірки; $n_x$ — кількість спостережень, при яких спостерігались ознаки $X$ менше $x$. Тоді відносна частота (або частість) події $X < x$ дорівнює $\frac{n_x}{n}$. Якщо $x$ змінюється, то може змінюватись відносна частота, тобто $\frac{n_x}{n}$ є функція від $x$. Ця функція знаходиться емпіричним (дослідним) шляхом, тому її називають емпіричною.

**Означення 1.5** Емпіричною функцією розподілу (або функцією розподілу вибірки) називають функцію $F^*(x)$, яка визначає для кожного значення $x$ частість події $X < x$.

Математично це означення має вигляд

$$F^*(x) = \frac{n_x}{n},$$

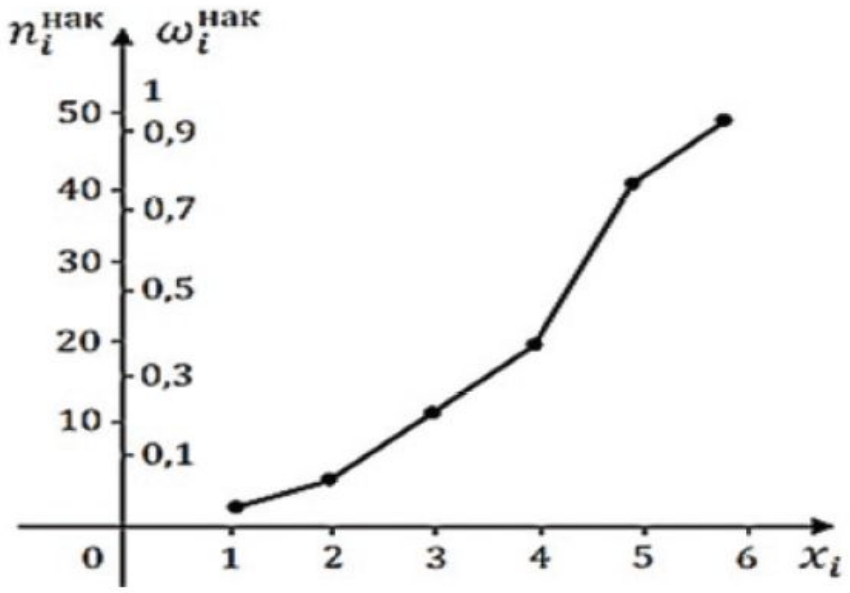


Рис. 1.3 — Кумулята накопичених частот та накопичених відносних частот

де $n_x$ — кількість варіант, які менше від $x$, $n$ — об'єм вибірки.

Таким чином, щоб знайти, наприклад, $F^*(x_3)$, треба кількість варіант, що менше $x_3$, поділити на об'єм вибірки, тобто

$$F^*(x_3) = \frac{n_1 + n_2}{n}.$$

Інтегральну функцію розподілу $F(x)$ генеральної сукупності у математичній статистиці називають теоретичною функцією розподілу. Вона відрізняється від емпіричної функції розподілу $F^*(x)$ тим, що визначає імовірність події $X < x$, а не частість цієї події.

З теореми Бернуллі випливає, що частість

$$F^*(x) = \frac{n_x}{n} \text{ події } X < x$$

прямує до імовірності

$$F(x) = P(X < x)$$

цієї події. Тому $F(x)$ та $F^*(x)$ мало відрізняються одна від одної.

Доцільно використовувати $F^*(x)$ для наближеного представлення функції розподілу $F(x)$ генеральної сукупності.

Емпірична функція розподілу $F^*(x)$ має такі властивості

1. $0 \leq F^*(x) \leq 1$;
2. $F^*(x)$ — зростаюча функція.

3.

$$F^*(x) = \begin{cases} 0 \text{ при } x \le x_1; \\ 1 \text{ при } x > x_m, \end{cases}$$

де $x_1$ — найменша варіанта, $x_m$ — найбільша варіанта.

**Приклад 1.8** Знайти емпіричну функцію розподілу за статистичним розподілом вибірки

Табл. 1.12

| $x_i$ | 2 | 6 | 10 |
|---|---|---|---|
| $n_i$ | 12 | 18 | 30 |

та побудувати її графік.

**Розв'язок задачі.** Об'єм цієї вибірки буде $n = 12 + 18 + 30 = 60$. Найменша варіанта дорівнює 2, тому $F^*(x) = 0$ для $x \le 2$ Найбільша варіанта дорівнює 10, тому $F^*(x) = 1$ для $x \ge 10$. Значення $x \le 6$, тобто $X = (x_1 = 2)$, спостерігалось 12 разів, тому $F^*(x) = \dfrac{12}{60} = 0.2$ при $2 < x \ge 6$.

Значення $X < 10$, тобто $X = (x_1 = 2)$ та $X = (x_2 = 6)$ спостерігались 12+18=30 разів, тому $F^*(x) = \dfrac{30}{60} = 0.5$ при $6 < x \le 10$. Тобто, простий статистичний розподіл частоти, що заданий Таблицею 1.12, замінюється згрупованим розподілом частоти (див. Таблицю 1.13).

Табл. 1.13

| Варіанта $x_i$ | Частота $n_i$ | Варіанта $x_i$ | Накоп.частота $F_i$ |
|---|---|---|---|
| $x \le 2$ | 0 | менше ніж 2 | 0 |
| $2 < x \le 6$ | 12 | менше ніж 6 | 12 |
| $6 < x \le 10$ | 18 | менше ніж 10 | 30 |
| $10 < x$ | 30 | більше ніж 10 | 60 |
| Разом | 60 | | |

Тут же побудований розподіл накопиченої частоти. Таким чином, одержали

емпіричну функцію розподілу вигляду

$$F^*(x) = \begin{cases} 0 \text{ при } x \leq 2; \\ 0.2 \text{ при } x < 2 \leq 6; \\ 0.5 \text{ при } 6 < 2 \leq 10; \\ 1 \text{ при } x > 10, \end{cases}$$

Графік цієї функції зображено на рис. 1.4. Цей графік можна розуміти як

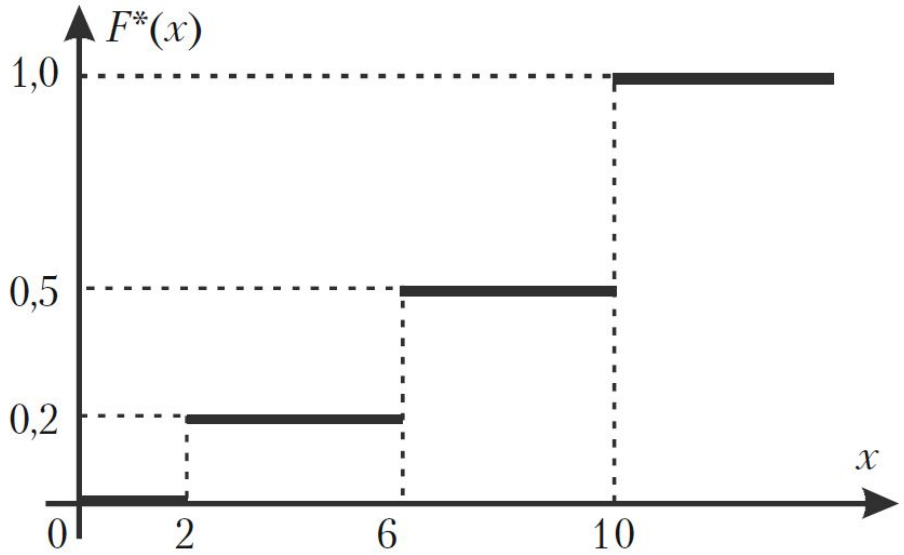


Рис. 1.4 —

наближений графік теоретичної функції розподілу $F(x)$.

Встановимо зв'язок між емпіричною функцією розподілу $F^*(x)$ і функцією накопичених частот $F_i$. Як слідує з Таблиці 1.12 емпірична функція розподілу $F^*(x)$ визначається як кусково-постійна функція, рівна значенню накопиченої відносної частоти

$$F^*(x) = \frac{n_x}{n} = \frac{F_x}{n},$$

на кожному класі інтервалів

$$J_i = (x_{imin}, x_{imax}), \quad 1 = 1, 2, ..., k.$$

## 1.5. Питання для самоперевірки

- Який об'єкт в математичній статистиці називається генеральною сукупністю?
- Яким чином будується вибіркова сукупність?

- Що таке варіаційним ряд?
- Що таке статистичний ряд вибірки і чим він відрізняється від варіаційного рядн?
- Як будується інтервальний статистичний ряд?
- Які графічні зображення статистичних рядів використовуються математичній статистиці?
- Що таке емпірична функція розподілу та які її головні властивості?
- Який існує зв'язок між емпіричною функцією розподілу і функцією накопичених частот?

# 2. Статистичні оцінки параметрів розподілу

## 2.1. Основні вимоги до статистичних оцінок

У багатьох випадках треба дослідити кількісну ознаку $X$ генеральної сукупності, використовуючи результати вибірки. Часто для цього достатньо знати наближені значення математичного сподівання $M(X)$, дисперсію $D(X)$, середньоквадратичне відхилення $\sigma(X)$, початкові або центральні моменти випадкової величини $X$.

Іноді з деяких міркувань вдається встановити закон розподілу $X$. Тоді треба вміти оцінювати параметри цього закону розподілу. Наприклад, відомо, що випадкова величина $X$ розподілена рівномірно; треба по даних вибірки наближено знайти відрізок, в якому знаходяться значення випадкової величини $X$.

Якщо $X$ розподілена у генеральній сукупності за нормальним законом, то її щільність імовірностей має вигляд

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

Необхідно оцінити (знайти наближені значення) параметра $a$, який дорівнює $M(X)$, та $\sigma$, який дорівнює $\sigma(X)$. Ці параметри повністю визначають нормальний розподіл $X$.

Якщо $X$ розподілена за законом Пуассона, то необхідно оцінювати лише один параметр $a$, яким цей розподіл визначається.

Дослідник має у своєму розпорядженні лише дані вибірки, одержані в результаті спостережень. Саме через ці дані і треба виразити потрібний

параметр випадкової величини $X$ генеральної сукупності.

**Означення 2.1** Статистичною оцінкою невідомого параметра випадкової величини $X$ генеральної сукупності (теоретичного розподілу $X$) називають функцію від випадкових величин (результатів вибірки), що спостерігаються.

Щоб статистичні оцінки давали найкращі наближення параметрів, вони повинні задовольняти певним вимогам. Розглянемо ці вимоги.

Нехай $\theta^*$ є статистична оцінка невідомого параметра $\theta$ теоретичного розподілу.

Припустимо, що за вибіркою об'єму $n$ знайдена оцінка $\theta_1^*$. При інших вибірках того ж об'єму одержимо деякі інші оцінки $\theta_2^*, \theta_3^*, ..., \theta_m^*$. Саме оцінку $\theta^*$ можна розглядати як випадкову величину, а числа $\theta_2^*, \theta_3^*, ..., \theta_m^*$ як її можливі значення.

Якщо числа $\theta_k^*(k = 1, 2, ..., m)$ будуть більші значення $\theta$, тоді оцінка $\theta^*$ дає наближене значення $\theta$ з надлишком. У цьому випадку математичне сподівання випадкової величини $\theta^*$ буде більше $\theta$, $M(\theta^*) > \theta$. Якщо ж $\theta^*$ дає оцінку $\theta$ з недостачею, тоді $M(\theta^*) < \theta$.

Таким чином, використання статистичної оцінки, математичне сподівання якої не дорівнює параметру $\theta$, приводить до систематичних (одного знака) похибок.

Вимога $M(\theta^*) = \theta$ застерігає від систематичних похибок.

**Означення 2.2** Статистичну оцінку $\theta^*$ параметра $\theta$ називають незсунутою, якщо $M(\theta^*) = \theta$. Оцінку $\theta^*$ називають зсунутою, якщо ця рівність не виконується.

Вимога про незсунутість оцінки $\theta^*$ є недостатньою тому, що можливі значення $\theta^*$ можуть бути сильно розсіяні від свого середнього значення, дисперсія $D(\theta^*)$ може бути великою. Тоді знайдена за даними однієї вибірки оцінка, наприклад, $\theta_k^*$ може набагато відрізнятись від середнього значення $\theta^*$, отже і від параметра $\theta$.

Якщо $D(\theta^*)$ буде малою, тоді можливість допустити велику помилку буде виключена. Тому до статистичної оцінки виникає вимога про її ефективність.

**Означення 2.3** Ефективною називають таку статистичну оцінку $\theta^*$ яка при заданому об'єму $n$ має найменшу можливу дисперсію.

При розгляданні вибірки великого об'єму $n \to \infty$ до статистичних оцінок пред'являють вимогу їх обгрунтованості.

**Означення 2.4** Обгрунтованою називають статистичну оцінку, яка при $n \to \infty$ прямує за імовірністю до оцінюваного параметра.

Наприклад, якщо дисперсія незсунутої оцінки при $n \to \infty$ прямує до нуля,

то оцінка буде і обгрунтованою.

## 2.2. Числові характеристики вибіркової сукупності

### Вибіркові характеристики

У доповнення до табличних та графічних методів представлення даних наступним найважливішим засобом обробки даних є обчислення їх числових характеристик. Найважливіші з них: середнє значення, дисперсія, середнє квадратичне відхилення (стандартне відхилення).

Ці характеристики можуть бути обчислені за даними, що знаходяться у вибірці або за даними, що входять у кінцеву генеральну сукупність.

Числові характеристики, обчислені по вибірці або ті, що використовуються для опису даних вибірки, називають статистиками.

Числові характеристики, обчислені по генеральній сукупності або ті, що використовуються для опису даних генеральної сукупності, називають параметрами.

По аналогії з математичним сподіванням, дисперсією та середнім квадратичним відхиленням дискретної випадкової величини визначають вибіркові характеристики, замінюючи при цьому імовірності $p_i$ частостями вибірки $\frac{n_i}{n}$.

Але у статистиці застосовують ще інші числові характеристики.

**Означення 2.5** Простою середньоарифметичною вибірки називають суму варіант вибірки, поділену на об'єм вибірки. Її позначають

$$\overline{x}_a = \frac{1}{n}\sum_{i=1}^{m} x_i,$$

де $x_i$, $i = 1, 2, ..., m$ — варіанти вибірки, $n$ — об'єм вибірки.

**Означення 2.6** Вибірковою середньою або зваженою середньоарифметичною називають середню арифметичну варіант вибірки з врахуванням їх частостей і позначають

$$\overline{x}_B = \frac{1}{n}\sum_{i=1}^{m} n_i x_i, \tag{2.1}$$

де $n$ — об'єм вибірки, $m$ — число різних варіант, $n_1, n_2, ..., n_m$ — частоти варіант ($n = n_1 + n_2 + ... + n_m$), $x_i$ — значення $i$-тої варіанти.

Вибіркова середня є аналогом математичного сподівання і використовується дуже часто. Вона може приймати різні числові значення при різних вибірках однакового об'єму.

Тому можна розглядати розподіли (теоретичний та емпіричний) вибіркової середньої та числові характеристики цього розподілу (цей розподіл називають вибірковим).

## Основні властивості вибіркової середньої

1. При множенні усіх варіант вибірки на однаковий множник вибіркова середня також множиться на цей множник

$$\frac{1}{n}\sum_{i=1}^{m} n_i(cx_i) = \frac{c}{n}\sum_{i=1}^{m} n_i x_i = c\overline{x}_B.$$

2. Якщо додати (відняти) до всіх варіант вибірки однакове число, то вибіркова середня зростає (зменшується) на це число

$$\frac{1}{n}\sum_{i=1}^{m} n_i(x_i + c) = \frac{1}{n}\sum_{i=1}^{m} n_i x_i + \frac{c}{n}\sum_{i=1}^{m} n_i = \overline{x}_B + c.$$

Ці властивості можна поєднати в одну формулу, яку називають формулою моментів

$$\mu = \frac{k}{n}\sum_{i=1}^{m} n_i \frac{x_i - c}{k} \tag{2.2}$$

і використовують у статистиці.

**Означення 2.7** Якщо ввести умовну варіанту

$$u_i = \frac{x_i - c}{k},$$

то формула моментів (5.5) приймає простий вигляд

$$\overline{x}_B = \frac{k}{n}\sum_{i=1}^{m} n_i u_i + c.$$

**Означення 2.8** Степеневою середньою вибірки називають таку середню, яку знаходять за формулою

$$\overline{x}_C = \frac{1}{n}\Big(\sum_{i=1}^{m} n_i {x_i}^{\alpha}\Big)^{\frac{1}{\alpha}}. \tag{2.3}$$

При $\alpha = 1$ одержимо формулу (9.1), тобто вибіркову середню.

При $\alpha = 2$ одержимо середньоквадратичну вибірки

$$\overline{x}_2 = \sqrt{\frac{1}{n}\sum_{i=1}^{m} n_i {x_i}^2}.$$

При $\alpha = -1$ одержуємо середню гармонічну

$$\overline{x}_{-1} = \sqrt{\frac{n}{\sum\limits_{i=1}^{m} \frac{n_i}{x_i}}}.$$

Середню гармонічну застосовують у тому випадку, коли шуканий показник є величина, що обернена середньому значенню ознаки.

При $\alpha = 0$ вираз (2.3) буде невизначеним. Застосовуючи логарифмування та правило Лопіталя розкриття невизначеності, одержимо середню геометричну

$$\overline{x}_g = ({x_1}^{n_1} \cdot {x_2}^{n_2} \cdot ... \cdot {x_m}^{n_m} \cdot)^{\frac{1}{n}}$$

Ця середня обчислюється лише при умові, що усі варіанти є додатні $x_I > 0$, $i = 1, 2, ..., m$.

Середня геометрична застосовується у статистиці для визначення темпу зростання при дослідженні змін ознаки з часом.

Обрання тієї чи іншої середньої для характеристики розподілу пов'язано з якісним аналізом цього розподілу.

Крім вказаних степеневих середніх, у статистиці застосовуються ще структурні середні, які не залежать від значень варіант, що розташовані на краях розподілу, а пов'язані з рядом частот.

До структурних середніх відносять моду та медіану. Нагадаємо, що модою називають значення варіанти, яка має найбільшу частоту.

**Означення 2.9** Вибірковою дисперсією $D_B$ називають середню квадратів відхилення варіант від вибіркової середньої з врахуванням відповідних частостей

$$D_B = \frac{1}{n}\sum_{i=1}^{m} n_i (x_i - \overline{x}_B)^2. \tag{2.4}$$

Обчислення вибіркової дисперсії спрощується, якщо її знаходити за формулою

$$D_B = \frac{1}{n}\sum_{i=1}^{m} n_i {x_i}^2 - \left(\frac{1}{n}\sum_{i=1}^{m} n_i x_i\right)^2 = \frac{1}{n}\sum_{i=1}^{m} n_i {x_i}^2 - (\overline{x}_B)^2. \tag{2.5}$$

**Означення 2.10** Вибірковим середньоквадратичним відхиленням (стандартом)

називають квадратний корінь із вибіркової дисперсії

$$\sigma_B = \sqrt{D_B}. \tag{2.6}$$

Вибіркова дисперсія дає занижені значення для дисперсії $D(X)$ генеральної сукупності, вона буде зсунутою оцінкою $D(X)$. Але математичне сподівання $D_B$ буде

$$M(D_B) = \frac{n-1}{n} D(X).$$

Тому вибіркову дисперсію доцільно виправити таким чином, щоб вона стала незсунутою оцінкою. Для цього достатньо $D_B$ помножити на дріб $\dfrac{n}{n-1}$.

Виправлену вибіркову дисперсію позначають

$$S^2 = \frac{n}{n-1} D_B = \frac{1}{n-1} \sum_{i=1}^{m} n_i (x_i - \overline{x}_B)^2. \tag{2.7}$$

Тоді виправленим середньоквадратичним відхиленням вибірки буде $S = \sqrt{S^2}$.

Із формул (2.4) та (2.7) випливає, що при досить великих $n$ (об'єм вибірки) вибіркова дисперсія $D_B$ та виправлена вибіркова дисперсія $S^2$ різняться дуже мало. Тому в практичних задачах виправлену дисперсію $S^2$ та виправлене середньоквадратичне відхилення вибірки $S$ використовують лише при об'ємі вибірки $n < 30$.

**Приклад 2.1** Вибіркова сукупність задана таблицею

Табл. 2.1

| $x_i$ | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| $n_i$ | 20 | 15 | 10 | 5 |

Знайти вибіркові характеристики.

**Розв'язок задачі.** У даному випадку об'єм вибірки дорівнює

$$n = 20 + 15 + 10 + 5 = 50.$$

За формулою (9.1) знаходимо вибіркову середню

$$\overline{x}_B = \frac{1}{50}(1 \cdot 20 + 2 \cdot 15 + 3 \cdot 10 + 4 \cdot 5) = \frac{100}{50} = 2.$$

За формулою (2.4) знаходимо вибіркову дисперсію

$$\overline{D}_B = \frac{1}{50}((1-2)^2 \cdot 20 + (2-2)^2 \cdot 15 + (3-2)^2 \cdot 10 + (4-2)^2 \cdot 5) = \frac{50}{50} = 1.$$

За формулою (2.6) знаходимо вибіркове середньоквадратичне відхилення (стандарт)

$$\sigma_B = \sqrt{1} = 1.$$

## Обчислення вибіркових характеристик методом добутків

Як правило, обчислення $\overline{x}_B$ та $D_B$ за формулами (9.1) та (2.4) або (2.5) проводиться з використанням комп'ютерної техніки. Часто розрахунки можна спростити, використовуючи метод добутків, в основі якого лежать рівновіддалені варіанти та наступна розрахункова таблиця

Табл. 2.2

| 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|
| $x_k$ | $n_k$ | $u_k$ | $n_k \cdot u_k$ | $n_k \cdot {u_k}^2$ | $n_k \cdot (u_k + 1)^2$ |

Дамо необхідні пояснення до цього методу

## Алгоритм методу добутків

1. У перший стовпчик таблиці записують рівновіддалені варіанти $x_k$ вибірки, розміщуючи їх у зростаючому порядку.

2. У другий стовпчик таблиці записують відповідні частоти $n_k$ варіант. Суму усіх елементів цього стовпчика (об'єм вибірки $n$) записують у останню клітинку цього стовпчика.

