МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

ТЕРНОПІЛЬСЬКИЙ НАЦІОНАЛЬНИЙ ЕКОНОМІЧНИЙ УНІВЕРСИТЕТ

МАТЕМАТИКА ДЛЯ ЕКОНОМІСТІВ

ч. II

«ТЕОРІЯ ЙМОВІРНОСТЕЙ І МАТЕМАТИЧНА СТАТИСТИКА»

(***тексти лекцій і приклади розв'язування задач***)

Для студентів заочної форми навчання

Тернопіль — 2008

УДК 519.21..519.22
М 33

Рецензент **Кривень Василь Андрійович**
доктор технічних наук, професор, завідувач кафедри математичних методів в інженерії Тернопільського державного технічного університету ім. І. Пулюя.
Недашковський Микола Олександрович ,
доктор фізико-математичних наук, професор, завідувач кафедри автоматизованих систем і програмування Тернопільського національного економічного університету

Затверджено на засіданні кафедри економіко-математичних методів протокол №7 від 31.01.2008 р.

Єрьоменко В. О., Шинкарик М. І., Бабій Р. М., Мартинюк О. М., Мигович Ф. М.

М 33Математика для економістів. Ч. ІІ. Теорія ймовірностей і математична статистика (тексти лекцій і приклади розв'язування задач). Для студентів заочної форми навчання. — Тернопіль, 2008. — 144 с.

Посібник містить теоретичний матеріал з основних розділів теорії імовірностей та математичної статистики, а також приклади розв'язання типових задач.

Для студентів заочної форми навчання всіх економічних спеціальностей ВНЗ.

УДК 519.21..519.22

ПЕРЕДМОВА

Існують два погляди на математику і її роль серед інших наук. Згідно першого вважають, що математика — це щось самостійне. Другий це також визнає, але в основному вважає математику інструментом, оволодіння яким корисне і необхідне. Безперечно, математика має певне світоглядне значення, але для економіста-менеджера математика — це інструмент аналізу, організації, управління. В економіці у зв'язку з наближеним, випадковим характером даних більшість задач моделюються за допомогою імовірносних і статистичних методів. Тому виникає необхідність детального вивчення питань розділу «Теорія імовірностей і математична статистика» (ТІМС) курсу «Математика для економістів».

Пропоновані тексти лекцій охоплюють всі питання програми і допоможуть студентам-заочникам краще засвоїти основи розділу ТІМС. У кожній з лекцій наведено план, висвітлено теоретичні питання. Крім цього, дано зразки розв'язування типових задач. Слід відмітити, що більшість задач даного посібника мають економічний зміст, що допоможе студентам в майбутньому використовувати одержані знання .

Посібник складається з двох частин: І «Теорія імовірностей» і ІІ «Математична статистика». Нумерація формул та задач здійснюється по наростанню в межах параграфа і має дві позиції, перша з яких вказує на номер параграфа. При цьому нумерація параграфів для кожної із частин автономна. У випадку використання в другій частині посилання на формулу з першої частини до її нумерації добавляється ЧІ, що усуває можливу двозначність. Символи ○ та ● означають відповідно початок і завершення розв'язування задачі.

ЧАСТИНА ПЕРША
ТЕОРІЯ ІМОВІРНОСТЕЙ

§ 1. ВИЗНАЧЕННЯ ІМОВІРНОСТІ

1. *Події та їх види.*
2. *Класичне означення імовірності випадкової події. Властивості імовірностей.*
3. *Елементи комбінаторики в теорії імовірностей.*
4. *Відносна частота випадкової події. Статистична імовірність.*
5. *Геометрична імовірність.*

1. Під **випробуванням** будемо розуміти здійснення намічених дій і отримання результату при виконанні певного комплексу умов S. При цьому припускається, що ці умови є фіксованими; вони або об'єктивно існують, або створюються штучно і можуть бути відтворені необмежене число разів.

Результатом випробування є подія. Розрізняють події **достовірні**, **неможливі** та **випадкові**.

Достовірною називають подію, яка при випробуванні обов'язково відбувається. **Неможливою** називають подію, яка при випробуванні обов'язково не відбувається. **Випадкова** — це та подія, яка при випробуванні може як відбутися, так і не відбутися.

Достовірну подію позначимо літерою Ω, а неможливу — $\varnothing$.

Розглянемо деякі **властивості випадкових подій**.

Дві події називаються **несумісними** (**сумісними**), якщо при випробуванні відбуття однієї **виключає (не виключає)** відбуття іншої.

Сукупність випадкових подій утворює **повну групу**, якщо одна з них при випробуванні обов'язково відбувається, а ***будь-які*** дві події є несумісними.

Елементарними будемо називати найпростіші випадкові події, які можуть відбутися при випробуванні.

Події, «породжені» одним випробуванням, назвемо **рівноможливими**, якщо є підстави вважати, що жодна з них не є більш можливою, ніж інші.

2. Чисельну міру можливості відбуття випадкової події дає ймовірність цієї події.

Означення. Класичною імовірністю події A називається відношення числа елементарних рівноможливих подій, що сприяють появі події A, до загального числа елементарних

рівноможливих подій, що утворюють повну групу.

Кожна з елементарних випадкових подій, по суті, є одним із наслідків випробування. Такий підхід є корисним при аналізі задач.

З врахуванням цього зауваження аналітичний вираз класичного означення набере такого виду:

$$P(A) = \frac{m}{n}, \tag{1.1}$$

де $P(A)$ — класична імовірність події A;

m — число елементарних рівноможливих подій, що сприяють появі події A (число наслідків випробування, в яких відбувається подія А);

n — число елементарних рівноможливих подій, що утворюють повну групу (загальне число рівноможливих наслідків випробування).

Дане означення дозволяє сформулювати **основні властивості класичної імовірності**:

1) $P(\Omega) = 1$ (імовірність достовірної події дорівнює 1);

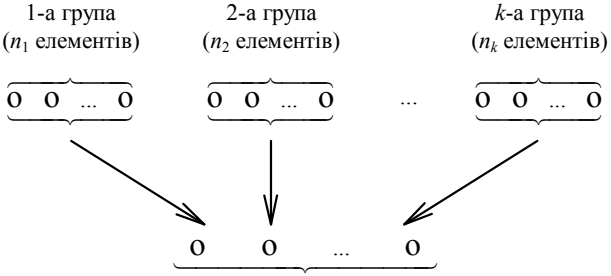
2) $P(\varnothing) = 0$ (імовірність неможливої події дорівнює 0);

3) якщо A — випадкова подія, тоді:

$$0 < P(A) < 1. \tag{1.2}$$

3. Для того, щоб мати деякі стандартні методи при розрахунках по схемі класичної імовірності, наведемо **основну формулу комбінаторики**, а також розглянемо поняття **комбінацій**, **розміщень** та **перестановок**.

Нехай є k груп елементів, чисельність кожної з яких відповідно дорівнює $n_1, n_2, ..., n_k$. Виберемо довільним чином по **одному** елементу з кожної групи:



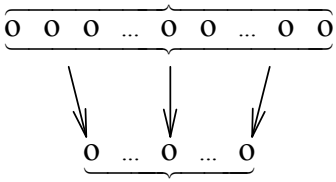
Тоді загальне число N способів, якими можна здійснити такий відбір, визначається співвідношенням

$$N = n_1 \cdot n_2 \cdot \ldots \cdot n_k, \tag{1.3}$$

яке називається **основною формулою комбінаторики**.

Розглянемо сукупність **різних** елементів довільної природи, чисельністю n. Будемо утворювати групи по m ($m \le n$) **різних** елементів із цієї сукупності:

n елементів



m елементів

Такі групи в теорії імовірностей часто називаються **вибірками**.

Нехай $m < n$. **Комбінаціями** називаються такі групи, які відрізняються одна від одної хоча б одним елементом. Загальне число комбінацій C_n^m (читається: це з n по m) знаходиться за формулою

$$C_n^m = \frac{n(n-1)\cdot(n-2)\cdot\ldots\cdot(n-m+1)}{m!}, \tag{1.4}$$

де $m! = m(m-1)\cdot\ldots\cdot 2\cdot 1$ (читається: m факторіал).

Зауваження. В чисельнику (1.4) є m співмножників. Якщо чисельник і знаменник помножити на $(n-m)!$, тоді отримається така рівність:

$$C_n^m = \frac{n!}{(n-m)!m!}. \tag{1.4*}$$

Нехай $m \le n$. **Розміщеннями** називаються такі групи, які відрізняються одна від іншої або хоча б одним елементом, або порядком розташування цих елементів в групі. Число розміщень A_n^m (читається: а з n по m) знаходиться за формулою:

$$A_n^m = \underbrace{n(n-1)\cdot(n-2)\cdot\ldots\cdot(n-m+1)}_{m \text{ співмножників}}. \tag{1.5}$$

Якщо в формулі (1.5) $m = n$, то A_n^m — число таких розміщень, які відрізняються тільки порядком розташування елементів, а не самими елементами. Такі розміщення називаються **перестановками**. Їх число P_n за формулою (1.5)

$$A_n^n = n(n-1)(n-2)\ldots 2\cdot 1 = n! = P_n,$$

тобто

$$P_n = A_n^n = n!. \tag{1.6}$$

Число n може набирати не тільки натуральні значення, воно може також дорівнювати нулю. Порожня множина (вибірка) є підмножиною довільної множини і природно вважати, що вона може бути впорядкована тільки одним способом. Тому вважається, що **0! = 1**.

Число комбінацій володіє такими властивостями:

1) $C_n^0 = C_n^n = 1$; 2) $C_n^1 = C_n^{n-1} = n$; 3) $C_n^m = C_n^{n-m}$;

4) $C_n^0 + C_n^1 + C_n^2 + \ldots + C_n^n = 2^n$.

Відмітимо, що числа розміщень, перестановок і комбінацій пов'язані рівністю

$$A_n^m = P_m C_n^m.$$

При розв'язуванні задач комбінаторики використовуються такі правила:

Правило суми. Якщо деякий об'єкт α може бути відібраний із сукупності об'єктів k способами, а другий об'єкт β може бути відібраний s способами, то відібрати або α, або β можна $k + s$ способами.

Правило добутку. Якщо об'єкт α можна вибрати із сукупності об'єктів k способами і після кожного такого відбору об'єкт β можна вибрати s способами, то пара об'єктів (α, β) у вказаному порядку може бути вибрана $k \cdot s$ способами.

4. Разом із імовірністю до основних понять теорії імовірностей належить відносна частота.

Відносною частотою випадкової події називається відношення числа випробовувань, в яких подія відбулася, до загального числа фактично проведених випробовувань. Тобто, відносна частота події A визначається формулою:

$$W(A) = \frac{M}{N}, \tag{1.7}$$

де M — число появ події A, N— загальне число випробовувань.

Співставлення означень імовірності і відносної частоти дозволяє зробити висновок: імовірність обчислюють до випробування (тобто вона є апріорною величиною), а відносну частоту — після випробування (апостеріорна величина).

Із означення (1.7) для **випадкової** події A випливає така подвійна нерівність (порівняйте з (1.2)!):

$$0 \le W(A) \le 1.$$

При невеликій кількості випробовувань відносна частота випадко-

вої події може помітно змінюватися від однієї серії випробовувань до іншої. Проте тривалі спостереження показали, що коли в однакових умовах проводиться достатньо велике число випробовувань, то відносна частота виявляє властивість **стійкості**. Ця властивість полягає у тому, що в різних серіях випробовувань відносна частота **змінюється мало** (тим менше, чим більше число випробовувань), коливаючись навколо деякого постійного числа. Виявилось, що це стале число є імовірністю випадкової події. Математичне формулювання цієї властивості стійкості буде дано в темі «Закон великих чисел» (теорема Я. Бернуллі).

Властивість стійкості відносної частоти, а також можливість, хоча б принципово, проводити необмежене число випробовувань відносно випадкової події дозволяють сформулювати **статистичне означення імовірності**: в якості статистичної імовірності випадкової події береться відносна частота цієї події.

5. Один із недоліків класичного означення імовірності, пов'язаний із неможливістю використання такого означення для випадку випробувань із нескінченним числом наслідків випробування, може бути усунутий з допомогою **геометричної імовірності** — імовірності попадання точки в область (відрізок, частину площини, частину тіла тощо).

Припустимо, що здійснюється випробування – на прямокутник Ω, в якому міститься довільна фігура ***A***, навмання кидається точка (рис. 1.1).

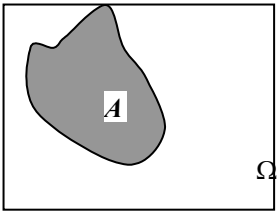


Рис.1.1.

При цьому вважається, що виконуються такі припущення: вона може опинитися в будь-якій точці прямокутника Ω, імовірність (можливість) попадання точки на фігуру ***A*** пропорційна площі цієї фігури і не залежить ні від розташування ***A*** відносно Ω, ні від форми ***A***. Хай випадкова подія A – точка попала в фігуру ***A***. Тоді **геометричне означення імовірності** події A дається рівністю:

$$P(A) = \frac{S(\boldsymbol{A})}{S(\Omega)} \quad , \tag{1.8}$$

де $S(\boldsymbol{A})$ — площа фігури $\boldsymbol{A}$, $S(\Omega)$ – площа прямокутника Ω.
Означення (1.8) є частинним випадком загального означення геометричної імовірності. Для $\Omega \subset \mathbb{R}^1$ *або* $\Omega \subset \mathbb{R}^3$.

$$P(A) = \frac{l(\boldsymbol{A})}{l(\Omega)} \text{ або } P(A) = \frac{V(\boldsymbol{A})}{V(\Omega)}. \tag{1.9}$$

РОЗВ'ЯЗУВАННЯ ТИПОВИХ ЗАДАЧ

Задача 1.1. Банк протягом місяця може видати в кредит позику п'ятьом своїм клієнтам, в той час як поступили замовлення на кредит від 15 клієнтів першого району і 10 клієнтів другого району. Для збереження клієнтів банк розглядає як тимчасову вимушену міру — розігрування випадковим чином п'яти позик серед тих, від кого поступило замовлення. Знайти імовірність того, що число клієнтів першого району, яким дістанеться позика, дорівнює: а) 5; б) 0; в) 3.

○ Випробування — проведення розігрування серед клієнтів банку. Наслідок випробування — п'ятірка клієнтів. Число наслідків випробування дорівнює числу всеможливих п'ятірок, які можна утворити із сукупності чисельністю 25 елементів (клієнтів банку).

Для таких груп елементів характерним є те, що вони відрізняються одна від іншої хоча б одним елементом. Крім того, як основна сукупність елементів (з 25 клієнтів), так і утворені групи по 5 елементів, складаються з **різних** елементів (відсутність повторів). Це дає підстави зробити висновок, що такі групи є комбінаціями. І їх число для всіх трьох випадків дорівнює:

$$n = C_{25}^5 = \frac{25 \cdot 24 \cdot 23 \cdot 22 \cdot 21}{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} = 53130.$$

а) A — власниками кредиту є п'ять клієнтів першого району. Число наслідків випробування, для яких відбувається подія A, дорівнює числу всеможливих п'ятірок клієнтів, котрі можна утворити із загального числа 15 клієнтів першого району. Ці п'ятірки знову є комбінаціями і їх загальне число

$$m = C_{15}^5 = \frac{15 \cdot 14 \cdot 13 \cdot 12 \cdot 11}{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} = 3003.$$

За класичним означенням імовірності

$$P(A) = \frac{m}{n} = \frac{3003}{53130} \approx 0,057.$$

б) B — власниками кредиту є п'ять клієнтів другого району. Число наслідків випробування m, для яких відбувається подія B, дорівнює

числу всеможливих п'ятірок клієнтів, що можна утворити із 10 клієнтів другого району. Знову ж ці п'ятірки є комбінаціями і

$$m = C_{10}^5 = \frac{10 \cdot 9 \cdot 8 \cdot 7 \cdot 6}{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} = 252,$$

а $$P(B) = \frac{m}{n} = \frac{252}{53130} \approx 0,005.$$

в) C — власниками кредиту є три клієнти першого району і два клієнти другого району. Всеможливі трійки клієнтів першого району можна утворити C_{15}^3 способами, а всеможливі двійки клієнтів другого району — C_{10}^2. Використавши **правило добутків**, отримаємо:

$$m = C_{15}^3 \cdot C_{10}^2 = \frac{15 \cdot 14 \cdot 13}{3 \cdot 2 \cdot 1} \cdot \frac{10 \cdot 9}{2 \cdot 1} = 455 \cdot 45 = 20475,$$

звідки $$P(C) = \frac{20475}{53130} \approx 0,385.$$ ●

Задача 1.2. На чотирьох картках написані літери *Е*, *Н*, *Т,У*. Картки перемішуються і розкладаються в ряд. Знайти імовірність того, що в результаті буде отримане слово «*ТНЕУ*».

○ Випробування — розкладання карток в ряд після перемішування. Подія A — отримання слова «*ТНЕУ*». Тільки один наслідок випробування сприяє появі цієї події, тобто $m = 1$. Наслідки випробування — це всеможливі групи по 4 елементи (літери), для яких суттєвий порядок розташування елементів, і їх число $n = P_4 = 4! = 4 \cdot 3 \cdot 2 \cdot 1 = 24$. Отже, $P(A) = 1/24$. ●

Задача 1.3. На книжковій полиці розташовано 26 книг, вартості яких відповідно рівні: 8 по 5 грн., 12 по 4 грн. і 6 по 3 грн. Навмання беруться дві книги. Яка імовірність того, що їх сумарна вартість складає 8 грн.?

○ Подія A — сумарна вартість відібраних двох книг дорівнює 8 грн. Випробування — відбір двох книг. Число наслідків дорівнює числу всеможливих способів відбору двох книг. Такі групи по дві книги є комбінаціями, бо для них несуттєвий порядок розташування в групі елементів (книг). Тому $n = C_{26}^2 = \frac{26 \cdot 25}{2 \cdot 1} = 325.$

Подія A відбувається або тоді, коли обидві книги коштують по 4 грн. (таких наслідків випробування є C_{12}^2), або одна книга коштує 5 грн., а друга — 3 грн. (таких наслідків є $8 \cdot 6 = C_8^1 \cdot C_6^1$). Згідно з

правилом суми комбінаторики $m = C_{12}^2 + C_8^1 \cdot C_6^1 = 66 + 48 = 114.$

Остаточно P(A) = 114/325 = 0,351. ●

В деяких випадках при знаходженні числа комбінацій доцільніше користуватися формулою (1.4*), а не (1.4). При цьому корисним є використання основних властивостей числа комбінацій.

Задача 1.4. Після буревію з'ясувалося, що телефонна лінія пошкоджена на ділянці між 20-м і 40-м кілометрами. Яка імовірність того, що пошкодження сталося між 25-м і 30-м кілометрами лінії?

○ Припустимо, що імовірність знаходження точки розриву лінії на відрізку l пропорційна довжині цього відрізка і не залежить від його розташування відносно відрізка L. Тоді згідно із формулою (1.9)

$$P(A) = \frac{\textit{довжина} \quad l}{\textit{довжина} \quad L},$$

де випадкова подія $A = \{x \in l\}$, x – координата точки розриву на числовій осі.

В даному випадку

$$P(A) = P(25 \le x \le 30) = \frac{30 - 25}{40 - 20} = \frac{5}{20} = 0{,}25\,.$$ ●

§ 2. ТЕОРЕМИ МНОЖЕННЯ І ДОДАВАННЯ ІМОВІРНОСТЕЙ ТА ЇХ НАСЛІДКИ

1. *Дії над подіями (алгебра подій). Діаграми В'єнна.*
2. *Умовна імовірність. Теорема множення імовірностей.*
3. *Теореми додавання імовірностей.*
4. *Основна властивість подій, які утворюють повну групу. Імовірність появи хоча б однієї події. Імовірність відбуття тільки однієї події.*
5. *Алгоритм розв'язування задач з використанням теорем додавання та множення імовірностей.*
6. *Формула повної імовірності. Формули Байєса.*
7. *Алгоритм розв'язування задач з використанням формул повної імовірності та Байєса.*

1. Введемо в розгляд дії **над подіями**.

Дві події, які утворюють повну групу, називаються протилежними.

Означення. Під подією $\overline{A}$ (читається: не A) розуміється подія, яка полягає в тому, що при випробуванні подія A не відбувається і відбувається подія, протилежна до події A. Дія над подією, яка визначається рисочкою над цією подією, називається **запереченням** (цієї події).

Сумою подій $A_1, A_2, ..., A_k$ називається подія A, яка полягає в тому, що при випробуванні відбувається хоча б одна з подій $A_1, A_2, ..., A_k$. Символічний запис: $A = A_1 + A_2 + ... + A_k$. Зокрема, якщо $k = 2$, тоді подія $A = A_1 + A_2$ полягає в тому, що при випробуванні відбувається **або** подія A_1, **або** подія A_2, **або** A_1 та A_2 (якщо A_1 і A_2 є сумісними). В зв'язку з цим отримуємо таке **мнемонічне правило**: знак суми асоціюється із словом «або» («+» ↔ «або»). Якщо A_1 та A_2 несумісні, то $A_1 + A_2$ — це подія, яка полягає в тому, що при випробуванні відбувається або подія A_1, або A_2.

Добутком подій $A_1, A_2, ..., A_k$ називається подія A, яка полягає в тому, що при випробуванні відбуваються всі події $A_1, A_2, ..., A_k$.

Символічно: $A = A_1 \cdot A_2 \cdot ... \cdot A_k$ (або $A = A_1 A_2 ... A_k$).

Відбуття всіх k подій передбачає відбуття **і** події A_1, **і** події A_2, ..., **і** події A_k. Отже, мнемонічно дія множення асоціюється із «і» («×» ↔ «і»).

2. Нехай випробування, в результаті якого може відбутися випадкова подія B, доповнюється умовою про відбуття випадкової події A. Тоді **імовірність події B, знайдена при умові, що подія A відбулася, називається умовною імовірністю події** B і позначається $P_A(B)$ або $P(B|A)$.

Нехай в урні є 6 білих і 4 чорних куль,однакових за розміром і на дотик. З неї двічі навмання беруть по одній кулі без повернення відібраних. Знайти ймовірність того, що навмання взята друга куля буде чорного кольору, якщо перша куля виявилась білого кольору. Введемо в розгляд події: A — перша куля біла, B — друга куля чорна. Тоді за класичним означенням

$$P_A(B) = \frac{m}{n} = \frac{4}{9}.$$

При обчисленні умовних імовірностей рекомендується включати зміст випадкових подій, починаючи із читання умовної імовірності. Наприклад, шукаючи праву частину вище записаної формули, ліву слід прочитати: яка імовірність того, що друга витягнута куля **виявиться** чорною, якщо перша куля **виявилась** білою. Тобто, при використанні класичного означення імовірності обов'язково слід врахувати відбуття події A (зокрема, в цьому прикладі n=9, бо перша куля відібрана, m=4, оскільки відбір кулі не змінив число чорних).

Дві події називаються **незалежними**, якщо імовірність однієї з них не змінюється від того, відбулася чи ні інша. Аналітичним критерієм незалежності подій A та B є рівність

$$P_A(B) = P_{\bar{A}}(B). \tag{2.1}$$

У випадку виконання (2.1) можна вважати, що

$$P_A(B) = P_{\bar{A}}(B) = P(B), \tag{2.2}$$

де остання імовірність називається **безумовною**.

Дві події називаються **залежними**, якщо імовірність однієї з них змінюється внаслідок відбуття іншої. Тобто, якщо події A та B залежні, то рівність (2.2) порушується:

$$P_A(B) \neq P_{\bar{A}}(B). \tag{2.3}$$

Теорема множення імовірностей

Імовірність добутку двох сумісних подій дорівнює добутку імовірностей однієї з них на умовну імовірність іншої події, обчислену в припущенні, що перша подія відбулася:

$$P(AB) = P(A)P_A(B)\big(= P(B)P_B(A)\big). \tag{2.4}$$

Зауваження. Продовження рівності в дужках вказує на «рівноправність» подій A та B з врахуванням комутативності дії множення ($AB = BA$).

Наслідок 1. Імовірність добутку k подій дорівнює добутку імовірності однієї з них на умовні імовірності всіх решти, причому імовірність кожної наступної події знаходиться в припущенні, що всі попередні події вже відбулися:

$$P(A_1A_2A_3...A_k) = P(A_1)P_{A_1}(A_2)P_{A_1A_2}(A_3)...P_{A_1A_2...A_{k-1}}(A_k). \quad (2.5)$$

Наслідок 2. Якщо події А та В незалежні, тоді

$$\boldsymbol{P(AB) = P(A)\ P(B).} \quad (2.6)$$

Для узагальнення наслідка 2 на випадок довільного числа подій розглянемо такі означення.

Декілька подій називаються попарно незалежними, якщо **кожні** дві з них незалежні. Наприклад, події A_1, A_2, A_3 попарно незалежні, якщо незалежні події A_1 і A_2, A_1 і A_3, A_2 і A_3.

Декілька подій називаються незалежними в сукупності (або просто **незалежними**), якщо вони попарно незалежні, а також є незалежними кожна з них і всі можливі добутки інших. Наприклад, якщо події A, B, C незалежні в сукупності, то незалежні події A і B, A і C, B і C, A і BC, B і AC, C і AB.

Виявляється, що **попарна незалежність декількох подій ще не гарантує їх незалежність в сукупності**.

Наслідок 3. Імовірність добутку *k* подій, незалежних в сукупності, дорівнює добутку імовірностей цих подій:

$$\boldsymbol{P(A_1\,A_2\,...\,A_k) = P(A_1)\ P(A_2)\ ...\ P(A_k).} \quad (2.7)$$

3. **Теорема 1. Імовірність суми двох несумісних подій дорівнює сумі імовірностей цих подій**:

$$P(A + B) = P(A) + P(B). \quad (2.8)$$

Наслідок. **Імовірність суми *k* попарно несумісних подій дорівнює сумі імовірностей цих подій:**

$$\boldsymbol{P(A_1 + A_2 + ... + A_k) = P(A_1) + P(A_2) + ... + P(A_k).} \quad (2.9)$$

Теорема 2. Імовірність суми двох сумісних подій дорівнює сумі імовірностей цих подій без імовірності їх сумісної появи:

$$\boldsymbol{P(A + B) = P(A) + P(B) - P(AB).} \quad (2.10)$$

4. **Теорема 3. Сума імовірностей подій, які утворюють повну групу, дорівнює одиниці**:

$$\boldsymbol{P(A_1) + P(A_2) + ... P(A_n) = 1.} \quad (2.11)$$

Для $n = 2$ події A_1, A_2 є протилежними, тому рівність (2.11) набуває такого виду:

$$P(A) + P(\overline{A}) = 1 \quad \text{або} \quad p + q = 1, \quad (2.11^*)$$

де $p = P(A)$, $q = P(\overline{A})$.