3. Третій стовпчик містить умовні варіанти $u_k$ вибірки. Для знаходження умовних варіант вибірки треба:
   — значення варіанти вибірки з найбільшою частотою $C$ обрати за умовний нуль. Це значення варіанти називається модою;
   — знайти різницю $h$ між будь-якими двома сусідніми варіантами;
   — обчислити умовні варіанти вибірки за формулою

$$u_k = \frac{x_k - C}{h}. \tag{2.8}$$

   Відмітимо, що умовні варіанти завжди будуть цілими числами.

4. У четвертий стовпчик записують добутки частот та відповідних умовних варіант $n_k \cdot u_k$. Суму елементів стовпчика записують в останню клітинку цього стовпчика.

5. Знаходять добутки частот та квадратів умовних варіант $n_k \cdot {u_k}^2$ і записують їх у п'ятий стовпчик. Суму елементів стовпчика $\sum_{i=k}^{m} n_k {u_k}^2$ записують в останню клітинку цього стовпчика.

6. Знаходять добутки частот та квадратів умовних варіант, збільшених на одиницю, $n_k(u_k+1)^2$ і записують їх у шостий контрольний стовпчик. Суму елементів стовпчика $\sum_{i=k}^{m} n_k(u_k+1)^2$ записують в останню клітинку цього стовпчика.

7. Перевіряють обчислення так: сума елементів шостого стовпчика повинна задовольняти тотожність

$$\sum_{i=k}^{m} n_k(u_k+1)^2 = \sum_{i=k}^{m} n_k \cdot u_k^2 + 2\sum_{i=k}^{m} n_k \cdot u_k + n. \tag{2.9}$$

8. Обчислюють умовні моменти за формулами

$${M_1}^* = \frac{1}{n}\sum_{i=k}^{m} n_k \cdot u_k; \quad {M_2}^* = \frac{1}{n}\sum_{i=k}^{m} n_k \cdot {u_k}^2. \tag{2.10}$$

9. Обчислюють вибіркові середню та дисперсію за формулами

$$\overline{x}_B = {M_1}^* \cdot h + C; \quad D_B = \left({M_2}^* - ({M_1}^*)^2\right) \cdot h^2. \tag{2.11}$$

## 2.3. Статистичні моменти розподілу

Визначимо аналогічно початковому та центральному моментам розподілу із теорії імовірностей деякі числові характеристики вибірки.

**Означення 2.11** Моментом порядку k називають середнє значення $k$-го степеня різниці $x_i - C$.

При $C = 0$ одержимо початковий момент порядку $k$ вибірки

$${\nu^*}_k = \frac{1}{n}\sum_{i=1}^{m}(x_i - \overline{x}_B)^k.$$

Моменти порядку $k$ та умовні моменти ${M_1}^*$ та ${M_2}^*$ які обчислюють за формулами 2.10, часто використовують у статистиці.

У випадку згрупованої вибірки припускається, що всяке значення варіанти, що потрапило в даний клас інтервалів, дорівнює середньому значенню варіанти в цьому класі.

1. Вибіркове середнє обчислюється по формулі

$$x_B = \frac{1}{n}\sum_{i=1}^{k} n_i \widetilde{x}_i \tag{2.12}$$

де $n$ — об'єм вибірки, $\widetilde{x}$ — середнє значення варіанти на класі, і $n_i$ — частота $i$-того класу інтервалів, $k$ — кількість класів.

2. Для дисперсії маємо формулу

$$D_B = \frac{1}{n}\sum_{i=1}^{k} n_i(\widetilde{x}_i - \overline{x}_B)^2 \tag{2.13}$$

і формулу для обчислень

$$D_B = \frac{1}{n}\sum_{i=1}^{k} n_i \widetilde{x}_i{}^2 - (\overline{x}_B)^2. \tag{2.14}$$

Для виправленої дисперсії маємо формулу

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{k} n_i(\widetilde{x}_i - \overline{x}_B)^2$$

і формулу для обчислень

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{k} n_i \widetilde{x}_i{}^2 - \frac{1}{n(n-1)}\Big(\sum_{i=1}^{k} n_i \widetilde{x}_i\Big)^2$$

або

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{k} n_i \widetilde{x}_i{}^2 - \frac{n}{n-1}(\widetilde{x}_B)^2.$$

**Приклад 2.2** Обчислити числові характеристики вибірки згрупованого розподілу частот середньомісячної платні співробітників фірми $N$ (див. Таблицю 2.3).

Табл. 2.3

| Платня (класи інтервалів) | Частота $n_i$ |
|---|---|
| 280–290 | 1 |
| 290–300 | 10 |
| 300–310 | 14 |
| 310–320 | 14 |
| 320–330 | 25 |

| | |
|---|---|
| 330–340 | 16 |
| 340–350 | 7 |
| 350–360 | 4 |
| 360–370 | 7 |
| 370–380 | 0 |
| 380–390 | 2 |
| Разом: | 100 |

Табл. 2.4

| Платня (середнє в класі) $\widetilde{x}_i$ | Частота $n_i$ |
|---|---|
| 285 | 1 |
| 295 | 10 |
| 305 | 14 |
| 315 | 14 |
| 325 | 25 |
| 335 | 16 |
| 345 | 7 |
| 355 | 4 |
| 365 | 7 |
| 375 | 0 |
| 385 | 2 |
| Разом: | 100 |

Перейдемо від Таблиці 2.3 до Таблиці 2.4, замінивши класи інтервалів на середні значення варіант в класі. Користуючись даними з Таблиці 2.4, за формулою (2.12) отримуємо

$$\overline{x}_B = \frac{1}{100}(285 \cdot 1 + 295 \cdot 10 + ... + 385 \cdot 2) = 325.6 \text{ грн.}$$

Порівнюючи отримане значення вибіркової середньої $\overline{x}_B = 325.6$ грн. із точним значенням $\overline{x}_B = 324.3$ грн. бачимо що вони відрізняються незначним чином. Помилка виникла за рахунок округлення всіх варіант у класі до середнього значення. Проте, як показує практика, помилка, що вноситься при цьому, незначна.

Далі, за формулами (2.13), (2.14) отримаємо

$$D_B = 439.64 \text{ грн}^2.$$

Для середньоквадратичного відхилення маємо

$$\sigma_B = \sqrt{439.64} = 20.07 \text{ грн.}$$

Згідно із запровадженої вище термінології $\overline{x}_B$, $D_B$, $\sigma_B$, $S^2$ є статистиками.

Як видно з формул (2.13) і (2.14) дисперсія є мірою розсіювання варіант у вибірці навколо їхнього середнього значення. Пояснимо це на наступному прикладі.

**Приклад 2.3** Обстежені по 65 випадків виплати страхових сум двома страховими компаніями $N$ і $R$ за деякий період часу. За одиницю виплати прийнята деяка стандартна сума. Виплата може приймати будь-які значення від 0 до 4. Знак «мінус» перед числом означає, що виплату робить страхова компанія, а знак «плюс» — що компанія отримує страховий внесок. Треба підрахувати числові характеристики цих вибірок.

Розподіл частот виплат страхових сум обох компаній наведено у Таблицях 2.5 та 2.6.

Табл. 2.5

| Страховий внесок | Частота |
|---|---|
| -4 | 1 |
| -3 | 4 |
| -2 | 9 |
| -1 | 16 |
| 0 | 25 |
| 1 | 16 |
| 2 | 9 |
| 3 | 4 |
| 4 | 1 |
| Разом: | 65 |

Табл. 2.6

| Страховий внесок | Частота |
|---|---|
| -2 | 1 |
| -1.5 | 5 |
| -1 | 8 |
| -0.5 | 17 |
| 0 | 23 |
| 0.5 | 17 |
| 1 | 8 |

| | |
|---|---|
| 1.5 | 5 |
| 2 | 1 |
| Разом: | 65 |

Підрахуємо вибіркове середнє $\overline{x}_B(N)$ і $\overline{x}_B(R)$ для обох компаній

$$\overline{x}_B(N) = \frac{1}{65}(-4\cdot 1 - 3\cdot 4 - 2\cdot 9 - 1\cdot 16 + 0.25 + 1\cdot 16 + 2\cdot 9 + 3\cdot 4 + 4\cdot 1) = 0,$$

$$\overline{x}_B(R) = \frac{1}{65}(-4\cdot 1 - 3\cdot 5 - 2\cdot 8 - 1\cdot 17 + 0.23 + 1\cdot 17 + 2\cdot 8 + 3\cdot 4 + 4\cdot 1) = 0.$$

Підрахуємо вибіркові дисперсії $D_B(N)$ і $D_B(R)$

$$D_B(N) = \frac{1}{65}(1\cdot 4^2 + 4\cdot 3^2 + 9\cdot 2^2 + 16\cdot 1^2) = \frac{208}{65} = 3.2,$$

$$D_B(R) = \frac{1}{65}(1\cdot 4^2 + 5\cdot 3^2 + 4\cdot 8 + 1\cdot 17) = \frac{55}{65} = 0.85.$$

Полігон частот обох розподілів зображень на рис.2.1. З рис. 2.1 видно, що чим менша дисперсія, тим у більш вузькому інтервалі дані вибірки групуються навколо свого середнього значення (у нашому випадку $\overline{x}_B = 0$).

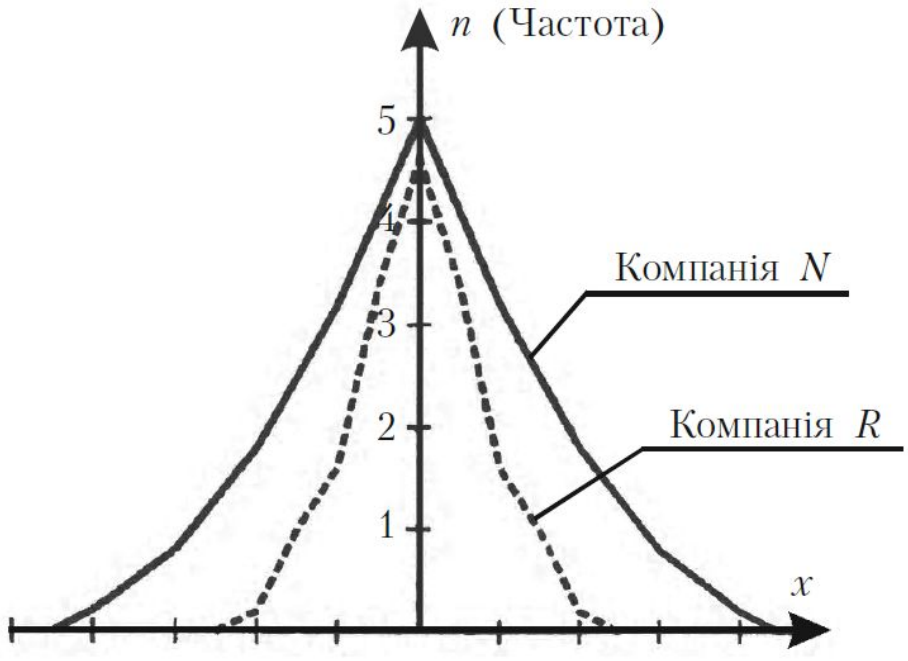


Рис. 2.1 — Полігон частот розподілів для компаній $N$ та $R$

## 2.4. Питання для самоперевірки

- Що називається статистичною оцінкою невідомого параметра випадкової величини?

- Яку статистичну оцінке невідомого параметра називають зсунутою?
- Яку статистичну оцінке невідомого параметра називають ефективною?
- Яку статистичну оцінке невідомого параметра називають обгрунтованою?
- Які числові характеристики, що використовуються для опису даних вибірки називають статистиками?
- Які основні властивості вибіркової середньої
- Яким чином у математичній мтатистиці визначають моду та медіану?
- Що називається у математичній мтатистиці виправленим середньоквадратичним відхиленням вибірки?
- Який алгоритм обчислення вибіркових характеристик методом добутків?
- Яким чином обчислюється моментом $k$—го порядку?
- Який алгоритм обчислення числових характеристик вибірки згрупованого розподілу частот?

# 3. Точкові та інтервальні оцінки

## 3.1. Загальні поняття

**Означення 3.1** Точковими оцінками параметрів розподілу генеральної сукупності називають такі оцінки, які визначаються одним числом.

Наприклад, вибіркова середня $x_B$, вибіркова дисперсія та вибіркове середньоквадратичне $\sigma_B$ — точкові оцінки відповідних числових характеристик генеральної сукупності.

Точкові оцінки параметрів розподілу є випадковими величинами, їх можна вважати первинними результатами обробки вибірки тому, що невідомо, з якою точністю кожна з них оцінює відповідну числову характеристику генеральної сукупності.

Якщо об'єм вибірки досить великий, то точкові оцінки задовольняють практичні потреби точності.

Якщо об'єм вибірки малий, то точкові оцінки можуть давати значні похибки, тому питання точності оцінок у цьому випадку дуже важливе і використовують інтервальні оцінки.

**Означення 3.2** Інтервальною називають оцінку, яка визначається двома числами — кінцями інтервалу.

Інтервальні оцінки дозволяють встановити точність та надійність оцінок.

Нехай знайдена за даними вибірки статистична оцінка $\theta^*$ буде оцінкою невідомого параметра $\theta$.

Ясно, що $\theta^*$ тим точніше визначає $\theta$ чим менше абсолютна величина різниці $\theta - \theta^*$.

Іншими словами, якщо $\delta > 0$ і $|\theta - \theta^*| < \delta$ тоді меншому $\delta$ відповідає більш точна оцінка. Тому число $\delta$ характеризує точність оцінки.

Але статистичні методи не дозволяють категорично стверджувати, що оцінка $\theta^*$ задовольняє нерівність $|\theta - \theta^*| < \delta$.

Таке твердження можна зробити лише з імовірністю $\gamma$.

**Означення 3.3** Надійністю (довірчою імовірністю) оцінки параметра $\theta$ за $\theta^*$ називають імовірність

$$\gamma = P(|\theta - \theta^*| < \delta), \tag{3.1}$$

з якою виконується нерівність $|\theta - \theta^*| < \delta$.

Найчастіше число $\gamma$ задається наперед і, залежно від обставин, воно дорівнює 0.95 або 0.99 або 0.999.

Формулу (3.1) можна записати у вигляді

$$P(\theta^* - \delta < \theta < \theta^* + \delta) = \gamma. \tag{3.2}$$

З цієї рівності випливає, що інтервал $(\theta^* - \delta, \theta^* + \delta)$ містить невідомий параметр $\theta$ генеральної сукупності.

**Означення 3.4** Інтервал $(\theta^* - \delta, \theta^* + \delta)$ називають довірчим, якщо він покриває невідомий параметр $\theta$ із заданою надійністю $\gamma$.

Кінці довірчого інтервалу є випадковими величинами.

## 3.2. Довірчий інтервал для оцінки математичного сподівання нормального розподілу

Нехай кількісна ознака $X$ генеральної сукупності розподілена за нормальним законом, середньоквадратичне відхилення $\sigma$ відомо. Треба знайти довірчий інтервал, що покриває математичне сподівання $a$ генеральної сукупності із заданою надійністю $\gamma$.

Згідно із властивістю нормально розподіленої випадкової величини $X$ маємо

$$\begin{aligned} &P(|\theta - \theta^*| < \delta) = P(\theta^* - \delta < \theta < \theta^* + \delta) = \\ &\Phi\Big(\frac{a + \delta - a}{\sigma}\Big) - \Phi\Big(\frac{a - \delta - a}{\sigma}\Big) = \Phi\Big(\frac{\delta}{\sigma}\Big) - \Phi\Big(-\frac{\delta}{\sigma}\Big). \end{aligned} \tag{3.3}$$

Оскільки інтегральна функція Лапласа $\Phi$ є непарна, то одержимо

$$P(|X-a|<\delta)=2\Phi\Big(\frac{\delta}{\sigma}\Big).$$

Але $\overline{x}_B$ випадкова величина, $M(\overline{x}_B)=a$, $\sigma(\overline{x}_B)=\dfrac{\sigma}{\sqrt{n}}$, тому при заміні $X$ на $\overline{x}_B$, $\sigma$ на $\dfrac{\sigma}{\sqrt{n}}$ одержимо

$$P(|X-a|<\delta)=2\Phi\Big(\frac{\delta\cdot\sqrt{n}}{\sigma}\Big)=2\Phi(t), \tag{3.4}$$

де

$$t=\frac{\delta\cdot\sqrt{n}}{\sigma},\quad \delta=\frac{t\cdot\sigma}{\sqrt{n}}.$$

Використовуючи формули (3.4) та (3.2), одержимо

$$P\Big(\overline{x}_B-\frac{t\cdot\sigma}{\sqrt{n}}<a<\overline{x}_B+\frac{t\cdot\sigma}{\sqrt{n}}\Big)=2\Phi(t)=\gamma, \tag{3.5}$$

тобто з надійністю $\gamma$ довірчий інтервал

$$\Big(\overline{x}_B-\frac{t\cdot\sigma}{\sqrt{n}},\overline{x}_B+\frac{t\cdot\sigma}{\sqrt{n}}\Big)$$

покриває невідомий параметр $a$. Точність оцінки буде

$$\delta=\frac{t\cdot\sigma}{\sqrt{n}}. \tag{3.6}$$

Число $t$ визначається із рівності

$$2\Phi(t)=\gamma,\quad \Phi(t)=\frac{\gamma}{2} \tag{3.7}$$

з використанням таблиці значень інтегральної функції Лапласа.

З формули (3.6) випливає, що при зростанні об'єму вибірки $n$ число $\delta$ зменшується, а це означає, що точність оцінки збільшується. Коли надійність $\gamma$ збільшується, функція $\Phi(t)$ зростає і згідно з її властивістю зростає $t$ і, як наслідок, зростає $\delta$. Отже, збільшення надійності оцінки зменшує її точність.

**Приклад 3.1** Випадкова величина розподілена за нормальним законом з параметром $\sigma=2$. Зроблена вибірка об'єму $n=25$. З надійністю $\gamma=0.95$ знайти довірчий інтервал невідомого параметра $a$ цього розподілу.

**Розв'язок задачі.** Із рівності

$$\Phi(t)=\frac{\gamma}{2},\quad \gamma=0.475.$$

З таблиці інтегральної функції Лапласа $\Phi$ знайдемо число $t = 1.96$. Тоді за формулою (3.6) точність оцінки буде

$$\delta = \frac{t \cdot \sigma}{\sqrt{n}} = \frac{2 \cdot 1.96}{\sqrt{25}} = 0.784.$$

Отже, довірчий інтервал буде $(\overline{x}_B - 0.784, \overline{x}_B + 0.784)$. Якщо $\overline{x}_B = 2.8$, то з надійністю 95% інтервал (2.016,3.584) покриває параметр $a = 2.8$ з точністю до 0.8.

**Знаходження об'єму вибірки.** Нехай ознака $X$ генеральної сукупності розподілена за нормальним законом з параметром $\sigma$ і треба знайти об'єм вибірки $n$, який із заданою точністю $\delta$ та надійністю $\gamma$ дозволить знайти оцінку параметра $a$. Із формули (3.6) одержуємо рівність

$$t = \frac{\delta\sqrt{n}}{\sigma}$$

з якої випливає

$$n = \left(\frac{t\sigma}{\delta}\right)^2. \tag{3.8}$$

Для надійності $\gamma$, використовуючи (3.7) та таблицю значень інтегральної функції Лапласа, знайдемо відповідне число $t$. Тепер $t$, $\delta$ та $\sigma$ відомі, тому за формулою (3.8) можна знайти потрібний об'єм вибірки.

**Приклад 3.2** Випадкова величина розподілена за нормальним законом з параметром $\sigma$. Знайти мінімальний об'єм $n$ вибірки, щоб з надійністю $\gamma$ та точністю $\delta$ виконувалась рівність $\overline{x}_B = a$, якщо $\sigma = 0.5$, $\gamma = 0.95$, $\delta = 0.1$.

**Розв'язок задачі.** Для $\gamma = 0.95$, згідно з формулою (3.7) маємо

$$\Phi(t) = 0.475, \quad t = 1.96.$$

Використовуючи формулу (3.8), знайдене $t$ та задані $\sigma$, $\delta$ одержуємо

$$n = \left(\frac{1.96 \cdot 0.5}{0.2}\right)^2 = 9.8^2 = 96.04.$$

Отже, мінімальний об'єм вибірки $n = 96$.

Якщо невідоме середньоквадратичне відхилення $\sigma$ ознаки $X$ генеральної сукупності, то використовують розподіл Стьюдента, можна також в формулах (3.4)–(3.5), (3.8) замість $\sigma$ використати $\sigma_B$.

## 3.3. Питання для самоперевірки

- Які оцінками називається точковими оцінками параметрів розподілу генеральної сукупності?

- Які точкові оцінки називають інтервальними?
- Як визначається надійність оцінки невідомого параметра?
- Який алгоритм визначення довірчого інтервалу?
- Який алгоритм знаходження об'єму вибірки, якщо генеральна сукупность розподілена за нормальним законом з відомим стандартним відхиленням?

# 4. Обробка вибірки методом найменших квадратів

## 4.1. Основні поняття

Припустимо, що нам відома функціональна залежність між випадковими величинами $Y$ та $X$ вигляду

$$Y = f(X, a_1, a_2, ..., a_m)$$

з невідомими параметрами $a_1, a_2, ..., a_m$.

Наприклад, можна розглядати залежність між собівартістю продукції (ознака $Y$) та об'ємом продукції (ознака $X$) деякої кількості однотипних підприємств.

Звичайно при зростанні об'єму продукції $X$ собівартість $Y$ повинна спадати. Але ця залежність не є однозначною. Внаслідок різних причин при випуску однакового об'єму продукції собівартість її на різних підприємствах буде неоднаковою.

Нехай внаслідок $n$ незалежних випробувань одержані варіанти ознак $Y$ та $X$, які оформлені у статистичній таблиці вигляду

Табл. 4.1

| Випробування | 1 | 2 | ... | k | ... | n |
|---|---|---|---|---|---|---|
| $X$ | $x_1$ | $x_2$ | ... | $x_k$ | ... | $x_n$ |
| $Y$ | $y_1$ | $y_2$ | ... | $y_k$ | ... | $y_n$ |

Для знаходження оцінок параметрів функціональної залежності $a_1, a_2, ..., a_m$ за даними вибірки застосуємо метод найменших квадратів. Цей метод базується на тому, що найімовірніші значення параметрів $a_1, a_2, ..., a_m$ повинні

давати мінімум функції

$$S = \sum_{k=1}^{n} \Big( y_k - f(x_k, a_1, a_2, ..., a_m) \Big)^2. \tag{4.1}$$

Якщо функція $f(x_k, a_1, a_2, ..., a_m)$ має неперервні частинні похідні відносно невідомих параметрів $a_1, a_2, ..., a_m$ то необхідною умовою існування мінімуму функції $S$ буде система $m$ рівнянь з $m$ невідомими

$$\frac{\partial S}{\partial a_k} = 0, \quad k = 1, 2, ..., m.$$

Знаходження функціональної залежності між випадковими величинами $X$ та $Y$ з використанням даних випробувань (або вибірки) називають вирівнюванням емпіричних даних вздовж кривої $y_k = f(x_k, a_1, a_2, ..., a_m)$.

Нижче розглянемо детальніше оцінки параметрів лінійної та параболічної функціональної залежностей, які використовуються найчастіше.

## 4.2. Оцінка параметрів лінійної функції

Нехай між випадковими величинами $X$ та $Y$ існує лінійна функціональна залежність

$$Y = aX + b, \tag{4.2}$$

параметри $a$ та $b$ якої невідомі. Згідно з формулою (4.2) маємо

$$S = \sum_{k=1}^{n} (y_k - (ax_k + b))^2. \tag{4.3}$$

Ця функція $S$ неперервно диференційовна, тому згідно з необхідними умовами існування мінімуму $S$ повинні виконуватись рівності

$$\frac{\partial S}{\partial a} = 0, \quad \frac{\partial S}{\partial b} = 0.$$

У нашому випадку ці рівності мають вигляд

$$\begin{cases} \sum\limits_{k=1}^{n} (y_k - (ax_k + b))x_k = 0; \\ \sum\limits_{k=1}^{n} (y_k - (ax_k + b)) = 0; \end{cases}$$

або

$$\begin{cases} a\sum\limits_{k=1}^{n} {x_k}^2 + b\sum\limits_{k=1}^{n} x_k = \sum\limits_{k=1}^{n} x_k y_k; \\ a\sum\limits_{k=1}^{n} x_k + bn = \sum\limits_{k=1}^{n} y_k. \end{cases}$$

Виписана система є неоднорідною лінійною системою двох рівнянь відносно двох невідомих $a$ та $b$. За правилом Крамера можна знайти єдиний розв'язок цієї системи у вигляді

$$a = \frac{\Big(\sum\limits_{k=1}^{n} x_k\Big)\cdot\Big(\sum\limits_{k=1}^{n} y_k\Big) - n\sum\limits_{k=1}^{n} x_k y_k}{\Big(\sum\limits_{k=1}^{n} x_k\Big)^2 - n\sum\limits_{k=1}^{n} x_k^2}; \tag{4.4}$$

$$b = \frac{\Big(\sum\limits_{k=1}^{n} x_k\Big)\cdot\Big(\sum\limits_{k=1}^{n} x_k y_k\Big) - \Big(\sum\limits_{k=1}^{n} {x_k}^2\Big)\cdot\Big(\sum\limits_{k=1}^{n} y_k\Big)}{\Big(\sum\limits_{k=1}^{n} x_k\Big)^2 - n\sum\limits_{k=1}^{n} x_k^2}; \tag{4.5}$$

Якщо кількість значень $x_k$ та $y_k$ велика, то обчислення параметрів $a$ та $b$ за формулами (4.4), (4.5) ускладнюється. Для спрощення обчислень початок розрахунків величин $x_k$ переносять у середнє значення усіх $x_k$, тобто у точку

$$\overline{x}_a = \frac{1}{n}\sum_{k=1}^{n} x_k.$$

Тоді після деяких проміжних викладок одержують

$$a = \frac{\sum\limits_{k=1}^{n}(x_k - \overline{x}_a)y_k}{\sum\limits_{k=1}^{n}(x_k - \overline{x}_a)^2}; \quad b = \frac{1}{n}\sum_{k=1}^{n} y_k - a\overline{x}_a. \tag{4.6}$$

Формули (4.6) дозволяють визначити параметри $a$ та $b$ лінійної функціональної залежності (4.2) шляхом обчислення простішим, ніж за формулами (4.4), (4.5).

## 4.3. Оцінка параметрів параболічної функціональної залежності

Нехай між випадковими величинами $X$ та $Y$ існує функціональна залежність вигляду

$$Y = aX^2 + bX + c. \tag{4.7}$$

Методом найменших квадратів на основі даних випробувань знайдемо значення невідомих параметрів $a$, $b$, $c$. Тепер формула (4.1) буде мати вигляд

$$S = \sum_{k=1}^{n} (y_k - (a{x_k}^2 + bx_k + c))^2. \tag{4.8}$$

Необхідні умови існування мінімуму функції $S$ є рівності нулю частиних похідних першого порядку

$$\begin{cases} \dfrac{\partial S}{\partial a} = -2\sum\limits_{k=1}^{n}(y_k - ax^2 - bx_k - c){x_k}^2 = 0; \\ \dfrac{\partial S}{\partial b} = -2\sum\limits_{k=1}^{n}(y_k - ax^2 - bx_k - c)x_k = 0; \\ \dfrac{\partial S}{\partial c} = -2\sum\limits_{k=1}^{n}(y_k - ax^2 - bx_k - c) = 0. \end{cases}$$

Цю систему можна записати у вигляді

$$\begin{cases} a\sum\limits_{k=1}^{n}{x_k}^4 + b\sum\limits_{k=1}^{n}{x_k}^3 + c\sum\limits_{k=1}^{n}{x_k}^2 = \sum\limits_{k=1}^{n}{x_k}^2 y_k; \\ a\sum\limits_{k=1}^{n}{x_k}^3 + b\sum\limits_{k=1}^{n}{x_k}^2 + c\sum\limits_{k=1}^{n}x_k = \sum\limits_{k=1}^{n}x_k y_k; \\ a\sum\limits_{k=1}^{n}{x_k}^2 + b\sum\limits_{k=1}^{n}x_k + cn = \sum\limits_{k=1}^{n}y_k. \end{cases} \tag{4.9}$$

Система (4.9) є неоднорідною лінійною системою трьох рівнянь з невідомими $a$, $b$, $c$. Розв'язок цієї системи можна знайти різними методами (матричним, за правилом Крамера, методом Гаусса-Жордана), а його вигляд буде громіздкий при досить великій кількості випробувань $n$.

Система (4.9) та її розв'язок значно спрощуються, коли значення $x_k$ рівновіддалені ($x_2 - x_1 = x_3 - x_2 = ... = x_n - x_{n-1}$) та виконується умова

$$\sum_{k=1}^{n} x_k = 0,$$

яку можна одержати за допомогою нового аргументу

$$z_k = x_k - \overline{x}_a, \quad \overline{x}_a = \frac{1}{n}\sum_{k=1}^{n} x_k.$$

Припустимо, що вказані умови виконуються, тоді замість системи (4.9) одержимо систему

$$\begin{cases} a\sum\limits_{k=1}^{n} {x_k}^4 + c\sum\limits_{k=1}^{n} {x_k}^2 = \sum\limits_{k=1}^{n} {x_k}^2 y_k; \\ b\sum\limits_{k=1}^{n} {x_k}^2 = \sum\limits_{k=1}^{n} x_k y_k; \\ a\sum\limits_{k=1}^{n} {x_k}^2 + cn = \sum\limits_{k=1}^{n} y_k. \end{cases}$$

Розв'язок цієї системи можна знайти за формулами

$$\begin{cases} a = \dfrac{n\sum\limits_{k=1}^{n} {x_k}^2 y_k - \Big(\sum\limits_{k=1}^{n} {x_k}^2\Big)\Big(\sum\limits_{k=1}^{n} y_k\Big)}{n\sum\limits_{k=1}^{n} {x_k}^4 - \Big(\sum\limits_{k=1}^{n} {x_k}^2\Big)^2}; \\ b = \dfrac{n\sum\limits_{k=1}^{n} x_k y_k}{\sum\limits_{k=1}^{n} {x_k}^2}; \\ c = \dfrac{\Big(\sum\limits_{k=1}^{n} {x_k}^4\Big)\Big(\sum\limits_{k=1}^{n} y_k\Big) - \Big(\sum\limits_{k=1}^{n} {x_k}^2\Big)\Big(\sum\limits_{k=1}^{n} {x_k}^2 y_k\Big)}{n\sum\limits_{k=1}^{n} {x_k}^4 - \sum\limits_{k=1}^{n} {x_k}^2}. \end{cases}$$

**Приклад 4.1** Застосовуючи метод найменших квадратів, скласти рівняння параболи (4.7), яка проходить найближче до точок

| $x_i$ | –4 | –3 | –2 | –1 | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|---|---|---|---|
| $y_i$ | 6 | 3 | 1 | 0.3 | –0.1 | –0.2 | 0 | 0.2 | 1 |

Табл. 4.2 — Параметри задачі

**Розв'язок задачі.** У даному випадку значення $x_i$ рівновіддалені

$$n = 9; \quad \sum_{k=1}^{n} x_k = -4 - 3 - 2 - 1 + 0 + 1 + 2 + 3 + 4 = 0$$

і виконуються умови, які дозволяють знаходити параметри $a$, $b$, $c$ за спрощеними формулами. Тому, підставивши значення $x_i$ та $y_i$ з таблиці у ці формули, одержимо

$$\begin{cases} a = \dfrac{9 \cdot 144.81 - 60 \cdot 11.2}{9 \cdot 708 - 3600} = \dfrac{1303 - 672}{6312 - 3600} = \dfrac{631.29}{2712} = 0.20003; \\ b = \dfrac{-24 - 9 - 2 - 0.3 - 0.2 + 0.6 + 4}{60} \dfrac{-30.9}{60} = -0.515; \\ c = \dfrac{708 \cdot 11.2 - 60 \cdot 144.81}{2712} = \dfrac{-759}{2712} = -0.27987. \end{cases}$$

тому що

$$\sum_{k=1}^{9} {x_k}^2 y_k = 144.81; \quad \sum_{k=1}^{9} {x_k}^4 = 708; \quad \sum_{k=1}^{9} {x_k}^2 = 60;$$

$$\sum_{k=1}^{9} y_k = 11.2; \quad \sum_{k=1}^{9} x_k y_k = -30.9.$$

Таким чином, рівнянням шуканої параболи буде

$$Y = 0.20003 \cdot X^2 - 0.515 \cdot X - 0.27987.$$

## 4.4. Питання для самоперевірки

- Як визначають статистичні оцінки числових характеристик та умови їх незсунутості, ефективності, обґрунтованості?
- Вказати числові характеристики вибірки та формули, за якими їх обчислюють.
- Які властивості має вибіркова середня?
- Як визначають степеневу середню вибірки, середню квадратичну, середню гармонічну та середню геометричну вибірки?
- Що називають вибірковим середньо-квадратичним відхиленням?
- В яких випадках використовують виправлену вибіркову дисперсію і як вона пов'язана із вибірковою дисперсією?
- В яких випадках обчислюють характеристики вибірки методом добутків? Який порядок дій при використанні цього методу?

- Які статистичні оцінки називають точковими, інтервальними? В яких випадках використовують інтервальні оцінки та що вони дозволяють встановити?
- Який порядок дій знаходження довірчого інтервалу для оцінки математичного сподівання нормального розподілу при відомому та невідомому ?
- Як знаходять об'єм вибірки, який із заданими точністю та надійністю дозволить знайти оцінку математичного сподівання нормально розподіленої випадкової величини?
- Маємо криву $y = f(x, a_1, a_2, ..., a_m)$. Що називають вирівнюванням емпіричних даних вздовж цієї кривої?
- Яка суть методу найменших квадратів знаходження невідомих параметрів функціональної залежності випадкових величин?

  Які дії та в якій послідовності треба виконати при застосуванні цього методу?

# 5. Статистична перевірка гіпотез

## 5.1. Статистичні гіпотези та їх різновиди

Часто необхідно знати закон розподілу генеральної сукупності. Якщо закон розподілу невідомий, але є міркування для припущення його певного вигляду $A$, наприклад, розподіл рівномірний, показниковий або нормальний, тоді висувають гіпотезу:

**Генеральна сукупність розподілена за законом** $A$.

У цій гіпотезі йде мова про вигляд невідомого розподілу.

Іноді закон розподілу генеральної сукупності відомий, але його параметри (числові характеристики) невідомі. Якщо є міркування припустити, що невідомий параметр $\theta$ дорівнює певному значенню $\theta_0$ то висувають гіпотезу: $\theta = \theta_0$. Ця гіпотеза вказує на припущену величину параметра відомого розподілу.

Можливі також інші гіпотези: про рівність параметрів двох різних розподілів, про незалежність вибірок, про те, що у листопаді 2031 року буде кінець світу, та багато інших.

**Означення 5.1** Статистичними називають гіпотези про вигляд розподілу генеральної сукупності або про параметри відомих розподілів.

Наприклад, статистичними будуть гіпотези:

- генеральна сукупність розподілена за нормальним законом;
- дисперсії двох сукупностей, розподілених за законом Пуассона, рівні між собою.

Відомо, що на творчі можливості людей впливають не тільки гени та умови життя, але й космос. Розглянемо гіпотези:

- значна частина народжених у першому півріччі має краще розвинену ліву частину мозку, яка здійснює логічне мислення;
- значна частина людей, народжених у другому півріччі, має краще розвинену праву частину мозку, яка здійснює образне мислення.

Ці гіпотези не статистичні, бо в них не йде мова ні про вигляд, ні про параметри розподілу. Але для вказаної ситуації можна сформулювати декілька статистичних гіпотез.

Разом з припущеною гіпотезою завжди можна розглядати протилежну їй гіпотезу. Якщо припущена гіпотеза була відхилена, тоді має місце протилежна гіпотеза. Отже, ці гіпотези доцільно відрізняти.

**Означення 5.2** Основною (нульовою) називають припущену гіпотезу і позначають $H_0$.

**Означення 5.3** Альтернативною (конкурентною) називають гіпотезу, що суперечить основній гіпотезі, її позначають $H_1$.

Наприклад, якщо $H_0 : M(X) = 6$, то $H_1 : M(X) \neq 6$.

Гіпотези можуть містити тільки одне або більше одного припущення.

**Означення 5.4** Гіпотезу звуть простою, якщо вона містить лише одне припущення.

Наприклад, якщо $\lambda$ — параметр показникового розподілу, то гіпотеза $H_0 : \lambda = 5$ буде проста.

**Означення 5.5** Гіпотезу називають складною, якщо вона складається із скінченої або нескінченої кількості простих гіпотез.

Наприклад, гіпотеза $H_0$ : математичне сподівання нормального розподілу дорівнює 2 — складна гіпотеза тому, що середнє квадратичне відхилення $\sigma$ невідоме і може приймати будь-яке значення.

Гіпотеза $H$ : показниковий розподіл має параметр $\lambda > 2$ складається із нескінченої множини гіпотез $H_k : \lambda = c_k$, де $c_k > 2, k = 1, 2, ....$

## 5.2. Похибки перевірки гіпотез

Статистична гіпотеза, яка висунута, може бути правильною або неправильною, тому виникає необхідність її перевірки.

Перевірка гіпотези здійснюється за даними вибірки, тобто статистичними методами. Тому перевірку гіпотези за даними вибірки називають статистичною.

При перевірці статистичної гіпотези за даними випадкової вибірки можна зробити хибний висновок. При цьому можуть бути похибки першого та другого роду.

**Означення 5.6** Якщо за висновком буде відкинута правильна гіпотеза, то кажуть, що це похибка першого роду.

**Означення 5.7** Якщо за висновком буде прийнята неправильна гіпотеза, то кажуть, що це похибка другого роду.

Відмітимо, що наслідки цих похибок можуть бути різними. Наприклад, якщо відкинути правильну гіпотезу «продовжити будівництво авіазаводу», то ця похибка першого роду буде сприяти матеріальним витратам.

Якщо прийняти неправильну гіпотезу «продовжити будівництво, не враховуючи можливості обвалу об'єкту будівлі», то внаслідок цієї похибки другого роду можуть загинути люди.

**Означення 5.8** Імовірність здійснити похибку першого роду позначають $a$ і називають рівнем значущості.

Найчастіше рівень значущості приймають рівним 0.05 або 0.01. Якщо прийнято рівень значущості рівним 0.05, то це означає, що в п'яти випадках із 100 ми ризикуємо одержати похибку першого роду (відкинути правильну гіпотезу).

При контролі якості продукції імовірність признати нестандартними стандартні вироби називають ризиком виробника, а імовірність признати придатними браковані вироби називають ризиком споживача.

## 5.3. Критерії узгодження для перевірки гіпотез

### Статистичний критерій перевірки основної гіпотези

Перевірку статистичної гіпотези можна здійснити лише з використанням даних вибірки. Для цього слід обрати, деяку випадкову статистичну характеристику (вибіркову функцію), точний або наближений розподіл якої відомий, і за допомогою цієї характеристики перевірити основну гіпотезу.

**Означення 5.9** Статистичним критерієм узгодження перевірки гіпотези

(або просто критерієм) називають випадкову величину K, розподіл якої (точний або наближений) відомий і яка застосовується для перевірки основної гіпотези.

В цьому означенні не враховується вид розподілу статистичної характеристики. Якщо статистична характеристика розподілена нормально, то критерій позначають не літерою $K$, а літерами $U$ або $Z$. Якщо статистична характеристика розподілена за законом Фішера - Снедекора, то її позначають $F$.

У випадку розподілу статистичної характеристики за законом Стьюдента її позначають $T$, а у випадку закону "$\chi$-квадрат" — $\chi^2$..

Наприклад, для перевірки гіпотез про рівність дисперсії двох нормальних генеральних сукупностей за статистичну характеристику $K$ вибирають відношення виправлених вибіркових дисперсій

$$F = \frac{{S_1}^2}{{S_2}^2}.$$

В різних дослідах дисперсія буде приймати різні, наперед невідомі значення, тому ця величина випадкова. Вона розподілена за законом Фішера - Снедекора.

**Означення 5.10** Спостереженим значенням критерію узгодження називають значення відповідного критерію, обчислене за даними вибірки.

Наприклад, якщо за даними вибірок із двох нормальних генеральних сукупностей знайдено виправлені вибіркові дисперсії ${S_1}^2 = 18$ та ${S_2}^2 = 6$, тоді спостереженим значенням критерію узгодження буде

$$F_{exp} = \frac{18}{3} = 3.$$

Існує багато критеріїв узгодження. Наприклад, найбільш точний (асимптотично) критерій Неймана - Пірсона використовує нерівності або відношення функцій правдоподібності.

## Критична область

Після обрання певного критерію узгодження, множину усіх його можливих значень поділяють на дві підмножини, що не перетинаються: одна з них містить значення критерію, при яких основна гіпотеза відхиляється, а друга - при яких вона приймається.

**Означення 5.11** Критичною областю називають сукупність значень

критерію, при яких основна гіпотеза відхиляється.

**Означення 5.12** Областю прийняття гіпотези (областю допустимих значень) називають множину значень критерію, при яких гіпотезу приймають.

Критерій узгодження $K$ - одновимірна випадкова величина, усі її можливі значення належать деякому інтервалу. Тому критична область та область прийняття гіпотези також будуть інтервалами, а це означає, що існують точки, які ці інтервали відокремлюють.

**Означення 5.13** Критичними точками (межами) критерію $K$ називають точки $k_{kp}$, які відокремлюють критичну область від області прийняття гіпотези.

Розрізняють однобічну (правобічну та лівобічну) та двобічну критичні області.

**Означення 5.14** Правобічною називають критичну область, що визначається нерівністю $K > k_{kp}$.

**Означення 5.15** Лівобічною називають критичну область, що визначається нерівністю $K < k_{kp}$.

### Знаходження критичних областей

Щоб знайти однобічну критичну область, треба знайти критичну точку $k_{kp}$. Для цього задають достатньо малу імовірність - рівень значущості $\alpha$, а потім шукають критичну точку з врахуванням вимоги

$$P(K > k_{kp}) = \alpha$$

у випадку правобічної критичної області або

$$P(K < k_{kp}) = \alpha$$

у випадку лівобічної критичної області.

У випадку двобічної критичної області повинно виконуватись тотожність

$$P(K < k_{kp}) + P(K > k_{kp}) = \alpha$$

Для кожного критерію узгодження є відповідні таблиці, які дозволяють знайти таку точку $k_{kp}$ яка задовольняє потрібну умову.

При знаходженні критичної області доцільно враховувати потужність критерію.

**Означення 5.16** Потужністю критерію називають імовірність належності

критерію критичній області при умові, що правильна альтернативна гіпотеза.

Іншими словами, потужність критерію є імовірність того, що основна гіпотеза буде відхилена, якщо альтернативна гіпотеза правильна.

Якщо рівень значущості $\alpha$ вже обрано, то критичну область доцільно будувати так, щоб потужність критерію була максимальною. Виконання цієї вимоги забезпечує мінімальну імовірність похибки другого роду.

Єдиний спосіб одночасного зменшення імовірностей похибок першого та другого роду це є збільшення об'єму вибірки.

## Порядок дій при перевірці статистичних гіпотез

Для перевірки правильності основної статистичної гіпотези $H_0$ необхідно:

1. визначити гіпотезу $H_1$, альтернативну до гіпотези $H_0$;
2. обрати статистичну характеристику перевірки;
3. визначити допустиму імовірність похибки першого роду, тобто рівень значущості $\alpha$;
4. знайти за відповідною таблицею критичну область (критичну точку) для обраної статистичної характеристики.

До критичної області належать такі значення статистичної характеристики, при яких гіпотеза $H_0$ відхиляється на користь альтернативної гіпотези $H_1$. Підкреслимо, що між рівнем значущості $\alpha$ та критичною областю існує такий зв'язок: якщо гіпотеза $H_0$ правильна, то з імовірністю значення вибіркової функції будуть належати критичній області.

Так, при перевірці гіпотези про рівність дисперсій двох нормальних сукупностей при альтернативній

$$H_1 : D(X) > D(Y)$$

треба знайти спостережене значення критерія Фішера-Снедекора, тобто

$$F_{exp} = \frac{{S_1}^2}{{S_2}^2},$$

а потім з таблиці критичних точок цього розподілу по заданому рівню значущості $\alpha$ та степенях вільності $k_1 = n_1 - 1$ та $k_2 = n_2 - 1$ знайти $F_{kp}(\alpha; k_1, k_2)$.