Імовірність появи хоча б однієї події

Нехай в результаті випробування можуть відбутися події $A_1, A_2, ..., A_k$, імовірності появи кожної з яких відомі і які є

незалежними в сукупності. Позначимо A — поява хоча б однієї із цих подій у випробуванні. Тоді згідно з означенням суми подій $A = A_1 + A_2 + ... + A_k$. Оскільки події $A_1, A_2, ..., A_k$ є сумісні, то теорема додавання імовірностей не «працює» для $k \geq 3$. Використавши (2.11*), (2.7) і $\overline{A} = \overline{A}_1 \overline{A}_2 ... \overline{A}_k$, одержимо:

$$P(A) = 1 - q_1 q_2 ... q_k, \qquad (2.12)$$

де $q_1 = P(\overline{A}_1), q_2 = P(\overline{A}_2), \ldots, q_k = P(\overline{A}_k)$.

В частинному випадку, коли $P(A_1) = P(A_2) = ... = P(A_k) = p$, отримується така формула

$$P(A) = 1 - q^k, \qquad (2.12^*)$$

де $q = 1 - p$.

Нарешті, якщо події $A_1, A_2, ..., A_k$ є **залежними** (не володіють властивістю незалежності в сукупності), тоді

$$P(A) = 1 - P(\overline{A}_1) P_{\overline{A}_1}(\overline{A}_2) \ldots P_{\overline{A}_1 \overline{A}_2 \ldots \overline{A}_{k-1}}(\overline{A}_k). \qquad (2.13)$$

5. Алгоритм розв'язування задач з використанням теорем додавання та множення імовірностей.

1) Вводяться в розгляд подія, імовірність якої треба знайти, а також більш простіші події, імовірності яких відомі або можуть бути знайдені за класичним означенням.

2) «Шукана» випадкова подія (імовірність якої потрібно знайти) виражається через простіші події за допомогою алгебри подій, тобто операцій суми, добутку, заперечення (протилежної події). При цьому потрібно керуватися мнемонічними правилами: «+» ↔ **або**, «×» ↔ **і**.

3) В залежності від виду отриманого виразу використовуються теореми додавання імовірностей або(і) теорема множення імовірностей та її наслідки. При реалізації цього пункту необхідно з'ясовувати властивості подій (сумісність, несумісність, залежність, незалежність, протилежність або повноту пари чи групи подій).

Зауваження. Слід мати на увазі те, що в багатьох задачах реалізація пункту 2) неєдина. В таких випадках бажано вибрати найкомпактнішу, переконавшись у співпаданні остаточних результатів після виконання пункту 3). Якщо ж результати не співпадають, то необхідно перевірити правильність побудови в п. 2) або коректність виконання п. 3).

6. Формула повної імовірності

Нехай подія A може відбутися тільки при умові появи однієї із подій $B_1, B_2, ..., B_n$, які утворюють повну групу. Нехай відомі

імовірності $P(B_k)$, $P_{B_k}(A)$, $k = (\overline{1,n})$. Відповідь на питання: як знайти $P(A)$? — дає

Теорема. Імовірність події *A*, яка може відбутися тільки після появи однієї із подій $B_1, B_2, ..., B_n$, які утворюють повну групу, знаходиться за так званою формулою повної імовірності:

$$P(A) = P(B_1)P_{B_1}(A) + P(B_2)P_{B_2}(A) + \ldots + P(B_n)P_{B_n}(A). \quad (2.14)$$

Зауваження. Оскільки до випробування невідомо, після якої із подій $B_1, B_2, ..., B_n$ відбудеться подія ***A***, то ці події називаються **гіпотезами (припущеннями)**.

Формули Байєса

Нехай виконуються умови теореми відносно подій $B_1, B_2, ..., B_n$ та A. Припустимо, що проведено випробування, в результаті якого відбулася подія A. Виникає питання: як «переоцінити» імовірності гіпотез $B_1, B_2, ..., B_n$ із врахуванням відбуття події A, тобто знайти умовні імовірності $P_A(B_k)$, $(k = \overline{1,n})$? Відповідь дають так звані формули Байєса:

$$P_A(B_k) = \frac{P(B_k)P_{B_k}(A)}{\sum_{i=1}^{n} P(B_i)P_{B_i}(A)}, k = 1,2,\ldots,n. \quad (2.15)$$

7. Алгоритм розв'язування задач з використанням формул повної імовірності та Байєса.

Рекомендується така послідовність розв'язування задач.

1) Формулюють гіпотези $B_1, B_2, ..., B_n$ і подію A. При цьому слід перевірити **повноту групи гіпотез**, а також те, що **подія *A* може відбутися тільки після появи однієї із гіпотез**.

2) Знаходяться імовірності гіпотез. Правильність розрахунків контролюється виконанням рівності $P(B_1) + P(B_2) + ... + P(B_n) = 1$. Обчислюються умовні імовірності $P_{B_1}(A), P_{B_2}(A), \ldots, P_{B_n}(A)$.

3) Вибирається формула повної імовірності або формули Байєса. Останні використовуються тоді, коли є інформація про відбуття випадкової події.

РОЗВ'ЯЗУВАННЯ ТИПОВИХ ЗАДАЧ

Задача 2.1. В кейсі є 5 акцій першого виду, 6 — другого і 3 — третього. Знайти імовірність того, що три навмання взяті акції виявляться одного і того ж виду.

○ Позначимо: A — три навмання взяті акції є одного виду, B_i — три

взяті акції i-го виду ($i = 1, 2, 3$). Подія A відбувається тоді, коли відбуваються або подія B_1, або B_2, або B_3. Інших можливостей для появи A немає. Тому мають місце такі рівності: $A = B_1 + B_2 + B_3$, $P(A) = P(B_1 + B_2 + B_3)$. Події B_1, B_2, B_3 попарно несумісні і згідно з наслідком теореми 1 (рівність (2.9))

$$P(A) = P(B_1) + P(B_2) + P(B_3).$$

За класичним означенням імовірності

$$P(B_1) = \frac{C_5^3}{C_{14}^3} = \frac{5}{182}, \quad P(B_2) = \frac{C_6^3}{C_{14}^3} = \frac{5}{91}, \quad P(B_3) = \frac{C_3^3}{C_{14}^3} = \frac{1}{364}.$$

Остаточно

$$P(A) = \frac{5}{182} + \frac{5}{91} + \frac{1}{364} = \frac{31}{364}.$$

Покажемо, як можна знайти імовірності подій B_1, B_2 та B_3, використовуючи теорему множення імовірностей. З цією метою позначимо: $B_j^{(i)}$ — j-та відібрана акція ($j = 1, 2, 3$) є акцією i-го виду ($i = 1, 2, 3$). Тоді, зокрема, B_1 відбудеться, якщо всі акції (**і** перша, **і** друга, **і** третя) є акціями першого виду, тобто коли відбуваються всі події $B_1^{(1)}, B_2^{(1)}, B_3^{(1)}$.

Отже, $B_1 = B_1^{(1)} B_2^{(1)} B_3^{(1)}$, $P(B_1) = P\left(B_1^{(1)} B_2^{(1)} B_3^{(1)}\right)$ і згідно з рівністю (2.5) для випадку $k = 3$

$$P(B_1) = P\left(B_1^{(1)}\right) P_{B_1^{(1)}}\left(B_2^{(1)}\right) P_{B_1^{(1)} B_2^{(1)}}\left(B_3^{(1)}\right) = \frac{5}{14} \cdot \frac{4}{13} \cdot \frac{3}{12} = \frac{5}{182}.$$

Відмітимо, що вибір рівності (2.5) на противагу рівності (2.7) зумовлений залежністю подій $B_1^{(1)}, B_2^{(1)}, B_3^{(1)}$.

Аналогічно

$$P(B_2) = P\left(B_1^{(2)}\right) P_{B_1^{(2)}}\left(B_2^{(2)}\right) P_{B_1^{(2)} B_2^{(2)}}\left(B_3^{(2)}\right) = \frac{6}{14} \cdot \frac{5}{13} \cdot \frac{4}{12} = \frac{5}{91},$$

$$P(B_3) = P\left(B_1^{(3)}\right) P_{B_1^{(3)}}\left(B_2^{(3)}\right) P_{B_1^{(3)} B_2^{(3)}}\left(B_3^{(3)}\right) = \frac{3}{14} \cdot \frac{2}{13} \cdot \frac{1}{12} = \frac{1}{364}.$$

Отже, другий метод знаходження імовірностей подій B_1, B_2 та B_3 дає ті ж самі результати. ●

Задача 2.2. Підприємство планує здійснювати поставки двох видів виробів. Імовірність зриву поставок для першого виду виробів складає 0,05, а для другого 0,08. Згідно із проектом контракту при порушенні термінів поставок хоча б одного виду продукції до виробника застосовуються штрафні санкції, які приводять до

нерентабельності виробництва обох видів виробів. Знайти імовірність нерентабельності виробництва цих виробів.

○ Позначимо: C — нерентабельність виробництва обох видів продукції, A та B — зрив поставки виробів першого та другого видів відповідно. Згідно з умовою задачі подія C відбувається, якщо відбувається **або** подія A, **або** подія B, **або** події A та B разом, тобто хоча б одна з них. Тому $C = A + B$, де A та B сумісні події. Згідно з теоремою додавання для сумісних подій

$$P(C) = P(A + B) = P(A) + P(B) - P(AB).$$

За умовою $P(A) = 0{,}05$, $P(B) = 0{,}08$ і можна вважати, що події A та B незалежні, тобто $P(AB) = P(A)\,P(B)$.

Остаточно $P(C) = 0{,}05 + 0{,}08 - 0{,}05 \cdot 0{,}08 = 0{,}126$.

Другий метод розв'язування. Подію C можна представити через більш прості події A та B ще й таким чином:

$$C = A\overline{B} + \overline{A}B + AB.$$

Доданки-події справа є попарно несумісними випадковими подіями. Справді, припустимо, що події $A\overline{B}$ та $\overline{A}B$ є сумісними, тобто можуть відбутися у випробуванні. Тоді отримується суперечність: подія A, зокрема, в одному випробуванні і відбувається, і не відбувається. Це протиріччя вказує на хибність припущення. Переконайтеся в попарній несумісності цих подій, використавши діаграми В'єнна. Використавши рівності (2.9) та (2.6) і незалежність подій A та $\overline{B}$, $\overline{A}$ та B, отримаємо:

$$P(C) = P(A\overline{B} + \overline{A}B + AB) = P(A\overline{B}) + P(\overline{A}B) + P(AB) =$$

$$= P(A)P(\overline{B}) + P(\overline{A})P(B) + P(A)P(B) =$$

$$= 0{,}05 \cdot 0{,}92 + 0{,}95 \cdot 0{,}08 + 0{,}05 \cdot 0{,}08 = 0{,}046 + 0{,}076 + 0{,}004 = 0{,}126,$$

де згідно із рівністю (2.11*)

$$P(\overline{A}) = 1 - P(A) = 1 - 0{,}05 = 0{,}95,$$

$$P(\overline{B}) = 1 - P(B) = 1 - 0{,}08 = 0{,}92.$$

Третій метод. Протилежною до події C є подія $\overline{C}$, яка полягає в тому, що внаслідок здійснення поставок продукції виробництво обох видів виробів є рентабельним. Подія $\overline{C}$ відбувається тоді, коли **і** першого виду вироби вчасно поставляються, **і** другого, тобто відбуваються події і $\overline{A}$, і $\overline{B}$. Тому $\overline{C} = \overline{A}\,\overline{B}$, $P(\overline{C}) = P(\overline{A})P(\overline{B}) = 0{,}95 \cdot 0{,}92 = 0{,}874$. Але згідно з (2.11*) $P(C) + P(\overline{C}) = 1$, звідки $P(C) = 1 - P(\overline{C}) = 1 - 0{,}874 = 0{,}126$.

Висновки. Перевага третього методу, зокрема, полягає в тому, що він дозволяє розв'язати задачу для довільного скінченного

числа видів продукції. Що ж до отриманої відповіді, то знайдену імовірність слід інтерпретувати таким чином: в середньому в 126 випадках з кожної тисячі (12,6%) очікується нерентабельність виробництва обох видів продукції. Оскільки така імовірність достатньо велика, то керівництву підприємства потрібно подбати про зменшення імовірностей зривів поставок або(і) зменшення штрафних санкцій. ●

Задача 2.3. У зв'язці шість різних ключів, з яких тільки одним можна відкрити замок. Навмання вибирається ключ і робиться спроба відкрити ним замок. Ключ, що не підійшов, більше не використовується. Знайти імовірність того, що для відкривання буде використано не більше трьох ключів.

○ Позначимо: A_k (k = 1, 2, 3) — замок буде відкрито k-тим за порядку відбору ключем, B — замок відкривається після використання не більше трьох ключів. Подія B відбудеться, якщо до замка підійде **або** перший ключ (відбувається A_1), **або** другий (при цьому перший ключ не підійшов — відбувається подія $\overline{A}_1 A_2$), **або** третій (перший і другий ключі не підійшли — відбувається подія $\overline{A}_1 \overline{A}_2 A_3$). Тобто вираз B через простіші події A_1, A_2, A_3 має такий вид:

$$B = A_1 + \overline{A}_1 A_2 + \overline{A}_1 \overline{A}_2 A_3.$$

Для знаходження $P(B)$ потрібно використати рівність (2.9), оскільки доданки є попарно несумісними подіями, а потім теорему множення імовірностей для залежних подій (обчисливши $P_{\overline{A}_1}(A_2)$ і $P_{A_1}(A_2)$, переконайтеся у тому, що події $\overline{A}_1$ і A_2 є залежними):

$$P(B) = P(A_1 + \overline{A}_1 A_2 + \overline{A}_1 \overline{A}_2 A_3) = P(A_1) + P(\overline{A}_1 A_2) + P(\overline{A}_1 \overline{A}_2 A_3) =$$
$$= P(A_1) + P(\overline{A}_1) P_{\overline{A}_1}(A_2) + P(\overline{A}_1) P_{\overline{A}_1}(\overline{A}_2) P_{\overline{A}_1 \overline{A}_2}(A_3) =$$
$$= \frac{1}{6} + \frac{5}{6} \cdot \frac{1}{5} + \frac{5}{6} \cdot \frac{4}{5} \cdot \frac{1}{4} = 0{,}5. \quad ●$$

Задача 2.4. Однотипні деталі виготовляються трьома автоматами, продуктивності яких відносяться як 3:2:5. Із нерозсортованої партії деталей навмання беруться дві. Знайти імовірність того, що: а) одна із них виготовлена першим автоматом; б) обидві виготовлені одним автоматом; в) виготовлені різними автоматами.

○ Випробування – відбір двох деталей.

а) Позначимо:

B — одна із двох відібраних деталей виготовлена першим

автоматом;

A_i — деталь виготовлена i-тим автоматом (i=1, 2, 3).

Подія B відбудеться, якщо одна із двох деталей виготовлена першим автоматом, а друга – іншими (другим або третім). Проте, зокрема, події A_1A_2 і A_2A_1 – різні, хоча добуток володіє властивістю комутативності $A_1A_2=A_2A_1$, оскільки A_1A_2 – **перша** взята виготовлена І-м і **друга** взята деталь виготовлена ІІ-м автоматом, а A_2A_1 – **перша** деталь з ІІ-го автомата і **друга** – з І-го. Отже,

$B = A_1A_2 + A_2A_1 + A_1A_3 + A_3A_1$.

Згідно з умовою A_1, A_2, A_3 є незалежними в сукупності і за класичним означенням імовірності

$$P(A_1) = \frac{3}{3+2+5} = 0{,}3\,;\; P(A_2) = \frac{2}{3+2+5} = 0{,}2\,;\; P(A_3) = \frac{5}{3+2+5} = 0{,}5\,.$$

Використавши теорему суми для попарно несумісних подій, а потім добутку (для незалежних подій), отримаємо

$$P(B) = P(A_1A_2 + A_2A_1 + A_1A_3 + A_3A_1) =$$
$$= P(A_1A_2) + P(A_2A_1) + P(A_1A_3) + P(A_3A_1) =$$
$$= P(A_1)\cdot P(A_2) + P(A_2)\cdot P(A_1) + P(A_1)\cdot P(A_3) + P(A_3)\cdot P(A_1) =$$
$$= 0{,}3\cdot 0{,}2 + 0{,}2\cdot 0{,}3 + 0{,}3\cdot 0{,}5 + 0{,}5\cdot 0{,}3 = 0{,}42\,.$$

б) Подія C – обидві відібрані деталі виготовлені одним автоматом. Тоді $C = A_1A_1 + A_2A_2 + A_3A_3$ і

$P(C) = 0{,}3\cdot 0{,}3 + 0{,}2\cdot 0{,}2 + 0{,}5\cdot 0{,}5 = 0{,}38$.

в) Подія D – обидві відібрані деталі виготовлені різними автоматами. Протилежною до події D є подія C, тобто $\overline{D} = C$, звідси $P(\overline{D}) = P(C) = 0{,}38$.

За формулою (2.11*)

$P(D) = 1 - P(\overline{D}) = 1 - 0{,}38 = 0{,}62$. ●

Задача 2.5. Підприємство отримує від суміжників продукцію $П_1, П_2, П_3, П_4$ і виконає план, якщо вчасно отримає якісну продукцію в потрібній кількості. Відсоток зриву поставок (як по кількості, так і по якості) по кожній із цих компонент технологічного процесу відповідно дорівнює 2%; 1%; 10%; 5%. Знайти імовірність невиконання заводом плану.

○ Позначимо: A — невиконання заводом плану, A_i — зрив поставки компоненти $П_i$ (i = 1, 2, 3, 4). Якщо відбувається хоча одна із подій A_1, A_2, A_3, A_4, тоді відбувається подія A. За умовою задачі $P(A_1) = 0{,}02$; $P(A_2) = 0{,}01$; $P(A_3) = 0{,}1$; $P(A_4) = 0{,}05$ і ці події можна

вважати незалежними в сукупності. Використовуючи формулу (2.12), отримаємо:

$$q_1 = 1 - 0{,}02 = 0{,}98; \quad q_2 = 1 - 0{,}01 = 0{,}99;$$
$$q_3 = 1 - 0{,}1 = 0{,}9; \quad q_4 = 1 - 0{,}05 = 0{,}95,$$
$$P(A) = 1 - 0{,}98 \cdot 0{,}99 \cdot 0{,}9 \cdot 0{,}95 = 1 - 0{,}829521 \approx 0{,}171. \bullet$$

Задача 2.6. На чотирнадцяти окремих картках написано по одній літері: 5 карток з літерою «А», 4 — з літерою «І», 3 — з літерою «Р» і 2 – «Т». Навмання витягуються шість карток і розкладаються зліва направо. Яка імовірність того, що отримається слово «АРАРАТ»?

○ Позначимо випадкові події: B — отримання слова «АРАРАТ»,
A_1 — перша картка має літеру «А»,
A_2 — друга картка має літеру «Р»,
A_3 — третя картка має літеру «А»,
A_4 — четверта картка має літеру «Р»,
A_5 — п'ята картка має літеру «А»,
A_6 — шоста картка має літеру «Т»,

Подія B відбудеться, якщо відбудуться всі події $A_1, A_2, ..., A_6$, тобто $B = A_1A_2A_3A_4A_5A_6$. Використавши теорему множення імовірностей для залежних подій, отримаємо:

$$P(B) = P(A_1A_2A_3A_4A_5A_6) = P(A_1) \cdot P_{A_1}(A_2) \cdot P_{A_1A_2}(A_3) \cdot P_{A_1A_2A_3}(A_4) \times$$
$$\times P_{A_1A_2A_3A_4}(A_5) \cdot P_{A_1A_2A_3A_4A_5}(A_6) = \frac{5}{14} \cdot \frac{3}{13} \cdot \frac{4}{12} \cdot \frac{2}{11} \cdot \frac{3}{10} \cdot \frac{2}{9} = \frac{1}{3003} \approx 0{,}00033. \bullet$$

Задача 2.7. Гральний кубик кидають до тих пір, поки два рази підряд на верхній грані не випаде шість очок. Знайти ймовірність того, що дослід закінчиться до п'ятого кидання.

○ Позначимо випадкові події: B — дослід закінчиться до п'ятого кидання,
A_1 — випала грань з шістьма очками за першим разом,
A_2 — випала грань з шістьма очками за другим разом,
A_3 — випала грань з шістьма очками за третім разом,
A_4 — випала грань з шістьма очками за четвертим разом.

Оскільки, за умовою задачі кубик кидають до тих пір, поки два рази підряд на верхній грані не випаде шість очок, то принаймні двічі необхідно кинути кубик і дослід закінчиться, якщо відбудеться подія $C_1 = A_1A_2$. Проте, якщо при такій спробі двічі не випало шість очок, то кубик доведеться кидати вже тричі і дослід закінчиться в цьому випадку, якщо відбудеться подія $C_2 = \overline{A}_1A_2A_3$. Якщо ж і в цьому

випадку двічі не випала грань з шістьма очками, то кубик кидатимуть чотири рази. Таким чином, при трьох киданнях кубика подія B відбудеться, якщо відбудеться одна із подій: $C_3 = \overline{A}_1 \overline{A}_2 A_3 A_4$ або $C_4 = A_1 \overline{A}_2 A_3 A_4$. (Здавалося б, що два рази грань з шістма очками випаде при настанні події $A_1 \overline{A}_2 A_3$ (перший і третій рази). Проте за умовою задачі грань з шістьма очками повинна випасти два рази підряд, тому подія B в цьому випадку не відбудеться).

Таким чином,

$$B = C_1 + C_2 + C_3 + C_4 = A_1 A_2 + \overline{A}_1 A_2 A_3 + \overline{A}_1 \overline{A}_2 A_3 A_4 + A_1 \overline{A}_2 A_3 A_4 .$$

$$P(B) = P(A_1 A_2 + \overline{A}_1 A_2 A_3 + \overline{A}_1 \overline{A}_2 A_3 A_4 + A_1 \overline{A}_2 A_3 A_4)$$

Оскільки події C_1, C_2, C_3, C_4 попарно несумісні використаємо (2.9):

$$P(B) = P(A_1 A_2) + P(\overline{A}_1 A_2 A_3) + P(\overline{A}_1 \overline{A}_2 A_3 A_4) + P(A_1 \overline{A}_2 A_3 A_4) .$$

В свою чергу події A_1, A_2, A_3, A_4 незалежні в сукупності, тому використовуючи (2.7). одержимо

$$P(B) = P(A_1)P(A_2) + P(\overline{A}_1)P(A_2)P(A_3) + P(\overline{A}_1)P(\overline{A}_2)P(A_3)P(A_4) +$$

$$+ P(A_1)P(\overline{A}_2)P(A_3)P(A_4) = \frac{1}{6}\cdot\frac{1}{6} + \frac{5}{6}\cdot\frac{1}{6}\cdot\frac{1}{6} + \frac{5}{6}\cdot\frac{5}{6}\cdot\frac{1}{6}\cdot\frac{1}{6} + \frac{1}{6}\cdot\frac{5}{6}\cdot\frac{1}{6}\cdot\frac{1}{6} =$$

$$= \frac{961}{1296} \approx 0,074 . \bullet$$

Задача 2.8. Три верстати-автомати штампують однотипні деталі, що потрапляють на спільний конвейєр. Продуктивність другого автомату на 40% вища від продуктивності першого і вдвічі — від третього. Відсоток браку для кожного з автоматів дорівнює відповідно 3, 6, 2. а). Яка імовірність того, що навмання взята з конвейєра деталь виявиться бракованою? б). Навмання взята деталь виявилася бракованою. Що імовірніше: ця деталь виготовлена першим чи третім автоматом?

○ а) Першу частину задачі розв'язуємо за формулою повної імовірності, оскільки відсутня інформація про відбуття випадкової події.

При відборі довільним чином деталі з конвейєра невідомо, яким автоматом вона виготовлена. Цю **невизначеність не можна трактувати як відбуття випадкової події** (типова помилка студентів). Можна зробити припущення: будь-яка деталь (стандартна чи бракована) виготовлена або першим автоматом (подія B_1), або другим (B_2), або третім (B_3). Неважко переконатися в тому, що події

B_1, B_2, B_3 утворюють повну групу (доведіть!).

Позначимо: A — навмання взята деталь бракована. Ясно, що подія A може відбутися після появи однієї із гіпотез, адже щоб говорити про бракованість деталі, потрібно її спочатку виготовити.

Знайдемо імовірності гіпотез. Для цього позначимо через a кількість деталей, що виготовить перший автомат за деякий проміжок часу. Тоді за умовою задачі за цей же проміжок часу другий автомат виготовить $1{,}4a$ деталей, а третій — $0{,}7a$. За класичним означенням імовірності

$$P(B_1) = \frac{a}{3{,}1a} = \frac{10}{31};\ P(B_2) = \frac{1{,}4a}{3{,}1a} = \frac{14}{31};\ P(B_3) = \frac{0{,}7a}{3{,}1a} = \frac{7}{31}.$$

Перевіримо обчислення. Оскільки гіпотези утворюють повну групу, то сума їх імовірностей повинна дорівнювати одиниці:

$$P(B_1) + P(B_2) + P(B_3) = \frac{10}{31} + \frac{14}{31} + \frac{7}{31} = 1.$$

Згідно з умовою задачі і за класичним означенням

$$P_{B_1}(A) = \frac{3}{100} = 0{,}03;\ P_{B_2}(A) = \frac{6}{100} = 0{,}06;\ P_{B_3}(A) = \frac{2}{100} = 0{,}02.$$

Прочитайте ліві частини цих рівностей з урахуванням змісту подій A, B_1, B_2, B_3, а також переконайтеся у правильності цих рівностей.

Формула повної імовірності (2.14) у цьому випадку має такий вид:

$$P(A) = P(B_1)P_{B_1}(A) + P(B_2)P_{B_2}(A) + P(B_3)P_{B_3}(A).$$

Підставивши обчислені дані в праву частину, отримаємо

$$P(A) = \frac{10}{31} \cdot 0{,}03 + \frac{14}{31} \cdot 0{,}06 + \frac{7}{31} \cdot 0{,}02 \approx 0{,}0401.$$

б) У другій частині задачі використаємо формули Байєса, оскільки є інформація про відбуття випадкової події (A — деталь виявилася бракованою). При цьому ми повинні знайти $P_A(B_1)$, $P_A(B_3)$ і порівняти їх. Але більшою з них буде та імовірність, чисельник в правій частині формули Байєса для якої буде більшим. Тобто, достатньо порівняти $P(B_1)P_{B_1}(A)$ і $P(B_3)P_{B_3}(A)$:

$P(B_1)P_{B_1}(A) = \frac{0{,}3}{31}$, $P(B_3)P_{B_3}(A) = \frac{0{,}14}{31}$. Отже, $P_A(B_1) > P_A(B_3)$, і можна зробити висновок: більш імовірніше, що бракована деталь виготовлена першим автоматом. ●

Задача 2.9. В кожній із двох урн захояться кульки чорного і білого

кольору, причому в першій урні з 20 куль — 8 чорного кольору, а в другій із 15 куль — 5 білого кольору. Із першої урни навмання взяли кульку переклали в другу, після чого, перемішавши її вміст, дістали навмання одну кульку. Знайти ймовірність того, що ця куля біла.