Якщо $F_{exp} < F_{kp}$ то гіпотеза $H_0$ приймається. Якщо $F_{exp} > F_{kp}$ то $H_0$ відхиляють.

## 5.4. Деякі критерії перевірки статистичних гіпотез

### Перевірка гіпотези про рівність математичних сподівань нормальних генеральних сукупностей

Нехай дві нормально розподілені генеральні сукупності мають рівні дисперсії, а математичні сподівання можуть бути різними.

Із сукупностей зробили вибірку об'єму $n_1$ та $n_2$ і знайшли вибіркові середні $\overline{x}_1$ та $\overline{x}_2$, а також виправлені дисперсії $S_1$ та $S_2$ відповідно.

Потрібно перевірити гіпотезу $H_0$ : різниця математичних сподівань цих сукупностей дорівнює числу $c_0$

$$H_0 : a_1 - a_2 = c_0$$

Альтернативна гіпотеза буде

$$H_1 : a_1 - a_2 \neq c_0.$$

Для перевірки гіпотези в якості статистичної характеристики (вибіркової функції) візьмемо функцію

$$\nu_{exp} = \frac{\overline{x}_1 - \overline{x}_2 - c_0}{\sqrt{\dfrac{n_1 \cdot {S_1}^2 + n_2 \cdot {S_2}^2}{n_1 + n_2 - 2} \cdot \left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}, \tag{5.1}$$

яка розподілена за законом Стьюдента з степенями вільності, що дорівнюють $n_1 + n_2 - 2$.

Для заданого рівня значущості $\alpha$ можна знайти критичну область для статистичної характеристики $\nu$ з врахуванням альтернативної гіпотези $H_1$.

**Приклад 5.1** Підприємство виготовляє однакові деталі двома способами. Першим способом виготовлено 10 деталей, витрати сировини були такими $1.4, 1.6, 1.2, 1.5, 1.4, 1.6, 1.5, 1.8, 1.1, 1.4$. Другим способом виготовлено 6 деталей, витрати сировини були такими $1.8, 1.7, 1.9, 1.3, 1.6, 1.5$.

Припускаючи, що дисперсія витрат сировини однакова, при рівні значущості $\alpha = 0.02$ перевірити гіпотезу

$$H_0 : a_1 - a_2 = 0$$

при альтернативній гіпотезі

$$H_1 : a_1 - a_2 \neq 0.$$

**Розв'язок задачі.** Треба перевірити гіпотезу про рівність математичних

сподівань двох нормальних генеральних сукупностей. Згідно з гіпотезою $H_1$ критична область буде двобічною і визначається умовою

$$P\Big(\nu > \nu_{\frac{\alpha}{2}}\Big) = P\Big(\nu < -\nu_{\frac{\alpha}{2}}\Big),$$

де статистична характеристика $\nu$ визначена формулою (5.1).

Степінь вільності дорівнює

$$n_1 + n_2 - 2 = 10 + 6 - 2 = 14.$$

З таблиці критичних значень $\nu_{\frac{\alpha}{2}}$ для 14 степеней вільності одержимо

$$\Big(\nu_{\frac{\alpha}{2}}\Big)_{kp} = 2.6.$$

За даними вибірки можна знайти

$$\overline{x}_1 = 1.45; \quad {S_1}^2 = 0.04; \quad \overline{x}_2 = 1.63; \quad {S_4}^2 = 0.05.$$

Тепер за формулою (5.1) одержимо

$$\nu_{exp} = \frac{1.45 - 1.63}{\sqrt{\dfrac{10 \cdot 0.04 + 6 \cdot 0.05}{10 + 6 - 2} \cdot \Big(\dfrac{1}{10} + \dfrac{1}{6}\Big)}} = \frac{-0.18}{0.115} = -1.58.$$

Отже, $\nu_{exp}$ не належить до критичної області, тому гіпотеза $H_0$ може бути прийнята.

## Критерій дисперсійного аналізу

Нехай є $N$ нормально розподілених генеральних сукупностей з рівними дисперсіями та, можливо, з різними математичними сподіваннями.

Із кожної сукупності робимо вибірку об'єму

$$\{n_i\}, \quad i = 1, 2, ..., N,$$

тоді $\sum\limits_{i=1}^{N} n_i = n$ — об'єм усієї вибірки.

Позначимо $j$ варіанту випадкової величини $X$ з $i$-тої сукупності $x_{ij}$. Тоді середня арифметична вибірки із $i$-тої сукупності буде

$$x_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}, \tag{5.2}$$

а середня усієї вибірки буде

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{N} x_i n_i. \tag{5.3}$$

При рівні значущості $\alpha$ треба перевірити основну гіпотезу про рівність математичних сподівань розглядаємих сукупностей

$$H_0 : a_1 = a_2 = ... = a_N.$$

При рівності дисперсій статистична характеристика буде мати розподіл Фішера з $N-1$ та $n-N$ степенів вільності. Тому в якості статистичної характеристики для перевірки цієї гіпотези візьмемо функцію

$$F = \frac{\dfrac{1}{N-1}\displaystyle\sum_{i=1}^{N}(x_i - \overline{x})^2 n_i}{\dfrac{1}{n-N}\displaystyle\sum_{i=1}^{N}\sum_{j=1}^{n_i}(x_{ij} - \overline{x})^2}. \tag{5.4}$$

Критичну область у цьому випадку знаходять з урахуванням умови

$$P(F > f_a) = a,$$

де $f_a$ — критичне (і табульоване) значення розподілу Фішера.

**Приклад 5.2** Є дані про вартість (в тис. гривень) проданих трьох видів виробів певним магазином в окремі дні тижня

Табл. 5.1 — Продажа трьох видів виробів

| вівторок | середа | четвер | п'ятниця | субота |
|---|---|---|---|---|
| 10.2 | 10.8 | 10.7 | 13.0 | 12.0 |
| 11.5 | 9.8 | 11.5 | 13.2 | 11.5 |
| 12.0 | 12.1 | 12.0 | 11.5 | 11.8 |

Припускаючи нормальний закон розподілу одержаної суми кожного дня та рівність дисперсій, перевірити гіпотезу $H_0 : a_1 = a_2 = a_3 = a_4 = a_5$ при рівні значущості $\alpha = 0.05$.

**Розв'язок задачі.** Умови прикладу дозволяють застосувати до розв'язання задачі критерій дисперсійного аналізу.

У цьому випадку маємо

$$N = 5; \quad n_i = 3, i = 1, 2, ..., 5; \quad n = 15.$$

За формулами (2) та (3) знаходимо

$$x_1 = 11.2, \quad x_2 = 10.8, \quad x_3 = 11.4, \quad x_4 = 12.6, \quad x_5 = 11.8, \quad \overline{x} = 11.6.$$

Зробимо обчислення сум, що входять до формули (5.4)

$$\sum_{i=1}^{5}(x_i - \overline{x})^2 n_i = 0.48 + 1.92 + 0.12 + 3 + 0.12 = 5.64, \tag{5.5}$$

$$\sum_{i=1}^{5}\sum_{j=1}^{3}(x_{ij} - \overline{x})^2 = 1.92 + 0.01 + 0.16 + 1 + 3.24 + 0.25 + 0.81 + \\ 0.01 + 0.16 + 1.96 + 2.56 + 0.01 + 0.15 + 0.01 + 0.04 = 12.34. \tag{5.6}$$

Тепер за формулою (5.4) знайдемо значення статистичної характеристики

$$F = \frac{\frac{1}{4} \cdot 5.64}{\frac{1}{10} \cdot 12.34} = 1.14.$$

Із таблиці критичних значень розподілу Фішера із степенями вільності $N - 1 = 5 - 1 = 4$ та $n - N = 15 - 5 = 10$ і рівнем значущості $\alpha = 0.05$ знаходимо

$$F_{kp} = f_{0.05} = 3.48.$$

Одержали, що $F_{exp} = 1.14 < F_{kp} = 3.48$, тому гіпотеза $H_0$ може бути прийнята.

## Критерій узгодження Пірсона ($\chi^2$)

Критерій узгодження Пірсона ($\chi$-квадрат) ефективно використовують для перевірки гіпотези про розподіл генеральної сукупності, тобто що розподіл випадкової величини має певний функціональний вираз.

Обмежимось застосуванням цього критерію для перевірки гіпотези про нормальний розподіл генеральної сукупності.

Нехай вибірка має такий розподіл об'єму $n$

Табл. 5.2 — Нормальний розподіл генеральної сукупності

| Варіанти $x_k$ | $x_1$ | $x_2$ | ... | $x_m$ |
|---|---|---|---|---|
| Частоти $n_k$ | $n_1$ | $n_2$ | ... | $n_m$ |

або

Табл. 5.3 — Нормальний розподіл інтервалів

| Інтервали $x_k$ | $(x_1, x_2)$ | $(x_2, x_3)$ | ... | $(x_m, x_{m+1})$ |
|---|---|---|---|---|
| Частоти $n_k$ | $n_1$ | $n_2$ | ... | $n_m$ |

Потрібно з рівнем значущості $\alpha$ перевірити основну гіпотезу $H_0$ : генеральна сукупність розподілена нормально.

Критерієм перевірки цієї гіпотези беруть випадкову величину $\chi^2$, яка у різних випробуваннях приймає різні, наперед невідомі значення.

Критичне значення цієї випадкової величини залежить від рівня значущості $\alpha$ та степенів вільності її розподілу $k$

$$\chi^2{}_{kp} = \chi^2(\alpha, k.)$$

Ці критичні значення табульовані для різних $\alpha$ та $k$.

Для розподілу генеральної сукупності за нормальним законом степінь вільності буде

$$k = m - 3, \tag{5.7}$$

де $m$ — кількість варіант вибірки або часткових інтервалів варіант.

Правило Пірсона. Щоб при заданому рівні значущості $\alpha$ перевірити основну гіпотезу $H_0$ : генеральна сукупність розподілена нормально, треба

1. обчислити теоретичні частоти $n_k^{'}$ для варіант вибірки;
2. обчислити спостережене значення критерія $\chi^2$ за формулою

$$\chi^2{}_{exp} = \sum_{k=1}^{m} \frac{n_k - n_k^{'}}{n_k^{'}} \tag{5.8}$$

3. знайти степінь вільності $\chi^2$ за формулою (5.7);
4. знайти з таблиці критичну точку $\chi^2{}_{kp}$ яка відповідає заданому рівню значущості $\alpha$ та степені вільності $k$;
5. порівняти $\chi^2{}_{exp}$ та $\chi^2{}_{kp}$ зробити висновок
   — якщо $\chi^2{}_{exp} < \chi^2{}_{kp}$, то гіпотезу $H_0$ треба прийняти;
   — якщо $\chi^2{}_{exp} > \chi^2{}_{kp}$ то гіпотезу треба відхилити.

**Приклад 5.3** При рівні значущості $\alpha = 0.05$ перевірити гіпотезу про нормальний розподіл генеральної сукупності, якщо відомі емпіричні та теоретичні частоти

Табл. 5.4 — Емпіричні та теоретичні частоти

| $n_k$ | 6 | 13 | 38 | 74 | 106 | 85 | 30 | 14 |
|---|---|---|---|---|---|---|---|---|
| $n_k^{'}$ | 3 | 14 | 42 | 82 | 99 | 76 | 37 | 13 |

**Розв'язок задачі.** У даному випадку теоретичні частоти $n_k^{'}$ задані, кількість варіант вибірки $m = 8$. Тому за формулою (5.7) знаходимо $k = 8 - 3 = 5$.

З таблиці критичних точок розподілу $\chi^2(\alpha, k)$ $\alpha = 0.05$ та $k = 5$ знаходимо ${\chi^2}_{kp} = 11.1$.

Для обчислення ${\chi^2}_{exp}$ за формулою (5.8) використаємо розрахункову таблицю

Табл. 5.5 — розрахункову таблиця для обчислення ${\chi^2}_{exp}$

| $n_k$ | $n_k^{'}$ | $n_k - n_k^{'}$ | $(n_k - n_k^{'})^2$ | $\frac{(n_k - n_k^{'})^2}{n_k^{'}}$ |
|---|---|---|---|---|
| 6 | 3 | 3 | 9 | 3 |
| 13 | 14 | -1 | 1 | 0.07 |
| 38 | 42 | -4 | 16 | 0.38 |
| 74 | 82 | -8 | 64 | 0.78 |
| 106 | 99 | 7 | 49 | 0.49 |
| 85 | 76 | 9 | 81 | 1.07 |
| 30 | 37 | -7 | 49 | 1.32 |
| 14 | 13 | 1 | 1 | 0.08 |

В результаті отримаємо ${\chi^2}_{exp} = 7.19$. Таким чином, ${\chi^2}_{exp} = 7.19 < {\chi^2}_{kp}$ тому за правилом Пірсона гіпотезу $H_0$ слід прийняти. Отже, дані вибірки узгоджуються з гіпотезою $H_0$, тому що розбіжність емпіричних та теоретичних частот незначна.

## Знаходження теоретичних частот нормального розподілу

Згідно з класичним означенням імовірності

$$p_k = \frac{n_k^{'}}{n}, \quad n_k^{'} = p_k \cdot n, \quad k = 1, 2, ..., m.$$

Отже, для знаходження теоретичних частот $n_k^{'}$ треба знайти імовірність

$$p_k = P(X = x_k) \quad \text{або} \quad p_k = P(x_k < X < x_{k+1})$$

відповідно.

Імовірність $p_k = P(X = x_k)$ можна знайти, використовуючи локальну

функцію Лапласа $\varphi(x)$ та дані вибірки за формулою

$$p_k = P(X = x_k) = \frac{h}{\sigma_B}\varphi(u_k),$$

де

$$h = x_{k+1} - x_k; \quad u_k = \frac{x_k - \overline{x}_B}{\sigma_B}; \quad \varphi(u) = \frac{1}{2\pi}e^{-\frac{u^2}{2}},$$

варіанти $x_k$ рівновіддалені.

Імовірність $p_k = P(x_k < X < x_{k+1})$ можна знайти, використовуючи інтегральну функцію Лапласа Ф(x) за формулою

$$p_k = P(x_k < X < x_{k+1}) = \Phi\left(\frac{x_{k+1} - \overline{x}_B}{\sigma_B}\right) - \Phi\left(\frac{x_k - \overline{x}_B}{\sigma_B}\right).$$

## 5.5. Питання для самоперевірки

- Які гіпотези називають статистичними, основною та альтернативною, простою та складною?
- Що таке похибки першого та другого роду перевірки статистичної гіпотези?
- Який смисл рівня значущості $\alpha$?
- Що називають статистичним критерієм, критичною областю та критичною точкою перевірки гіпотези?
- Як перевіряють гіпотезу про рівність дисперсій двох нормальних сукупностей?
- Який смисл потужності критерія перевірки гіпотези?
- Вказати порядок дій при перевірці гіпотез.
- Як здійснюють перевірку гіпотези про рівність математичних сподівань?
- За яким критерієм здійснюють перевірку гіпотези про рівність математичних сподівань $N$ нормально розподілених сукупностей?
- Коли застосовують критерій узгодження Персона ($\chi$-квадрат)?
- Як формулюють правило Пірсона?
- Як знаходять теоретичні частоти нормального розподілу для перевірки гіпотези за правилом Пірсона?

# 6. Задачі математичної статистики і їх розв'язок з використанням мови програмування Python

## 6.1. Функції математичної статистики у мові програмування Python

Для розв'яку задач математичної статистики з використанням мови програмування Python [7] необхідно підкючити модуль **stats**. Для цього, обмеживщись роботою в режимі калькулятора, потрібно виконати таку інструкцію

>>> from scipy import stats

Модуль **stats** надає функції для обчислення математичної статистики числових (зі значенням Real) даних.

Модуль не є конкурентом сторонніх бібліотек, таких як NumPy, SciPy, або пропрієтарним повнофункціональним статистичним пакетам, призначеним для професійних статистиків, таких як Minitab, SAS і Matlab. Він орієнтований на рівень графічних і наукових калькуляторів.

Якщо явно не вказано інше, ці функції підтримують типи даних int, float, Decimal (числа с фіксованою точністю) і Fraction (раціональні числа (звичайні дроби)). Поведінка з іншими типами (у числовій вежі чи ні) наразі не підтримується. Колекції зі змішаними типами також не визначені і залежать від реалізації. Якщо ваші вхідні дані складаються зі змішаних типів, ви можете використовувати map() для забезпечення узгодженого результату, наприклад: map(float, input_data).

Цей модуль надає функції для обчислення математичної статистики числових (зі значенням Real) даних.

**Середні та вимірювання центрального розташування**. Ці функції обчислюють середнє або типове значення для генеральної сукупності або вибірки.

- mean() values Середнє арифметичне ('середнє') даних.
- fmean() values Швидке арифметичне з плаваючою точкою.
- geometric_mean() values Середнє геометричне даних.
- harmonic_mean() values Гармонійне середнє даних.
- median() values Медіана (середнє значення) даних.
- median_low() values Низька медіана даних.
- median_high() values Висока медіана даних.

- median_grouped() values Медіана або 50-а процентиль згрупованих даних.
- mode() values Одиночний режим (найпоширеніше значення) дискретних або номінальних даних.
- multimode() values Список режимів (найпоширеніших значень) дискретних або номінальних даних.
- quantiles() values Поділ даних на інтервали з рівною ймовірністю.

Ці функції обчислюють міру того, наскільки генеральна сукупність або вибірка схильні відхилятися від типових або середніх значень.

- pstdev() values Популяційне стандартне відхилення даних.
- pvariance() values Популяційна дисперсія даних.
- stdev() values Стандартне відхилення вибірки даних.
- variance() values Вибіркова дисперсія даних.

## 6.2. Властивості функцій математичної статистики у мові програмування Python

**Примітка**. Функції не вимагають відсортованих даних. Однак для зручності читання в більшості прикладів показано відсортовані послідовності.

**statistics.mean(data)** values Повертає вибіркове середнє арифметичне data, яке може бути послідовним або ітерованим.

Середнє арифметичне - це сума даних, поділена на кількість точок даних. Його зазвичай називають "середнім", хоча це лише одне з безлічі різних математичних середніх. Це показник центрального розташування даних.

Якщо data порожній, буде викликано StatisticsError.

### Деякі приклади використання функцій математичної статистики в режимі калькулятора:

```
>>> mean([1, 2, 3, 4, 4, 4])
2.8

>>> mean([-1.0, 2.5, 3.25, 5.75])
2.625

>>> from decimal import Decimal as D
```

```
>>> mean([D("0.5"), D("0.75"), D("0.625"), D("0.375")])
Decimal("0.5625")
```

**Примітка**. Середнє значення сильно залежить від викидів і не є надійною оцінкою для центрального місця розташування: середнє значення не обов'язково є типовим прикладом точок даних. Для надійніших заходів при центральному розташуванні (див. median() і mode()).

Середнє значення вибірки дає об'єктивну оцінку істинного середнього значення сукупності, тож у разі взяття середнього значення за всіма можливими вибірками mean(sample) сходиться до істинного середнього значення всієї сукупності. Якщо data представляє всю генеральну сукупність, а не вибірку, то mean(data) еквівалентний обчисленню істинного середнього значення генеральної сукупності ...

**statistics.fmean(data)** values Перетворює data в числа з плаваючою комою і обчислює середнє арифметичне. Працює швидше, ніж функція mean(), і завжди повертає float. data може бути послідовністю або ітерованим об'єктом. Якщо вхідна множина даних порожня, виникає StatisticsError.

```
>>> fmean([3.5, 4.0, 5.25])
4.25
```

**statistics.geometric_mean(data)** values Перетворює data в числа з плаваючою комою і обчислює середнє геометричне.

Середнє геометричне вказує центральну тенденцію або типове значення data з використанням добутку значень (на відміну від середнього арифметичного, яке використовує їхню суму).

Викликає StatisticsError, якщо вхідна множина даних порожня, якщо вона містить нуль або від'ємне значення. data може бути послідовним або повторюваним.

Ніяких особливих зусиль для досягнення точних результатів не докладається. (Однак це може змінитися в майбутньому.)

```
>>> round(geometric_mean([54, 24, 36]), 1)
36.0
```

**statistics.harmonic_mean(data)** values Повертає гармонійне середнє data, послідовність або ітерований об'єкт дійсних чисел.

Гармонійне середнє, іноді зване середнім субсуперечливим, є зворотною величиною арифметичного mean() зворотних величин даних. Наприклад,

гармонійне середнє трьох значень a, b і c буде еквівалентним

$$\frac{3}{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}}.$$

Якщо одне зі значень дорівнює нулю, результат буде нульовим.

Гармонійне середнє - це тип середнього, міра центрального розташування даних. Це часто доречно при усередненні показників або співвідношень, наприклад швидкостей.

Припустимо, автомобіль їде 10 км зі швидкістю 40 км/год, а потім ще 10 км зі швидкістю 60 км/год. Яка середня швидкість?

```
>>> harmonic_mean([40, 60])
48.0
```

Припустимо, інвестор купує акції кожної з трьох компаній рівної вартості з коефіцієнтами P/E (ціна/прибуток) 2.5, 3 і 10. Який середній коефіцієнт P/E для портфеля інвестора?

```
>>> harmonic_mean([2.5, 3, 10]) # Для рівного інвестиційного портфеля.
3.6
```

StatisticsError виникає, якщо data порожній або будь-який елемент менший за нуль.

Поточний алгоритм має ранній вихід, коли він зустрічає нуль на вході. Це означає, що наступні вхідні дані не перевіряються на достовірність. (Ця поведінка може змінитися в майбутньому.)

**statistics.median(data)** values Повертає медіанне значення (середнє значення) числових даних, використовуючи загальний метод "середнього двох середніх". Якщо data порожній, викликається StatisticsError. data може бути послідовним або ітерованим.

Медіана є надійним показником центрального місця розташування і меншою мірою залежить від наявності викидів. Якщо кількість точок даних непарна, повертається середня точка даних:

```
>>> median([1, 3, 5])
3
```

Коли кількість точок даних парна, медіана інтерполюється шляхом взяття середнього з двох середніх значень:

```
>>> median([1, 3, 5, 7])
4.0
```

Це підходить для випадків, коли ваші дані дискретні, і ви не заперечуєте, що медіана може не бути фактичною точкою даних.

Якщо дані є порядковими (підтримують операції сортування), але не числовими (не підтримує додавання), розгляньте можливість використання замість них median_low() або median_high().

**statistics.median_low(data)** values Повертає низьку медіану числових даних. Якщо data порожній, викликається StatisticsError. data може бути послідовним або ітерованим.

Низька медіана завжди входить у множину даних. Якщо кількість точок даних непарна, повертається середнє значення. Коли вона парна, повертається менше з двох середніх значень.

```
>>> median_low([1, 3, 5])
3

>>> median_low([1, 3, 5, 7])
3
```

Використовуйте низьку медіану, коли ваші дані дискретні і ви віддаєте перевагу, щоб медіана була фактичною точкою даних, а не інтерпольованою.

**statistics.median_high(data)** values Повертає високе середнє значення даних. Якщо data порожні, викликається StatisticsError. data може бути послідовним або ітерованим.

Висока медіана завжди входить у множину даних. Якщо кількість точок даних непарна, повертається середнє значення. Коли вона парна, повертається більше з двох середніх значень.

```
>>> median_high([1, 3, 5])
3

>>> median_high([1, 3, 5, 7])
5
```

Використовуйте високу медіану, коли ваші дані дискретні і вам потрібно, щоб медіана була фактичною точкою даних, а не інтерпольованою.

**statistics.median_grouped(data, interval=1)** values Повертає медіанне значення згрупованих безперервних даних, розраховане як 50-й процентиль, з використанням інтерполяції. Якщо data порожній, викликається StatisticsError. data може бути послідовним або ітерованим.

```
>>> median_grouped([52, 52, 53, 54])
52.5
```

У наступному прикладі дані округлені, так що кожне значення являє собою середню точку класів даних, наприклад 1 - це середня точка класу 0.5-1.5, 2 - середня точка 1.5-2.5, 3 - середня точка 2.5-3.5 тощо. За наведених даних середнє значення потрапляє десь у клас 3.5-4.5, і для його оцінки використовується інтерполяція:

>>> median_grouped([1, 2, 2, 3, 4, 4, 4, 4, 4, 4, 5])
3.7

Необов'язковий аргумент interval представляє інтервал класу і за замовчуванням дорівнює 1. При зміні інтервалу класу природним чином змінюється інтерполяція:

>>> median_grouped([1, 3, 3, 5, 5, 7], interval=1)
3.25

>>> median_grouped([1, 3, 3, 5, 7], interval=2)
3.5

Функція не перевіряє, чи рознесені точки даних як мінімум на interval.

Деталі реалізації CPython: За деяких обставин median_grouped() може перетворювати точки даних на плаваючі. Ця поведінка, ймовірно, зміниться в майбутньому.

**statistics.mode(data)** values Повертає єдину найпоширенішу точку даних з дискретного або номінального data. Режим (якщо він існує) є найбільш типовим значенням і слугує мірою центрального розташування.

Якщо є кілька режимів з однаковою частотою, повертає перший із виявлених у data. Якщо замість цього потрібно найменший або найбільший з них, використовувати min(multimode(data)) або max(multimode(data)). Якщо вхідне data порожнє, викликається StatisticsError.

mode приймає дискретні дані і повертає одне значення. Це стандартне поводження з режимом, якому зазвичай вчать у школах:

>>> mode([1, 1, 2, 3, 3, 3, 3, 3, 4])
3

Цей режим унікальний тим, що це єдина статистика в цьому пакеті, яка також застосовується до номінальних (нечислових) даних:

>>> mode(["red "blue "blue "red "green "red "red "red"])
'red'

Змінено у версії 3.8: Тепер обробляє мультимодальні набори даних, повертаючи перший виявлений режим. Раніше він видавав StatisticsError, коли виявлялося більше одного режиму.

**statistics.multimode(data)** values Повертає список значень, які найчастіше зустрічаються, в тому порядку, в якому вони вперше були виявлені в data. Поверне більше одного результату, якщо є кілька режимів, або порожній список, якщо data порожній:

```
multimode('aabbbbbbbbccddddddeeffffgg')
['b', 'd', 'f']

>>> multimode('')
[{}]
```

**statistics.pstdev(data, mu=None)** values Повертає стандартне відхилення генеральної сукупності (квадратний корінь із дисперсії генеральної сукупності). Див. аргументи та інші подробиці в pvariance().

```
>>> pstdev([1.5, 2.5, 2.5, 2.5, 2.75, 3.25, 4.75])
0.986893273527251
```

**statistics.pvariance(data, mu=None)** values Повертає дисперсію генеральної сукупності data, непорожню послідовність або ітерацію дійсних чисел. Дисперсія, або другий момент відносно середнього, є мірою мінливості (розкиду або дисперсії) даних. Великий розкид вказує на те, що дані рознесені; невелика дисперсія вказує на те, що вона згрупована близько до середнього значення.

Якщо вказано необов'язковий другий аргумент mu, зазвичай це середнє значення data. Його також можна використовувати для обчислення другого моменту навколо точки, яка не є середнім значенням. Якщо він відсутній або None (за замовчуванням), автоматично розраховується середнє арифметичне.

Використовуйте цю функцію, щоб обчислити дисперсію для всієї генеральної сукупності. Для оцінки дисперсії за вибіркою зазвичай краще використовувати функцію variance().

Викликає StatisticsError, якщо data порожній.

**Приклади:**

```
>>> data = [0.0, 0.25, 0.25, 1.25, 1.5, 1.75, 2.75, 3.25]
>>> pvariance(data)
1.25
```

Якщо ви вже обчислили середнє значення ваших даних, ви можете передати його як необов'язковий другий аргумент mu, щоб уникнути перерахунку:

```
>>> mu = mean(data)
>>> pvariance(data, mu)
1.25
```

Підтримуються десяткові розряди та дроби:

```
>>> from decimal import Decimal as D
>>> pvariance([D("27.5"), D("30.25"), D("30.25"), D("34.5"), D("41.75")])
Decimal('24.815'))
```

```
>>> from fractions import Fraction as F
>>> pvariance([F(1, 4), F(5, 4), F(1, 2)])
Fraction(13, 72)
```

**Примітка**. При виклику для всієї генеральної сукупності це дає дисперсію сукупності. Коли замість цього викликається вибірка, це зміщена дисперсія вибірки s?, також відома як дисперсія з N ступенями свободи.

Якщо ви якимось чином знаєте справжнє середнє значення генеральної сукупності ?, ви можете використовувати цю функцію для обчислення дисперсії вибірки, задавши відоме середнє значення генеральної сукупності як другий аргумент. Якщо точки даних являють собою випадкову вибірку генеральної сукупності, результатом буде об'єктивна оцінка дисперсії генеральної сукупності.

**statistics.stdev(data, xbar=None))** values Повертає стандартне відхилення вибірки (квадратний корінь із дисперсії вибірки). Див. аргументи та інші подробиці в variance().

```
>>> stdev([1.5, 2.5, 2.5, 2.5, 2.75, 3.25, 4.75])
1.0810874155219827
```

**statistics.variance(data, xbar=None))** values Повертає приблизну дисперсію data, ітерацію щонайменше двох дійсних чисел. Дисперсія, або другий момент відносно середнього, є мірою мінливості (розкиду або дисперсії) даних. Великий розкид вказує на те, що дані рознесені; невелика дисперсія вказує на те, що вона згрупована близько до середнього значення.

Якщо вказано необов'язковий другий аргумент xbar, він має бути середнім значенням data. Якщо він відсутній або None (за замовчуванням), середнє значення розраховується автоматично.

Використовуйте цю функцію, якщо ваші дані є вибіркою з генеральної сукупності. Щоб розрахувати дисперсію для всієї генеральної сукупності, див. pvariance().

Підвищує StatisticsError, якщо у data менше двох значень.

**Приклади:**

```
>>> data = [2.75, 1.75, 1.25, 0.25, 0.5, 1.25, 3.5]
```

```
>>> variance(data)
1.3720238095238095
```

Якщо ви вже обчислили середнє значення ваших даних, ви можете передати його як необов'язковий другий аргумент xbar, щоб уникнути перерахунку:

```
>>> m = mean(data)
>>> variance(data, m)
1.3720238095238095
```

Функція не намагається перевірити, що ви пройшли фактичне середнє значення як xbar. Використання довільних значень для xbar може призвести до недійсних або неможливих результатів.

Підтримуються десяткові та дробові значення:

```
>>> from decimal import Decimal as D
>>> variance([D("27.5"), D("30.25"), D("30.25"), D("34.5"), D("41.75")])
Decimal('31.01875')

>>> from fractions import Fraction as F
>>> variance([F(1, 6), F(1, 2), F(5, 3)])
Fraction(67, 108)
```

**Примітка**. Це вибіркова дисперсія $s^2$ з поправкою Бесселя, також відома як дисперсія з $N - 1$ ступенями вільності. За умови, що точки даних є репрезентативними (наприклад, незалежними й однаково розподіленими), результатом має бути об'єктивна оцінка істинної дисперсії сукупності.

Якщо ви якимось чином знаєте фактичне середнє значення генеральної сукупності ?, ви повинні передати його у функцію pvariance() як параметр $\mu$, щоб отримати дисперсію вибірки.

**statistics.quantiles(data, *, n=4, method='exclusive'))** values Розділити data на безперервні інтервали $n$ з рівною ймовірністю. Повертає список $n - 1$ точок відсікання, що розділяють інтервали.

Встановити $n$ на 4 для квартилів (за замовчуванням). Встановити для n значення 10 для децилей. Встановити для $n$ значення 100 для процентилів, що дає 99 точок відсікання, які розділяють data на 100 груп рівного розміру. Підвищує StatisticsError, якщо $n$ не менше 1.

data може бути будь-яким ітеративним, що містить зразки даних. Для отримання значущих результатів кількість точок даних у data має бути більшою, ніж $n$. Викликає StatisticsError, якщо немає хоча б двох точок даних.

Точки розрізу лінійно інтерполюються з двох найближчих точок даних.

Наприклад, якщо точка відсікання потрапляє на одну третину відстані між двома значеннями вибірки, 100 і 112, точка відсікання обчислюватиметься як 104.

method для обчислення квантилів може змінюватись залежно від того, чи включає data найнижчі та найвищі можливі значення з генеральної сукупності, чи виключає їх.

За замовчуванням method є "exclusive"і використовується для даних, взятих із сукупності, яка може мати більше екстремальних значень, ніж у вибірках. Частка популяції, що потрапляє нижче $i-th$ з $m$ відсортованих точок даних, обчислюється як $\frac{i}{m+1}$. Враховуючи дев'ять значень вибірки, метод сортує їх і призначає такі процентилі: 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%.

Встановлення method на "inclusive"використовується для опису даних про сукупність або для вибірок, які, як відомо, включають найбільш екстремальні значення від сукупності. Мінімальне значення в data розглядається як 0-й процентиль, а максимальне значення розглядається як 100-й процентиль. Частка популяції, що потрапляє нижче $i-th$ з m відсортованих точок даних, обчислюється як $\frac{i-1}{m-1}$. Враховуючи 11 значень вибірки, метод сортує їх і призначає такі процентилі: 0%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, 100%.

# Децильні точки для емпірично обраних даних

```
>>> data = [
105, 129, 87, 86, 111, 111, 89, 81, 108, 92, 110,
100, 75, 105, 103, 109, 76, 119, 99, 91, 103, 129,
106, 101, 84, 111, 74, 87, 86, 103, 103, 106, 86,
111, 75, 87, 102, 121, 111, 88, 89, 101, 106, 95,
103, 107, 101, 81, 109, 104]

>>> [round(q,1) for q in quantiles(data, n=10)]
[81.0, 86.2, 89.0, 99.4, 102.5, 103.6, 106.0, 109.8, 111.0]
```

**Винятки**. Визначено єдиний виняток:
exception statistics.StatisticsError values Підклас ValueError для винятків, пов'язаних зі статистикою.

NormalDist - це інструмент для створення нормальних розподілів випадкової змінної. Це клас, який розглядає середнє значення і стандартне відхилення вимірювань даних як єдине ціле.

Нормальні розподіли виникають із центральної граничної теореми і мають

широкий спектр застосування в статистиці.

- class statistics.NormalDist($\mu = 0.0$, $\sigma = 1.0$) values Повертає новий об'єкт NormalDist, де mu представляє середнє арифметичне, а $\sigma$ представляє стандартне відхилення. Якщо $\sigma$ негативна, викликає StatisticsError.
- mean values Властивість тільки для читання для середнього арифметичного нормального розподілу.
- median values Властивість тільки для читання для медіани нормального розподілу.
- mode values Властивість тільки для читання для режиму нормального розподілу.
- stdev values Властивість тільки для читання для стандартного відхилення нормального розподілу.
- variance values Властивість тільки для читання для дисперсії нормального розподілу. Дорівнює квадрату стандартного відхилення.
- classmethod from_samples(data) values Створює екземпляр нормального розподілу з параметрами mu і $\sigma$, обчисленими з data з використанням fmean() і stdev(). data може бути будь-яким ітерованим і має складатися зі значень, які можна перетворити на тип float. Якщо data не містить принаймні двох елементів, викликається StatisticsError, тому що потрібна принаймні одна точка для оцінки центрального значення і принаймні дві точки для оцінки дисперсії.
- samples(n, *, seed=None) values Створює випадкові вибірки n для заданого середнього і стандартного відхилення. Повертає list зі значень float.

Якщо задано seed, створює новий екземпляр базового генератора випадкових чисел. Корисно для створення відтворюваних результатів навіть у багатопотоковому контексті.

- pdf(x) values Використовуючи функцію щільності ймовірності (pdf), обчислює відносну ймовірність того, що випадкова величина $X$ буде близька до заданого значення $x$. Математично це межа відношення $\frac{P(x <= X < x + dx)}{dx}$, оскільки $dx$ наближається до нуля. Відносну ймовірність обчислюють як імовірність того, що вибірка перебуває у вузькому діапазоні, поділену на ширину діапазону (звідси й слово "густина"). Оскільки ймовірність відноситься до інших точок, її значення може бути більшим за "1.0".

- cdf(x) values Використовуючи кумулятивну функцію розподілу (cdf), обчислює ймовірність того, що випадкова величина X буде меншою або дорівнюватиме x. Математично це написано $P(X <= x)$.

- inv_cdf(p) values Обчислює зворотну кумулятивну функцію розподілу, також відому як квантильна функція або відсоткові точки. Математично це написано $x : P(X <= x) = p$. Знаходить значення $x$ випадкової величини $X$, таке, що ймовірність того, що змінна буде меншою або дорівнюватиме цьому значенню, дорівнює заданій ймовірності $p$.

- overlap(other) values Вимірює відповідність між двома нормальними розподілами ймовірностей. Повертає значення від 0.0 до 1.0, отримуючи область перекриття для двох функцій щільності ймовірності.

- quantiles(n=4) values Розділяє нормальний розподіл на $n$ безперервних інтервалів з рівною ймовірністю. Повертає список $(n - 1)$ точок розрізу, що розділяють інтервали.

Встановити $n$ на 4 для квартилів (за замовчуванням). Встановити для $n$ значення 10 для децилів. Встановити для $n$ значення 100 для процентилів, що дає 99 точок відсікання, які поділяють нормальний розподіл на 100 груп рівного розміру.

Екземпляри NormalDist підтримують додавання, віднімання, множення і ділення на константу. Ці операції використовуються для переведення і масштабування. Наприклад:

```
>>> temperature_february = NormalDist(5, 2.5) # Цельсій
>>> temperature_february*(9/5)+32 # Фаренгейт
NormalDist(mu=41.0, sigma=4.5)
```

Ділення константи на екземпляр NormalDist не підтримується, тому що результат не буде нормально розподілений.

Оскільки нормальні розподіли виникають унаслідок адитивних ефектів незалежних змінних, скласти і відняти дві незалежні нормально розподілені випадкові величини можна представити як екземпляри NormalDist. Наприклад:

```
>>> birth_weights = NormalDist.from_samples([2.5, 3.1, 2.1, 2.4, 2.7, 3.5])
>>> drug_effects = NormalDist(0.4, 0.15)
>>> combined = birth_weights + drug_effects
>>> round(combined.mean, 1)
3.1

>>> round(combined.stdev, 1)
```

0.5

**NormalDist Приклади та рецепти** NormalDist легко розв'язує класичні ймовірнісні задачі.

Наприклад, з огляду на історичні дані для іспитів SAT, які свідчать, що бали зазвичай розподіляються із середнім значенням 1060 і стандартним відхиленням 195, після округлення до найближчого цілого числа визначте відсоток учнів із результатами тестів від 1100 до 1200:

```
>>> sat = NormalDist(1060, 195)
>>> fraction = sat.cdf(1200 + 0.5) - sat.cdf(1100 - 0.5)
>>> round(fraction * 100.0, 1)
18.4
```

Знайти квартилі та децилі для результатів SAT:

```
>>> list(map(round, sat.quantiles()))
[928, 1060, 1192]
>>> list(map(round, sat.quantiles(n=10)))
[810, 896, 958, 1011, 1060, 1109, 1162, 1224, 1310]
```

Щоб оцінити розподіл для моделі, яку нелегко вирішити аналітично, NormalDist може згенерувати вхідні вибірки для Симуляції Монте-Карло:

```
>>> def model(x, y, z): return (3*x+7*x*y-5*y)/(11*z)
>>> n = 100 000
>>> X = NormalDist(10, 2.5).samples(n, seed=3652260728)
>>> Y = NormalDist(15, 1.75).samples(n, seed=4582495471)
>>> Z = NormalDist(50, 1.25).samples(n, seed=6582483453)
>>> quantiles(map(model, X, Y, Z))
[1.4591308524824727, 1.8035946855390597, 2.175091447274739]
```

Нормальні розподіли можна використовувати для апроксимації Біноміального розподілу, коли розмір вибірки великий і коли ймовірність успішного випробування близька до 50%.

Наприклад, конференція відкритого вихідного коду налічує 750 учасників і дві зали місткістю 500 осіб. Є розмова про Python і ще одна про Ruby. На попередніх конференціях 65% учасників воліли слухати виступи Python. Якщо припустити, що вподобання людей не змінилися, яка ймовірність того, що кімната Python залишиться в межах своєї місткості?

```
>>> n = 750 # Розмір зразка
>>> p = 0.65 # Перевага Python
>>> q = 1.0 - p # Перевага Ruby
>>> k = 500 # Місткість кімнати
```

```
>>> # Апроксимація з використанням кумулятивного нормального розподілу
>>> from math import sqrt
>>> round(NormalDist(mu=n*p, sigma=sqrt(n*p*q)).cdf(k + 0.5), 4)
0.8402

>>> # Розв'язання з використанням кумулятивного біноміального розподілу
>>> from math import comb, fsum
>>> round(fsum(comb(n, r) * p**r * q**(n-r) for r in range(k+1)),4)
0.8402

>>> # Апроксимація з використанням симуляції
>>> from random import seed, choices
>>> seed(8675309)
>>> def trial(): return choices(('Python', 'Ruby'), (p, q), k=n).count('Python')
>>> mean(trial() <= k for i in range(10 000))
0.8398
```

Нормальні розподіли зазвичай виникають у задачах машинного навчання.

У Вікіпедії є гарний приклад наївного байєсівського класифікатора. Завдання полягає в тому, щоб передбачити стать людини на основі вимірювань нормально розподілених характеристик, включно зі зростом, вагою і розміром стопи.

Нам дано безліч тренувальних даних із вимірюваннями для восьми осіб. Передбачається, що вимірювання мають нормальний розподіл, тому ми підсумовуємо дані з NormalDist:

```
>>> height_male = NormalDist.from_samples([6, 5.92, 5.58, 5.92])
>>> height_female = NormalDist.from_samples([5, 5.5, 5.42, 5.75])
>>> вага_чоловіка = NormalDist.from_samples([180, 190, 170, 165])
>>> вага_жінки = NormalDist.from_samples([100, 150, 130, 150])
>>> size_foot_male = NormalDist.from_samples([12, 11, 12, 10])
>>> foot_size_female = NormalDist.from_samples([6, 8, 7, 9])
```

Потім ми зустрічаємо нову людину, розміри рис якої відомі, але стать невідома:

```
>>> ht = 6.0 # зріст
>>> wt = 130 # вага
>>> fs = 8 # розмір ноги
```

Починаючи з 50% апріорної ймовірності чоловіків або жінок, ми обчислюємо апостеріорне значення як попередній час як добуток імовірностей для вимірів ознак з урахуванням статі:

>>> prior_male = 0.5
>>> prior_female = 0.5
>>> posterior_male = (prior_male*height_male.pdf(ht)* weight_male.pdf(wt*foot_si
>>> posterior_female = (prior_female*height_female.pdf(ht)*weight_female.pdf(wt)*
foot_size_female.pdf(fs))
Остаточне передбачення відноситься до найбільшої задньої частини. Це відомо як апостеріорний максимум або MAP:

>>> 'male' if posterior_male > posterior_female else 'female'
'female'

## 6.3. Способи запису та зчитування великого обсягу даних у файл у середовищі Python

Ми будемо обговорювати технологію розв'язання основних задач математичної статистики в середовищі мови програмування Python спочатку без використання спеціалізованого модуля **stats** на прикладі реальних даних, які за сучасними поняттями належать до галузі науки про великі дані (BigData).

Як приклади джерел виникнення великих даних наводяться дані, що безперервно надходять з вимірювальних пристроїв, події від радіочастотних ідентифікаторів, потоки повідомлень із соціальних мереж, метеорологічні дані, дані дистанційного зондування Землі, потоки даних про місцезнаходження абонентів мереж мобільного зв'язку, пристроїв аудіо- та відеореєстрації.

Для освоєння технології аналізу великих даних насамперед необхідно розглянути способи запису і зчитування в середовищі Python великого обсягу даних у текстові файли csv формату. csv (comma-separated value) - це формат текстового представлення табличних даних (наприклад, це можуть бути дані з таблиці або дані з БД). У стандартній бібліотеці Python є модуль csv, який дає змогу працювати з файлами в csv форматі.

У цьому форматі кожен рядок файлу - це рядок таблиці. Незважаючи на назву формату, роздільником може бути не тільки кома.