○ Подія A — куля навмання взята із другої урни, яку перед цим доповнили однією кулею з першої урни, виявиться білого кольору — може відбутися тільки після появи однієї із гіпотез: навмання взята куля з першої урни може бути або білого кольору (гіпотеза B_1), або чорного кольору (гіпотеза B_2).

За умовою

$$P(B_1)=\frac{12}{20},\ P(B_2)=\frac{8}{20},\ P_{B_1}(A)=\frac{5+1}{15+1}=\frac{6}{16},\ P_{B_2}(A)=\frac{5}{16}.$$

Відсутня також інформація про відбуття випадкової події. Тому за формулою повної імовірності (2.14) при n=2 отримаємо

$$P(A)=\frac{12}{20}\cdot\frac{6}{16}+\frac{8}{20}\cdot\frac{5}{16}=\frac{40}{320}=0{,}125.\ \bullet$$

Задача 2.10. Відомо, що в період між переналадками обладнання в середньому 96% випущеної підприємством продукції задовільняє умовам стандарту. На підприємстві в якості експерименту почали використовувати спрощену систему контролю якості продукції. Результати показали, що ця система визнає придатним виріб з ймовірністю 0,99, якщо він справді придатний, і з ймовірністю 0,07, якщо він бракований.

1) Знайти імовірність того, що навмання відібраний виріб буде визнано стандартним згідно цієї спрощеної системи контролю.

2) Яка ймовірність того, що виріб справді стандартний, якщо він: визнаний стандартним за спрощеним контролем;

○ 1) Навмання взятий виріб може бути або справді стандартним (гіпотеза B_1), або бракованим (гіпотеза B_2). Тоді A — виріб визнано стандартним згідно спрощеної схеми – може відбутися тільки після появи однієї із гіпотез.

За умовою

$$P(B_1)=0{,}96,\ P(B_2)=0{,}04,\ P_{B_1}(A)=0{,}99,\ \ P_{B_2}(A)=0{,}07.$$

Відсутня також інформація про відбуття випадкової події. Тому за формулою повної імовірності (2.14) при n=2 отримаємо

$$P(A)=0{,}96\cdot 0{,}99+0{,}04\cdot 0{,}07=0{,}9532.$$

2. а) Згідно з умовою подія A відбулася. Тому за формулою Байєса

$$P_A(B_1) = \frac{P(B_1) \cdot P_{B_1}(A)}{P(A)} = \frac{0.96 \cdot 0{,}99}{0{,}9532} \approx 0{,}9971.$$

§ 3. ПОВТОРНІ НЕЗАЛЕЖНІ ВИПРОБУВАННЯ

1. *Формула Бернуллі.*
2. *Найімовірніше число появи події.*
3. *Локальна формула Лапласа.*
4. *Формула Пуассона.*
5. *Інтегральна формула Лапласа.*
6. *Імовірність відхилення відносної частоти події від її постійної імовірності.*
7. *Алгоритм розв'язування задач для повторних незалежних випробувань.*

На практиці часто зустрічаються випадки, коли проводиться не одне випробування, а декілька (можливо, дуже велике число). Такі випробування називаються **повторними**, а їх сукупність — **схемою повторних випробувань**, або **схемою Бернуллі**. В кожному із таких випробувань може відбутися одна і та ж випадкова подія A. Якщо $P(A)$ залишається незмінною для **кожного** випробування, то такі випробування будемо називати **незалежними** (**відносно події A**).

1. Формула Бернуллі

Теорема. Імовірність того, що в n повторних незалежних випробуваннях випадкова подія A відбудеться рівно m разів, знаходиться за формулою Бернуллі:

$$P_n(m) = C_n^m p^m q^{n-m}, \tag{3.1}$$

де C_n^m — число комбінацій, визначене формулою (1.4),

p — імовірність появи події A в **одному** випробуванні ($p = P(A)$),

q — імовірність **непояви** події A в одному випробуванні ($q = P(\overline{A})$).

2. Найімовірніше число появи події.

В n повторних незалежних випробуваннях подія A може відбутися число разів

$$0, 1, ..., m_0-1, m_0, m_0+1, ..., n \tag{3.2}$$

з відповідними ймовірностями (які можна знайти за формулою Бернуллі):

$$P_n(0), P_n(1), ..., P_n(m_0-1), P_n(m_0), P_n(m_0+1), ..., P_n(n). \tag{3.3}$$

Означення. Число m_0 в послідовності (3.2) називається найімовірнішим числом появи події в n незалежних випробуваннях (або модою), якщо йому відповідає найбільша імовірність в послідовності (3.3).

Найімовірніше число можна знайти з такої подвійної нерівності:

$$np - q \le m_0 \le np + p. \qquad (3.4)$$

Оскільки різниця між правою і лівою частинами цієї нерівності дорівнює 1, то можна зробити висновок, що максимальне число найімовірніших подій рівне двом.

3. Локальна формула Лапласа.

Користуватися формулою Бернуллі для великих значень n досить важко. Наприклад, при $p = 0{,}2$, $q = 0{,}8$ $P_{50}(30) = C_{50}^{30}(0{,}2)^{30}(0{,}8)^{20}$, де $C_{50}^{30} = 4712921 \cdot 10^7$.

Одну із реалізацій наближеного знаходження правої частини формули (3.1) дає

Локальна теорема Лапласа (Муавра—Лапласа).

Якщо імовірність p появи випадкової події A в кожному із n повторних випробувань залишається незмінною (причому $0 < p < 1$), а число випробувань достатньо велике, то імовірність того, що в n випробуваннях подія відбудеться m разів, знаходиться за наближеною формулою

$$P_n(m) \approx \frac{1}{\sqrt{npq}}\varphi(x), \qquad (3.5)$$

де $\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ — функція Гаусса, $x = \frac{m - np}{\sqrt{npq}}$, $q = 1 - p$.

Зауваження. Формула (3.5) називається **локальною формулою Лапласа**. Її точність зростає при збільшенні n.

Функція Гаусса протабульована і її значення наведені в табл. 1 додатків. Для правильного користування цією таблицею слід враховувати такі властивості функції Гаусса: 1) $\varphi(x)$ визначена для всіх $x \in R$; 2) $\varphi(x)$ — парна функція $\big(\varphi(-x) = \varphi(x)\big)$; 3) для додатних x $\varphi(x)$ дуже швидко прямує до 0 при збільшенні x, зокрема $\varphi(3{,}99) = 0{,}0001$. Згідно із цими властивостями в таблиці наведені значення функції Гаусса для x з проміжку $[0; 5]$.

Практично можна вважати, що локальна формула Лапласа дає добре наближення, якщо $npq > 9$. Якщо ж вимоги до точності значення вищі, то слід вимагати виконання нерівності $npq \ge 25$.

4. Формула Пуассона.

Нерівність $npq > 9$ навіть для великих n може не виконуватися (а отже, похибка при використанні локальної формули Лапласа буде дуже великою) у випадку рідкісних (малоімовірних) подій A, тобто таких подій, для яких p значно менше 0,1 ($p << 0{,}1$). В таких випадках слід користуватися іншим наближенням правої частини формули Бернуллі. Одне із них дається таким твердженням.

Теорема Пуассона.

Якщо в кожному із n повторних випробувань імовірність p появи події A стала і мала ($p << 0{,}1$), а число випробувань n досить велике, то імовірність того, що подія A настане в цих випробуваннях рівно m разів, знаходиться за формулою

$$P_n(m) \approx \frac{\lambda^m e^{-\lambda}}{m!}, \qquad (3.6)$$

де $\lambda = np$.

Зауваження. Похибка в наближеній рівності (3.6), яка називається **формулою Пуассона**, тим менша, чим більше число випробувань n.

Значення функції $P(m) = \frac{\lambda^m e^{-\lambda}}{m!}$ двох змінних λ та m для деяких m та λ наведені в табл. 2 додатків.

Відмітимо, що формула Пуассона використовується також до числа невідбуття події A, якщо $q << 0{,}1$, а nq невелике.

5. Інтегральна формула Лапласа.

Інтегральна теорема Лапласа (Муавра—Лапласа).

Якщо імовірність p появи події A в кожному із n повторних випробувань є сталою ($0 < p < 1$), а число випробувань досить велике, то імовірність того, що подія A в цих випробуваннях відбудеться не менше m_1 разів і не більше m_2 разів, знаходиться за такою наближеною рівністю (інтегральною формулою Лапласа):

$$P_n(m_1 \le m \le m_2) \approx \Phi(x_2) - \Phi(x_1), \qquad (3.7)$$

де $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-t^2/2} dt$ — функція Лапласа,

$$x_1 = \frac{m_1 - np}{\sqrt{npq}}, x_2 = \frac{m_2 - np}{\sqrt{npq}}.$$

Функція Лапласа протабульована і її значення наведені в табл. 3 додатків. При користуванні цією таблицею потрібно враховувати такі

властивості функції Лапласа:

1) $\Phi(x)$ визначена для всіх $x \in R$;
2) $\Phi(x)$ непарна функція ($\Phi(-x) = -\Phi(x)$);
3) $\Phi(x)$ монотонно зростає для всіх $x \in R$, при цьому $y = -0{,}5$ — лівостороння асимптота, а $y = 0{,}5$ — правостороння;
4) швидкість зростання $\Phi(x)$ на проміжку $[0; 5]$ дуже висока (зокрема $\Phi(5) = 0{,}499997$), тому для всіх $x > 5$ з мізерною похибкою $\Phi(x) \approx 0{,}5$.

Точність інтегральної формули Лапласа тим більша, чим більше число випробувань n. Як і у випадку локальної, **інтегральна формула дає добрі наближення, якщо $npq > 9$**. Якщо ж вимоги до точності значно вищі, то потрібно вимагати виконання нерівності $npq \geq 25$.

6. Імовірність відхилення відносної частоти події від її постійної імовірності.

Теорема. Якщо імовірність p появи випадкової події A в кожному з n повторних випробувань стала, а число випробувань досить велике, то імовірність того, що відхилення відносної частоти m/n події A від її імовірності p по абсолютній величині не перевищить заданого числа $\varepsilon > 0$, знаходиться за формулою

$$P\left(\left|m/n - p\right| \leq \varepsilon\right) \approx 2\Phi\left(\varepsilon\sqrt{n/pq}\right). \qquad (3.8)$$

7. Алгоритм розв'язування задач для повторних незалежних випробувань

В цьому параграфі, як і для попередніх параграфів, залишається актуальним питання вибору тієї чи іншої формули при розв'язуванні конкретних задач. Це зумовлено, по-перше, тим, що у всіх трьох формулах (Бернуллі, локальній формулі Лапласа та Пуассона) ліві частини однакові. З другого боку, при знаходженні імовірності $P_n(m_1 \leq m \leq m_2)$ зовсім не обов'язково (а деколи й помилково) використовувати інтегральну формулу Лапласа.

1. Обчислення $P_n(m)$.

а) Якщо n мале ($n \leq 15$), то використовується формула Бернуллі для будь-яких значень p та q.

б) Якщо n велике, а p та q немалі, тобто при виконанні нерівності $npq > 9$, тоді використовується локальна формула Лапласа.

в) Якщо ж n велике, а p дуже мале (значно менше 0,1) і $\lambda = np \leq 9$, то застосовується формула Пуассона. При великому n, дуже малому q ($q << 0{,}1$) і при виконанні нерівності $\lambda' = nq \leq 9$ слід перейти до числа невиконання події A.

2. Знаходження $P_n(m_1 \le m \le m_2)$.

а) Якщо n мале ($n \le 15$), тоді потрібно використати спочатку теорему додавання імовірностей, а потім формулу Бернуллі.

б) Для великих n і немалих p та q, тобто при виконанні нерівності $npq > 9$ використовується інтегральна формула Лапласа.

в) Для великих n і малих p використовується або теорема додавання імовірностей з наступним застосуванням формули Пуассона, або здійснюється перехід до протилежної події з наступним використанням теореми додавання імовірностей і формули Пуассона. При виборі однієї із альтернатив слід керуватися мінімізацією числа доданків в теоремі додавання імовірностей. Якщо n велике, а q мале і $\lambda' = nq \le 9$, тоді потрібно перейти до числа невиконання події A, а потім виконати рекомендації початку цього підпункту.

Зауваження. Якщо в n повторних незалежних випробуваннях потрібно знайти імовірності $P(m \le k)$ та $P(m \ge k)$, де k ціле число, що не перевищує n, тоді потрібно скористатися рівностями $P(m \le k) = P_n(0 \le m \le k)$, $P(m \ge k) = P_n(k \le m \le n)$ і перейти до п. 2 алгоритму.

Розглянемо реалізацію наведених вище рекомендацій.

РОЗВ'ЯЗУВАННЯ ТИПОВИХ ЗАДАЧ

Задача 3.1. Компанія володіє мережею дилерів на біржі. Імовірність того, що діл ер буде грати вдало, становить 0,7. 1) Знайти імовірність того, що з п'яти дилерів будуть у збитках: а) два; б) хоча б два (вважається, що дії дилерів на біржі є незалежними). 2) Знайти найімовірніше число дилерів, які будуть грати вдало, а також імовірність такої кількості.

○ Випробування — гра діл ера. Оскільки дилерів є 5, то $n = 5$. 1) Подія A — збиткова гра діл ера. За умовою $P(\overline{A}) = q = 0{,}7$, тоді $p = 1 - q = 0{,}3$. Для випадку а) число появи події A $m = 2$, і потрібно знайти $P_5(2)$. Так як $n = 5$ — мале, то використаємо формулу Бернуллі:

$$P_5(2) = C_5^2 (0{,}3)^2 \cdot (0{,}7)^3 = 10 \cdot 0{,}09 \cdot 0{,}343 = 0{,}3087.$$

б) $P_5(m \ge 2) = P_5(2 \le m \le 5)$ — шукана імовірність. Хоча вона візуально нагадує ліву частину інтегральної формули Лапласа, останню використовувати не можна, бо $npq = 5 \cdot 0{,}3 \cdot 0{,}7 = 1{,}05 << 9$. Випадкова подія $(2 \le m \le 5)$ може відбутися тоді, коли або $(m = 2)$, або $(m = 3)$, або $(m = 4)$, або $(m = 5)$, тобто $(2 \le m \le 5) = (m = 2) + (m = 3) + (m = 4) + (m = 5)$. Випадкові події-доданки справа є попарно несумісними, тому згідно з теоремою

додавання ймовірностей

$$P_5(2 \le m \le 5) = P(m = 2) + P(m = 3) + P(m = 4) + P(m = 5) =$$
$$= P_5(2) + P_5(3) + P_5(4) + P_5(5),$$

де в останній рівності було враховано, зокрема, те, що випадкова подія ($m = 2$) — в п'яти випробуваннях подія A відбудеться рівно два рази, тобто $P(m = 2) = P_5(2)$. Використавши формулу Бернуллі, отримаємо:

$$P_5(3) = C_5^3 (0{,}3)^3 \cdot (0{,}7)^2 = 0{,}1323;$$
$$P_5(4) = C_5^4 (0{,}3)^4 \cdot 0{,}7 = 0{,}02835;$$
$$P_5(5) = C_5^5 (0{,}3)^5 \cdot 0{,}7^0 = 0{,}00243.$$

Остаточно

$$P_5(2 \le m \le 5) = 0{,}3087 + 0{,}1323 + 0{,}028535 + 0{,}00243 = 0{,}47178.$$

Другий метод. Події ($m \ge 2$) та ($m < 2$) протилежні, тому

$$P_5(m \ge 2) = 1 - P_5(m < 2) = 1 - [P_5(0) + P_5(1)] =$$
$$= 1 - [C_5^0 (0{,}3)^0 (0{,}7)^5 + C_5^1 (0{,}3)^1 (0{,}7)^4] =$$
$$= 1 - (0{,}16807 + 0{,}36015) = 0{,}47178.$$

Висновок: другий метод значно швидше веде до мети, його ефективність ще більш відчутна при збільшенні числа доданків.

2) Подія A — вдала гра дилерів. За умовою задачі $n = 5$, $p = 0{,}7$, $q = 0{,}3$. Найімовірніше число m_0 дилерів, які будуть грати вдало, знайдемо за подвійною нерівністю (3.4)

$$np - q \le m_0 \le np + p.$$

Підставивши значення в ліву та праву частини, знайдемо $3{,}2 \le m_0 \le 4{,}2$, звідки з врахуванням того, що m_0 — ціле число, остаточно отримаємо: $m_0 = 4$. Нарешті,

$$P_5(m_0) = P_5(4) = C_5^4 (0{,}7)^4 \cdot 0{,}3 = 0{,}36015. \quad \bullet$$

Задача 3.2. Цех отримав замовлення на термінове виготовлення 105 виробів. Проте кожні три вироби із десяти виготовлених вимагають тривалої доводки і тому не можуть бути включені у партію виробів термінового замовлення.

1) Знайти, скільки потрібно виготовити цеху виробів, щоб число 105 було найімовірнішим. 2) Враховуючи знайдене число виробів, оцінити можливість виконання замовлення. При цьому вважати, що продукція є високоліквідною.

○ Випробування – виготовлення виробу. Подія A — виріб стандартний (не вимагає доводки). За умовою $p=P(A)=0{,}7$, тоді $q=0{,}3$.

1) Знайдемо число випробувань n, використавши інформацію про те, що найімовірніше число появи події A дорівнює 105, тобто

m_0=105. Згідно із співвідношенням (3.4)

$np - q \le m_0 \le np + p$.

Підставляючи дані задачі, отримаємо систему нерівностей для знаходження n:

$$\begin{cases} 0,7n - 0,3 \le 105, \\ 0,7n + 0,7 \ge 105, \end{cases} \Leftrightarrow \begin{cases} 0,7n \le 105,3 \\ 0,7n \ge 104,3 \end{cases} \Leftrightarrow 149 \le n \le 150,429.$$

Оскільки n – ціле, то остаточно отримаємо, що n=149 або n=150.

2) Знайдемо імовірність виконання замовлення при виготовленні 149 виробів (при цьому врахуємо можливість випуску більшої кількості виробів від замовленої, оскільки за умовою понадпланова продукція буде реалізована згодом):

$P_{149}(m \ge 105) = P_{149}(105 \le m \le 149)$.

Використаємо інтегральну формулу Лапласа

$(npq = 149 \cdot 0,7 \cdot 0,3 = 31,29 >> 9)$:

$$x_1 = \frac{105 - 149 \cdot 0,7}{\sqrt{31,29}} = 0,125\ ,\ x_2 = \frac{149 - 149 \cdot 0,7}{\sqrt{31,29}} = 7,99\ ,$$

$$P_{149}(105 \le m \le 149) \approx \Phi(7,99) - \Phi(0,125) = 0,5 - 0,0474 = 0,45026\ .$$

Якщо ж n=150, тоді

$$x_1 = \frac{105 - 150 \cdot 0,7}{\sqrt{150 \cdot 0,7 \cdot 0,3}} = 0\ ,\ x_2 = \frac{150 - 150 \cdot 0,7}{5,6125} = 8,02\ ,$$

$$P_{150}(105 \le m \le 150) \approx \Phi(8,02) - \Phi(0) = 0,5\ .$$ ●

Задача 3.3. При штампуванні деталей продукція вищої якості можлива у 35 випадках із 100. Знайти імовірність того, що з 400 навмання відібраних деталей вищої якості виявилось: а) 130 деталей; б)) не більше 130.

○ Випробування — відбір деталі, за умовою n = 400. Подія A — відібрана деталь має вищу якість. За умовою $P(A)$ = 0,35.

а) Потрібно знайти $P_{400}(130)$. Оскільки n = 400 — велике, p та q немалі і виконується нерівність $npq = 400 \cdot 0,35 \cdot 0,65 = 91 > 9$, то потрібно вибрати локальну формулу Лапласа, яка в даному випадку дасть високу точність наближення. Згідно із (3.5)

$$x = \frac{m - np}{\sqrt{npq}} = \frac{130 - 400 \cdot 0,35}{\sqrt{400 \cdot 0,35 \cdot 0,65}} = -1,05,$$

$$\varphi(-1,05) = \varphi(1,05) = 0,2299 \text{ (за табл. 1 додатків)},$$

$$P_{400}(130) \approx \frac{1}{\sqrt{400 \cdot 0,35 \cdot 0,65}} \varphi(-1,05) = \frac{0,2299}{9,54} = 0,0241.$$

б) Імовірність $P_{400}(m \le 130) = P_{400}(0 \le m \le 130)$ знову обчислюємо за інтегральною формулою Лапласа:

$$x_1 = \frac{0 - 400 \cdot 0,35}{9,54} = -14,68; \quad x_2 = \frac{130 - 400 \cdot 0,35}{9,54} = -1,05,$$

$$P_{400}(0 \le m \le 130) \approx \Phi(-1,05) - \Phi(-14,68) = \Phi(14,68) - \Phi(1,05) =$$

$$= 0,5 - 0,3531 = 0,1469.$$

В останніх рівностях використана непарність $\Phi(x)$. ●

Задача 3.4. При скануванні текстового матеріалу в середньому на кожну тисячу символів два помилкові. Знайти імовірність того, що після сканування тексту обсягом в 2 500 символів виявиться помилкових: а) шість символів; б) хоча б шість.

○ Випробування — сканування символу тексту, подія A — отримання помилкового символу. За умовою $n = 2500$, $p = P(A) = 0,002$.

а) Число появи події $m = 6$. Для знаходження $P_{2500}(6)$ скористаємося формулою Пуассона, оскільки n — велике, $p = 0,002 << 0,1$, $\lambda = np = 5 < 9$.

$$P_n(m) \approx \frac{\lambda^m e^{-\lambda}}{m!}, \quad P_{2500}(6) \approx \frac{5^6 \cdot e^{-5}}{6!} = 0,146223.$$

Останнє значення знайдене для функції $\dfrac{\lambda^m e^{-\lambda}}{m!}$ двох змінних λ та m в табл. 2 додатків для значень $\lambda = 5$, $m = 6$.

б) Випадкова подія ($m \ge 6$), імовірність якої треба знайти, зображається через прості події таким чином:

$$(m \ge 6) = (m = 6) + (m = 7) + ... + (m = 2500).$$

Використовувати теорему додавання ймовірностей в такому випадку практично неможливо в зв'язку з тим, що в правій частині є 2495 доданків (!). З другого боку, протилежною до події ($m \ge 6$) є подія ($m < 6$), для якої виконується рівність

$$(m < 6) = (m = 0) + (m = 1) + (m = 2) + (m = 3) + (m = 4) + (m = 5),$$

звідки після використання теореми додавання ймовірностей отримуємо

$$P_{2500}(m < 6) = P_{2500}(0) + P_{2500}(1) + P_{2500}(2) + P_{2500}(3) + P_{2500}(4) + P_{2500}(5).$$

Кожний із доданків обчислюється за формулою Пуассона. В даному випадку ці імовірності знаходяться за табл. 2 додатків для $\lambda = 5$ та $m = 0, 1, ..., 5$. Тобто

$$P_{2500}(m < 6) = 0,00674 + 0,03369 + 0,08422 + 0,14037 + 0,17547 + 0,17547 = 0,61596.$$

З врахуванням протилежності подій

$$P_{2500}(m \geq 6) = 1 - P_{2500}(m < 6) = 1 - 0{,}61596 = 0{,}38404. \bullet$$

Задача 3.5. Імовірність того, що виріб задовольняє вимогам вищого сорту, дорівнює 0,8.

1) За місяць ВТК заводу перевірено 400 виробів. Знайти імовірність того, що відносна частота виготовлення виробу вищого сорту відхилиться від його імовірності по модулю не більше від 0,09.

2) Скільки виробів треба перевірити, щоб з імовірністю 0,95 можна було очікувати відхилення відносної частоти виготовлення виробу вищого сорту від його імовірності по модулю не більше від 0,04?

3) За наступні два місяці ВТК перевірив 900 виробів. Знайти з імовірністю 0,95 межі, в яких буде знаходитися число m виробів вищого ґатунку серед перевірених.

○ Випробування — перевірка виробу. Подія A — виріб задовольняє якостям вищого сорту.

За умовою задачі $n = 400$, $p = P(A) = 0{,}8$, $q = 0{,}2$. Використовуючи формулу (3.8)

$$P\left(|m/n - p| \leq \varepsilon\right) \approx 2\Phi\left(\varepsilon\sqrt{n/(pq)}\right),$$

де за умовою $\varepsilon = 0{,}09$, отримаємо

$$P\left(|m/400 - 0{,}8| \leq 0{,}09\right) \approx 2\Phi\left(0{,}09\sqrt{400/(0{,}8 \cdot 0{,}2)}\right) =$$

$$= 2\Phi(4{,}5) = 2 \cdot 0{,}499997 = 0{,}999994.$$

2) Згідно з умовою задачі $p = 0{,}8$, $q = 0{,}2$, $\varepsilon = 0{,}04$ і $P\left(|m/n - 0{,}8| \leq 0{,}04\right) = 0{,}95$. Потрібно знайти n.

З формули (3.8) знайдемо рівність

$$2\Phi\left(0{,}04\sqrt{n/(0{,}8 \cdot 0{,}2)}\right) = 0{,}95,$$

звідки

$$\Phi\left(0{,}1\sqrt{n}\right) = 0{,}475.$$

За таблицею значень функції Лапласа $\left(\Phi(1{,}96) = 0{,}475\right)$ відтворимо аргумент, для якого виконується остання рівність: $0{,}1\sqrt{n} = 1{,}96$. Далі $\sqrt{n} = 19{,}6, \quad n = 384{,}16$.

Враховуючи те, що n — ціле, а також поведінку похибки в наближеній рівності (3.8), остаточно отримуємо: $n \geq 385$.