Приклад коду для запису даних data у форматі списку списків у csv-файл

```
import csv

data = [['hostname', 'vendor', 'model', 'location'],
        ['sw1', 'Cisco', '3750', 'London, Best str'],
```

```
        ['sw2', 'Cisco', '3850', 'Ліверпуль, Better str'],
        ['sw3', 'Cisco', '3650', 'Ліверпуль, Better str'],
        ['sw4', 'Cisco', '3650', 'London, Best str']]

file_csv=r'E:\Edu_box\PythonBox\pythonProject\data.csv';
with open(file_csv, mode='w') as f:
    writer = csv.writer(f,delimiter=',',lineterminator='\r')
    for row in data:
        writer.writerow(row)
f.close()
```

де delimiter: вказує роздільник, який розділяє значення даних у CSV-файлі, а dtype: вказує тип даних для списку NumPy. Використовуючи None, ми дозволяємо одночасно імпортувати кілька типів даних у список.

У результаті виконання цього коду Python створить data.csv файлу з таким наповненням

```
hostname,vendor,model,location
sw1,Cisco,3750, "London, Best str"
sw2,Cisco,3850, "Liverpool, Better str"
sw3,Cisco,3650, "Ліверпуль, кращий str"
sw4,Cisco,3650, "London, Best str"
```

Видно, що рядки в останньому стовпчику взяті в лапки, а інші значення - ні. Лапки вказують на те, що кому, яка міститься в лапках, модуль csv не повинен сприймати як роздільник.

Код, представлений нижче, завантажує дані з файлу data.csv у змінну dataset у вигляді списку списків

```
import csv

file_csv=r'E:\Edu_box\PythonBox\pythonProject\data.csv';
with open(file_csv, mode='r') as f:
        salary = list(csv.reader(f))
print(salary)
f.close()
```

Інструкція print(dataset) поверне дані у вихідному вигляді вигляді.

Найчастіше заголовки стовпців зручніше отримати окремим об'єктом. Це можна зробити таким чином

```
import csv

file_csv=r'E:\Edu_box\PythonBox\pythonProject\data.csv';
```

```
with open(file_csv, mode='r') as f:
    reader = csv.reader(f)
    headers = next(reader)
    print('Headers: ', headers)
    for row in reader:
        print(row)
f.close()
```

## 6.4. Основні задачі математичної статистики в середовищі Python

Ми будемо обговорювати технологію розв'язання основних задач математичної статистики в середовищі мови програмування Python [7], на прикладі реальних даних, узятих із платформи Kaggle (https://www.kaggle.com/datasets) із датасету ds_salary (Зарплати в галузі науки про дані у 2023 році).

Код, представлений нижче, завантажує датасет ds_salary.csv у змінну salary у вигляді списку списків. Далі ми будемо розглядати завдання математичної статистики на прикладі цієї змінної

```
import csv

file_csv=r'E:\Edu_box\PythonBox\pythonProject\ds_salary.csv'
with open(file_csv, mode='r') as f:
        salary = list(csv.reader(f))
f.close()
```

Набір даних про заробітну плату в галузі обробки та аналізу даних містить 11 стовпців, кожен з яких має такий зміст

- 01 work_year: Рік, коли було виплачено зарплату.
- 02 experience_level: Рівень досвіду роботи протягом року
- 03 employment_type: Тип зайнятості для цієї ролі
- 04 job_title: Ця роль працювала протягом року.
- 05 salary: загальна виплачена сума брутто-заробітної плати.
- 06 salary_currency: Валюта виплачуваної заробітної плати у вигляді коду валюти ISO 4217.
- 07 salaryinusd: Зарплата в доларах США

- 08 employee_residence: Основна країна проживання працівника протягом робочого року у вигляді коду країни ISO 3166.
- 09 remote_ratio: Загальний обсяг роботи, виконаної віддалено
- 10 company_location: Країна головного офісу або філії роботодавця
- 11 company_size: Середня кількість людей, які працювали в компанії протягом року.

Коли в нас є набір спостережень, корисно звести ознаки наявних даних в одне визначення. Цим займається описова статистика. Як випливає з назви, описова статистика описує конкретну властивість даних, які вона узагальнює. Таку статистику можна розділити на дві категорії

- міри центральної тенденції
- міри розкиду

### Ключові поняття

- описова статистика використовується для систематизації та кількісного опису даних;
- середнє значення вказує на типове значення в нашому наборі даних. Воно не робастне;
- медіана є центральним значенням у ряді даних. Вона робастна;
- мода values значення, яке з'являється найчастіше;
- розмах values це різниця між максимальним і мінімальним значеннями в наборі даних;
- стандартне відхилення і дисперсія є середньою відстанню від середнього арифметичного значення.

### Заходи центральної тенденції

Заходи центральної тенденції — показники, що являють собою відповідь на запитання: "На що схожа середина даних?". Слово "середина"звучить неточно, оскільки існує безліч визначень для її опису. Далі ми обговоримо, як кожна нова міра змінює наше визначення "середини".

### Середнє значення в наборі даних

Ця характеристика описує середнє значення в наборі даних. Обчислити її досить просто: складіть усі значення і розділіть отриману суму на кількість значень.

У випадку із середнім значенням "серединою"датасету буде середнє арифметичне його значень. Середнє значення відображає типовий показник у наборі даних. Якщо ми випадково виберемо один із показників, то, найімовірніше, отримаємо значення, близьке до середнього.

Обчислити середнє значення на Python просто. Давайте з'ясуємо, чому дорівнює середня заробітна плата (долари США) в нашому датасеті:

```
import csv

file_csv=r'E:\Edu_box\PythonBox\pythonProject\ds_salary.csv'
# Витягуємо датасет ds_salary.csv у змінну salary
with open(file_csv, mode='r') as f:
        salary = list(csv.reader(f))
# Витягуємо зарплати з датасету
x_values = [int(X[6]) for X in salary if X[6].isdigit()]
sum_x_values = sum(x_values) # Складаємо всі зарплати
num_x_values = len(x_values) # Шукаємо кількість зарплат
# Рахуємо середнє значення
avg_x_values = round(sum_x_values/num_x_values)
print(avg_x_values)
f.close()
```

Це середнє значення говорить нам, що "типова"зарплат у датасеті дорівнює приблизно 137570. Відповідно, більшість зарплат мають досить високий рейтинг, якщо припустити, що оцінюють за шкалою від 5132 до 450000.

Є різні типи середнього значення, але тут використана є найпоширеніша форма. Воно називається середнім арифметичним, оскільки значення, які нас цікавлять, складаються.

### Медіана в наборі даних

Наступна міра центральної тенденції, про яку піде мова, — медіана. Медіана, як і середнє значення, потрібна для визначення типового значення в наборі даних, але при цьому не потребує обчислень.

Щоб знайти медіану, дані потрібно розташувати в порядку зростання. Медіаною буде значення, яке збігається із серединою набору даних.

Якщо кількість значень парна, то береться середнє двох значень, які "оточують"середину.

Спробуємо знайти медіану зарплат

```
import csv

file_csv=r'E:\Edu_box\PythonBox\pythonProject\ds_salary.csv'
# Витягуємо датасет ds\salary.csv у змінну salary
with open(file_csv, mode='r') as f:
        salary = list(csv.reader(f))
# Витягуємо зарплати з датасету
x_values = [int(X[6]) for X in salary if X[6].isdigit()]
num_salary = len(x_values) /# Знаходимо їхню кількість
/# Сортуємо в порядку зростання
sorted_x_values = sorted(x_values)
/# Шукаємо індекс середнього елемента
middle = (num_salary / 2) + 0.5
print(sorted_x_values[middle]) /# Знаходимо медіану
f.close()
```

Медіанна зарплати становить 135000. Це передбачає, що щонайменше у половини співробітників зарплата в датасеті дорівнює або нижча за 135000. З огляду на те, що і медіана, і середнє значення відображають типове значення, можна припустити, що вони повинні бути приблизно однакові:

```
import csv

file_csv=r'E:\Edu_box\PythonBox\pythonProject\ds_salary.csv'
# Витягуємо датасет ds_salary.csv у змінну salary
with open(file_csv, mode='r') as f:
        salary = list(csv.reader(f))
# Витягуємо зарплати з датасету
x_values = [int(X[6]) for X in salary if X[6].isdigit()]
print(sum(x_values)/len(x_values)) /# 137570
f.close()
```

Середня зарплата в 137570 вища за медіанну 135000. Різниця між медіаною і середнім значенням існує через робастність (викидостійкість).

### Проблема викидів

Середнє значення можна знайти, склавши всі значення і розділивши суму на їхню кількість, тоді як медіану шукають простою перестановкою значень.

Якщо в даних є викиди значень, які набагато вищі або нижчі за інші, то це може негативно вплинути на середнє значення. Таким чином, середнє значення не робастне, а медіана - навпаки, вибросостійка.

Максимальна і мінімальна зарплата в наших даних:

```
import csv

file_csv=r'E:\Edu_box\PythonBox\pythonProject\ds_salary.csv'
# Витягуємо датасет ds_salary.csv у змінну salary
with open(file_csv, mode='r') as f:
        salary = list(csv.reader(f))
# Витягуємо зарплати з датасету
x_values = [int(X[6]) for X in salary if X[6].isdigit()]
min_values = min(x_values)
max_values = max(x_values)
print(min_values, max_values) # 5132, 450000
f.close()
```

Тепер ми знаємо, що в даних є викиди. Викиди можуть відображати цікаві події або помилки в нашому наборі даних, тому важливо вміти визначати їхню наявність. Порівняння медіани і моди є один зі способів визначити наявність викидів, хоча візуалізація зазвичай дає змогу зробити це швидше.

### Мода в наборі даних

Це остання міра центральної тенденції, про яку піде мова. Мода визначається як значення, яке найчастіше зустрічається в наборі даних. Мода не так очевидно відповідає поняттю "середини"як середнє значення або медіана, але ця відповідність абсолютно обґрунтована: якщо значення з'являється в даних неодноразово, воно наблизить середнє значення до моди. Що частіше з'являється значення, то сильніше воно впливає на середнє. Таким чином, мода показує найбільш значущий фактор, що формує середнє значення.

Моду можна обчислити порахувавши кількість повторень різних зарплат і вибравши найчастішу:

```
import csv

file_csv=r'E:\Edu_box\PythonBox\pythonProject\ds_salary.csv'
# Витягуємо датасет ds_salary.csv у змінну salary
with open(file_csv, mode='r') as f:
        salary = list(csv.reader(f))
# Витягуємо зарплати з датасету
```

```
x_values = [int(X[6]) for X in salary if X[6].isdigit()]
# Створюємо порожній словник, у якому будемо
# рахувати кількість появ зарплат
values_counts = {}
for m in x_values:
    if m not in values_counts:
        values_counts[m] = 1
    else:
        values_counts[m] += 1
# Проходимося по словнику і шукаємо
# максимальну кількість повторень
maxp = 0
mode_values = None
for k, v in values_counts.items():
    if maxp < v:
        maxp = v
        mode_values = k
print(mode_values, maxp) # 100000 99
f.close()
```

Мода відносно близька до медіани, тому можна впевнено сказати, що і мода, і медіана відображають середні значення зарплат.

Центральної тенденції корисні для опису середнього значення даних. Проте вони не показують, наскільки великий розкид присутній у даних. Тут на допомогу приходять міри розкиду даних.

## Міри розкиду даних

Міри розкиду відповідають на запитання: "Як сильно варіюються дані?". У світі існує не так багато речей, які залишаються в одному і тому ж стані при кожному спостереженні. Ця мінливість робить світ нечітким і невизначеним, тому корисно мати показники, які можуть узагальнити цю "нечіткість".

## Розмах у наборі даних

Наша перша міра розкиду буде розмах. З усіх вимірів, які ми розглянемо далі, його обчислити найпростіше. Для цього потрібно просто відняти від найбільшого значення в наборі даних найменше.

Ми знайшли максимальну і мінімальну зарплату, коли шукали медіану, тож зараз можемо використовувати ці значення:

```
import csv

file_csv=r'E:\Edu_box\PythonBox\pythonProject\ds_salary.csv'
# Витягуємо датасет ds_salary.csv у змінну salary
with open(file_csv, mode='r') as f:
        salary = list(csv.reader(f))
# Витягуємо зарплати з датасету
x_values = [int(X[6]) for X in salary if X[6].isdigit()]
min_values = min(x_values)
max_values = max(x_values)
values_range = max_values - min_values
print(values_range) # 444868
f.close()
```

Отже, розмах дорівнює 444868, але що це означає? Коли ми розглядаємо результати різних вимірювань, дуже важливо робити це в контексті наших даних. Наша медіанна зарплата була 135000, а розмах дорівнює 444868. Тут розмах на багато більший за медіану, що вказує на сильний розкид даних. Можливо, якби у нас був ще один зарплатний датасет, ми могли б порівняти розмахи, щоб зрозуміти, як вони відрізняються. В іншому разі сам по собі розмах не надто корисний.

Ми радше хотіли б дізнатися, як сильно дані відрізняються від типового значення. Тут нам допоможуть стандартне відхилення і дисперсія випадкової величини.

## Стандартне відхилення в наборі даних

Стандартне відхилення теж є мірою розкиду даних. Воно допомагає дізнатися, як сильно дані відрізняються від типового значення. Іншими словами, воно говорить про те, як сильно дані відрізняються від середнього арифметичного. Відношення до середнього арифметичного добре видно під час розрахунку відхилення:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}}$$

Розглянемо конструкцію рівняння. Як ви пам'ятаєте, середнє арифметичне розраховується шляхом додавання всіх значень і ділення на їхню кількість. Рівняння стандартного відхилення схоже, але використовується, щоб знайти, на скільки в середньому значення відхиляються від типового, і включає

додаткову операцію з добуванням кореня.

Ми хочемо порахувати стандартне відхилення, щоб більш повно описати зарплати та їхні оцінки, тому напишемо свою функцію. Пошук кумулятивної суми вручну мав би доволі громіздкий вигляд, але цикли for у Python все спрощують. Ми пишемо свою функцію, щоб показати, що на Python легко займатися такою статистикою.

```
import csv

file_csv=r'E:\Edu_box\PythonBox\pythonProject\ds_salary.csv'
# Витягуємо датасет ds_salary.csv у змінну salary
with open(file_csv, mode='r') as f:
        salary = list(csv.reader(f))
# Витягуємо зарплати з датасету
x_values = [int(X[6]) for X in salary if X[6].isdigit()]
# Розрахунок стандартного відхилення
def stdev(nums):
    diffs = 0
    avg = sum(nums)/len(nums)
    for n in nums:
        diffs += (n - avg)**(2)
    return (diffs/(len(nums)-1))**(0.5)

print(round(stdev(x_values))) # 63056
f.close()
```

Такі результати цілком очікувані. Зарплати варіюються від 5132 до 450000, тому можна припустити, що стандартне відхилення буде суттєвим. Таке значне відхилення в зарплатах пояснюється через викиди. Чим більше стандартне відхилення, тим більше розсіяні дані навколо середнього значення, і навпаки.

Середнє значення і стандартне відхилення використовується для побудови гістограми та графіка нормального розподілу зарплат (див.рис. 6.1) Далі наведено код для побудови графічного відображення відносного розподілу зарплат у залузі BigData у 2023 році.

```
import numpy as np
import matplotlib.pyplot as plt
import csv

file_csv=r'E:\Edu_box\PythonBox\pythonProject\ds_salary.csv'
with open(file_csv, mode='r') as f:
        salary = list(csv.reader(f))
```

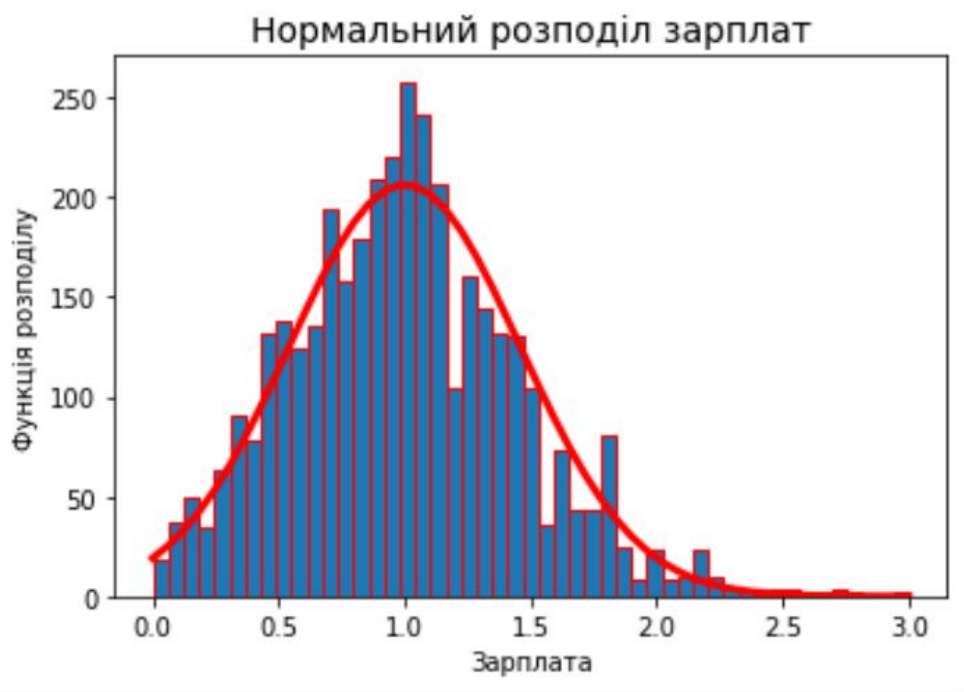


Рис. 6.1 — Гістограма та графік нормального розподілу зарплат

```
x_values = [int(X[6]) for X in salary if X[6].isdigit()]

sum_x_values = sum(x_values)
num_x_values = len(x_values)
avg_x_values = round(sum_x_values/num_x_values)

x =sorted([X/avg_x_values for X in x_values])

x_sum = sum(x)
x_num = len(x)
x_avg = x_sum/x_num

def stdev(nums):
    diffs = 0
    avg = sum(nums)/len(nums)
    for n in nums:
        diffs += (n-avg)**(2)
    return (diffs/(len(nums)-1))**(0.5)
x_sgm = stdev(x)

def f(x):
    a = 1/(x_sgm*np.sqrt(2*np.pi))
    b = np.exp(-(x-x_avg)**2/(2*x_sgm**2))
    return a*b
```

```
z = np.linspace(0, 3, 50)
count, bins, ignored = plt.hist(x, z, edgecolor='red')

scale = count.max()/f(x_avg)
scale = scale*0.8

plt.plot(bins, f(bins)*scale, linewidth=3, color='r')

plt.ylabel('Функція розподілу')
plt.xlabel('Зарплата')
plt.title('Нормальний розподіл зарплат', fontsize= 14)

plt.show()
```

Для того щоб знайти значення функції щільності розподілу ймовірності (висоту стовбчика гістогами) в тому чи іншому інтервалі (корзині) можна скористатися таким кодом

```
hist_values, bin_edges, _ = plt.hist(data, bins=50),
```

де bins — кількість стовбчиків у гістограмі. Повернуті значення зберігаються в масивах hist_values, bin_edges та _ (ігнорований параметр). Потім ми зможемо використовувати індекс інтервалу, який нас цікавить, щоб отримати відповідне значення висоти стовпчика з масиву hist_values.

Значення bin_edges представляють межі інтервалів гістограми, а значення hist_values містять висоту кожного стовпчика. Якщо вам також потрібно отримати межі інтервалів для певного стовпця, ви можете використовувати bin_edges для цього.

Звернемо увагу на той факт, що розмірність масиву bin_edges на одиницю більша за розмірність масиву hist_values.

Із Рис. 6.1 можна зробити висновок, що розподіл зарплат є несиметричним, а тому використання нормального розподілу тут є не зовсім вдалим рішенням. Більш вдалим буде використання несиметричних розподілів (див. далі).

## Дисперсія в наборі даних

Часто стандартне відхилення і дисперсію пов'язують між собою і роблять це не без причини. Ось вираз дисперсії

$$D = \sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$$

Дисперсія і стандартне відхилення майже одне й те саме. Дисперсія це просто квадрат стандартного відхилення. Більше того, обидві величини відображають одну й ту саму річ — міру розкиду, хоча варто зазначити, що одиниці виміру різні. Хоч би в яких одиницях вимірювалися дані, одиниці виміру відхилення будуть такими самими, а в дисперсії їх буде піднесено до квадрата.

Виникає запитання: "Навіщо зводити відхилення у квадрат? Хіба не можна позбутися негативних доданків за допомогою модуля?". Позбавлення від від'ємних значень - це хороша причина для піднесення до квадрата, але не єдина. Як і на середнє значення, на дисперсію і стандартне відхилення впливають викиди. Дуже часто нас цікавлять викиди, тому піднесення до квадрата дає змогу виділити цю особливість.

Найчастіше під час статистичного аналізу нам знадобляться тільки середнє значення і стандартне відхилення, однак дисперсія, як і раніше, важлива в інших академічних галузях. Міри центральної тенденції та розкиду дають нам змогу систематизувати дані та витягти з них знання для їх аналізу.

## 7. Нормальний розподіл

Нормальний закон розподілу трапляється в природі досить часто, тому для нього розроблено окремі ефективні методи моделювання.

Випадкова величина $X$ має нормальний розподіл, якщо ймовірність $P$ того, що $X$ прийме значеннння із діапазону від $-\infty$ до $x$ визначається інтегралом Лапласа

$$F = P(X < x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(t-\overline{x})^2}{2\sigma^2}} dt. \tag{7.1}$$

Як видно, нормальний розподіл має два параметри: математичне сподівання $\overline{x}$ і середньоквадратичне відхилення $\sigma$ величини $x$ від цього математичного сподівання $\overline{x}$.

Очевидно, що ймовірність того, що випадкова величина $X$ попадає в інтервал $a \leq X \leq b$ визначається інтегралом $P(a \leq X \leq b) = F(b) - F(a)$.

Формула щільності розподілу ймовірності значень випадкової величини $x$ за нормальним законом має вигляд

$$f = \frac{1}{\sigma\sqrt{2\pi}} \exp\Big( - \frac{(x-\overline{x})^2}{2\sigma^2}\Big). \tag{7.2}$$

Дослідити властивості нормального розподілу випадкової величини $x$ можна за допомогою програмного коду написаного на мові програмування Python.