3) За умовою $n = 900$, $p = 0{,}8$, $q = 0{,}2$, $P\left(|m/n - 0{,}8| \leq \varepsilon\right) = 0{,}95$. Потрібно знайти межі для числа m. Знайдемо спочатку ε, використавши згідно із формулою (3.8) рівність

$$2\Phi\left(\varepsilon\sqrt{900/(0{,}8\cdot 0{,}2)}\right)=0{,}95 \quad \text{або} \quad \Phi(75\varepsilon)=0{,}475.$$

За табл. 3 додатків знайдемо $\Phi(1{,}96)=0{,}475$, тому $75\varepsilon=1{,}96$, звідки $\varepsilon\approx 0{,}03$.

Таким чином, з імовірністю 0,95 відхилення відносної частоти числа виробів вищої якості від імовірності 0,8 задовольняє нерівності

$$\left|m/900-0{,}8\right|\le 0{,}03 \quad \text{або} \quad 0{,}77\le m/900\le 0{,}83,$$

і остаточно $693\le m\le 747$. ●

Задача 3.6. Підприємство відправило замовнику 20000 стандартних виробів. Середнє число виробів, які пошкоджуються при транспортуванні, складає 0,02%. Знайти імовірність того, що замовник отримає непошкоджених виробів: а) 19997; б) хоча б 19997.

○ Випробування – транспортування одного виробу. Подія A – виріб залишиться стандартним після транспортування. За умовою n=20000, p=$P(A)$=0,9998.

а) Потрібно знайти $P_{20000}(19997)$. Оскільки $npq=20000\cdot 0{,}9998\cdot 0{,}0002=3{,}99992<<9$, то використання локальної формули Лапласа призведе до великої похибки. З другого боку, для партії 20000 виробів випадкові події "19997 стандартних (непошкоджених) виробів" і "3 пошкоджені вироби" – рівносильні. Тому $P_{20000}(19997)$= $P_{20000}(3)$, де в останній імовірності вже в якості події A фігурує подія $\overline{A}$ і те, що $\lambda=np=4<9$, а n велике, дозволяє використати формулу Пуассона (3.6):

$$P_{20000}(3)\approx\frac{4^3\cdot e^{-4}}{3!}.$$

За табл.2 додатків знайдемо значення функції $P(m;\lambda)$ при m=3, $\lambda=4$; $P(3;\ 4)$=0,19537. Отже, шукана імовірність дорівнює 0,19537.

б) Використаємо наведений вище перехід від події A до події $\overline{A}$ - пошкодження виробу внаслідок транспортування. Тоді шукана імовірність позначиться $P_{20000}(m\le 3)$, де p=0,0002.

$$P_{20000}(m\le 3)=P_{20000}(0)+P_{20000}(1)+P_{20000}(2)+P_{20000}(3).$$

Кожний із доданків справа обчислимо за формулою Пуассона, використавши табл.2 додатків (знаходимо значення функції $P(m;\lambda)$ для $\lambda=4$ і m=0, 1, 2, 3):

$$P_{20000}(m\le 3)=0{,}01832+0{,}07326+0{,}14653+0{,}19537=0{,}43348.$$ ●

Зауваження. Використання локальної формули Лапласа в п. а) дає імовірність 0,17604, а інтегральної в п. б) — 0,28579.

§ 4. ДИСКРЕТНІ ВИПАДКОВІ ВЕЛИЧИНИ ТА ЇХ ЧИСЛОВІ ХАРАКТЕРИСТИКИ

1. *Випадкові величини та їх види. Закон розподілу ймовірностей дискретної випадкової величини.*
2. *Основні розподіли дискретних (цілочисельних) випадкових величин (рівномірний, біноміальний, пуассонівський, геометричний, гіпергеометричний). Найпростіший потік подій.*
3. *Дії над випадковими величинами.*
4. *Числові характеристики дискретних випадкових величин та їх властивості (математичне сподівання, дисперсія, середнє квадратичне відхилення).*
5. *Числові характеристики основних законів розподілу.*

1. Випадковою називається величина, яка при випробуванні набирає єдиного значення із всіх можливих з деякою імовірністю, тобто наперед невідомо, яке конкретне можливе значення вона набере, оскільки це залежить від випадкових причин.

Дискретною (**перервною**) називається випадкова величина, можливі значення якої є ізольованими числами. Число можливих значень дискретної випадкової величини може бути скінченним або зчисленним. В останньому випадку можна встановити взаємно-однозначну відповідність між можливими значеннями і натуральними числами 1, 2, 3, ..., n, ...

Неперервною називається випадкова величина, можливі значення якої заповнюють суцільно деякий скінченний або нескінченний проміжок. Очевидно, що число можливих значень кожної неперервної величини нескінченне.

Зауваження. Означення неперервної випадкової величини має попередній характер. Уточнення буде зроблено в наступному параграфі.

Інформації про множину можливих значень недостатньо для повного описання випадкової величини (різні величини можуть мати однакові можливі значення). Потрібно ще знати, з якими ймовірностями набираються можливі значення випадковою величиною.

Законом розподілу ймовірностей дискретної випадкової величини називається відповідність між можливими значеннями та ймовірностями, з якими вони набираються випадковою величиною.

Таблична форма задання закону розподілу має такий вид

$$\begin{array}{c|cccc} X & x_1 & x_2 & \dots & x_n \\ \hline P & p_1 & p_2 & \dots & p_n \end{array}, \qquad (4.1)$$

де $p_i = P(X = x_i)$, $i = \overline{1,n}$. Оскільки в одному випробуванні випадкова

величина набирає тільки одне із своїх можливих значень, то випадкові події $(X = x_1)$, $(X = x_2)$, ..., $(X = x_n)$ утворюють повну групу. Тому сума їх імовірностей дорівнює одиниці:

$$p_1 + p_2 + \dots p_n = 1. \tag{4.2}$$

Ця рівність називається **умовою нормування**.

Якщо множина можливих значень дискретної випадкової величини зчисленна: $x_1, x_2, \dots, x_n, \dots$, то ряд $\sum_{t=1}^{\infty} p_i$ збігається і його сума дорівнює одиниці.

2. Основні розподіли дискретних (цілочисельних) випадкових величин.

Закон розподілу дискретної випадкової величини X можна задати також **аналітично**, тобто з допомогою формули $p_i = P(X = x_i) = g(x_i)$, $i = \overline{1,n}$. Всі нижче наведені закони задаються аналітично.

- Цілочисельна випадкова величина **розподілена за рівномірним законом** (**рівномірно розподілена**), якщо імовірності в законі розподілу мають такий вид: $p_k = P(X = k) = 1/n$, $k = \overline{1,n}$.
- Закон розподілу цілочисельної випадкової величини, імовірності якого знаходяться за формулою Бернуллі, називається **біноміальним:**

$$p_{m+1} = P(X = m) = C_n^m p^m q^{n-m}, \tag{4.3}$$

де $m = 0, 1, 2, \dots, n$; $q = 1 - p$.

- Закон розподілу цілочисельної випадкової величини, імовірності якого знаходяться за формулою Пуассона:

$$p_{m+1} = P(X = m) = \frac{\lambda^m e^{-\lambda}}{m!}, \quad m = 0,\ 1,\ 2,\ \dots,\ n,$$

називається **пуассонівським** (n — велике, $p << 0{,}1$, $\lambda = np \le 9$).

- Закон розподілу ймовірностей, який визначається послідовністю ймовірностей :

$$p_1 = p,\ p_2 = qp,\ p_3 = q^2 p,\ \dots,\ p_n = q^{n-1} p,\ \dots \tag{4.4}$$

називається **геометричним**.

- Закон розподілу ймовірностей, який визначається співвідношенням

$$P(X = m) = C_M^m \cdot C_{N-M}^{n-m} \big/ C_N^n, \quad m = 0,\ 1,\ 2,\ \dots,\ n. \tag{4.5}$$

називається **гіпергеометричним,** де N — загальна кількість об'єктів,

серед яких M володіють певною ознакою ($M < N$), при цьому навмання вибирається n виробів ($n \le M$) без повернення кожного із них.

Найпростіший потік подій

Потоком подій називається послідовність подій, які відбуваються у випадкові моменти часу.

Найпростішим називається **потік подій**, який має властивості **стаціонарності**, **відсутності післядії** та **ординарності**.

Властивість стаціонарності полягає в тому, що імовірність появи m подій потоку за будь-який проміжок часу довжиною t залежить тільки від m і t, і не залежить від початку відліку часу; при цьому різні проміжки часу не повинні перетинатись.

Властивість відсутності післядії визначає те, що ймовірність появи m подій на довільному проміжку часу не залежить від появи чи непояви подій потоку в моменти часу, що передують початку цього проміжку.

Інтенсивністю λ потоку називається середнє число подій, які відбуваються за одиницю часу.

Математичною моделлю найпростішого потоку подій є формула Пуассона

$$P_t(m) = (\lambda t)^m e^{-\lambda t} / m!, \tag{4.6}$$

з допомогою якої можна знайти імовірність появи m подій найпростішого потоку за проміжок часу довжиною t.

3. Дії над випадковими величинами.

Визначимо **добуток сталої величини C на дискретну випадкову величину X**, задану законом розподілу (4.1), як дискретну випадкову величину CX, закон розподілу якої має вид

$$\begin{array}{c|cccc} CX & Cx_1 & Cx_2 & \dots & Cx_n \\ \hline P & p_1 & p_2 & \dots & p_n \end{array}. \tag{4.7}$$

Квадрат випадкової величини X, тобто X^2 — це нова дискретна випадкова величина, яка описується законом розподілу

$$\begin{array}{c|cccc} X^2 & x_1^2 & x_n^2 & \dots & x_n^2 \\ \hline P & p_1 & p_2 & \dots & p_n \end{array}. \tag{4.8}$$

Нехай випадкова величина X задана розподілом (4.1), а Y — законом розподілу $\begin{array}{c|cccc} Y & y_1 & y_2 & \dots & y_m \\ \hline P & g_1 & g_2 & \dots & g_m \end{array}$. Ці величини називаються **незалежними**, якщо випадкові події $(X = x_i)$, $(Y = y_j)$ незалежні при довільних $i = \overline{1,n}$ та $j = \overline{1,m}$. Поняття незалежності випадкових величин поши-

рюється на довільне скінченне число випадкових величин. В протилежному випадку випадкові величини називаються **залежними**.

Дискретні випадкові величини $X_1, X_2, ..., X_k$ називаються **незалежними у сукупності** (**взаємно незалежними**), якщо закон розподілу **кожної** із них не змінюється, якщо довільна випадкова величина або будь-які групи цих величин наберуть яке завгодно із своїх можливих значень.

Визначимо **суму випадкових величин** X та Y як випадкову величину $X + Y$, можливі значення якої рівні сумам кожного можливого значення X з кожним можливим значення Y, а імовірності можливих значень $X + Y$ для незалежних величин X та Y дорівнюють добуткам імовірностей доданків; для залежних величин — добуткам імовірностей одного доданку на умовну імовірність другого. Якщо деякі суми $x_i + y_j$ виявляються рівними між собою, тоді імовірність можливого значення суми дорівнює сумі відповідних імовірностей.

Визначимо **добуток незалежних випадкових величин** X та Y як випадкову величину XY, можливі значення якої рівні добуткам кожного можливого значення X на кожне можливе значення Y; імовірності можливих значень добутку XY дорівнюють добуткам імовірностей можливих значень співмножників. У випадку рівності добутків x_iy_j імовірність можливого значення XY рівна сумі відповідних імовірностей.

4. Числові характеристики дискретних випадкових величин та їх властивості.

Закон розподілу ймовірностей дає повну інформацію про дискретну випадкову величину. Проте в багатьох практично важливих випадках економісту-досліднику буває достатньо знати одне або кілька чисел, пов'язаних із випадковою величиною, які дають менш повне, але більш наочне описання випадкової величини. Такі числа, які сумарно описують випадкову величину, називаються її **числовими характеристиками**.

Математичне сподівання

Математичним сподіванням дискретної випадкової величини X називається сума добутків всіх її можливих значень на відповідні імовірності:

$$M(X) = x_1p_1 + x_2p_2 + ... + x_np_n. \qquad (4.9)$$

Якщо величина X може набирати зчисленну множину значень (наприклад, як у випадку геометричного або пуассонівського розподілу), то при умові, що цей ряд абсолютно збіжний,

$$M(X) = \sum_{i=1}^{\infty} x_i p_i . \qquad (4.10)$$

Імовірносний зміст математичного сподівання: ***M*(*X*) приблизно дорівнює** (тим точніше, чим більше число випробувань) **середньому арифметичному спостережених значень випадкової величини**.

Властивості математичного сподівання.

1. **Математичне сподівання сталої величини дорівнює цій сталій:**

$$M(C) = C.$$

2. **Сталий множник можна виносити за знак математичного сподівання:** $M(CX) = CM(X)$.

3. **Математичне сподівання суми двох випадкових величин дорівнює сумі математичних сподівань доданків:**

$$M(X + Y) = M(X) + M(Y).$$

Наслідок. Математичне сподівання суми кількох випадкових величин дорівнює сумі математичних сподівань доданків.

4. **Математичне сподівання добутку двох незалежних випадкових величин дорівнює добутку їх математичних сподівань:**

$$M(XY) = M(X) \cdot M(Y).$$

Наслідок. Математичне сподівання добутку кількох незалежних у сукупності випадкових величин дорівнює добутку їх математичних сподівань.

Дисперсія

Для оцінки розкиду можливих значень випадкової величини навколо середнього значення (математичного сподівання) використовують, зокрема, числову характеристику, яку називають **дисперсією**.

Дисперсією (розкидом) випадкової величини X називається математичне сподівання квадрату відхилення X від $M(X)$:

$$D(X) = M[X - M(X)]^2. \qquad (4.11)$$

Згідно із цим означенням дисперсія характеризує середню величину розкиду можливих значень випадкової величини навколо її математичного сподівання (середньої) в квадратних одиницях.

Якщо X розподілена за законом (4.1), то з врахуванням закону (4.8) випадкова величина $[X - M(X)]^2$ має такий закон розподілу:

$[X - M(X)]^2$	$[x_1 - M(X)]^2$	$[x_2 - M(X)]^2$	...	$[x_n - M(X)]^2$
P	p_1	p_2	...	p_n

.

Використавши означення дисперсії та математичного сподівання, отримаємо формулу для обчислення $D(X)$:

$$D(X) = M[X - M(X)]^2 = [x_1 - M(X)]^2 p_1 + [x_2 - M(X)]^2 p_2 + \\ + \ldots + [x_n - M(X)]^2 p_n = \sum_{i=1}^{n} [x_i - M(X)]^2 p_i. \quad (4.11^*)$$

Зменшення об'єму обчислень досягається за рахунок використання **розрахункової формули** для обчислення дисперсії:

$$D(X) = M(X^2) - [M(X)]^2. \quad (4.12)$$

Застосувавши означення математичного сподівання до закону розподілу (4.8), отримаємо **числову реалізацію розрахункової формули**:

$$D(X) = \sum_{i=1}^{n} x_i^2 p_i - [M(X)]^2. \quad (4.12^*)$$

Властивості дисперсії.

1. **Дисперсія сталої дорівнює нулю:** $D(C) = 0$.

2. **Сталий множник можна виносити за знак дисперсії, піднісши його до квадрату:** $D(CX) = C^2 D(X)$.

3. **Дисперсія суми двох незалежних випадкових величин дорівнює сумі дисперсій цих величин:** $D(X + Y) = D(X) + D(Y)$.

Наслідок. Дисперсія суми кількох незалежних у сукупності випадкових величин дорівнює сумі дисперсій цих величин.

4. **Дисперсія різниці двох незалежних випадкових величин дорівнює сумі їх дисперсій:** $D(X - Y) = D(X) + D(Y)$.

Зауваження. Навіть якщо X та Y незалежні, то в загальному випадку виконується нерівність $D(XY) \neq D(X)\, D(Y)$.

Середнє квадратичне відхилення

Незручність використання дисперсії в деяких випадках зумовлена тим, що вона має розмірність, яка дорівнює квадрату розмірності випадкової величини. Наприклад, якщо можливі значення X вимірюються в кг, то $D(X)$ в $(\text{кг})^2$.

Середнім квадратичним відхиленням випадкової величини X називається квадратний корінь із дисперсії:

$$\sigma(X) = \sqrt{D(X)}. \quad (4.13)$$

$\sigma(X)$ характеризує середню величину розкиду можливих значень X навколо $M(X)$ (середньої) в лінійних одиницях.

5. Числові характеристики основних законів розподілу

Числові характеристики випадкової величини, розподіленої за **біноміальним законом:** $M(X) = np$, $D(X) = npq$. (4.14)

Числові характеристики випадкової величини, розподіленої за **законом Пуассона:** $M(X) = D(X) = np = \lambda$. (4.15)

РОЗВ'ЯЗУВАННЯ ТИПОВИХ ЗАДАЧ

Задача 4.1. Складальник навмання бере дві деталі із контейнера, в якому знаходиться 20 деталей, серед яких 4 нестандартні. Знайти закон розподілу числа нестандартних деталей серед відібраних.

○ Нехай X — число нестандартних деталей серед двох відібраних. Можливі значення X — 0, 1, 2. Знайдемо відповідні імовірності, використовуючи класичне означення:

$$p_1 = P(X = 0) = m/n = C_{16}^2 / C_{20}^2 = 12/19,$$

$$p_2 = P(X = 1) = C_{16}^1 C_4^1 / C_{20}^2 = 32/95,$$

$$p_3 = P(X = 2) = C_4^2 / C_{20}^2 = 3/95.$$

Перевірка: $p_1 + p_2 + p_3 = 12/19 + 32/95 + 3/95 = 1$.

Шуканий закон розподілу має такий вид:

X	0	1	2
P	12/19	32/95	3/95

. ●

Задача 4.2. Випадкові величини X та Y задані законами розподілу

X	2	5
P	0,8	0,2

, | Y | 1 | 3 | 6 |
|---|---|---|---|
| P | 0,3 | 0,5 | 0,2 |

. Знайти $X + Y$, якщо X та Y незалежні величини.

○ Якщо X та Y задані законами розподілу відповідно

X	x_1	x_2
P	p_1	p_2

,

Y	y_1	y_2	y_3
P	g_1	g_2	g_3

, то за означенням випадкова величина $X + Y$ описується законом розподілу

$X+Y$	x_1+y_1	x_1+y_2	x_1+y_3	x_2+y_1	x_2+y_2	x_2+y_3
P	p_1g_1	p_1g_2	p_1g_3	p_2g_1	p_2g_2	p_2g_3

.

Підставивши дані, отримуємо:

$X+Y$	$2+1$	$2+3$	$2+6$	$5+1$	$5+3$	$5+6$
P	$0{,}8\cdot 0{,}3$	$0{,}8\cdot 0{,}5$	$0{,}8\cdot 0{,}2$	$0{,}2\cdot 0{,}3$	$0{,}2\cdot 0{,}5$	$0{,}2\cdot 0{,}2$

Можливі значення 2 + 6 та 5 + 3 рівні, тому відповідні імовірності 0,16 та 0,1 додаємо, залишивши в остаточному законі розподілу $X + Y$ одне можливе значення 8:

$X+Y$	3	5	6	8	11
P	0,24	0,4	0,06	0,26	0,04

Закон розподілу XY має такий вид:

XY	2	5	6	12	15	30
P	0,24	0,06	0,4	0,16	0,1	0,04

●

Задача 4.3. Дослідження роботи блоків чотирьох типів в умовах перепаду напруги показало, що імовірність безвідмовної роботи на протязі часу T кожного із типів складає відповідно 0,8; 0,9; 0,7 та 0,75. Відбираються чотири блоки кожного типу. Скласти закон розподілу випадкової величини X – числа блоків, які безвідмовно працюватимуть у вказаних умовах.

○ Задачу можна розв'язати не одним способом.

Перший спосіб.

Нехай, події A_i – блок i-го типу працює безвідмовно, $i=\overline{1,4}$,

$\overline{A}_i$ – блок i-го типу вийшов з ладу.

За умовою $P(A_1)$=0,8, $P(A_2)$=0,9, $P(A_3)$=0,7, $P(A_4)$=0,75, $P(\overline{A}_1)=0{,}2$, $P(\overline{A}_2)=0{,}1$, $P(\overline{A}_3)=0{,}3$, $P(\overline{A}_4)=0{,}25$.

Можливі значення X – 0, 1, 2, 3, 4.

Знайдемо відповідні імовірності:

$P(X=0)=P(\overline{A}_1\overline{A}_2\overline{A}_3\overline{A}_4)=0{,}2\cdot 0{,}1\cdot 0{,}3\cdot 0{,}25=0{,}0015$;

$P(X=1)=P(A_1\overline{A}_2\overline{A}_3\overline{A}_4+\overline{A}_1A_2\overline{A}_3\overline{A}_4+\overline{A}_1\overline{A}_2A_3\overline{A}_4+\overline{A}_1\overline{A}_2\overline{A}_3A_4)=$

$0{,}8\cdot 0{,}1\cdot 0{,}3\cdot 0{,}25+0{,}2\cdot 0{,}9\cdot 0{,}3\cdot 0{,}25+0{,}2\cdot 0{,}1\cdot 0{,}7\cdot 0{,}25+0{,}2\cdot 0{,}1\cdot 0{,}3\cdot 0{,}75=$

$=0{,}006+0{,}0135+0{,}0035+0{,}0045=0{,}0275;$

$P(X=2)=P(A_1A_2\overline{A}_3\overline{A}_4+A_1\overline{A}_2A_3\overline{A}_4+A_1\overline{A}_2\overline{A}_3A_4+\overline{A}_1A_2A_3\overline{A}_4+\overline{A}_1A_2\overline{A}_3A_4+$

$\overline{A}_1\overline{A}_2A_3A_4)=0{,}1685;$

$P(X=3)=P(A_1A_2A_3\overline{A}_4+A_1A_2\overline{A}_3A_4+A_1\overline{A}_2A_3A_4+\overline{A}_1A_2A_3A_4)=0{,}4245;$

$P(X=4)=P(A_1A_2A_3A_4)=0{,}378$

Остаточно отримаємо шуканий закон розподілу:

X	0	1	2	3	4
P	0,0015	0,0275	0,1685	0,4245	0,378

(*)

Другий спосіб. Позначимо випадкові виличини: X_k – число блоків k-го типу, які безвідмовно працюватимуть в умовах перепаду напруги ($k = \overline{1,4}$).

Оскільки тип представлений тільки одним блоком, то X_k може набирати два можливих значення: 0 (блок вийшов з ладу) і 1 (блок безвідмовно працюватиме). Із врахуванням умови задачі отримаємо закони розподілу.

X_1	0	1
P	0,2	0,8

, | X_2 | 0 | 1 |
|---|---|---|
| P | 0,1 | 0,9 |

, | X_3 | 0 | 1 |
|---|---|---|
| P | 0,3 | 0,7 |

, | X_4 | 0 | 1 |
|---|---|---|
| P | 0,25 | 0,75 |

.

Число (**загальне**) блоків, які безвідмовно працюватимуть у вказаних умовах, складається із числа блоків кожного типу, що витримають випробування:

$$\boldsymbol{X=X_1+X_2+X_3+X_4.}$$

Позначимо $Y=X_1+X_2$; $Z=Y+X_3$; $X=Z+X_4$.

За аналогією із розв'язуванням задачі 4.2 послідовно отримаємо:

$Y = X_1 + X_2$	0	1	2
P	0,02	0,26	0,72

,

$Z = Y + X_3$	0	1	2	3
P	0,006	0,092	0,398	0,504

і, нарешті, закон розподілу для $X=Z+X_4$, який співпадає із (*). ●

Задача 4.4. Існує три методи експрес-контролю партії виробів. Кожний із них безпомилково виявляє якісні вироби, а при ідентифікації некондиційних допускає помилки, тобто вони визнаються якісними. Число таких виробів для кожного із методів є відповідно випадковими величинами X, Y та Z, закони розподілу яких мають такий вид:

X	0	1	3	4
P	0,5	0,4	0,06	0,04

, | Y | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| P | 0,6 | 0,2 | 0,1 | 0,1 |

, | Z | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| P | 0,8 | 0,1 | 0,04 | 0,03 | 0,03 |

.

Який із методів експрес-контролю кращий?

○ Нехай a — кількість стандартних виробів у партії. Тоді згідно з умовою задачі кожен із методів контролю відповідно дасть $a + X$, $a + Y$, $a + Z$ стандартних деталей після перевірки всієї партії, де кожний другий доданок у сумах — це помилка при ідентифікації бракованих виробів. Мінімальне значення її (у середньому) визначить кращий метод. Характеристикою середнього значення є мате-

матичне сподівання. Тому знайдемо

$$M(X) = 0 \cdot 0{,}5 + 1 \cdot 0{,}4 + 3 \cdot 0{,}06 + 4 \cdot 0{,}04 = 0{,}74;$$
$$M(Y) = 0 \cdot 0{,}6 + 1 \cdot 0{,}2 + 2 \cdot 0{,}1 + 3 \cdot 0{,}1 = 0{,}7;$$
$$M(Z) = 0 \cdot 0{,}8 + 1 \cdot 0{,}1 + 2 \cdot 0{,}04 + 3 \cdot 0{,}03 + 4 \cdot 0{,}03 = 0{,}39.$$

Отже, кращим методом експрес-контролю є третій. ●

Задача 4.5. Знайти середнє квадратичне відхилення випадкової величини $Z = 2X - 5Y - 30$, якщо X та Y — незалежні випадкові величини, $D(X) = 0{,}25$, $D(Y) = 0{,}2$.

○ Рівність випадкових величин зумовлює рівність їх дисперсій: $D(Z) = D(2X - 5Y - 30)$. Використання властивостей дисперсії та умов задачі дає такий ланцюжок рівностей:

$$D(Z) = D(2X + (-5Y) + (-30)) = D(2X) + D(-5Y) + D(-30) =$$
$$= 2^2 D(X) + (-5)^2 D(Y) + 0 = 4 \cdot 0{,}25 + 25 \cdot 0{,}2 = 6.$$

Тоді згідно із (4.13) $\sigma(Z) = \sqrt{D(Z)} = \sqrt{6} \approx 2{,}45$. ●

Задача 4.6. Середнє число обривів нитки на прядильному верстаті за 1 хв. дорівнює двом. Знайти імовірність того, що за 3 хв. число обривів нитки становитиме: а) 4; б) менше чотирьох; в) не менше чотирьох. Припускається, що потік обривів нитки найпростіший.

○ а) За умовою $\lambda = 2$, $t = 3$, $m = 4$. За формулою (4.6) імовірність того, що за 3 хв. число обривів нитки становитиме 4, $P_3(4) = 6^4 e^{-6}/4! = 0{,}13385$. Числове значення знаходиться з допомогою таблиці значень для формули Пуассона (табл. 2 додатків) для $\boldsymbol{\lambda = 6}$, $m = 4$ (порівняйте формулу (4.6) із формулою Пуассона!).

б) Випадкова подія $(m < 4) = (m = 0) + (m = 1) + (m = 2) + (m = 3)$, звідки із використанням теореми додавання імовірностей попарно несумісних подій

$$P_3(m < 4) = P_3(0) + P_3(1) + P_3(2) + P_3(3) =$$
$$= 0{,}00248 + 0{,}01487 + 0{,}04462 + 0{,}08924 = 0{,}15121.$$

в) Події $(m \geq 4)$ та $(m < 4)$ протилежні, тому

$$P_3(m \geq 4) = 1 - P_3(m < 4) = 1 - 0{,}15121 = 0{,}84879.$$ ●

Задача 4.7. Імовірність того, що виготовлений виріб вимагає додаткового регулювання, дорівнює p. Контролер перевіряє якість партії виробів, навмання вибираючи виріб. Якщо він вимагає додаткового регулювання, то наступні випробування припиняються, а вся партія відправляється на доробку. Якщо ж виріб стандарт-

ний, то контролер бере наступний виріб, тощо. Згідно із інструкцією контролер перевіряє не більше п'яти виробів.

1) Скласти закон розподілу числа перевірених контролером виробів. 2) Знайти імовірність доробки всієї партії виробів.

○ 1) Позначимо: X – число виробів, перевірених контролером, A_i — i-тий відібраний виріб вимагає додаткового регулювання ($i=\overline{1,5}$). За умовою $P(A_i)=p$, $i=\overline{1,5}$. Тоді $\overline{A_i}$ – i-тий виріб стандартний. $P(\overline{A_i})=1-p=q$, $i=\overline{1,5}$.

Можливі значення X: 1, 2, 3, 4, 5. Знайдемо імовірності, з якими X набирає ці значення. Випадкова подія (X=1) відбудеться тоді, коли перший відібраний прилад вимагатиме додатково регулювання, тобто (X=1)=A_1, звідки $p_1=P(X=1)=P(A_1)=p$. Відбуття події (X=2) означає, що перший виріб стандартний і другий вимагає регулювання: $(X=2)=\overline{A}_1\cdot A_2$. Використовуючи теорему множення ймовірностей, отримаємо: $p_2=P(X=2)=P(\overline{A}_1\cdot A_2)=qp$. Аналогічно знаходимо, що $p_3=P(X=3)=q^2p$, $p_4=P(X=4)=q^3p$.

Нарешті, подія (X=5) відбудеться або тоді, коли чотири перші вироби стандартні, а п'ятий вимагає регулювання (партія відправляється на доробку), або коли всі п'ять виробів стандартні (партія пропускається), тобто

$$(X=5)=\overline{A}_1\overline{A}_2\overline{A}_3\overline{A}_4A_5+\overline{A}_1\overline{A}_2\overline{A}_3\overline{A}_4\overline{A}_5,$$

звідки $p_5=P(X=5)=q^4p+q^5=q^4(p+q)=q^4$, бо $p+q$=1.

Остаточно шуканий закон розподілу має такий вид:

X	1	2	3	4	5
P	p	qp	q^2p	q^3p	q^4

.

Для перевірки з'ясуємо, чи виконується умова нормування:

$$\sum_1^5 p_i=p+qp+q^2p+q^3p+q^4=p+qp+q^2p+q^3(p+q)=$$

$$=p+qp+q^2(p+q)=p+q(p+q)=p+q=1.$$

2) Партія пропускається контролером в єдиному випадку, коли всі п'ять перевірених виробів є стандартними, імовірність цієї випадкової події дорівнює q^5. Тому імовірність протилежної події (партія виробів відправляється на доробку) дорівнює $1-q^5$ або $1-(1-p)^5$. ●

§ 5. НЕПЕРЕРВНІ ВИПАДКОВІ ВЕЛИЧИНИ ТА ЇХ ЧИСЛОВІ ХАРАКТЕРИСТИКИ

1. *Функція розподілу ймовірностей і її властивості.*
2. *Густина розподілу ймовірностей та її властивості.*
3. *Числові характеристики неперервних випадкових величин (математичне сподівання, дисперсія та середнє квадратичне відхилення, мода та медіана випадкової величини).*

1. Функцією розподілу ймовірностей називається функція $F(x)$ детермінованого (невипадкового) аргументу x, яка чисельно дорівнює імовірності того, що в результаті випробування випадкова величина X набере значення, менше від x, тобто

$$F(x) = P(X < x). \tag{5.1}$$

Іноді функцію розподілу називають ще інтегральною.

Уточнимо означення неперервної випадкової величини: випадкову величину називають **неперервною**, якщо її функція розподілу ймовірностей є неперервною на області її визначення, а похідна від функції розподілу неперервна у всіх точках, за виключенням, можливо, скінченого числа точок на довільному скінченому інтервалі.

Властивості функції розподілу ймовірностей

1. Область визначення функції розподілу — $R^1 = (-\infty; \infty)$, а область значень — відрізок [0; 1].

2. $F(x)$ — неспадна функція, тобто для будь-якої пари чисел x_1, x_2 з нерівності $x_2 > x_1$ випливає нерівність $F(x_2) \geq F(x_1)$.

Наслідок 1. Імовірність того, що випадкова величина при випробуванні набере можливого значення з проміжку $[a; b)$, дорівнює приросту функції розподілу на цьому проміжку:

$$P(a \leq X < b) = F(b) - F(a). \tag{5.2}$$

Наслідок 2. Імовірність того, що неперервна випадкова величина X набере при випробуванні одне конкретне можливе значення, дорівнює нулю.

Зауваження. Рівність $P(X = x_1) = 0$, де x_1 — конкретне можливе значення величини X, не означає, що подія $X = x_1$ є неможливою (на відміну від класичного означення імовірності). Нагадайте геометричне означення імовірності.

Наслідок 2 з використанням формули (5.2) та теореми додавання ймовірностей дозволяє отримати такий ланцюжок рівностей:

$$P(a < X < b) = \big(P(a \le X \le b) = P(a \le X < b) = P(a < X \le b)\big) = \\ = F(b) - F(a). \tag{5.3}$$

3. Якщо всі можливі значення випадкові величини належать відрізку $[a; b]$, то $F(x) = 0$ для всіх $x \le a$ і $F(x) = 1$ для всіх $x > b$.

Наслідок. Якщо можливими значеннями неперервної випадкової величини є всі дійсні числа, тоді мають місце такі граничні співвідношення:

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

2. Густиною розподілу ймовірностей неперервної випадкової величини X називається функція $f(x)$, яка дорівнює першій похідній від функції розподілу $F(x)$:

$$f(x) = F'(x). \tag{5.4}$$

Із означення (5.4) слідує, що $F(x)$ є первісною для густини розподілу. Нижче буде наведено формулу для знаходження $F(x)$, якщо відома функція $f(x)$. Таким чином, задання однієї із функцій $f(x)$ та $F(x)$ дозволяє знайти іншу. Тому в літературі ці функції називають **законами розподілу неперервної випадкової величини**.

Властивості густини розподілу

1. Область визначення функції $f(x)$ — R^1, а область значень проміжок $[0, \infty)$.

2. Невласний інтеграл від густини розподілу в межах від $-\infty$ до ∞ дорівнює одиниці:

$$\int_{-\infty}^{\infty} f(x)dx = 1. \tag{5.5}$$

Рівність (5.5) називають **умовою нормування** неперервної випадкової величини.

Зауваження. Якщо $[\alpha, \beta]$ - мінімальний проміжок, в якому містяться **всі** можливі значення неперервної випадкової величини, тоді умова нормування набирає такого виду:

$$\int_{\alpha}^{\beta} f(x)dx = 1. \tag{5.5*}$$

Мають місце такі формули:

$$P\left(\begin{matrix}(\le)\\ a\end{matrix} < X \begin{matrix}(\le)\\ <\end{matrix} b\right) = \int_a^b f(x)dx, \tag{5.6}$$

$$F(x) = P(X < x) = P(-\infty < X < x) = \int_{-\infty}^{x} f(x)dx. \tag{5.7}$$

Відмітимо, що рівність (5.5) є аналогом рівності одиниці суми ймовірностей із закону розподілу дискретної випадкової величини.

З'ясуємо **ймовірносний зміст** густини розподілу ймовірностей. Означення $f(x)$ (5.4) та похідної функції, а також рівність (5.3) дають такі рівності

$$f(x) = \lim_{\Delta x \to 0} \frac{F(x+\Delta x) - F(x)}{\Delta x} = \lim_{\Delta x \to 0} \frac{P(x < X < x + \Delta x)}{\Delta x}, \tag{5.8}$$

які дозволяють записати наближену рівність

$$P(x < X < x + \Delta x) \approx f(x)\Delta x. \tag{5.9}$$

з точністю до нескінченно малих вищого порядку відносно Δx.

Права частина (5.8) пояснює назву функції $f(x)$ за аналогією зі змістом густини маси, розподіленої на відрізку, а наближена рівність (5.9) — «вклад» величини $f(x)$ у величину імовірності події $(x < X < x + \Delta x)$ для кожного відрізка довжиною Δx.

3. Числові характеристики неперервних випадкових величин.

Математичним сподіванням неперервної випадкової величини X, всі можливі значення якої належать скінченому відрізку $[a; b]$, називається визначений інтеграл

$$M(X) = \int_a^b xf(x)dx. \tag{5.10}$$

Якщо можливі значення X належать R^1, тоді

$$M(X) = \int_{-\infty}^{\infty} xf(x)dx, \tag{5.10*}$$

де за припущенням невласний інтеграл збігається абсолютно.

Як і для випадку дискретної величини, **дисперсією неперервної випадкової величини** називається математичне сподівання квадрату її відхилення від математичного сподівання. Із врахуванням означення $M(X)$:

$$D(X) = \int_a^b \left[x - M(X)\right]^2 f(x)dx, \tag{5.11}$$

якщо всі можливі значення X належать скінченому відрізку $[a; b]$, і

$$D(X) = \int_{-\infty}^{\infty} \left[x - M(X)\right]^2 f(x)dx, \qquad (5.11^*)$$

якщо можливі значення X заповнюють R^1.

Розрахункові формули для обчислення дисперсії:

$$D(X) = \int_{a}^{b} x^2 f(x)dx - \left[M(X)\right]^2, \qquad (5.12)$$

$$D(X) = \int_{-\infty}^{\infty} x^2 f(x)dx - \left[M(X)\right]^2. \qquad (5.12^*)$$

Середнє квадратичне відхилення неперервної випадкової величини визначається, як і для випадку дискретної, рівністю

$$\sigma(X) = \sqrt{D(X)}. \qquad (5.13)$$

Модою $Mo(X)$ дискретної випадкової величини X називається те її можливе значення, якому відповідає найбільша імовірність її появи. Для неперервної випадкової величини X модою $Mo(X)$ називається можливе значення X, якому відповідає локальний максимум густини розподілу. Випадкова величина може мати кілька мод. У цьому випадку вона називається **многомодальною**. Зустрічаються також розподіли, що не мають максимуму. Такі розподіли називаються **антимодальними**.

Медіаною $Me(X)$ неперервної випадкової величини X називається те її можливе значення, для якого виконується рівність $P(X < Me(X)) = P(X > Me(X))$.

РОЗВ'ЯЗУВАННЯ ТИПОВИХ ЗАДАЧ

Задача 5.1. Дискретна випадкова величина X задана законом розподілу

X	1	3	6	8
P	0,2	0,4	0,3	0,1

Знайти функцію розподілу і побудувати її графік.

○ Якщо $x \le 1$, то у відповідності із властивістю 3 $F(x) = 0$.

Нехай $x \in (1; 3]$. Тоді випадкова подія $(X < x) = (X = 1)$, оскільки 1 — єдине можливе значення, яке менше від x. А тому згідно із (5.1) для $(1; 3]$ $F(x) = P(X < x) = P(X = 1) = 0{,}2$.

Якщо $x \in (3; 6]$, тоді подія $(X < x)$ відбувається тоді, і тільки тоді, коли або $X = 1$, або $X = 3$, тобто має місце така рівність $(X < x) = (X = 1) + (X = 3)$, де доданки справа є несумісними випад-

ковими подіями. Використання теореми додавання імовірностей і означення (5.1) дозволяє знайти $F(x)$ на цьому проміжку:

$$F(x) = P(X < x) = P(X = 1) + P(X = 3) = 0{,}2 + 0{,}4 = 0{,}6.$$

Якщо $x \in (6; 8]$, то за аналогією із попереднім проміжком

$$F(x) = P(X < x) = P(X = 1) + P(X = 3) + P(X = 6) = 0{,}9.$$

Нарешті якщо $x > 8$, то подія $(X < x)$ — достовірна, і $F(x) = 1$.

Отже, функція розподілу має такий вид:

$$F(x) = \begin{cases} 0, & \text{якщо } x \le 1, \\ 0{,}2, & \text{якщо } 1 < x \le 3, \\ 0{,}6, & \text{якщо } 3 < x \le 6, \\ 0{,}9, & \text{якщо } 6 < x \le 8, \\ 1, & \text{якщо } x > 8, \end{cases}$$

а її графік зображено на рис. 5.1.

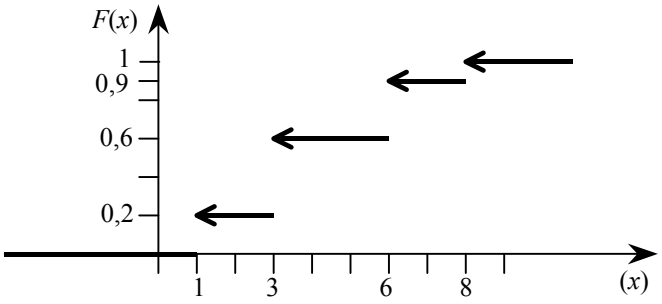


Рис. 5.1.

де стрілками відмічені правосторонні розриви. ●

Задача 5.2. Випадкова величина X задана функцією розподілу

$$F(x) = \begin{cases} 0 & \text{при } x \le -2, \\ x^2/4 + x + 1 & \text{при } -2 < x \le 0, \\ 1 & \text{при } x > 0. \end{cases}$$

Знайти: 1) імовірність того, що в результаті випробування X набере значення з інтервалу $(-1; 0)$; 2) числові характеристики даної випадкової величини; 3) побудувати графіки функцій $F(x)$ та $f(x)$.

○ 1). Для знаходження $P(-1 < X < 0)$ скористаємося однією із формул (5.3), поклавши $a = -1$, $b = 0$. Тоді

$$P(-1 < X < 0) = F(0) - F(-1) =$$

$= 0^2/4 + 0 + 1 - ((-1)^2/4 - 1 + 1) = 1 - 1/4 = 0{,}75.$

Зауваження. Типова помилка студентів при використанні формули (5.3) — неправильний вибір аналітичного виразу, за яким обчислюється $F(x)$ при знаходженні $F(b)$ та (або) $F(a)$. Тобто попередньо не враховується, до якого інтервалу (в задачі 5.2 їх є три) належить a та b.

2). Щоб обчислити числові характеристики неперервної випадкової величини необхідно знайти густину розподілу ймовірностей. Врахувавши (5.4), знайдемо $f(x)$:

$$f(x) = F'(x) = \begin{cases} 0 & при \ \ x \le -2, \\ \dfrac{x}{2} + 1 & при \ -2 < x \le 0, \\ 0 & при \ \ x > 0. \end{cases}$$

Для обчислення $\sigma(X)$ знайдемо спочатку $M(X)$ та $D(X)$ за формулами (5.10) та (5.12):

$$M(X) = \int_{-2}^{0} xf(x)dx = \int_{-2}^{0} x(\frac{x}{2}+1)dx = \int_{-2}^{0} (\frac{x^2}{2}+x)dx =$$

$$= (\frac{x^3}{6} + \frac{x^2}{2})\Big|_{-2}^{0} = 0 - (\frac{(-2)^3}{6} + \frac{(-2)^2}{2}) = -\frac{2}{3};$$

$$M(X^2) = \int_{-2}^{0} x^2 f(x)dx = \int_{-2}^{0} x^2(\frac{x}{2}+1)dx =$$

$$= \int_{-2}^{0} (\frac{x^3}{2} + x^2)dx = (\frac{x^4}{8} + \frac{x^3}{3})\Big|_{-2}^{0} = 0 - \left(\frac{(-2)^4}{8} + \frac{(-2)^3}{3}\right) = \frac{2}{3};$$

$$D(X) = M(X^2) - [M(X)]^2 = \frac{2}{3} - \left(-\frac{2}{3}\right)^2 = \frac{2}{9};$$

$$\sigma(X) = \sqrt{D(X)} = \sqrt{\frac{2}{9}} \approx 0{,}471.$$

Зауваження. В даній задачі всі можливі значення випадкової величини розміщені у проміжку $[-2;0]$, а тому і $M(X) = -\frac{2}{3} \in [-2;0]$ як приблизне середнє арифметичне спостережених значень випадкової величини. При виконанні контрольних завдань слід здійснювати наведену перевірку включення $M(X)$ у

проміжок , в якому знаходяться всі спостережені значення конкретної випадкової величини.

3). Графіки функції розподілу та густини розподілу ймовірності випадкової величини наведені на рис. 5.2 та рис. 5.3 відповідно.

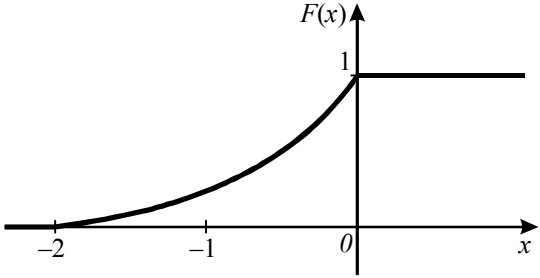


Рис. 5.2

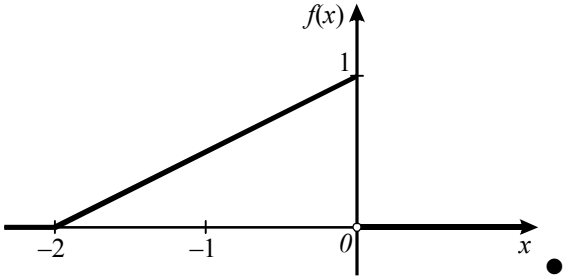


Рис. 5.3

Задача 5.3. Випадкова величина X задана густиною розподілу

$$f(x)=\begin{cases}0 & \text{при } x \le -\pi/2,\\ a\cos x & \text{при } -\pi/2 < x \le \pi/2,\\ 0 & \text{при } x > \pi/2.\end{cases}$$

Знайти параметр a і функцію розподілу ймовірностей, а також імовірність того, що при випробуванні величина X набере значення з інтервалу $(-\pi/4, \pi/3)$.

○ Параметр a знайдемо з рівності (5.6):

$$\int_{-\infty}^{\infty} f(x)dx = 1.$$

Розіб'ємо невласний інтеграл на три інтеграли у відповідності із інтервалами задання густини розподілу:

$$\int_{-\infty}^{\infty} f(x)dx = \int_{-\infty}^{-\pi/2} f(x)dx + \int_{-\pi/2}^{\pi/2} f(x)dx + \int_{\pi/2}^{\infty} f(x)dx =$$

$$= \int_{-\infty}^{-\pi/2} 0dx + \int_{-\pi/2}^{\pi/2} a\cos x dx + \int_{\pi/2}^{\infty} 0dx = a\int_{-\pi/2}^{\pi/2} \cos x dx =$$

$$= a\sin x\Big|_{-\pi/2}^{\pi/2} = a(\sin(\pi/2) - \sin(-\pi/2)) = 2a.$$

Отже, параметр a задовольняє рівнянню $2a = 1$, звідки $a = 0{,}5$. При знаходженні $F(x)$ врахуємо це значення параметра і використаємо рівність (5.7). Відмітимо, що вид густини розподілу в даній задачі вказує на те, що функція розподілу на трьох проміжках також буде задаватися різними аналітичними виразами. В зв'язку із цим нехай $x \le -\pi/2$. Тоді

$$F(x) = \int_{-\infty}^{x} f(x)dx = \int_{-\infty}^{x} 0dx = 0.$$

Для $x \in (-\pi/2; \pi/2]$

$$F(x) = \int_{-\infty}^{x} f(x)dx = \int_{-\infty}^{-\pi/2} f(x)dx + \int_{-\pi/2}^{x} f(x)dx = \int_{-\infty}^{-\pi/2} 0dx + \int_{-\pi/2}^{x} 0{,}5\cos x dx =$$

$$= 0{,}5\sin x\Big|_{-\pi/2}^{x} = 0{,}5(\sin x - \sin(-\pi/2)) = 0{,}5(1 + \sin x).$$

Нарешті, якщо $x > \pi/2$, то

$$F(x) = \int_{-\infty}^{x} f(x)dx = \int_{-\infty}^{-\pi/2} 0dx + \int_{-\pi/2}^{\pi/2} 0{,}5\cos x dx + \int_{\pi/2}^{x} 0dx = 0{,}5\sin x\Big|_{-\pi/2}^{\pi/2} = 1.$$

Таким чином, функція розподілу для даної величини X має такий вигляд:

$$F(x) = \begin{cases} 0 & \text{при } x \le -\pi/2, \\ 0{,}5(1 + \sin x) & \text{при } -\pi/2 < x \le \pi/2, \\ 1 & \text{при } x > \pi/2. \end{cases}$$

Використаємо формулу (5.6) для знаходження шуканої імовірності:

$$P(-\pi/4 < X < \pi/3) = \int_{-\pi/4}^{\pi/3} f(x)dx = \int_{-\pi/4}^{\pi/3} 0{,}5\cos x dx = 0{,}5\sin x\Big|_{-\pi/4}^{\pi/3} =$$

$$= 0{,}5(\sin(\pi/3) - \sin(-\pi/4)) = 0{,}5\left(\sqrt{3}/2 + \sqrt{2}/2\right) = \left(\sqrt{3} + \sqrt{2}\right)/4. \quad \bullet$$

§ 6. ОСНОВНІ ЗАКОНИ НЕПЕРЕРВНИХ ВИПАДКОВИХ ВЕЛИЧИН

1. *Нормальний закон (імовірносний зміст параметрів розподілу; нормальна крива; імовірність попадання у заданий інтервал; знаходження імовірності заданого відхилення).*
2. *Закон рівномірного розподілу.*
3. *Показниковий закон.*

1. Випадкова величина називається **нормально розподіленою** (**розподіленою за нормальним законом**), якщо її густина розподілу має такий вид

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-a)^2}{2\sigma^2}}. \qquad (6.1)$$

Густина розподілу (6.1) повністю визначається двома параметрами: a та σ. Імовірносний зміст цих параметрів визначається такими рівностями:

$$a = M(X), \qquad \sigma = \sqrt{D(X)}. \qquad (6.2)$$

Нормальний закон розподілу випадкової величини з параметрами a=0, σ=1 називається **стандартним** або **нормованим**.

Графік густини нормального розподілу називається нормальною кривою (крива Гаусса) (рис.6.1).

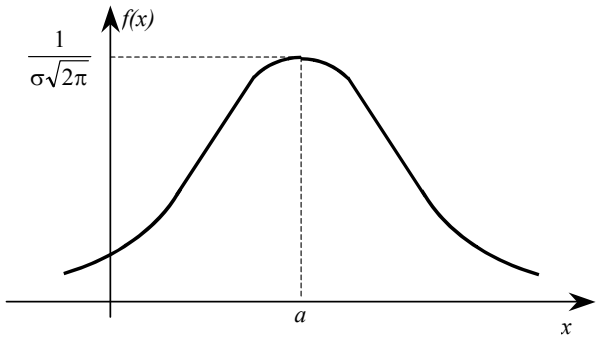


Рис. 6.1.

Імовірність того, що нормально розподілена випадкова величина при випробуванні набере значення з інтервалу (α, β), обчислюється за формулою

$$P(\alpha < X < \beta) = \Phi\left(\frac{\beta - a}{\sigma}\right) - \Phi\left(\frac{\alpha - a}{\sigma}\right). \qquad (6.3)$$

В практично важливих задачах виникає необхідність в знаходженні імовірності того, що відхилення нормально розподіленої випадкової величини від a по абсолютній величині менше від заданого додатного числа ε, тобто $P\left(|X - a| < \varepsilon\right)$.

Ця імовірність обчислюється за формулою

$$P\left(|X - a| < \varepsilon\right) = 2\Phi\left(\frac{\varepsilon}{\sigma}\right), \qquad (6.4)$$

яка, зокрема, показує, що імовірність відхилення для одного і того ж ε буде тим більша, чим меншим буде σ.

2. Випадкова величина називається **розподіленою за рівномірним законом (рівномірно розподіленою)**, якщо її густина розподілу має такий вид

$$f(x) = \begin{cases} C = \text{const}, & \text{якщо } x \in [a, b], \\ 0, & \text{якщо } x \notin [a, b]. \end{cases}$$

Знайдемо значення сталої C, використавши другу властивість густини розподілу. Геометричний зміст цієї властивості означає рівність одиниці площі, обмеженої кривою розподілу, тобто $C(b - a) = 1$, звідки $C = 1/(b - a)$. Отже, густина рівномірно розподіленої випадкової величини має такий вид:

$$f(x) = \begin{cases} 1/(b - a), & \text{при } x \in [a, b], \\ 0, & \text{при } x \notin [a, b]. \end{cases} \qquad (6.5)$$

Відрізок $[a, b]$ називається **відрізком концентрації рівномірного розподілу**.

3. Випадкова величина X розподіляється за **показниковим (експоненціальним) законом (показниково розподілена)**, якщо її густина розподілу ймовірностей має такий вид

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{якщо } x \geq 0, \\ 0, & \text{якщо } x < 0, \end{cases} \qquad (6.6)$$

де стала $\lambda > 0$ називається параметром експоненціального розподілу.

РОЗВ'ЯЗУВАННЯ ТИПОВИХ ЗАДАЧ

Задача 6.1. Коробки із шоколадом упаковуються автоматично, при цьому середня маса однієї коробки становить 1,04 кг. Відомо, що тільки 2,5% коробок мають масу, меншу від 1 кг. Припускаючи, що маса коробок розподілена нормально, знайти середнє квадратичне відхилення.

○ Позначимо: X — маса навмання взятої коробки із шоколадом. За умовою X — нормально розподілена випадкова величина, $M(X) = 1{,}04$ і $P(X < 1) = 0{,}025$. Протилежною до події $(X < 1)$ є подія $(1 \le X < \infty)$, тому $P(1 \le X < \infty) = 1 - P(X < 1) = 1 - 0{,}025 = 0{,}975$. Для знаходження σ використаємо формулу (6.3), де $a = 1{,}04$, $\alpha = 1$, $\beta = \infty$, тобто

$$P(1 \le X < \infty) = 0{,}975 = \Phi\left(\frac{\infty - 1{,}04}{\sigma}\right) - \Phi\left(\frac{1 - 1{,}04}{\sigma}\right), \text{ або}$$

$$0{,}975 = 0{,}5 + \Phi\left(\frac{0{,}04}{\sigma}\right),$$

оскільки $\Phi(x)$ — непарна функція і $\Phi(x) = 0{,}5$ для $x > 5$.