Використаємо можливості мови програмування Python для герерації нормального розподілу деякої випадкової величини norm_values та запису його до файлу norm.csv.

```
import numpy as np
import csv

# Середнє значення та стандартне відхилення
mu, sigma = 2.5, 0.1
# Генерація нормально розподілених випадкових величин
norm_values = np.random.normal(mu, sigma, 1000)

file_csv=r'E:\Edu_box\PythonBox\pythonProject\norm.csv'
with open(file_csv, mode='w') as f:
    csvwriter = csv.writer(f,delimiter=',',
        lineterminator='\r')
    csvwriter.writerows(map(lambda x: [x], norm_values))
f.close()
```

Далі будемо вважати, що у файлі norm.csv зосереджені дані про деякий випадковий процес і нам потрібно побудувати його гістограму та щільність нормального розподілу (див.рис. (7.1)). Для цього використаємо такий код

```
import numpy as np
import matplotlib.pyplot as plt
import csv

file_csv=r'E:\Edu_box\PythonBox\pythonProject\norm.csv'
with open(file_csv, 'r') as f:
  csvreader = csv.reader(f, delimiter=',')
  this_list = []
  for row in csvreader:
      this_list.append(float(row[0]))

hist_values, bin_edges, _ = plt.hist(this_list, 30,
density=True, edgecolor='red')
# bins це list даних, які відновідають
# висоті стовбчиків у гістограмі
this_list = np.array(this_list)
mu, sigma = this_list.mean(), this_list.std()

plt.plot(bin_edges, 1/(sigma*np.sqrt(2*np.pi))*
np.exp(-(bin_edges-mu)**2/(2*sigma**2)), linewidth=2, color='r')
```

```
plt.ylabel('Функція розподілу')
plt.xlabel('Випадкова величина $norm\_values$')
plt.title('Нормальний розподіл випадкової величини', fontsize= 1
plt.legend(['mu = 2.5, sigma = 0.1'])
plt.show()
f.close()
```

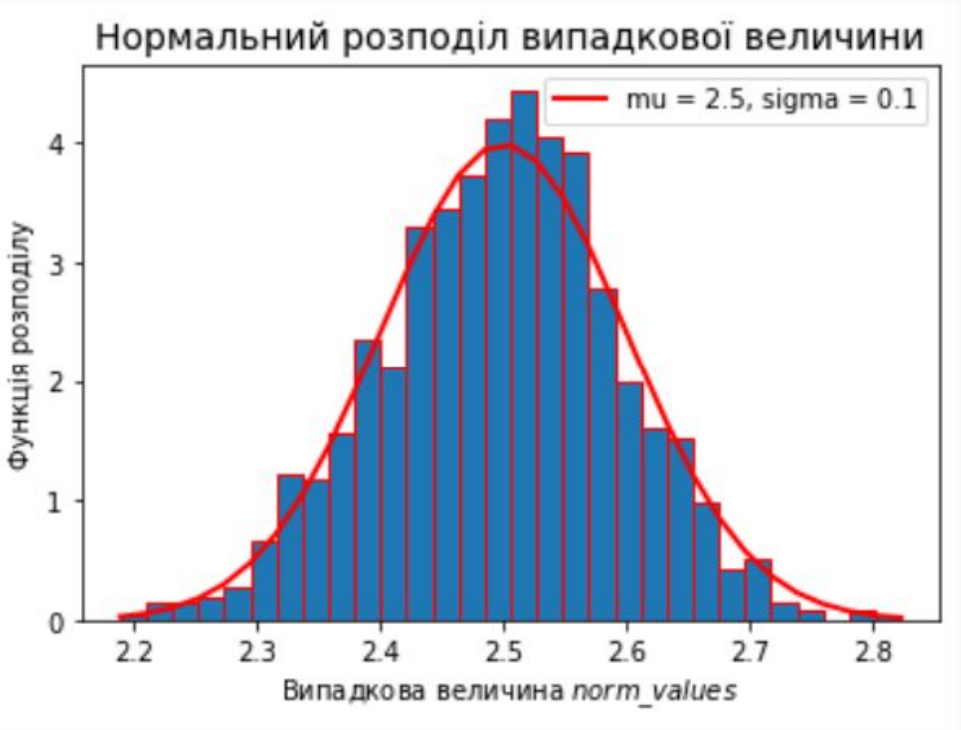


Рис. 7.1 — Приклад нормального розподілу випадкової величини $x$

Змінюючи у програмному коді параметри нормального розподілу *mu* та *sigma* для випадкової величини $x$ можна буде детально розглянути властивості нормального розподілу.

## 8. Логарифмічно нормальний розподіл і розподіл Вейбулла—Гнеденко

Логарифмічно нормальний (логнормальний) розподіл трапляється під час опису довговічності виробів у режимі зносу — старіння, місячної зарплати, розподілу доходів, банківських вкладів, посівних площ під різні культури тощо. т.п.

Розподіл Вейбулла-Гнеденко - це узагальнення експоненціального розподілу на випадок нестаціонарної інтенсивності подій. Використовується в теорії надійності, моделюванні процесів у техніці, у прогнозуванні погоди, в описі процесу подрібнення тощо.

Для практичного використання логнормального розподілу і розподілу Вейбулла—Гнеденко доцільно використовувати спеціалізовані модулі (див. далі) мови програмування Python.

## 8.1. Логнормальний розподіл

Логнормальний розподіл має успішне застосування в теорії надійності для описування: напрацювання на відмову складних технічних систем (наприклад, електронної техніки, засобів авоніки, транспортних систем та інших виробів); процесів відновлення; відмов, які виникають у результаті зношування; напрацювання при швидкому вигоранні "ненадійних"елементів; відмов, які викликані втомленістю матеріалів.

**Визначення**. Випадкова величина T (T $>$ 0) має логарифмічно нормальний (логнормальний) розподіл, якщо її натуральний логарифм $\ln(T)$ підпорядкований нормальному закону:

$$F = P(\ln(T) < \ln(t)) = \begin{cases} 0 & \text{if } t < 0; \\ \dfrac{1}{\sigma\sqrt{2\pi}} \displaystyle\int_0^{\ln(t)} \exp\Big(-\frac{(x-\mu)^2}{2\sigma^2}\Big)dx & \text{if } t \geq 0. \end{cases},$$

де $\mu$ і $\sigma$ – параметри, які оцінюються за статистичними даними. У даному випадку $\mu$ — математичне сподівання логарифму випадкової величини; $\sigma$ – середнє квадратичне відхилення логарифму випадкової величини.

Функцію розподілу $F = F(t)$ та імовірність $P = P(t)$ безвідмовної роботи об'єкту можна визначити через значення квантиля

$$U = \frac{\ln(t) - \mu}{\sigma}$$

таким чином

$$F = \frac{1}{2} + \Phi\Big(\frac{\ln(t) - \mu}{\sigma}\Big), \quad P = \frac{1}{2} - \Phi\Big(\frac{\ln(t) - \mu}{\sigma}\Big),$$

де $\Phi = \Phi(t)$ — інтеграл Лапласа.

Функція щільності розподілу ймовірностей логнормального розподілу $f = f(x)$ має вигляд (за правилом диференціювання інтеграла, що залежить від параметра)

$$f = \begin{cases} 0 & \text{if } t < 0; \\ \dfrac{1}{\sigma\sqrt{2\pi}t} \exp\Big(-\dfrac{(\ln(t) - \mu)^2}{2\sigma^2}\Big) & \text{if } t \geq 0. \end{cases}$$

Параметри $\mu$ та $\sigma$ — логнормального закону пов'язані з математичним

сподіванням $M$ та дисперсією $D$ випадкової величини $T$ такими виразами:

$$M = \exp\left(\mu + \frac{\sigma^2}{2}\right), \quad D = \left(\exp(\sigma^2) - 1\right)\exp(2\mu + \sigma^2)$$

Величина $M$ у середнє напрацювання до відмови, середній ресурс, середній термін служби, середній час відновлення тощо.

Для позитивних випадкових величин логнормальний розподіл може приймати різноманітні форми, які можна отримати варіацією прарметрів $\mu$ та $\sigma$.

Логарифмічно-нормальний розподіл має такі властивості

- логнормальний розподіл має одну моду, якщо $t = \exp(\mu \breve{} \sigma)$ та медіану, якщо $t = \exp(\mu)$. Розподіл має позитивну асиметрію;
- інтенсивність відмов логнормального розподілу має немонотонний характер зі зменшенням у кінці розподілу.

У модулі scipy є бібліотека scipy.stats здатна генерувати розподіли випадкових величин того чи іншого походження, зокрема lognorm. Використаємо її для геренації логнормального розподілу деякої випадкової величини lognorm_values та запису його до файлу lognorm.csv.

```
import numpy as np
from scipy.stats import lognorm
import csv

#make this example reproducible
np.random.seed(1)
#generate log-normal distributed
#random variable with 1000 values

init = 10
sigma = 0.25
mu = 0.75
#mu = np.exp(mu)

lognorm_values = lognorm.rvs(sigma, scale=mu,
loc=init, size=1000, random_state=None)

file_csv=r'E:\Edu_box\PythonBox\pythonProject\lognorm.csv'
with open(file_csv, mode='w') as f:
    csvwriter = csv.writer(f,delimiter=',',
        lineterminator='\r')
```

```
    csvwriter.writerows(map(lambda x: [x], lognorm_values))
f.close()
```

У функції *lognorm.rvs*() *sigma* — стандартне відхилення, а значення *mu* — середнє значення логнормального розподілу, який згенерується.

Далі будемо вважати, що у файлі lognorm.csv зосереджені дані про деякий випадковий процес і нам потрібно побудувати його гістограму та графік щілності логнормального розподілу (див.рис. (8.1)). Для цього використаємо такий код

```
import numpy as np
import matplotlib.pyplot as plt
import csv

init = 10
sigma = 0.25
mu = 0.75
#mu = np.exp(mu)

file_csv=r'E:\Edu_box\PythonBox\pythonProject\lognorm.csv'
with open(file_csv, 'r') as f:
  csvreader = csv.reader(f, delimiter=',')
  this_list = []
  for row in csvreader:
      this_list.append(float(row[0]))

count, bins, ignored = plt.hist (this_list, density= True,
edgecolor='black', bins=50)

def fn_pdf(x):
    return lognorm.pdf(x,sigma,scale=mu,loc=init)

plt.plot(bins,fn_pdf(bins), linewidth=2, color='r')
plt.ylabel('Функція розподілу')
plt.xlabel('Випадкова величина $lognorm\_values$')
plt.title('Нормальний розподіл \
випадкової величини', fontsize= 14)
plt.legend(['init = 10, mu = 0.2, sigma = 0.25'])
f.close()
```


Рис. 8.1 — Приклад логнормального розподілу випадкової величини $x$

## 8.2. Розподіл Вейбулла—Гнеденко

В інженерній практиці розподіл Вейбулла—Гнеденко займає одне з центральних місць при дослідженнях характеристик надійності складних технічних систем і машин. Він є найбільш загальним розподілом безвідмовної роботи елементів, тривалості роботи машини до граничного стану, для опису розподілів термінів служби машин і характеристик втомленості металів, які призводять до відмови.

Розподіл Вейбулла—Гнеденко може бути трипараметричним або двопараметричним.

Розглянемо далі трипараметричний розподіл. Інтегральна та диференціальна функції цього розподілу визначаються такими залежностями

$$F = \begin{cases} 0 & \text{if } t < 0; \\ 1 - \exp\left(-\left(\dfrac{t-\mu}{\sigma}\right)^{\alpha}\right) & \text{if } t \geq 0. \end{cases},$$

$$f = \begin{cases} 0 & \text{if } t < 0; \\ \dfrac{\alpha}{\sigma}\left(\dfrac{t-\mu}{\sigma}\right)^{\alpha-1} \exp\left(-\left(\dfrac{t-\mu}{\sigma}\right)^{\alpha}\right) & \text{if } t \geq 0. \end{cases}$$

де $\sigma$ — параметр масштабу, який характеризує ступінь розтягнення кривої розподілу вздовж осі $t$ і пов'язаний із середнім значенням випадкової величини; $\alpha$ — параметр форми; $\mu > 0$ — параметр зміщення, який є мінімально можливим значенням випадкової величини .

Для цього розподілу, доповнення інтегральної функції (імовірність безвідмовної

роботи) має вигляд

$$P(t) = \exp\Big(-\Big(\frac{t-\mu}{\sigma}\Big)^{\alpha}\Big),$$

Інтенсивність подій визначається через диференціальну функцію, доповнення до інтегральної функції має вигляд

$$\lambda = \frac{f(t)}{P(t)} = \frac{\frac{\alpha}{\sigma}\Big(\frac{t-\mu}{\sigma}\Big)^{\alpha-1}\exp\Big(-\Big(\frac{t-\mu}{\sigma}\Big)^{\alpha}\Big)}{\exp\Big(-\Big(\frac{t-\mu}{\sigma}\Big)^{\alpha}\Big)} = \frac{\alpha}{\sigma}\Big(\frac{t-\mu}{\sigma}\Big)^{\alpha-1}$$

Математичне сподівання для трипараметричного розподілу Вейбула визначається інтегралом

$$T_{avg} = \int_0^{\infty} P(t)dt = \int_0^{\infty} \exp\Big(-\Big(\frac{t-\mu}{\sigma}\Big)^{\alpha}\Big)dt$$

Розрахунок цього інтегралу дає можливість записати такий вираз

$$T_{avg} = \sigma\Gamma\Big(1+\frac{1}{\alpha}\Big) + \mu,$$

де $\Gamma = \int_0^{\infty} x^{t-1}e^{-x}dx$ — гама-функція.

Розподіл Вейбула займає проміжне положення між нормальним і експоненціальним розподілами.

Розподіл Вейбулла широко використовується в аналізі великих даних різного походження завдяки своїй універсальності. Залежно від значень параметрів, розподіл Вейбулла може бути використаний для моделювання як в теорії надійності технічних систем так і різноманітної життєвої поведінки.

Різні значення параметрів розподілу можуть мати помітний вплив на поведінку розподілу. Насправді, деякі значення параметра форми призводять до того, що рівняння розподілу зводяться до рівнянь інших розподілів.

Побудуємо гістограму та графік щілності розподілу Вейбулла (див.рис. (8.2)). Для цього використаємо такий код

```
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import weibull_min

mu = 2; sigma = 1; b = 5.

wb_values = weibull_min.rvs(b, loc=mu,
scale=sigma, size=1000)
```

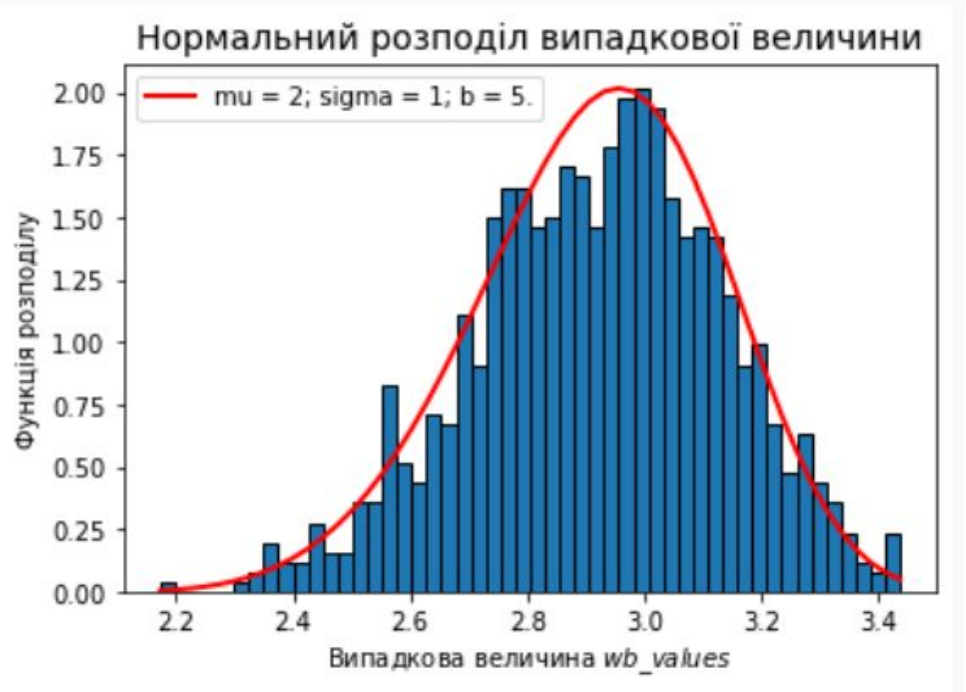


Рис. 8.2 — Приклад розподілу Вейбулла—Гнеденко випадкової величини $x$

```
count, bins, ignored = plt.hist(wb_values, density= True,
edgecolor='black', bins=50)

def weib(x,mu,a,b):
    return (b/a)*((x-mu)/a)**(b-1)*np.exp(-((x-mu)/a)**b)

scale = count.max()/weib(bins,mu,sigma,b).max()
plt.plot(bins, weib(bins,mu,sigma,b)*scale,
linewidth=2, color='r')

plt.ylabel('Функція розподілу')
plt.xlabel('Випадкова величина $wb\_values$')
plt.title('Нормальний розподіл випадкової величини',
fontsize= 14)
plt.legend(['mu = 2; sigma = 1; b = 5.'])
plt.show()
```

**Завдання**. Побудувати графіки функцій $P(t)$, $f(t)$, $\lambda(t)$ для наброру знань параметрів $\mu = 1$, $\sigma = 1$, $\alpha = (1, 2, 3, 0.5)$.

## Експоненціальний закон розподілу

Експоненціальний закон розподілу є окремим випадком закону розподілу Вейбулла при параметрі $\alpha = 1$. Він часто вживається в теорії масового

обслуговування, теорії надійності та в інших галузях науки.

Експоненціальний розподіл визначається одним параметром $\lambda = \dfrac{1}{T_{avg}}$ інтенсивністю подій (відмов, відновлення працездатності та ін.), де $T_{avg} = \bar{t}$ — середнє значення (математичне сподівання) випадкової величини $t$.

Інтенсивністю подій називається середня кількість подій, що з'явилися в одиницю часу. При $\lambda = const$ час появи подій має експоненціальний розподіл. Час появи раптових відмов має також експоненціальний розподіл.

Експоненціальний закон розподілу визначений в області додатних дійсних чисел $\geq 0$. Аналітичні вирази інтегральної та щільності функції експоненціального розподілу мають вигляд

$$F = 1 - e^{-\lambda t}, \quad f = \lambda e^{-\lambda t},$$

де $\lambda = \dfrac{1}{T_{avg}} = \dfrac{1}{\bar{t}} = \dfrac{1}{\sigma}$ — параметр, обернено пропорційний середньому значенню та стандарту відхиленню.

Характерною особливістю експоненціального розподілу є рівність середнього значення та стандартого відхилення, а також незмінні значення коефіцієнта варіації $v = \dfrac{\sigma}{\bar{t}} = 1$ та коефіцієнта асиметрії $A = \dfrac{1}{\sigma^3}\int\limits_0^\infty (t - \bar{t})^3 f(t)dt = 2$.

Побудуємо гістограму та графік щілності експоненціального розподілу (див.рис. (8.3)). Для цього використаємо такий код

```
import numpy as np
from scipy.stats import expon
import matplotlib.pyplot as plt

mu = 40 # mean
#generate exponential distribution with sample size 10000
exp_values = expon.rvs(scale=mu, size=10000)

#create plot of exponential distribution
count, bins, ignored = plt.hist(exp_values, density= True,
edgecolor='black', bins=50)

def fexp(x,mu):
    return (1/mu)*np.exp(-x/mu)

scale = count.max()/fexp(bins,mu).max()
plt.plot(bins, fexp(bins,mu)*scale, linewidth=3, color='r')
```

```
plt.ylabel('Функція розподілу')
plt.xlabel('Випадкова величина $exp_values$')
plt.title('Експоненціальний розподіл випадкової величини',
fontsize= 14)
plt.legend(['mu = 40'])
plt.show()
```

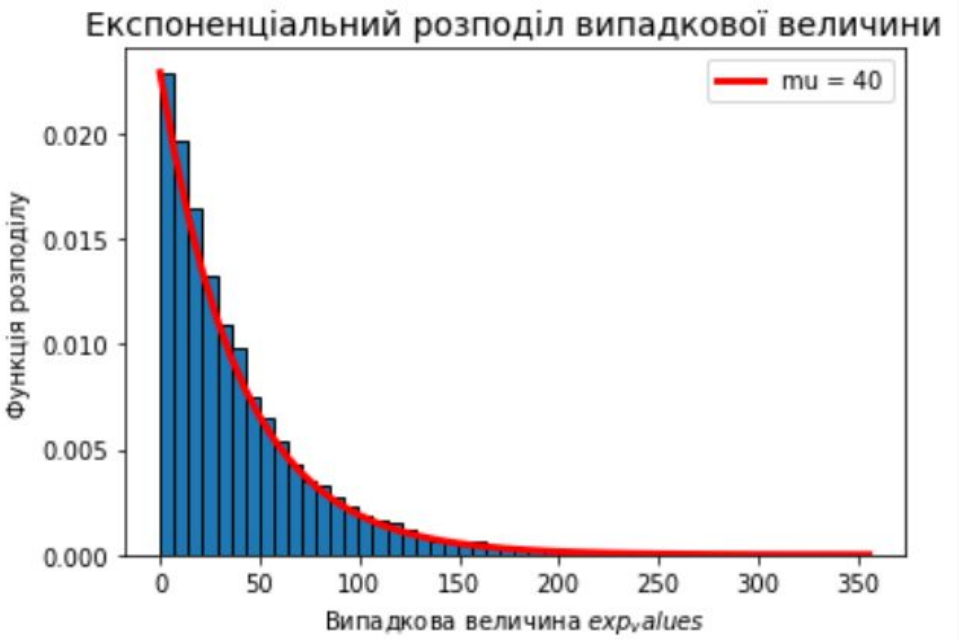


Рис. 8.3 — Приклад експоненціального розподілу випадкової величини *exp_values*

## 8.3. Функції щілності ймовірності для емпіричного набору випадкових величин

Підбір функції щілності ймовірності для емпіричного набору випадкових величин охоплює кілька методів і підходів. Ось деякі з них:

- **Аналітичний підхід**: У деяких випадках можна використовувати знання про систему або процес, що породжує випадкові величини, щоб аналітично вивести або припустити форму функції щільності ймовірності. Наприклад, якщо відомо, що випадкові величини розподілені нормально або експоненціально, можна використовувати відповідні аналітичні функції щільності ймовірності.
- **Використання відомих розподілів**: У деяких випадках можна використовувати відомі розподіли, які найкращим чином відповідають набору випадкових величин. Наприклад, якщо дані схожі на нормальний розподіл, можна використовувати нормальну функцію щільності ймовірності для моделювання даних.