Остаточно рівняння набуває такого вигляду

$$\Phi\left(\frac{0{,}04}{\sigma}\right) = 0{,}475.$$

За табл. 3 додатків $\Phi(1{,}96) = 0{,}475$, тому

$0{,}04/\sigma = 1{,}96$, звідки $\sigma = 0{,}04/1{,}96 \approx 0{,}0204$. ●

Задача 6.2. Випадкова величина X розподілена за нормальним законом із середнім значенням 10. Знайти $P(8<X<20)$, якщо $P(6<X<14)=0{,}5$.

○ За умовою задачі параметр $a=10$, а параметр σ - невідомий. Для знаходження його використаємо рівність $P(8<X<20)=0{,}5$, формулу (6.3) і непарність функції Лапласа:

$$P(6 < X < 14) = \Phi\left(\frac{14-10}{\sigma}\right) - \Phi\left(\frac{6-10}{\sigma}\right) = \Phi\left(\frac{4}{\sigma}\right) - \Phi\left(\frac{-4}{\sigma}\right) = 2\Phi\left(\frac{4}{\sigma}\right);$$

$$2\Phi\left(\frac{4}{\sigma}\right) = 0{,}5 \Rightarrow \Phi\left(\frac{4}{\sigma}\right) = 0{,}25 \Rightarrow \frac{4}{\sigma} = 0{,}675 \Rightarrow \sigma = \frac{4}{0{,}675} \approx 5{,}93.$$

Тоді за формулою (6.3)

$$P(8 < X < 20) = \Phi\left(\frac{20-10}{5{,}93}\right) - \Phi\left(\frac{8-10}{5{,}93}\right) = \Phi(1{,}69) - \Phi(-0{,}34) =$$

$$= \Phi(1{,}69) + \Phi(0{,}34) = 0{,}45449 + 0{,}13307 = 0{,}58756. \quad \bullet$$

Задача 6.3. Швейна фабрика виготовляє костюми, орієнтуючись на покупців конкретного регіону. Покладаючи, що ріст чоловіків певної вікової групи цього регіону є нормально розподіленою випадковою величиною X із параметрами a=174см і σ=5см, знайти: 1) густину розподілу імовірностей величини X; 2) частки костюмів третього росту (171-176см) і четвертого росту (176-181см), які потрібно передбачити в загальному обсязі виробництва для даної вікової групи.

○ 1) Згідно з умовою задачі густина розподілу імовірностей даної випадкової величини має вид (6.1):

$$f(x) = \frac{1}{5\sqrt{2\pi}} e^{-(x-174)^2/50}.$$

2) Нехай одиниця – загальний обсяг виробництва костюмів (всіх ростів). Тоді частка костюмів третього росту визначиться як імовірність $P(171<X<176)$, яку знайдемо за формулою (6.3):

$$P(171 < X < 176) = \Phi\left(\frac{176-174}{5}\right) - \Phi\left(\frac{171-174}{5}\right) = \Phi(0{,}4) + \Phi(0{,}6) =$$

$$= 0{,}15542 + 0{,}22575 = 0{,}38117.$$

Аналогічно знаходимо частку костюмів четвертого росту:

$$P(176 < X < 181) = \Phi\left(\frac{181-174}{5}\right) - \Phi\left(\frac{176-174}{5}\right) = \Phi(1{,}4) - \Phi(0{,}4) =$$

$$= 0{,}41924 - 0{,}15542 = 0{,}26382.$$

Отримані результати можна інтерпретувати таким чином: костюми третього росту повинні складати 38,1% всієї продукції, а четвертого – 26,4%. ●

§ 7. ЗАКОН ВЕЛИКИХ ЧИСЕЛ

1. *Лема та нерівність Чебишева.*
2. *Теорема Чебишева (стійкість середніх).*
3. *Теорема Бернуллі (стійкість відносних частот).*
4. *Центральна гранична теорема Ляпунова.*

1. Лема Чебишева. Якщо всі можливі значення випадкової величини Y невід'ємні, тоді імовірність того, що вона при випробуванні набере значення, більше від додатнього числа b, не більша від дробу, чисельник якого — математичне сподівання від Y, а знаменник — число b:

$$P(Y > b) \le M(Y)/b. \tag{7.1}$$

Нерівність Чебишева. Імовірність того, що відхилення випадкової величини X від її математичного сподівання за абсолютною величиною не більше від додатного числа ε, не менша, ніж $1 - D(X)/\varepsilon^2$:

$$P\left(|X - M(X)| \le \varepsilon\right) \ge 1 - D(X)/\varepsilon^2. \tag{7.2}$$

Зауваження. Якщо випадкова величина X розподілена за біноміальним законом (X — число появи події в n повторних незалежних випробуваннях), тоді нерівність Чебишева набере такого виду:

$$P\left(|m - np| \le \varepsilon\right) \ge 1 - \frac{npq}{\varepsilon^2} \tag{7.3}$$

або

$$P\left(\left|\frac{m}{n} - p\right| \le \varepsilon\right) \ge 1 - \frac{pq}{n\varepsilon^2}, \tag{7.4}$$

де m — число появи події A в n повторних незалежних випробуваннях, $p = P(A)$, m/n — відносна частота появи події A.

2. Теорема Чебишева. Якщо $X_1, X_2, ..., X_n$ попарно незалежні випадкові величини, дисперсії яких рівномірно обмежені $\left(D(X_i) \le C, i = \overline{1,n}\right)$, тоді для довільного $\varepsilon > 0$ і достатньо великого n імовірність події

$$\left|\frac{X_1 + X_2 + ... + X_n}{n} - \frac{M(X_1) + M(X_2) + ... + M(X_n)}{n}\right| \le \varepsilon \tag{7.5}$$

буде як завгодно близькою до одиниці.

При доведені теореми встановлюється правильність нерівності

$$P\left(\left|\frac{X_1+X_2+\ldots+X_n}{n}-\frac{M(X_1)+M(X_2)+\ldots+M(X_n)}{n}\right|\le\varepsilon\right)\ge 1-C/(n\varepsilon^2). \qquad (7.6)$$

3. Теорема Бернуллі. Якщо в кожному із n повторних випробувань імовірність p появи події A стала, тоді як завгодно близька до одиниці імовірність того, що відхилення відносної частоти події A від імовірності p за абсолютною величиною буде як завгодно малим, якщо число випробувань достатньо велике.

З нерівності (7.6) можна отримати нерівність

$$P\left(\left|m/n-p\right|\le\varepsilon\right)\ge 1-\frac{0{,}25}{n\varepsilon^2}. \qquad (7.7)$$

В теоремі Чебишева не використовувалася інформація про закони розподілу випадкових величин. Разом з тим для задач теорії і практики важливим є таке питання: за яким законом розподіляється сума достатньо великого числа випадкових величин?

Випадкову величину

$$Z_n=\left[\sum_{i=1}^{n}X_i-\sum_{i=1}^{n}M(X_i)\right]\Big/\sqrt{\sum_{i=1}^{n}D(X_i)}. \qquad (7.8)$$

будемо називати **нормованою сумою** або **центрованою випадковою величиною**.

4. Центральна гранична теорема Ляпунова. Якщо $X_1, X_2, \ldots, X_n$ — незалежні випадкові величини, які мають математичні сподівання m_i, дисперсії σ_i^2 і скінченні абсолютні центральні моменти третього порядку $|\mu_3|$, що задовольняють умовам

$$\lim_{n\to\infty}\left(\sum_{i=1}^{n}\left|\mu_3\left(X_i\right)\right|\right)\Big/\left(\sum_{i=1}^{n}\sigma_i^2\right)^{3/2}=0, \qquad (7.9)$$

тоді при необмеженому збільшенні n закон розподілу нормованої суми (7.8) збігається за імовірністю до нормального закону з параметрами $a = 0$, $\sigma = 1$.

Зазначимо, що $\mu_3\left(X_i\right)=M[\left|X_i-m\right|^3]$ називають центральним моментом третього порядку випадкової величини X_i.

Зміст умови (7.9) полягає в тому, що дисперсія кожної випадкової величини $X_i, i=\overline{1,n}$, складає лише малу частину в загальній дисперсії

суми $\sum_{i=1}^{n} X_i$.

Наслідок теореми Ляпунова. Якщо виконуються умови теореми Ляпунова, тоді випадкова величина $\sum_{i=1}^{n} X_i$ для великих n з достатнім ступенем точності розподілена за нормальним законом з параметрами

$$a = \sum_{i=1}^{n} m_i, \quad \sigma^2 = \sum_{i=1}^{n} \sigma_i^2.$$

В практичних задачах центральну граничну теорему Ляпунова часто використовують для обчислення імовірності того, що сума кількох випадкових величин набере значення, яке належить вказаному інтервалу.

Підставою такого використання є так званий

Частковий наслідок теореми Ляпунова. Якщо X_1, X_2, ..., X_n – незалежні випадкові величини, у яких існують рівні математичні сподівання $M(X_i)=m$, дисперсії $D(X_i)=\sigma^2$ і абсолютні центральні моменти третього порядку $M[|X_i - m|^3] = \mu_3 \quad i = \overline{1,n}$, то закон розподілу суми $Y_n=X_1+X_2+...+X_n$ при $n \to \infty$ необмежено наближається до нормального закону з параметрами $a = \sum_{i=1}^{n} m$, $\sigma^2 = \sum_{i=1}^{n} \sigma_i^2$.

РОЗВ'ЯЗУВАННЯ ТИПОВИХ ЗАДАЧ

Задача 7.1. Випадкова величина задана законом розподілу

X	0,3	0,6
P	0,2	0,8

Використовуючи нерівність Чебишева, оцінити імовірність того, що X відхилиться від свого математичного сподівання на величину, яка за абсолютною величиною не перевищить 0,2. Перевірити точність отриманої оцінки.

○ Знайдемо $M(X)$ та $D(X)$

$$M(X) = 0,3 \cdot 0,2 + 0,6 \cdot 0,8 = 0,54,$$

$$D(X) = M(X^2) - [M(X)]^2 = 0,3^2 \cdot 0,2 + 0,6^2 \cdot 0,8 - (0,54)^2 = 0,0144.$$

Використавши нерівність (7.2) з $\varepsilon = 0,3$, отримаємо

$$P\left(\left|X-0{,}54\right|\le 0{,}2\right)\ge 1-0{,}0144/0{,}04=0{,}64.$$

Випадкова подія $\left|X-0{,}54\right|\le 0{,}2$ рівносильна події $0{,}34\le X\le 0{,}74$, яка відбувається із врахуванням закону розподілу, тоді і тільки тоді, коли X=0,6. Тому

$$P\left(\left|X-0{,}54\right|\le 0{,}2\right)=P(0{,}34\le X\le 0{,}74)=P(X=0{,}6)=0{,}8.$$

Отже, нижня межа оцінки імовірності дорівнює 0,64, а справжнє значення імовірності дорівнює 0,8. ●

Задача 7.2. Імовірність появи події в кожному випробуванні дорівнює 0,25. Використовуючи нерівність Чебишева, оцінити імовірність того, що число X появи події міститься в межах від 150 до 250 включно, якщо буде проведено 800 випробувань. Перевірити точність отриманої оцінки.

○ Використаємо нерівність (7.3), де $X=m$, $np=800\cdot 0{,}25=200$ і попередньо треба знайти ε. Випадкову подію $(150\le m\le 250)$ можна записати в такій рівносильній формі: $\left(\left|m-200\right|\le 50\right)$. А тому в (7.3) ε = 50 і

$$P(150\le m\le 250)=P\left(\left|m-200\right|\le 50\right)\ge 1-\frac{800\cdot 0{,}25\cdot 0{,}75}{50^2}=0{,}94.$$

Тобто, остаточно $P(150\le m\le 250)\ge 0{,}94$.

Перевіримо точність отриманого результату, використавши інтегральну формулу Лапласа (3.7)

($npq=800\cdot 0{,}25\cdot 0{,}75=150>>9$):

$$P_n(m_1\le m\le m_2)\approx \Phi(x_2)-\Phi(x_1),$$

$$x_2=\frac{m_2-np}{\sqrt{npq}}=\frac{250-800\cdot 0{,}25}{\sqrt{150}}=\frac{50}{12{,}247}=4{,}08,$$

$$x_1=\frac{m_1-np}{\sqrt{npq}}=\frac{150-200}{\sqrt{150}}=\frac{-50}{12{,}247}=-4{,}08,$$

$$P_n(150\le m\le 250)\approx \Phi(4{,}08)-\Phi(-4{,}08)=2\Phi(4{,}08)=2\cdot 0{,}4999992=$$
$$=0{,}9999984.$$

Отже, нижня межа оцінки імовірності дорівнює 0,94, а справжнє значення імовірності дорівнює 0,9999984. ●

Задача 7.3. Дисперсія кожної із 2500 незалежних випадкових величин не перевищує 5. Оцінити імовірність того, що абсолютна величина відхилення середнього арифметичного цих випадкових величин від середнього арифметичного їх математичних сподівань не перевищить 0,4.

○ Для задачі виконуються обидві умови теореми Чебишева, а тому можна використати нерівність (7.6), де $n = 2500$, $\varepsilon = 0{,}4$, $C = 5$. Тоді

$$P\left(\left|\left(\sum_{i=1}^{2500} X_i - \sum_{i=1}^{2500} M(X_i)\right)\Big/ 2500\right| \le 0{,}4\right) \ge 1 - \frac{5}{2500(0{,}4)^2} = \frac{79}{80}.$$

Отже, шукана імовірність оцінюється знизу числом 79/80. ●

Задача 7.4. Відомо, що 80% виробів механічного цеху є першосортними. Оцінити імовірність того, що відносна частоту виробів першого сорту серед 20 000 виготовлених відрізнятиметься від імовірності виготовлення виробу першого сорту не більше, ніж на 0,02 в той чи інший бік. Перевірити точність отриманої оцінки.

○ Подія A — виготовлений виріб першосортний. Тоді $P(A) = 0{,}8$, $n = 20\,000$, $\varepsilon = 0{,}02$. Використаємо нерівність (7.7):

$$P\left(\left|m/n - 0{,}8\right| \le 0{,}02\right) \ge 1 - \frac{0{,}25}{20000 \cdot 0{,}0004} = 31/32.$$

Таким чином, шукана імовірність оцінюється знизу числом 31/32.

Для перевірки точності отриманої оцінки використаємо формулу (3.8):

$$P\left(\left|m/n - p\right| \le \varepsilon\right) \approx 2\Phi\left(\varepsilon\sqrt{n/(pq)}\right),$$

ε=0,02, n=20000, p=0,8, q=0,2,

$$P\left(\left|m/n - 0{,}8\right| \le 0{,}02\right) \approx 2\Phi\left(0{,}02\sqrt{\frac{20000}{0{,}8 \cdot 0{,}2}}\right) = 2\Phi(7{,}07) = 2 \cdot 0{,}5 = 1.$$

Оскільки npq=3200 і 7,07>5, то похибки у наближеній рівності (3.8) і $\Phi(7{,}07) \approx 0{,}5$ є надзвичайно малими, тобто подія $\left(\left|m/n - 0{,}8\right| \le 0{,}02\right)$ є майже достовірною. ●

ЧАСТИНА ДРУГА
МАТЕМАТИЧНА СТАТИСТИКА

§ 1.ВСТУП В МАТЕМАТИЧНУ СТАТИСТИКУ. ВИБІРКОВИЙ МЕТОД

1. *Задачі математичної статистики.*
2. *Генеральна та вибіркова сукупності. Способи утворення вибіркової сукупності.*
3. *Статистичний розподіл вибірки.*
4. *Емпірична функція розподілу та її властивості.*
5. *Графічне зображення статистичних розподілів (полігон та гістограма).*
6. *Числові характеристики вибірки.*

1. Перша задача математичної статистики — вказати метод відбору і групування статистичних даних, а також знаходження числа необхідних випробувань (статистичних даних).

Друга задача математичної статистики полягає в розробці методів аналізу статистичних даних в залежності від цілей дослідження. Одна з них — оцінка невідомих:

— імовірності випадкової події;

— функції розподілу ймовірностей (густини розподілу);

— параметрів розподілу, вид якого відомий;

— залежності випадкової величини від однієї або кількох випадкових величин.

Друга мета — це перевірка статистичних гіпотез про вид невідомого розподілу або про величину параметрів розподілу, вид якого відомий.

2. Генеральною називається вся сукупність однотипних об'єктів, яка підлягає вивченню. Множину цих об'єктів позначатимемо через Ω.

Об'єкти множини Ω можуть характеризуватися однією або кількома ознаками. Ці ознаки можуть бути кількісними та якісними.

Іноді проводять **суцільне** обстеження, тобто обстежують **кожний** елемент множини Ω відносно ознаки, яка досліджується. Проте на практиці суцільне обстеження застосовують порівняно рідко. Це зумовлено тим, що при перевірці об'єкт частково або повністю знищується, або дослідження об'єктів вимагає великих матеріальних витрат. Деколи провести суцільне обстеження фізично неможливо. В таких випадках природно відібрати із генеральної сукупності обмежене число

об'єктів, дослідити їх на кількісну чи якісну ознаку, а потім робити висновки про всю генеральну сукупність.

Вибірковою сукупністю або просто **вибіркою** називається сукупність **випадково** відібраних об'єктів із генеральної сукупності. Вибірка утворює підмножину V множини Ω ($V \subset \Omega$).

Обсягом сукупності (генеральної або вибіркової) називається число об'єктів цієї сукупності. Надалі обсяг генеральної сукупності позначатимемо літерою N, а вибіркової — n.

Для правомірності висновків про досліджувану ознаку об'єктів генеральної сукупності на підставі опрацювання вибірки необхідно, щоб об'єкти вибірки правильно представляли генеральну сукупність, тобто вибірка повинна володіти властивістю репрезентативності (представницькості). Випадковість відбору об'єктів у вибіркову сукупність і використання закону великих чисел дозволяють вирішити питання про репрезентативність вибірки.

Точність результатів вибіркового спостереження, в кінцевому підсумку, буде залежати від способу відбору об'єктів, ступеня коливання досліджуваної ознаки в генеральній сукупності та від обсягу вибірки.

На практиці використовуються різні способи утворення вибірки, які принципово розподіляються на два види:

1) відбір, що не вимагає розчленування генеральної сукупності на частини (**простий** (**власне випадковий**) **відбір**);

2) відбір, при якому генеральна сукупність розбивається на частини (**типовий** відбір, **механічний** відбір, **серійний** відбір, **комбінований** відбір).

Простим випадковим (**власне випадковим**) називається такий відбір, при якому об'єкти відбираються по одному випадковим чином із усієї генеральної сукупності. Проста випадкова вибірка може бути **повторною** або **безповторною**. **Повторною** називається вибірка, при утворенні якої відібраний об'єкт (перед відбором наступного) повертається в генеральну сукупність. **Безповторною** називається вибірка, в процесі утворення якої відібраний об'єкт в генеральну сукупність не повертається.

Нехай досліджується кількісна ознака X об'єктів генеральної сукупності Ω. Для скорочення припустимо, що вона є одновимірною випадковою величиною. Після опрацювання n об'єктів вибіркової сукупності отримуються n чисел $x_1, x_2, \ldots, x_n$, які називаються **варіантами** і утворюють **ряд варіант** або **простий статистичний ряд**.

Первинна обробка ряду варіант полягає у групуванні рівних варіант цього ряду. Ряд варіант розташуємо в порядку зростання і у відповідності з цим перенумеруємо їх. В результаті одержимо послідовність чисел $x_1 \le x_2 \le x_3 \le \ldots \le x_n$, яка називається варіаційним рядом. Якщо серед цієї послідовності є однакові варіанти, то ми їх знову перенумеруємо, залишаючи один і той самий номер однаковим варіантам. Нехай у варіаційному ряді варіанта x_1 повторюється n_1 разів, x_2 — n_2 разів, ..., x_k — n_k разів. Числа n_i називаються **частотами** (**абсолютними частотами**), а їх відношення до обсягу вибірки $n_i / n = w_i$ — **відносними частотами**. Із цих означень випливають рівності:

$$\sum_{i=1}^{k} n_i = n, \quad \sum_{i=1}^{k} w_i = 1. \tag{1.1}$$

3. Статистичним розподілом вибірки називається відповідність між варіантами та частотами або відносними частотами:

$$\begin{array}{c|cccc} x_i & x_1 & x_2 & \ldots & x_k \\ \hline n_i & n_1 & n_2 & \ldots & n_k \end{array}, \tag{1.2}$$

$$\begin{array}{c|cccc} x_i & x_1 & x_2 & \ldots & x_k \\ \hline w_i = n_i / n & w_1 & w_2 & \ldots & w_k \end{array}. \tag{1.3}$$

В більшості випадків статистичний розподіл вибірки у вигляді (1.2) або (1.3) використовується тоді, коли ряд варіант є реалізацією **дискретної** випадкової величини X. Якщо ж X є неперервною випадковою величиною, тоді статистичний розподіл вибірки задається у вигляді відповідності між інтервалами і частотами або відносними частотами тих варіант, які потрапляють у ці інтервали, тобто у вигляді таблиць:

$$\begin{array}{c|cccc} [x_i, x_{i+1}) & [x_1, x_2) & [x_2, x_3) & \ldots & [x_k, x_{k+1}] \\ \hline n_i & n_1 & n_2 & \ldots & n_k \end{array}, \tag{1.4}$$

$$\begin{array}{c|cccc} [x_i, x_{i+1}) & [x_1, x_2) & [x_2, x_3) & \ldots & [x_k, x_{k+1}] \\ \hline w_i = n_i / n & w_1 & w_2 & \ldots & w_k \end{array}. \tag{1.5}$$

Ці таблиці називаються **інтервальним статистичним розподілом вибірки**. При побудові інтервального статистичного розподілу на основі ряду варіант розглядається k інтервалів однакової довжини. Кількість інтервалів можна визначати наближено за формулою Стерджеса: $k = 1+3{,}322 \lg n$. При цьому $k \in [5;6]$ при $20 \le n \le 30$; $k \in [7;8]$ при $60 \le n \le 70$; $k \in [8;9]$ при $70 \le n \le 200$; $k \in [9;15]$ при $n > 200$.

Статистичні розподіли вибірки (1.3) та (1.5) називаються **емпіричними** (**дослідними**) **розподілами випадкової величини *X*** (кількісної ознаки об'єктів генеральної сукупності).

Одна із задач математичної статистики — оцінка (наближене знаходження) невідомої функції розподілу $F(x)$ імовірностей кількісної ознаки X об'єктів генеральної сукупності. За означенням $F(x) = P(X < x)$. В розпорядженні дослідника є статистичні дані, згруповані в статистичному розподілі частот або відносно частот. Тому, врахувавши властивість стійкості відносної частоти, доцільно імовірність події $(X < x)$ наближати відносною частотою цієї ж події.

4. Емпіричною функцією розподілу (**функцією розподілу вибірки**) називається функція $F^*(x)$ детермінованого аргумента x, яка дорівнює відносній частоті появи події $(X < x)$ для даної вибірки значень випадкової величини X, тобто

$$F^*(x) = W(X < x) = n_x/n, \qquad (1.6)$$

де n_x — сума частот тих варіант, які менші від x,

n — обсяг вибірки.

На противагу $F^*(x)$ функцію $F(x)$ називають теоретичною функцією розподілу.

Із означення емпіричної функції випливають такі її властивості, аналогічні властивостям теоретичної функції розподілу:

1) $D(F^*) = R$, $E(F^*) = [0, 1]$;
2) $F^*(x)$ — неспадна функція;
3) якщо x_1 — найменша варіанта, тоді $F^*(x) = 0$ для $x \le x_1$; якщо x_k — найбільша варіанта, тоді $F^*(x) = 1$ для $x > x_k$.

5. В процесі аналізу статистичних даних суттєву роль відіграє геометрична ілюстрація цих даних. Для наочності будують різні графіки статистичних розподілів, зокрема полігон і гістограму.

Нехай статистичний розподіл вибірки визначається таблицями (1.2) або (1.3).

Полігоном частот (**частотним многокутником**) називається ламана, прямолінійні відрізки якої з'єднують сусідні точки (x_1, n_1), (x_2, n_2), …, (x_k, n_k). Для побудови полігона на осі абсцис відкладають варіанти, а на осі ординат — відповідні їм частоти.

Полігоном відносних частот називається ламана, прямолінійні відрізки якої з'єднують сусідні точки (x_1, w_1), (x_2, w_2), …, (x_k, w_k). При

побудові полігона відносних частот на осі абсцис відкладають варіанти x_i, а на осі ординат — відповідні їм відносні частоти w_i.

У випадку неперервної ознаки доцільно будувати гістограму. Нехай розподіли (1.4) та (1.5) такі, що довжина кожного із частинних інтервалів дорівнює одному і тому ж числу h.

Гістограмою частот називається сходинкова фігура, що складається із прямокутників, основами яких є частинні інтервали довжиною h, а висоти дорівнюють відношенню n_i/h (**густина частоти**). Для побудови гістограми частот на осі абсцис відкладаються частинні інтервали, а над ними проводяться прямолінійні відрізки, паралельні осі абсцис на віддалі n_i/h. Площа i-го частинного прямокутника дорівнює $hn_i/h = n_i$, тобто сумі частот тих варіант, що потрапляють в i-ий інтервал. Тому площа гістограми частот дорівнює сумі всіх частот вибірки n.

Порівняння двох гістограм дозволяє зробити висновок про те, що виразність гістограми суттєво залежить від обрання довжини h частинних інтервалів.

Гістограмою відносних частот називається сходинкова фігура, що складається із прямокутників, основами яких є частинні інтервали довжиною h, а висоти дорівнюють відношенню w_i/h (**густини відносної частоти**). Для побудови гістограми відносних частот на осі абсцис відкладаються частинні інтервали, а над ними проводяться відрізки, паралельні осі абсцис на віддалі w_i/h. Площа i-ого частинного прямокутника дорівнює $hw_i/h = w_i$, тобто відносній частоті тих варіант, що потрапили в i-ий інтервал. Отже, площа гістограми відносних частот дорівнює одиниці.