- **Використання методу найменших квадратів**: Цей метод може використовуватися для апроксимації функції щільності ймовірності за допомогою комбінації базових функцій (наприклад, поліномів). Можна мінімізувати суму квадратів різниці між значеннями функції щільності ймовірності та фактичними спостережуваними значеннями.

- **Непараметричні методи**: Замість використання заздалегідь визначеної функції щільності ймовірності можна використовувати непараметричні методи, такі як ядерна оцінка щільності або згладжування даних. Ці методи дають змогу гнучко моделювати дані без необхідності припущень про конкретну форму функції щільності ймовірності.

- **Використання емпіричної функції розподілу**: У деяких випадках можна скористатися емпіричною функцією розподілу, яка будується на основі спостережуваних даних. Потім можна чисельно диференціювати емпіричну функцію розподілу для отримання наближеної функції щільності ймовірності.

Важливо зазначити, що вибір підходу і методу залежить від характеру даних і доступної інформації про систему або процес. У деяких випадках може знадобитися експериментальне дослідження й тестування різних функцій щільності ймовірності для досягнення найкращої відповідності даним.

## 8.4. Використання тесту Колмогорова-Смірнова для визначення функції щільності ймовірності

Тест Колмогорова-Смірнова є статистичним тестом, який використовується для перевірки гіпотези про те, що дві вибірки (або розподілу) були отримані з однієї і тієї ж генеральної сукупності. Він може використовуватися для перевірки відповідності емпіричної функції розподілу (ЕФР) спостережуваної вибірки з теоретичною функцією розподілу або для порівняння двох емпіричних функцій розподілу.

Суть тесту Колмогорова-Смірнова полягає в порівнянні максимального відхилення між емпіричною функцією розподілу спостережуваної вибірки та теоретичною функцією розподілу. Якщо відмінність між ними досить мала, то немає підстав відхилити нульову гіпотезу про те, що вибірки було отримано з однієї й тієї самої генеральної сукупності або вони мають один і той самий розподіл.

Результатом тесту Колмогорова-Смірнова є значення статистики Колмогорова (D-статистика) і відповідне йому p-значення. Якщо p-значення менше за обраний рівень значущості, зазвичай 0.05, то нульова гіпотеза відкидається

на користь альтернативної гіпотези про відмінності між вибірками або розподілами.

Тест Колмогорова-Смірнова широко застосовують у різних галузях статистики та аналізу даних для порівняння вибірок, оцінювання відповідності даних теоретичним моделям або перевірки гіпотез щодо розподілів даних.

**Приклад №1** використання тесту Колмогорова-Смірнова для визначення функції щільності ймовірності деякого набору випадкових величин на Python:

```
import numpy as np
from scipy.stats import kstest
"""Згенеруємо випадкові величини з відомого розподілу
для порівняння known_distribution =
np.random.normal(loc=0, scale=1, size=1000)"""
# Згенеруємо набір випадкових величин для аналізу
data = np.random.normal(loc=0, scale=1, size=500)
#data = np.random.exponential(scale=1, size=500)
"""Виконаємо тест Колмогорова–Смірнова
для порівняння з відомим розподілом"""
ks_statistic, p_value = kstest(data, 'norm', args=(0, 1))
# Виведемо результати тесту
print("Статистика теста Колмогорова–Смірнова:",
      ks_statistic)
print("P–значение:", p_value)
#Оцінимо результати тесту і зробимо висновки
alpha = 0.05 # Рівень значущості
if p_value > alpha:
    print("Немає підстав відкинути гіпотезу "+
          "про відповідність даних відомому розподілу.")
else:
    print("Гіпотеза про відповідність даних "+
          "відомому розподілу відкидається.")
```

У цьому прикладі ми згенерували випадкові величини з відомого нормального розподілу (known_distribution) і випадкові величини для аналізу (data). Потім ми використовували функцію kstest з модуля scipy.stats для виконання тесту Колмогорова-Смірнова.

Тут і далі параметр ks_statistic у тесті Колмогорова-Смірнова (отримані значення ks_statistic називають статистикою тесту) являє собою максимальну різницю між емпіричною функцією розподілу (ECDF) двох вибірок, які порівнюються, а величина p_value - відповідне p-значення, яке порівнюється з рівнем значущості alpha.

Коли ми виконуємо тест Колмогорова-Смірнова, ми порівнюємо дві вибірки даних та їхні розподіли. Емпірична функція розподілу (ECDF) використовується для оцінки їхніх розподілів. ECDF являє собою функцію, яка показує частку значень вибірки, що менші або дорівнюють даному значенню.

KS-статистика є максимальною різницею між значеннями ECDF двох вибірок. Вона вимірює найбільше відхилення між двома розподілами і слугує мірою схожості або відмінності між ними. Чим більше значення KS-статистики, тим більша відмінність між розподілами двох вибірок.

Під час виконання тесту Колмогорова-Смірнова ми порівнюємо значення KS-статистики з критичним значенням для певного рівня значущості. Якщо KS-статистика перевищує критичне значення, то гіпотеза про рівність розподілів відкидається.

Таким чином, значення ks_statistic дає нам змогу оцінити максимальне відхилення між розподілами двох вибірок і ухвалити рішення про рівність або відмінність цих розподілів.

Виводимо результати тесту і робимо висновки на основі рівня значущості alpha. Якщо p-значення більше рівня значущості, то немає підстав відкидати гіпотезу про відповідність даних відомому розподілу. В іншому разі, гіпотезу про відповідність даних відкидають.

Зверніть увагу, що в прикладі використано нормальний розподіл як відомий розподіл ('norm'), але ви можете адаптувати код для використання інших розподілів, залежно від ваших потреб.

**Приклад №2** використання тесту Колмогорова-Смірнова для визначення функції щільності ймовірності деякого набору випадкових величин на Python:

```
import numpy as np
from scipy.stats import kstest
"""Згенеруємо випадкові величини
з відомого розподілу для порівняння known_distribution =
np.random.exponential(scale=1, size=1000)"""
# Згенеруємо набір випадкових величин для аналізу
data = np.random.exponential(scale=1, size=500)
#data = np.random.normal(loc=0, scale=1, size=500)
"""Виконаємо двопараметричний тест Колмогорова–Смірнова
для порівняння з відомим розподілом"""
ks_statistic, p_value = kstest(data, 'expon', args=(0, 1))
# Виведемо результати тесту
print("Статистика теста Колмогорова–Смірнова:",
        ks_statistic)
```

```
print("P–значение:", p_value)
# Оцінимо результати тесту і зробимо висновки
alpha = 0.05 # Рівень значущості
if p_value > alpha:
    print("Немає підстав відкинути гіпотезу "+
        "про відповідність даних відомому розподілу.")
else:
    print("Гіпотеза про відповідність даних "+
        "відомому розподілу відкидається.")
```

У цьому прикладі ми згенерували випадкові величини з відомого експоненціального розподілу (known_distribution) і випадкові величини для аналізу (data). Потім ми використовували функцію kstest з модуля scipy.stats для виконання тесту Колмогорова-Смірнова.

Параметри args=(0, 1) передають у функцію kstest для вказівки параметрів відомого розподілу (у цьому випадку експоненціального розподілу).

Отримані значення ks_statistic являють собою статистику тесту, а p_value - відповідне p-значення. Потім ми виводимо результати тесту і робимо висновки на основі рівня значущості alpha. Якщо p-значення більше рівня значущості, то немає підстав відкидати гіпотезу про відповідність даних відомому розподілу. В іншому разі, гіпотезу про відповідність даних відкидають.

Зверніть увагу, що в прикладі використано експоненціальний розподіл як відомий розподіл ('expon'), але ви можете адаптувати код для використання інших розподілів, залежно від ваших потреб.

**Приклад №3** використання тесту Колмогорова-Смірнова для визначення функції щільності ймовірності деякого набору випадкових величин з розподілом Вейбулла–Гнеденко на Python:

```
from scipy.stats import weibull_min
from scipy.stats import kstest
""" Згенеруємо набір випадкових величин для аналізу з
бібліотеки weibull_min """

mu = 2; sigma = 1; b = 5.
data = weibull_min.rvs(b, loc=mu, \
scale=sigma, size=1000)
#data = np.random.exponential(scale=1, size=500)
"""Виконаємо тест Колмогорова–Смірнова
для порівняння з відомим розподілом"""
ks_statistic, p_value = kstest(data, weibull_min.rvs(5, loc=2, \
```

```
scale=1, size=5000))
# Виведемо результати тесту
print("Статистика теста Колмогорова–Смірнова:",
      ks_statistic)
print("P–значение:", p_value)
#Оцінимо результати тесту і зробимо висновки
alpha = 0.05 # Рівень значущості
if p_value > alpha:
    print("Немає підстав відкинути гіпотезу "+
          "про відповідність даних відомому розподілу.")
else:
    print("Гіпотеза про відповідність даних "+
          "відомому розподілу відкидається.")
```

**Приклад №4** використання функції ks_2samp з бібліотеки scipy.stats для порівняння двох наборів випадкових величин на Python:

```
import numpy as np
from scipy.stats import ks_2samp
# Згенеруємо два набори випадкових величин для порівняння
data1 = np.random.normal(loc=0, scale=1, size=1000)
data2 = np.random.normal(loc=0.5, scale=1, size=800)
"""Виконаємо двовибірковий тест Колмогорова–Смірнова
для порівняння двох наборів даних"""
ks_statistic, p_value = ks_2samp(data1, data2)
# Виведемо результати тесту
print("Статистика теста Колмогорова–Смірнова:",
      ks_statistic)
print("P–значение:", p_value)
# Оцінимо результати тесту і зробимо висновки
alpha = 0.05 # Рівень значущості
if p_value > alpha:
    print("Немає підстав відкинути "+
        "гіпотезу про рівність розподілів.")
else:
    print("Гіпотеза про рівність "+
        "розподілів відкидається.")
```

У цьому прикладі ми згенерували два набори випадкових величин data1 і data2 з нормального розподілу з різними параметрами. Потім ми використовували функцію ks_2samp з модуля scipy.stats для виконання двовибіркового тесту Колмогорова-Смірнова.

Функція ks_2samp повертає значення ks_statistic, що представляють собою

статистику тесту, і p_value, відповідне p-значення. Потім ми виводимо результати тесту і робимо висновки на основі рівня значущості alpha. Якщо p-значення більше рівня значущості, то немає підстав відкидати гіпотезу про рівність розподілів. В іншому разі, гіпотезу про рівність розподілів відкидають.

Зверніть увагу, що в прикладі використано нормальний розподіл для генерації даних, але ви можете адаптувати код для використання інших розподілів і різних наборів даних, залежно від ваших потреб.

### 8.5. Питання для самоперевірки

# 9. Метод Монте-Карло у наукових обчисленнях з використанням мови програмування Python

Уявімо, що ми провели $N$ експериментів і результат кожного експерименту випадкове число. У цих $N$ експериментах деяка подія трапляється $M$ разів. Оцінка ймовірності тоді $M/N$ і вона стає більш точною при прагненні числа експериментів $N$ до нескінченності (зауважте, що дріб при цьому не обертається на нуль, оскільки і число $M$ також прагне до нескінченності).

Математичний метод, що полягає у використанні великої кількості генерованих випадкових чисел, отримав назву метод Монте-Карло. Цей метод виявився надзвичайно корисним у науці та промисловості, там, де випадкову поведінку не можна не враховувати, або, наприклад, там, де не мається на увазі наявність випадкових чисел - у випадках складного інтегрування.

### 9.1. Звичайний алгоритм інтегрування Монте-Карло

Одним із найбільш ранніх застосувань генерованих випадкових чисел було обчислення інтегралів.

Припустимо, потрібно обчислити певний інтеграл

$$I = \int_a^b f(x)dx \tag{9.1}$$

Розглянемо випадкову величину $u$, рівномірно розподілену на відрізку інтегрування $[a, b]$. Тоді $f(u)$ також буде випадковою величиною, причому її математичне очікування виражається як

$$M(f(u)) = \int_a^b f(x)\varphi(x)dx$$

де $\varphi(x)$ — щільність розподілу випадкової величини $u$. А оскільки $u$ рівномірно розподілена на $[a, b]$, то

$$\varphi(u) = \frac{1}{b-a}$$

Таким чином, шуканий інтеграл виражається як

$$\int_a^b f(x)dx = (b-a)M(f(u))$$

але математичне очікування випадкової величини $f(u)$ можна легко оцінити, змоделювавши цю випадкову величину і вважаючи вибіркове середнє.

Отже, вибираємо N точок, рівномірно розподілених на $[a, b]$, для кожної точки $u$ обчислюємо $f(u)$. Потім знаходимо вибіркове середнє:

$$\overline{f(u)} = \frac{1}{N}\sum_{i=1}^{N} f(u_i)$$

Часто у методі Монте-Карло використовує випадкову величину $y$, рівномірно розподілену на відрізку $[0, 1]$. У разі вибіркове середнє трохи зміниться. Необхідно перетворити випадкову величину $y$ те щоб вона мала рівномірне розподіл на відрізку $[a, b]$. Тоді формула вище набуде наступного вигляду:

$$\overline{f(y)} = \frac{1}{N}\sum_{i=1}^{N} f(a + (b-a)y_i)$$

У результаті отримуємо оцінку інтегралу:

$$\int_a^b f(x)dx = \frac{(b-a)}{N}\sum_{i=1}^{N} f(u_i)$$

Точність оцінки залежить від кількості точок. Чим більша кількість $N$, тим точніше значення інтеграла. Цей метод зазвичай називається інтегрування за Монте-Карло. Ми можемо представити вираз (9.1) у вигляді невеликої програми

```
import random as random_number

def fn_MCint(f,a,b,n):
    s = 0
    for i in range(n):
        x = random_number.uniform(a, b)
        s += f(x)
    I = (float(b-a)/n)*s
```

```
    return I
```

Зазвичай, достатня точність методу задається більшим числом $n$, тому векторизована версія буде зручнішою:

```
from numpy import *

def fn_MCint_vec(f,a,b,n):
    x = random.uniform(a, b, n)
    s = sum(f(x))
    I = (float(b-a)/n)*s
    return I
```

Розглянемо інтегрування методом Монте-Карло на прикладі простої лінійної функції $f = 2x+3$, межі інтегрування - від 1 до 2. Було б цікаво подивитися, як метод справляється з розв'язанням задачі для різних $n$. Оцінку зробимо наступним трохи зміненим fn_MCint методом:

```
def fn_MCint2(f,a,b,n):
    s = 0

    I = zeros(n)
    for k in range(1, n+1):
        x = random_number.uniform(a, b)
        s += f(x)
        I[k-1] = (float(b-a)/k)*s
    return I
```

Зауважимо, що $k$ змінюється від 1 до $n$, тоді як індекси $i$ як і раніше йдуть від 0 до $n-1$. Оскільки $n$ може бути дуже великим, масив $I$ може переповнити пам'ять. Тому слід записувати лише кожне $N$ — те значення апроксимації. Це можливо за допомогою відомої нам функції визначення залишку

```
for k in range(1, n+1):
    ...
    if k % N == 0:
        # store
```

Отже, щоразу, коли $k$ ділиться на $N$ без залишку, ми записуємо значення (у нашому випадку кожне соте). Відповідна функція представлена нижче.

```
def fn_MCint3(f,a,b,n,N=100):
    '''Зберігає кожне N наближення інтеграла
    у масив I і записуємо відповідне значення k'''
    s = 0
```

```
    I_values = []
    k_values = []
    for k in range(1, n+1):
        x = random_number.uniform(a, b)
        s += f(x)
        if k % N == 0:
            I = (float(b-a)/k)*s
            I_values.append(I)
            k_values.append(k)
    return k_values, I_values
```

Тепер у нас є інструмент для того, щоб подивитися, як змінюється помилка в інтегруванні методом Монте-Карло зі зростанням $n$. Закінчена програма має такий вигляд

```
import random as random_number
import matplotlib.pyplot as plt
from numpy import array

def fn_MCint3(f,a,b,n,N=100):
    '''Зберігає кожне N наближення інтеграла
    у масив I і записуємо відповідне значення k'''
    s = 0

    I_values = []
    k_values = []
    for k in range(1, n+1):
        x = random_number.uniform(a, b)
        s += f(x)
        if k % N == 0:
            I = (float(b-a)/k)*s
            I_values.append(I)
            k_values.append(k)
    return k_values, I_values

def f1(x):
    return 2 + 3*x

k, I = fn_MCint3(f1,1,2,1000000,N=10000)
error = 6.5 - array(I)

plt.title('Інтегрування Монте-Карло')
```

```
plt.xlabel('n')
plt.ylabel('error')
plt.plot(k, error)
plt.show()
```

Результат роботи програми (може дещо відрізнятися через випадковість) подано на Рис. 9.1.

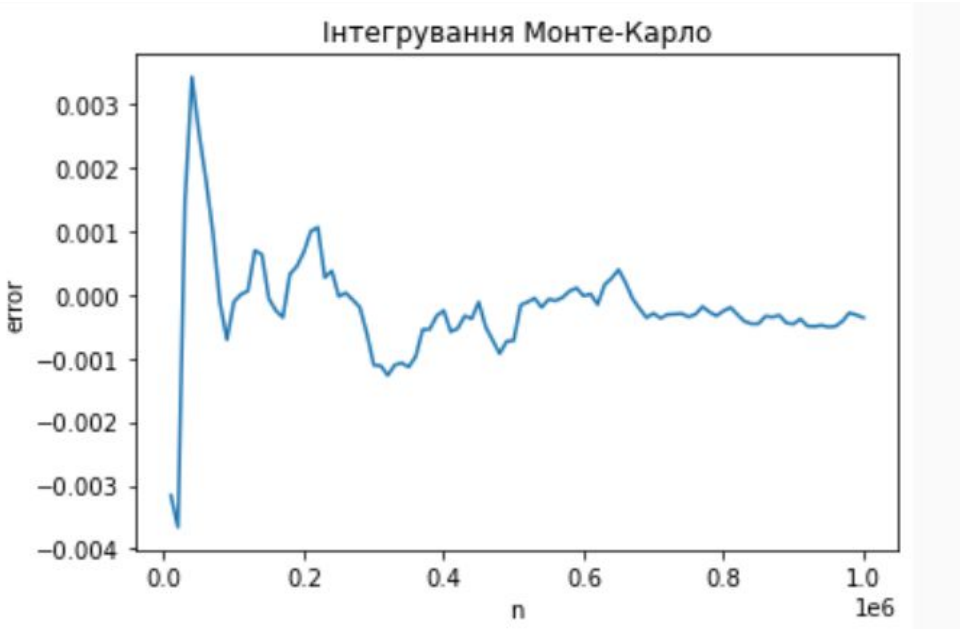


Рис. 9.1 —

## 9.2. Інтегрування через обчислення площі під кривою

Уявімо, у нас є деяка складна фігура $G$, площу якої ми хочемо обчислити. Цю фігуру ми можемо укласти в прямокутник $B$, який визначається координатами $[X1, X2] \times [Y1, Y2]$. Далі в цей прямокутник ми "кидаємо"випадкові числа із заданого масиву $[X, Y]$ і рахуємо скільки точок потрапило всередину контуру фігури $G$. Далі, знаючи відношення числа точок, що "впали"в контур $G$ до загального числа "кинутих"точок, і помноживши це відношення на площу простої фігури $B$, ми отримуємо площу складної фігури $G$.

Таким же чином ми могли б знаходити інтеграл від функції, оскільки він і визначає площу під нею. Уявімо, ми знаємо, як виглядає функція, область інтегрування лежить у межі від $a$ до $b$, а максимум функції в зазначеному інтервалі дорівнює $m$. Тоді аналогічно, попередньому розгляду, інтеграл буде виражений через число "викинутих"точок $N$ і число тих, що впали "під функцію"$M$ як $\frac{M}{N} m(b-a)$.

Векторизований розв'язок може в найпростішому випадку має такий вигляд

```
from numpy import random

def fn_MCint_area_vec(f,a,b,m,N):
    x = random.uniform(a,b,N)
    y = random.uniform(0,m,N)
    M = y[y < f(x)].size # у квадратних дужках умова,
        # size рахує число елементів, що їй задовольняють
    area = M/float(N)*m*(b-a)
    return area
```

Рішення можна написати і через стандартний random у звичайному вигляді за допомогою циклу, але швидкість його буде істотно меншою.

## Список використаної літератури


1. Курс теорії ймовірностей / Б.В.Гнєденко – К.: Київський університет, 2010. – 463 с.
2. Теорія ймовірностей та математична статистика / В.В.Барковський, Н.В.Барковська, О.К.Лопатін. – К.: ЦУЛ, 2002. – 448 с.
3. Теорія ймовірностей та математична статистика: навч. посіб. / О.І.Кушлик - Дивульська, Н.В.Поліщук, Б.П.Орел, П.І.Штабалюк. – К: НТУУ «КПІ», 2014. – 205 с.
4. Теорія ймовірностей. Збірник задач / А.Я.Дороговцев, Д.С.Сільвестров, А.В.Скороход, М.Й.Ядренко. – Київ: Вища школа, 1980. – 432 с.
5. Теорія ймовірностей у прикладах і задачах / І.Ю.Каніовська. – К.: ІВЦ "Видавництво «Політехніка» ТОВ "Фірма «Періодика» 2004. – 156 с.
6. Інженерний аналіз експериментальних даних. Методичні вказівки до самостійної роботи студентів / Пашинський В.А. – Кропивницький: ЦНТУ, 2017. – 82 с.
7. Програмування числових методів мовою Python: підруч. / А.В.Анісімов, А.Ю.Дорошенко, С.Д.Погорілий, Я.Ю.Дорогий; за ред. А.В.Анісімова. – К.: Видавничо-поліграфічний центр "Київський університет 2014. – 640с.