6. Побудова статистичних розподілів вибірки (1.2) або (1.4) та їх графічне зображення — це тільки перший крок на шляху розв'язування задач математичної статистики. Наступний крок передбачає знаходження числових характеристик, які у компактній формі виражають найбільш суттєві особливості статистичного розподілу вибірки і слугують оцінками (наближеними значеннями) невідомих параметрів розподілу кількісної ознаки генеральної сукупності.

Середня вибіркова (**середня арифметична варіант**) статистичного розподілу (1.2) визначається формулою

$$\overline{x}_в = \left(\sum_{i=1}^{k} x_i n_i \right) \Big/ n. \qquad (1.7)$$

Якщо всі n варіант різні, тоді (1.7) набирає такого виду:

$$\overline{x}_в = \left(\sum_{i=1}^{n} x_i \right) \Big/ n. \qquad (1.7^*)$$

Якщо вибірка задається інтервальним статистичним розподілом частот (1.4), тоді при знаходженні $\bar{x}_в$ потрібно перейти до дискретного розподілу (1.2), "нові" варіанти якого є серединами інтервалів, а потім використати формулу (1.7).

Крім вказаної середньої, у статистиці застосовують ще й **структурні середні**, які не залежать від значень варіант, що розташовані на краях розподілу, а пов'язані із рядом частот. До структурних середніх належать медіана та мода.

Медіаною Me^* дискретного статистичного розподілу вибірки (1.2) називається таке число, яке ділить варіаційний ряд, що "породжує" цей розподіл, на дві рівні за кількістю варіант частини. Якщо число варіант непарне, тобто $n = 2m + 1$, тоді $Me^* = x_{m+1}$. Якщо ж обсяг вибірки є парним числом, тобто $n = 2m$, тоді медіана дорівнює середньому арифметичному "середньої" (медіанної) пари варіант:

$$Me^* = (x_m + x_{m+1})/2.$$

Медіаною для інтервального статистичного розподілу називається таке число Me^*, для якого виконується рівність:

$$F^*(Me^*) = 0,5, \tag{1.8}$$

де $F^*(x)$ — емпірична функція цього розподілу.

Формула для обчислення медіани має такий вид:

$$Me^* = x_m + \frac{0,5 - F^*(x_m)}{F^*(x_{m+1}) - F^*(x_m)}(x_{m+1} - x_m), \tag{1.9}$$

де $[x_m, x_{m+1})$ – так званий медіанний частинний інтервал ($1 \le m \le k$) для якого виконуються нерівності $F^*(x_m) < 0,5$, $F^*(x_{m+1}) > 0,5$.

Модою Mo^* дискретного статистичного розподілу (1.2) називається варіанта, якій відповідає найбільша частота.

Мода для інтервального статистичного розподілу обчислюється таким чином. Спочатку визначається модальний інтервал $[x_m, x_{m+1})$, тобто такий інтервал, для якого $n_m/h_m = \max\limits_{1 \le i \le k}\{n_i/h_i\}$, де h_i — довжина частинного інтервалу $[x_m, x_{m+1})$, n_i — число варіант з цього інтервалу. Значення Mo^* міститься всередині модального інтервалу і обчислюється за інтерполяційною формулою

$$Mo^* = x_m + \frac{n_m - n_{m-1}}{2n_m - n_{m-1} - n_{m+1}} h_m. \tag{1.10}$$

Розглянемо деякі числові характеристики розсіювання варіант навколо середньої вибіркової.

Дисперсією вибірковою статистичного розподілу (1.2) називається середнє арифметичне квадратів відхилень варіант від середньої вибіркової:

$$D_в = \frac{1}{n}\sum_{i=1}^{k}(x_i - \bar{x}_в)^2 n_i. \qquad (1.11)$$

$D_в$ характеризує середню величину розкиду варіант навколо $\bar{x}_в$ в квадратних одиницях.

На практиці зручніше користуватися так званою **розрахунковою формулою для обчислення дисперсії**:

$$D_в = \overline{x^2} - (\bar{x}_в)^2 = \frac{1}{n}\sum_{i=1}^{k} x_i^2 n_i - (\bar{x}_в)^2. \qquad (1.12)$$

Недоліком $D_в$ є її розмірність. Для виправлення цього недоліку використовується інша числова характеристика:

середнє квадратичне відхилення вибіркове

$$\sigma_в = \sqrt{D_в}. \qquad (1.13)$$

Коливність окремих значень варіант характеризують показники варіації. Найпростішим із них є показник **розмаху варіації** R, який дорівнює різниці між найбільшою та найменшою варіантами розподілу: $R = x_{\max} - x_{\min}$. Розмах варіації використовується при статистичному вивченні якості продукції.

Якщо $\bar{x}_в$ відмінна від нуля, тоді для порівняння двох статистичних розподілів з точки зору їх розмірності відносно середньої вибіркової вводиться показник **коефіцієнт варіації**, який дорівнює відношенню середнього квадратичного відхилення до середньої вибіркової і виражений у відсотках:

$$V = \frac{\sigma_в}{\bar{x}_в} \cdot 100\%. \qquad (1.14)$$

Вибірковою часткою називається відношення числа m об'єктів вибірки з ознакою α до обсягу вибірки: $w = m/n$. Ознакою α може бути стандартність виробу, сортність продукції, стать людини тощо. За змістом w є відносною частотою випадкової події, яка полягає в тому, що навмання відібраний об'єкт із генеральної сукупності має ознаку α.

РОЗВ'ЯЗУВАННЯ ТИПОВИХ ЗАДАЧ

Задача 1.1. При вивченні питання про норми виробітку ткаль спостерігалася така частота обривів пряжі на однотипних ткацьких станках на різних проміжках часу однакової довжини t:

7, 1, 2, 1, 2, 5, 4, 4, 3, 2, 2, 6, 0, 1, 6,
5, 3, 2, 0, 1, 4, 3, 2, 1, 5, 3, 0, 4, 2, 3.

1) Скласти статистичний розподіл частот та відносних частот числа обривів пряжі на станках;

2) побудувати полігон частот та відносних частот;

3) знайти емпіричну функцію розподілу та побудувати її графік;

4) обчислити вибіркові: середню, дисперсію, середнє квадратичне відхилення, моду, медіану, розмах варіації, коефіцієнт варіації.

○ 1) Обсяг вибірки $n = 30$. Даний ряд варіант запишемо у вигляді варіаційного ряду:

0, 0, 0, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 4, 4, 4, 4, 5, 5, 5, 6, 6, 7.

Варіанти: $x_1 = 0$, $x_2 = 1$, $x_3 = 2$, $x_4 = 3$, $x_5 = 4$, $x_6 = 5$, $x_7 = 6$, $x_8 = 7$.

Частоти: $n_1 = 3$, $n_2 = 5$, $n_3 = 7$, $n_4 = 5$, $n_5 = 4$, $n_6 = 3$, $n_7 = 2$, $n_8 = 1$.

В підсумку одержимо статистичний розподіл частот:

x_i	0	1	2	3	4	5	6	7
n_i	3	5	7	5	4	3	2	1

.

Контроль: $\sum n_i = 3 + 5 + 7 + 5 + 4 + 3 + 2 + 1 = 30 = n$.

За формулою $w_i = n_i / n$ послідовно обчислюємо відносні частоти: $w_1 = 3/30$, $w_2 = 5/30$, $w_3 = 7/30$, $w_4 = 5/30$, $w_5 = 4/30$, $w_6 = 3/30$, $w_7 = 2/30$, $w_8 = 1/30$. Контроль: $\sum w_i = 3/30 + 5/30 + 7/30 + 5/30 + 4/30 + 3/30 + 2/30 + 1/30 = 1$. Отже, статистичний розподіл відносних частот числа обривів пряжі на станках має такий вид:

x_i	0	1	2	3	4	5	6	7
w_i	3/30	5/30	7/30	5/30	4/30	3/30	2/30	1/30

.

2) Полігон частот зобразимо на рис.1.1, а полігон відносних частот — на рис.1.2.

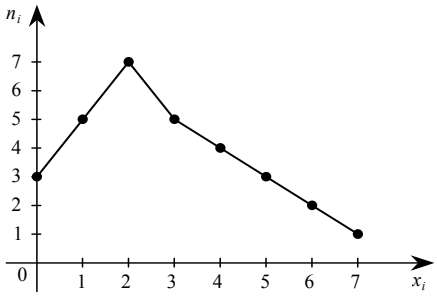


Рис.1.1.

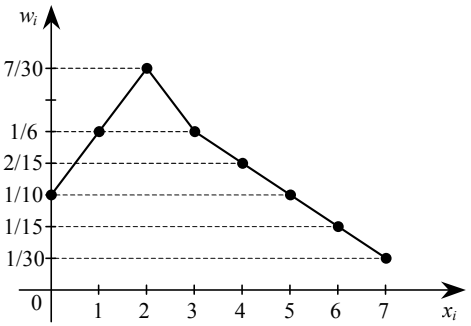


Рис.1.2.

3) Обсяг вибірки $n = 30$. Якщо $x \le 0$, тоді немає жодної варіанти, меншої від x, тобто $n_x = 0$, а тому і $F^*(x) = n_x/30 = 0$.

Нехай $x \in (0; 1]$. Тоді варіанта $x = 0$ є меншою від x, тому $n_x = 3$ і $F^*(x) = 3/30 = 0{,}1$.

Якщо x задовольняє подвійній нерівності $1 < x \le 2$, тоді меншими від x є варіанти 0 та 1, сума частот яких $n_x = 3 + 5 = 8$. Тому $F^*(x) = 8/30 = 4/15$ для $x \in (1; 2]$.

Якщо x таке, що виконується подвійна нерівність $2 < x \le 3$, тоді меншими від x є варіанти 0, 1, 2, суми частот яких $n_x = 3 + 5 + 7 = 15$. Тому для $x \in (2; 3]$ $F^*(x) = 15/30 = 0{,}5$.

Аналогічно знаходимо значення $F^*(x)$ для інтервалів $(3; 4]$, $(4; 5]$, $(5; 6]$, $(6; 7]$, $(7; \infty]$. В підсумку отримаємо шукану емпіричну функцію розподілу:

$$F^*(x)=\begin{cases} 0, & \text{якщо } x\le 0,\\ 1/10, & \text{якщо } 0<x\le 1,\\ 4/15, & \text{якщо } 1<x\le 2,\\ 1/2, & \text{якщо } 2<x\le 3,\\ 2/3, & \text{якщо } 3<x\le 4,\\ 4/5, & \text{якщо } 4<x\le 5,\\ 9/10, & \text{якщо } 5<x\le 6,\\ 29/30, & \text{якщо } 6<x\le 7,\\ 1, & \text{якщо } x>7. \end{cases}$$

Графік цієї функції наведено на рис. 1.3.

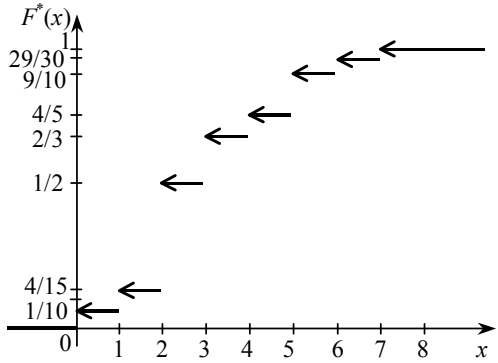


Рис. 1.3.

2) Для обчислення $\overline{x}_в$ та $D_в$ використаємо формули (1.7) і (1.12):

$$\overline{x}_в=\frac{\sum_{i=1}^{k}x_in_i}{n}=\frac{0\cdot 3+1\cdot 5+2\cdot 7+3\cdot 5+4\cdot 4+5\cdot 3+6\cdot 2+7\cdot 1}{30}=84/30=2,8;$$

$$D_в=\overline{x^2}-(\overline{x}_в)^2=\frac{\sum_{i=1}^{k}x_i^2n_i}{n}-(\overline{x}_в)^2=$$

$$=\frac{0^2\cdot 3+1^2\cdot 5+2^2\cdot 7+3^2\cdot 5+4^2\cdot 4+5^2\cdot 3+6^2\cdot 2+7^2\cdot 1}{30}-(2,8)^2=$$

$$=338/30-7,84=3,4267;$$

$$\sigma_в = \sqrt{D_в} = \sqrt{3,4267} = 1,8511.$$

Висновок: середнє число обривів пряжі на станках за проміжок часу складає 2,8, а середня величина розкиду чисел розривів пряжі навколо середньої 2,8 дорівнює 1,8511.

Мода $Mo^* = 2$, оскільки варіанті 2 відповідає найбільша частота 7.

Медіану Me^* можна знайти або за варіаційним рядом, отриманим в 1), або безпосередньо із статистичного розподілу. Сума частот перших трьох варіант цього розподілу дорівнює 15 (половині обсягу вибірки), а наступних п'яти — також 15. Тому медіана знаходиться між варіантами 2 та 3: $Me^* = (2+3)/2 = 2,5$. Неспівпадання $\bar{x}_в$, Mo^* та Me^* свідчить про відсутність строгої симетричності розподілу.

Коефіцієнт варіації (згідно із формулою (1.14))

$$V = \frac{\sigma_в}{\bar{x}_в} \cdot 100\% = \frac{1,8511}{2,8} \cdot 100\% = 66,11\%.$$

Задача 1.2. За період між черговими переналадками обладнання здійснено контрольні виміри товщини (в міліметрах) 200 вкладишів шатунних підшипників. Отримані дані наведені в табл. 1.2.

1) Скласти інтервальний статистичний розподіл частот та відносних частот вибірки.

На основі отриманого інтервального статистичного розподілу:

2) побудувати гістограми частот та відносних частот;

3) знайти емпіричну функцію розподілу та побудувати її графік;

4) обчислити $\bar{x}_в$, $D_в$, $\sigma_в$, моду та медіану.

Таблиця 1.2

1,754	1,739	1,743	1,764	1,733	1,736	1,743	1,742
1,728	1,732	1,731	1,752	1,737	1,747	1,758	1,737
1,724	1,737	1,733	1,713	1,740	1,740	1,729	1,740
1,751	1,739	1,747	1,748	1,730	1,750	1,740	1,732
1,740	1,730	1,748	1,757	1,741	1,733	1,743	1,745
1,748	1,723	1,737	1,748	1,741	1,751	1,714	1,750
1,744	1,748	1,758	1,756	1,727	1,731	1,738	1,753
1,735	1,738	1,743	1,729	1,743	1,737	1,731	1,734
1,741	1,742	1,744	1,756	1,744	1,752	1,739	1,740
1,729	1,745	1,742	1,753	1,743	1,734	1,731	1,734
1,732	1,732	1,746	1,748	1,755	1,738	1,742	1,729
1,731	1,725	1,729	1,745	1,739	1,754	1,752	1,720
1,750	1,734	1,749	1,738	1,747	1,757	1,751	1,746
1,723	1,736	1,746	1,744	1,759	1,728	1,751	1,750
1,746	1,759	1,748	1,740	1,735	1,745	1,740	1,746
1,737	1,726	1,743	1,755	1,740	1,726	1,745	1,744
1,735	1,746	1,739	1,732	1,758	1,744	1,754	1,724
1,742	1,750	1,761	1,758	1,753	1,757	1,720	1,733
1,738	1,728	1,758	1,732	1,763	1,733	1,745	1,766
1,745	1,743	1,734	1,733	1,755	1,756	1,769	1,750
1,740	1,762	1,738	1,742	1,740	1,740	1,760	1,752
1,746	1,728	1,743	1,718	1,738	1,762	1,728	1,734
1,753	1,751	1,748	1,735	1,739	1,729	1,754	1,736
1,762	1,748	1,738	1,726	1,757	1,738	1,726	1,720
1,751	1,734	1,724	1,741	1,752	1,732	1,738	1,739

○ 1). Всі варіанти знаходяться у проміжку [1,713; 1,769] довжиною 0,056 мм. Розіб'ємо його на 8 рівних за довжиною інтервалів:
[1,713; 1,720), [1,720; 1,727), [1,727; 1,734), [1,734; 1,741),
[1,741; 1,748), [1,748; 1,755), [1,755; 1,762), [1,762; 1,769].
Кожну із варіант табл. 1.2 віднесемо до одного із цих частинних інтервалів і просумуємо число варіант, що потрапляють в перший, другий, …, восьмий інтервали. В результаті отримаємо:
$n_1 = 3$, $n_2 = 13$, $n_3 = 32$, $n_4 = 49$, $n_5 = 42$, $n_6 = 35$, $n_7 = 19$, $n_8 = 7$,
$n = \sum_{i=1}^{8} n_i = 200$.

Підсумком є інтервальний статистичний розподіл частот, наведений в табл. 1.3.

Таблиця 1.3

$[x_i; x_{i+1})$	n_i	$[x_i; x_{i+1})$	n_i
[1,713; 1,720)	3	[1,741; 1,748)	42
[1,720; 1,727)	13	[1,748; 1,755)	35
[1,727; 1,734)	32	[1,755; 1,762)	19
[1,734; 1,741)	49	[1,762; 1,769]	7

Із врахуванням того, що обсяг вибірки $n = 200$, запишемо інтервальний статистичний розподіл відносних частот вибірки (табл. 1.4):

Таблиця 1.4

$[x_i; x_{i+1})$	w_i	$[x_i; x_{i+1})$	w_i
[1,713; 1,720)	3/200	[1,741; 1,748)	42/200
[1,720; 1,727)	13/200	[1,748; 1,755)	35/200
[1,727; 1,734)	32/200	[1,755; 1,762)	19/200
[1,734; 1,741)	49/200	[1,762; 1,769]	7/200

2). Гістограма частот розподілу із h=0,007 зображена на рис.1.4, а гістограма відносних частот на рис.1.5.

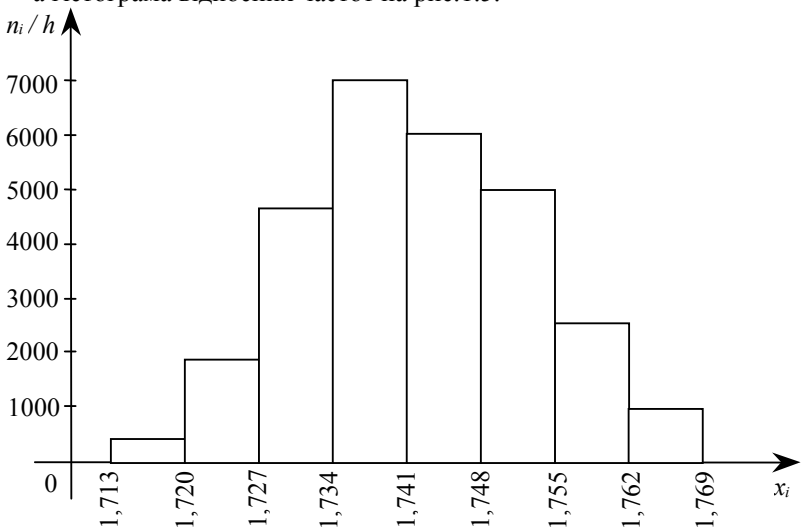


Рис.1.4.

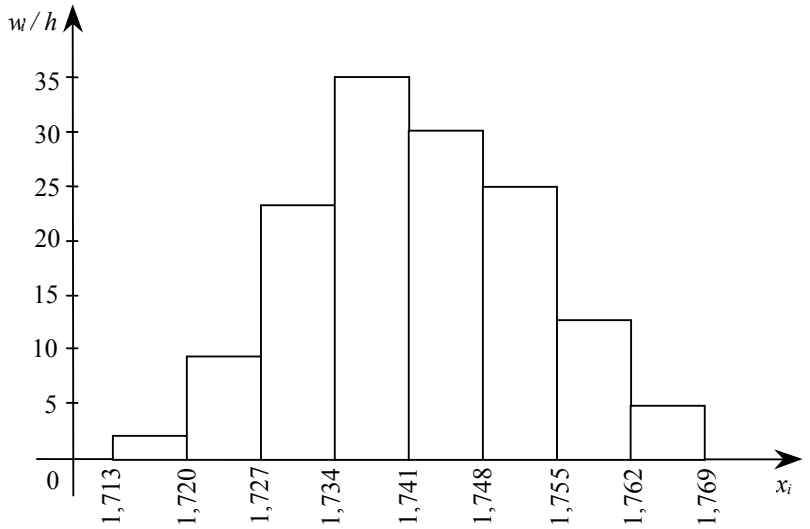


Рис.1.5.

3). Досліджувана кількісна ознака X є неперервною випадковою величиною. Тому і функція розподілу ймовірностей $F(x)$, і емпірична функція $F^*(x)$ є неперервними функціями детермінованого аргумента x.

Нехай $x \le 1{,}713$. Тоді $n_x = 0$, оскільки спостережених значень кількісної ознаки, менших від x, немає. Отже, і $F^*(x) = 0$ для всіх $x \le 1{,}713$.

Знайдемо значення емпіричної функції розподілу для $x = 1{,}720$ — лівого кінця другого інтервалу. При цьому вважатимемо, що ми не можемо зробити цього для кожної внутрішньої точки першого інтервалу. При $x = 1{,}720$ $n_x = 3$ і $F^*(1{,}720) = 3/200 = 0{,}015$.

Для $x = 1{,}727$ $n_x = 3 + 13 = 16$, $F^*(1{,}727) = 16/200 = 0{,}08$.

Для $x = 1{,}734$ $n_x = 16 + 32 = 48$, $F^*(1{,}734) = 48/200 = 0{,}24$.

Аналогічно знаходимо:

$F^*(1{,}741) = (48 + 49)/200 = 97/200 = 0{,}485$;

$F^*(1{,}748) = (97 + 42)/200 = 139/200 = 0{,}695$;

$F^*(1{,}755) = (139 + 35)/200 = 174/200 = 0{,}87$;

$F^*(1{,}762) = (174 + 19)/200 = 193/200 = 0{,}965$;

Нарешті, для $x > 1{,}769$ всі 200 спостережених значень кількісної ознаки менші від x, тобто $n_x = 200$. Тому $F^*(x) = 1$ для $x > 1{,}769$.

Побудуємо графік отриманої емпіричної функції розподілу: спочатку на інтервалах (–∞; 1,713] і (1,769; ∞), а потім у вказаних точках. Для того, щоб показати неперервність зміни $F^*(x)$, отримані сусідні точки з'єднаємо прямолінійними відрізками (рис. 1.6).

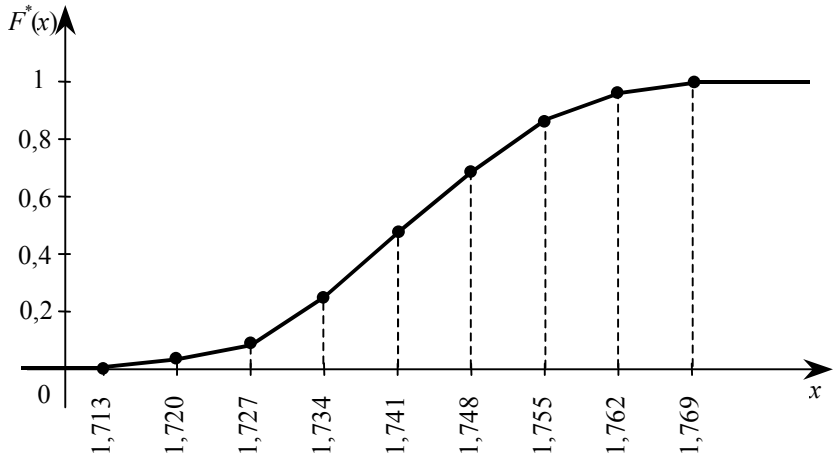


Рис 1.6.

4). Для знаходження числових характеристик вибірки, заданої інтервальним статистичним розподілом із п.1, перейдемо до дискретного розподілу :

x_i	1,7165	1,7235	1,7305	1,7375	1,7445	1,7515	1,7585	1,7655
n_i	3	13	32	49	42	35	19	7

При цьому "нові" варіанти є серединами частинних інтервалів.

$$\overline{x}_в = \frac{\sum_{i=1}^{k} x_i n_i}{n} =$$

$$= \frac{1,7165 \cdot 3 + 1,7235 \cdot 13 + 1,7305 \cdot 32 + 1,7375 \cdot 49 + 1,7445 \cdot 42 + 1,7515 \cdot 35 + 1,7585 \cdot 19 + 1,7655 \cdot 7}{200} =$$

$$= 1,7417;$$

$$D_в = \overline{x^2} - (\overline{x}_в)^2 = \frac{\sum_{i=1}^{k} x_i^2 n_i}{n} - (\overline{x}_в)^2 =$$

$$= \frac{(1,7165)^2 \cdot 3 + (1,7235)^2 \cdot 13 + (1,7305)^2 \cdot 32 + (1,7375)^2 \cdot 49}{200} +$$

$$+ \frac{(1,7445)^2 \cdot 42 + (1,7515)^2 \cdot 35 + (1,7585)^2 \cdot 19 + (1,7655)^2 \cdot 7}{200} - (1,7417)^2 =$$

$$= 0,0001283;$$

$$\sigma_в = \sqrt{D_в} = \sqrt{0,0001283} = 0,0113269.$$

Оскільки всі частинні інтервали даного розподілу мають однакову довжину, то модальним є інтервал [1,734; 1,741), якому відповідає найбільша частота 49. Значення Mo^* міститься всередині цього інтервалу і обчислюється за формулою (1.10).

$$Mo^* = x_m + \frac{n_m - n_{m-1}}{2n_m - n_{m-1} - n_{m+1}} h = 1,734 + \frac{49 - 32}{2 \cdot 49 - 32 - 42} \cdot 0,007 =$$

$$= 1,734 + \frac{17}{24} \cdot 0,007 \approx 1,73896.$$

Для обчислення медіани знайдемо спочатку медіанний частинний інтервал $[x_m; x_{m+1})$, для якого виконуються нерівності:

$$F^*(x_m) < 0,5, \qquad F^*(x_{m+1}) > 0,5.$$

Згідно із 3) $F^*(1,741)=0,485<0,5$, $F^*(1,748)=0,695>0,5$.

Отже, $x_m = 1,741$, $x_{m+1} = 1,748$. За формулою (1.9)

$$Mo^* = x_m + \frac{0,5 - F^*(x_m)}{F^*(x_{m+1}) - F^*(x_m)}(x_{m+1} - x_m) =$$

$$= 1,741 + \frac{0,5 - 0,485}{0,695 - 0,485}(1,748 - 1,741) =$$

$$= 1,741 + \frac{0,015}{0,21} \cdot 0,007 = 1,741 + 0,0005 = 1,7415.$$

Коефіцієнт варіації обчислюємо за формулою (1.14):

$$V = \frac{\sigma_в}{\overline{x}_в} \cdot 100\% = \frac{0,0113269}{1,7417} \cdot 100\% = 0,65\%.$$

§ 2. СТАТИСТИЧНЕ ОЦІНЮВАННЯ

1. *Точкові статистичні оцінки параметрів розподілу та їх властивості.*
2. *Оцінка середньої генеральної для простої вибірки (повторної та безповторної).*
3. *Оцінка генеральної частки для простої вибірки (повторної та безповторної).*
4. *Середні квадратичні помилки (СКП) простої вибірки. Виправлена дисперсія вибіркова.*
5. *Інтервальні статистичні оцінки (Довірчі інтервали для оцінок* $\bar{x}_г$ *та p для немалих вибірок.*
6. *Знаходження мінімального обсягу вибірки.*
7. *Довірчі інтервали для оцінки* $\bar{x}_г = a$ *для малої вибірки. Довірчі інтервали для* $D_г$ *та* $\sigma_г$ *у випадку малої вибірки).*

1. Точкові статистичні оцінки параметрів розподілу та їх властивості.

Нехай досліджується неперервна кількісна ознака X об'єктів генеральної сукупності з метою знаходження невідомого закону розподілу. В розпорядженні дослідника є статистичні дані вибірки. Припустимо, що з тих чи інших міркувань висунута гіпотеза про нормальний закон розподілу ознаки X (дослідження питання про правильність цієї гіпотези або хибність буде проведене в наступному параграфі). Оскільки нормальний розподіл повністю визначається двома параметрами a та σ, то виникає необхідність оцінити їх, тобто знайти наближені значення, використовуючи **тільки** спостережені варіанти $x_1, x_2, ..., x_n$ вибірки обсягом n. Параметр $a = M(X)$, математичне сподівання характеризує середнє арифметичне спостережених можливих значень випадкової величини X. З другого боку, $\bar{x}_в$ — це середнє арифметичне варіант. Тому (поки що інтуїтивно) доцільно наближати a середнім вибірковим:

$$a \approx \bar{x}_в = \left(\sum_{i=1}^{k} x_i n_i \right) \Big/ n. \qquad (2.1)$$

Аналогічно

$$\sigma \approx \sigma_в = \sqrt{\sum_{i=1}^{k} (x_i - \bar{x}_в)^2 n_i \Big/ n}. \qquad (2.2)$$

Ще раз відмітимо, що праві частини рівностей (2.1), (2.2) є **випадковими величинами**, оскільки об'єкти у вибірку потрапляють **випадковим чином**.

Нехай тепер кількісна ознака X об'єктів генеральної сукупності є дискретною випадковою величиною. Припустимо, що статистичні розподіли генеральної та вибіркової сукупності описуються таблицями:

$$\begin{array}{c|cccc} x_i & x_1 & x_2 & \dots & x_m \\ \hline N_i & N_1 & N_2 & \dots & N_m \end{array}, \qquad N = \sum_{i=1}^{m} N_i, \tag{2.3}$$

$$\begin{array}{c|cccc} x_i & x_1 & x_2 & \dots & x_m \\ \hline n_i & n_1 & n_2 & \dots & n_m \end{array}, \qquad n = \sum_{i=1}^{m} n_i, \tag{2.4}$$

де N та n — обсяги генеральної та вибіркової сукупностей відповідно. Розподіл (2.3) є гіпотетичним — він завжди буде для нас невідомим, бо в протилежному випадку відпала б необхідність у дослідженні вибіркової сукупності. Нарешті, зауважимо, що деякі із частот розподілу (2.4) можуть дорівнювати нулю, що відповідає ситуації, коли значення кількісної ознаки об'єктів генеральної сукупності не зустрілися серед варіант вибірки (порівняйте розподіл (2.4) з (1.2)).

За аналогією із числовими характеристиками вибірки наступні формули визначають **числові характеристики генеральної сукупності**:

$$\overline{x}_{\text{г}} = \left(\sum_{i=1}^{m} x_i N_i \right) \Big/ N, \tag{2.5}$$

$$D_{\text{г}} = \left(\sum_{i=1}^{m} (x_i - \overline{x}_{\text{г}})^2 N_i \right) \Big/ N, \tag{2.6}$$

$$\sigma_{\text{г}} = \sqrt{D_{\text{г}}}. \tag{2.7}$$

Ці **числа** невідомі, і оцінки (наближення) їх дають такі рівності:

$$\overline{x}_{\text{г}} \approx \overline{x}_{\text{в}}, \qquad D_{\text{г}} \approx D_{\text{в}}, \qquad \sigma_{\text{г}} \approx \sigma_{\text{в}}. \tag{2.8}$$

Наведені приклади дозволяють зробити деякі висновки. Ліві частини **наближених** рівностей (2.1), (2.2) та (2.8) є невідомими параметрами "відомого" закону розподілу або числовими характеристиками генеральної сукупності; вони є невідомими числами для конкретної генеральної сукупності. Праві частини цих рівностей є функціями випадкових величин, які для **фіксованої** вибірки статистичних даних набирають числові значення, що можна зобразити точками. Це дозволяє назвати їх **точковими статистичними оцінками** відповідних параметрів або числових характеристик генеральної сукупності.

Позначимо узагальнено символом Θ ліві частини наближених рівностей (2.1), (2.2), (2.8) (а також багатьох інших, що можна отримати для інших законів розподілу), а символом Θ^* — праві частини цих рівностей.

Точковою статистичною оцінкою, **вибірковою функцією** або **статистикою** числового параметра Θ називається функція вибіркових значень (варіант) $\Theta^* = \Theta^*(x_1, x_2, ..., x_n)$, яка в певному статистичному сенсі є близькою до справжнього значення цього параметра.

Незміщеною називається точкова статистична оцінка Θ^*, математичне сподівання якої дорівнює оцінюваному параметру Θ при довільному обсязі вибірки, тобто

$$M(\Theta^*) = \Theta. \qquad (2.9)$$

Зміщеною називається оцінка, для якої не виконується рівність (2.9).

Нехай для оцінювання параметра Θ можуть бути використані незміщені точкові оцінки $\Theta_1^*, \Theta_2^*, ..., \Theta_k^*$. Оцінка Θ_m^*, $1 \le m \le k$, називається **ефективною**, якщо при заданому обсязі n вибірки для неї виконується рівність

$$D(\Theta_m^*) = \min_{1 \le i \le k} D(\Theta_i^*).$$

Оцінка Θ^* називається **спроможною** оцінкою параметра Θ, якщо при $n \to \infty$ вона збігається по імовірності до Θ, тобто для як завгодно малого $\varepsilon > 0$ має місце граничний перехід

$$\lim_{n \to \infty} P\left(\left| \Theta - \Theta^* \right| < \varepsilon \right) = 1.$$

2. Оцінки середньої генеральної для простої вибірки.

Теорема 2.1. Для повторної вибірки обсягом n середня вибіркова $\bar{x}_в$ є незміщеною і спроможною оцінкою невідомої середньої генеральної $\bar{x}_г$. Якщо n досить велике, тоді $\bar{x}_в$ з достатнім ступенем точності розподілена за нормальним законом з параметрами:

$$a = \bar{x}_г, \qquad \sigma = \sqrt{D_г / n}. \qquad (2.10)$$

Теорема 2.2. Для безповторної вибірки обсягом n середня вибіркова $\bar{x}_в$ є незміщеною оцінкою невідомої середньої генеральної $\bar{x}_г$. Для досить великих n $\bar{x}_в$ з достатнім ступенем точності розподілена за нормальним законом з параметрами:

$$a = \bar{x}_{\text{г}}, \qquad \sigma = \sqrt{\frac{D_{\text{г}}}{n} \cdot \frac{N-n}{N-1}}, \tag{2.11}$$

де N — обсяг генеральної сукупності.

Зауваження. Обсяг генеральної сукупності N, як правило, дуже великий. Тому заміна в знаменнику другої рівності (2.11) $N-1$ на N невідчутна для σ. В зв'язку із цим надалі будемо користуватися рівністю

$$\sigma = \sqrt{\frac{D_{\text{г}}}{n} \cdot \left(1 - \frac{n}{N}\right)}. \tag{2.12}$$

3. Оцінки генеральної частки для простої вибірки.

Нехай досліджується якісна ознака об'єктів генеральної сукупності, число яких є скінченним. **Генеральною часткою** будемо називати відношення числа M об'єктів генеральної сукупності, що володіють ознакою α, до обсягу N генеральної сукупності:

$$p = M/N.$$

Теорема 2.3. Для повторної вибірки обсягом n середня вибіркова частка w є незміщеною і спроможною точковою оцінкою невідомої генеральної частки p. Якщо n є досить великим, тоді w з достатнім ступенем точності розподілена за нормальним законом з параметрами:

$$a = p, \qquad \sigma = \sqrt{pq/n}, \tag{2.13}$$

де $q = 1 - p$.

Теорема 2.4. Для безповторної вибірки обсягом n вибіркова частка w є незміщеною точковою оцінкою невідомої генеральної частки p. Якщо n є досить великим, тоді w з достатнім ступенем точності розподілена за нормальним законом з параметрами:

$$a = p, \qquad \sigma = \sqrt{\frac{pq}{n} \cdot \frac{N-n}{N-1}}, \tag{2.14}$$

Зауваження. Оскільки обсяг генеральної сукупності N в більшості випадків дуже великий, то заміна в знаменнику другої рівності (2.14) $N-1$ на N невідчутна для σ. З огляду на це надалі будемо користуватися рівністю

$$\sigma = \sqrt{\frac{pq}{n}\left(1 - \frac{n}{N}\right)}. \tag{2.15}$$

4. Середні квадратичні помилки (СКП) простої вибірки. Виправлена дисперсія вибіркова.

Випадкові величини $\overline{x}_{\text{в}}$ та w є незміщеними точковими статистичними оцінками невідомих $\overline{x}_{\text{г}}$ та p відповідно. Їх реалізація або можливі значення, знайдені на основі даних простої вибірки (повторної або безповторної), не співпадають із оцінюваними параметрами. І кожне таке неспівпадання або відхилення природно називати **помилкою репрезентативності оцінки**, зумовленою тим, що досліджується не вся генеральна сукупність, а лише її частина (вибіркова сукупність). Для статистики дуже важливою є інформація про середню величину таких помилок.

Середньою квадратичною помилкою (**СКП**) при оцінюванні невідомих середньої генеральної $\overline{x}_{\text{г}}$ та генеральної частки p називається середнє квадратичне відхилення середньої вибіркової $\overline{x}_{\text{в}}$ та вибіркової частки w відповідно.

Із врахуванням результатів теорем 2.1–2.4 можна вказати наступні формули для визначення середніх квадратичних помилок:

СКП середньої вибіркової повторної вибірки (див. (2.10))

$$\overline{\sigma}_{\overline{x}} = \sqrt{D_{\text{г}}/n}; \tag{2.16}$$

СКП середньої вибіркової безповторної вибірки (див. (2.12))

$$\overline{\sigma}'_{\overline{x}} = \sqrt{\frac{D_{\text{г}}}{n}\left(1-\frac{n}{N}\right)}; \tag{2.17}$$

СКП вибіркової частки повторної вибірки (див. (2.13))

$$\overline{\sigma}_{w} = \sqrt{pq/n}; \tag{2.18}$$

СКП вибіркової частки безповторної вибірки (див. (2.15))

$$\overline{\sigma}'_{w} = \sqrt{\frac{pq}{n}\left(1-\frac{n}{N}\right)}. \tag{2.19}$$

На практиці користуватися формулами (2.16)–(2.19) неможливо, оскільки для цього необхідно знати або дисперсію генеральну, або генеральну частку (проаналізуйте, наскільки обтяжливою є ця умова, згадавши початкову умову задачі оцінювання). Цю принципову трудність можна усунути за рахунок заміни у формулах СКП дисперсії генеральної $D_{\text{г}}$ на дисперсію вибіркову $D_{\text{в}}$, а генеральної частки p — на вибіркову частку w. Проте при такій заміні потрібно попередньо впевнитися, чи не з'явиться ще додаткова систематична помилка за рахунок зміщеності оцінок $D_{\text{в}}$ в обох задачах оцінювання. Якби вони виявилися зміщеними, то при заміні слід було б ввести "виправлені" оці-

нки невідомих дисперсій генеральних. На шляху реалізації вказаного вище підходу корисними є такі твердження.

Теорема 2.5. Математичне сподівання дисперсії вибіркової в задачі про оцінювання невідомої середньої генеральної для повторної вибірки визначається рівністю

$$M(D_{\text{в}}) = \frac{n-1}{n} D_{\text{г}}, \tag{2.20}$$

а для безповторної —

$$M(D_{\text{в}}) = \frac{n-1}{n} \cdot \frac{N}{N-1} D_{\text{г}}, \tag{2.21}$$

де n та N відповідно обсяги вибіркової та генеральної сукупностей.

Теорема 2.6. Математичне сподівання дисперсії вибіркової в задачі про оцінювання невідомої генеральної частки для повторної вибірки визначається рівністю

$$M(D_{\text{в}}) = \frac{n-1}{n} pq, \tag{2.22}$$

а для безповторної —

$$M(D_{\text{в}}) = \frac{n-1}{n} \cdot \frac{N}{N-1} pq. \tag{2.23}$$

Зміст теорем 2.5 та 2.6 полягає в тому, що дисперсія вибіркова є зміщеною оцінкою дисперсії генеральної в задачах оцінювання невідомих $\bar{x}_{\text{г}}$ та p у випадку простої вибірки (повторної та безповторної). При цьому формули (2.20)–(2.23) вказують на заниження значень дисперсії генеральної за рахунок наявності в кожній із них множника $(n-1)/n$.

Проте, цю зміщеність легко "виправити": достатньо дисперсію вибіркову помножити на дріб $n/(n-1)$.

Виправленою дисперсією називається числова характеристика S^2, яка визначається рівністю

$$S^2 = \frac{n}{n-1} D_{\text{в}}.$$

Для **немалих** вибірок (обсягом $n \geq 30$), замінивши у формулах (2.16)–(2.19) дисперсію генеральну $D_{\text{г}}$ вибірковою $D_{\text{в}}$, а генеральну частку p — вибірковою w, отримаємо використовувані на практиці формули для знаходження СКП:

СКП середньої вибіркової повторної вибірки

$$\overline{\sigma}_{\bar{x}} = \sqrt{D_{\text{в}}/n} = \sigma_{\text{в}}/\sqrt{n}\,; \tag{2.24}$$

СКП середньої вибіркової безповторної вибірки

$$\overline{\sigma}'_{\bar{x}} = \sqrt{\frac{D_{\text{в}}}{n}\left(1 - \frac{n}{N}\right)}; \qquad (2.25)$$

СКП вибіркової частки повторної вибірки

$$\overline{\sigma}_w = \sqrt{\frac{w(1-w)}{n}}; \qquad (2.26)$$

СКП вибіркової частки безповторної вибірки

$$\overline{\sigma}'_w = \sqrt{\frac{w(1-w)}{n}\left(1 - \frac{n}{N}\right)}. \qquad (2.27)$$

5. Інтервальні статистичні оцінки.

Точкова статистична оцінка Θ^* не співпадає (за виключанням рідкісних випадків) із справжнім значенням невідомого параметра Θ. Тому завжди виникає похибка при заміні невідомого параметра його оцінкою, тобто $\left|\Theta - \Theta^*\right| > 0$. Величина похибки при цьому невідома, хоча потрібно знати, до яких помилок може призвести вказана вище заміна.

Інтервальною називається статистична оцінка, яка визначається **двома числами** — кінцями інтервалу. Перевагою інтервальних оцінок є те, що вони дозволяють встановити точність і надійність оцінок.

Число $\delta > 0$, яке фігурує у нерівності

$$\left|\Theta - \Theta^*\right| < \delta, \qquad (2.28)$$

природно назвати **точністю** оцінки Θ^*. Проте говорити про виконання нерівності (2.28) можна тільки в імовірносному сенсі, бо Θ^* — випадкова величина. Тобто, для достатньо малого δ ця нерівність є випадковою подією.

Надійністю (довірчою імовірністю) оцінки Θ по Θ^* називається імовірність γ виконання нерівності (2.28):

$$P\left(\left|\Theta - \Theta^*\right| < \delta\right) = \gamma. \qquad (2.29)$$

Довірчим називається інтервал $(\Theta^* - \delta;\ \Theta^* + \delta)$, який із заданою надійністю γ покриває невідомий параметр Θ.

Довірчі інтервали для оцінок $\bar{x}_{\text{г}}$ та p для немалих вибірок

Найбільше відхилення середньої вибіркової (або вибіркової частки) від середньої генеральної (або генеральної частки), яке можливе для заданої довірчої імовірності γ, називається **граничною помилкою** Δ.

Гранична помилка знаходиться за формулою

$$\Delta = t\overline{\sigma}, \tag{2.30}$$

де t — корінь рівняння $2\Phi(t) = \gamma$.

Підставивши в рівність (2.30) вирази СКП (2.24)–(2.27), отримаємо **придатні для практики формули граничної помилки**:

середньої вибіркової повторної вибірки

$$\Delta = t\sqrt{D_{\text{в}}/n}; \tag{2.31}$$

середньої вибіркової безповторної вибірки

$$\Delta = t\sqrt{\frac{D_{\text{в}}}{n}\left(1-\frac{n}{N}\right)}; \tag{2.32}$$

частки вибіркової повторної вибірки

$$\Delta = t\sqrt{w(1-w)/n}; \tag{2.33}$$

частки вибіркової безповторної вибірки

$$\Delta = t\sqrt{\frac{w(1-w)}{n}\left(1-\frac{n}{N}\right)}. \tag{2.34}$$

В результаті $(\overline{x}_{\text{в}} - \Delta;\ \overline{x}_{\text{в}} + \Delta)$ є довірчим інтервалом, який з надійністю γ покриває невідому середню генеральну $\overline{x}_{\text{г}}$. Аналогічно $(w - \Delta;\ w + \Delta)$ — довірчий інтервал, який з тією ж надійністю покриває невідому генеральну частку p.

Зауваження. В деяких випадках може бути відомою числова характеристика $\sigma_{\text{г}}$ (наприклад, як інформація технологічного процесу). Тоді $D_{\text{в}}$ у формулах (2.31), (2.32) слід замінити на $\sigma_{\text{г}}^2$.

6. Знаходження мінімального обсягу вибірки.

Перед утворенням вибіркової сукупності необхідно з'ясувати, яким повинен бути її обсяг.

Наступні формули визначають мінімальні обсяги вибірки при оцінюванні невідомих:

а) середньої генеральної

$$n = \frac{t^2 D_{\text{г}}}{\Delta^2} \tag{2.35}$$

для повторної вибірки,

$$n' = \frac{nN}{n+N} \tag{2.36}$$

для безповторної вибірки (n визначається (2.35));

б) генеральної частки

$$n = \frac{t^2 pq}{\Delta^2} \tag{2.37}$$

для повторної вибірки,

$$n' = \frac{nN}{n+N} \tag{2.38}$$

для безповторної вибірки (n визначається (2.37));
де N — обсяг генеральної сукупності.

7. Довірчі інтервали для оцінки $\bar{x}_г = a$ для малої вибірки

Нехай про досліджувану кількісну ознаку X відомо тільки те, що вона розподілена за нормальним законом. Ставиться задача: побудувати довірчий інтервал для оцінки невідомого параметра $a = M(X)$ (або $\bar{x}_г$ у випадку скінченності обсягу генеральної сукупності) за даними повторної вибірки малого обсягу n, заданою довірчою імовірністю γ і якщо невідомий параметр $\sigma = \sigma_г$.

У [8] доведено, що довірчий інтервал для невідомого параметра a ($\bar{x}_г$) при невідомому $\sigma(\sigma_г)$ з надійністю γ має такий вид:

$$\bar{x}_в - t(\gamma, n-1)\frac{S}{\sqrt{n}} < a < \bar{x}_в + t(\gamma, n-1)\frac{S}{\sqrt{n}}, \tag{2.39}$$

де

$$S^2 = \sum_{i=1}^{n}(x_i - \bar{x}_в)^2 \Big/ (n-1), \tag{2.40}$$

параметр $t=t(\gamma, n\text{-}1)$ – корінь рівняння

$$P\left(\left|\frac{\bar{x}_в - a}{S/\sqrt{n}}\right| < t\right) = P\left(|T| < t\right) = 2\int_0^t g_k(t)dt = \gamma, \tag{2.41}$$

який можна знайти за табл.4 додатків в залежності від заданої довірчої імовірності γ і числа ступенів вільності $k = n - 1$; $g_k(t)$ – густина розподілу Ст'юдента.

Довірчі інтервали для $D_г$ та $\sigma_г$ у випадку малої вибірки.

При знаходженні мінімального обсягу вибірки необхідною є інформація про дисперсію генеральну (середнє квадратичне відхилення генеральне). Часто в розпорядженні дослідника є тільки вибірка малого обсягу.

Можна довести (див. [8]), що довірчий інтервал для оцінки невідомої дисперсії генеральної $D_г=\sigma^2$ з надійністю γ має такий вид:

$$\frac{(n-1)S^2}{\chi_2^2} < \sigma^2 < \frac{(n-1)S^2}{\chi_1^2}, \tag{2.42}$$

де S визначається рівністю (2.40), значення χ_1^2 і χ_2^2 знаходяться за табл.6 додатків з використанням рівнянь

$$P\left(\chi^2(k) > \chi_1^2(p;k)\right) = p\,,\ p = \frac{1+\gamma}{2}; \tag{2.43}$$

$$P\left(\chi^2(k) > \chi_2^2(p;k)\right) = p\,,\ p = \frac{1-\gamma}{2}; \tag{2.44}$$

$k = n - 1$ – число ступенів вільності закону розподілу χ^2.

Із подвійної нерівності (2.42) отримаємо рівносильну подвійну нерівність

$$\frac{S\sqrt{n-1}}{\chi_2} < \sigma < \frac{S\sqrt{n-1}}{\chi_1}, \tag{2.45}$$

яка визначає довірчий інтервал для оцінки невідомої $\sigma(\sigma_г)$.

Зауваження. В табл. 6 додатків наведені значення $\chi^2(p;k)$, що задовільняють рівняння $P\left(\chi^2(k) > \chi^2(p;k)\right) = p$ тільки для $k = n - 1 \le 30$, а також для $k = 40, 50, 100$. Це зумовлено тим, що при зростанні k закон розподілу χ^2 наближається до нормального.

РОЗВ'ЯЗУВАННЯ ТИПОВИХ ЗАДАЧ

Задача 2.1. Результати випробувань на міцність стальних дротів однакової довжини наведені в табл. 2.1 інтервальним розподілом.

Таблиця 2.1

Розривне зусилля, (кг/мм^2)	[30; 32)	[32; 34)	[34; 36)	[36; 38)	[38; 40)	[40; 42)	[42; 44]
Кількість дротів	6	12	17	29	19	13	4

1) Знайти довірчий інтервал, який з надійністю $\gamma = 0{,}95$ покриває середнє розривне всієї партії 4000 дротів.

2) Знайти довірчу імовірність того, що середня вибіркова відхилиться від середньої генеральної за абсолютною величиною не більше, ніж на 0,5 кг/мм^2.

○ 1) Утворена (немала, бо $n = 100 > 30$) вибірка є безповторною, оскільки після перевірки об'єкт (дріт) не може бути повернутий в генеральну сукупність (дріт розривається внаслідок перевірки його міцності). Для знаходження меж довірчого інтервалу граничну помилку Δ знайдемо за формулою (2.32), попередньо знайшовши числові характеристики вибірки $\bar{x}_в$ та $D_в$.

$x_i = \frac{x_k + x_{k+1}}{2}$	31	33	35	37	39	41	43
n_i	6	12	17	29	19	13	4

$$\bar{x}_в = \frac{\sum_{i=1}^{k} x_i n_i}{n} =$$

$$= \frac{31\cdot 6 + 33\cdot 12 + 35\cdot 17 + 37\cdot 29 + 39\cdot 19 + 41\cdot 13 + 43\cdot 4}{100} = 36{,}96;$$

$$D_в = \overline{x^2} - (\bar{x}_в)^2 = \frac{\sum_{i=1}^{k} x_i^2 n_i}{n} - (\bar{x}_в)^2 =$$

$$== \frac{31^2\cdot 6 + 33^2\cdot 12 + 35^2\cdot 17 + 37^2\cdot 29 + 39^2\cdot 19 + 41^2\cdot 13 + 43^2\cdot 4}{100} - (36{,}96)^2 =$$

$$= 9{,}0384$$

За таблицями значень функції Лапласа (табл. 3 додатків) знайдемо значення t з рівняння $\Phi(t) = \gamma/2 = 0{,}95/2 = 0{,}475$, $t = 1{,}96$. За формулою (2.32), де $N = 4000$,

$$\Delta = 1{,}96\sqrt{\frac{9{,}0384}{100}\left(1 - \frac{100}{4000}\right)} \approx 0{,}5818,$$

тоді ліва межа довірчого інтервалу

$$\bar{x}_{\text{в}} - \Delta = 36{,}96 - 0{,}5818 = 36{,}3782,$$

права —

$$\bar{x}_{\text{в}} + \Delta = 36{,}96 + 0{,}5818 = 37{,}5418.$$

Остаточно шуканий довірчий інтервал має такий вид (36,3782; 37,5418).

2) Шуканою є імовірність $P\left(|\bar{x}_{\text{в}} - \bar{x}_{\text{г}}| \le 0{,}5\right)$, яку можна знайти за формулою

$$P\left(|\bar{x}_{\text{в}} - \bar{x}_{\text{г}}| < \varepsilon\right) = 2\Phi\left(\frac{\varepsilon}{\bar{\sigma}}\right),$$

оскільки $\bar{x}_{\text{в}}$ є нормально розподіленою випадковою величиною (згідно з теоремою 2.2), $M(\bar{x}_{\text{в}}) = \bar{x}_{\text{г}}$. За формулою (2.25)

$$\bar{\sigma} = \bar{\sigma}'_{\bar{x}} = \sqrt{\frac{D_{\text{в}}}{n}\left(1 - \frac{n}{N}\right)} = \sqrt{\frac{9{,}0384}{100}\left(1 - \frac{100}{4000}\right)} \approx 0{,}2968.$$

Тоді $\varepsilon/\bar{\sigma} = 0{,}5/0{,}2968 = 1{,}69$ і за табл. 3 додатків

$$P\left(|\bar{x}_{\text{в}} - \bar{x}_{\text{г}}| \le 0{,}5\right) = 2\Phi\left(0{,}5/0{,}2968\right) =$$
$$= 2\Phi\left(1{,}69\right) = 2 \cdot 0{,}45449 = 0{,}90898. \quad \bullet$$

Задача 2.2. Із партії 9000 однотипних деталей перевірено 400 деталей. Серед них виявилося 360 першосортних деталей. Знайти межі, в яких з імовірністю 0,9542 міститься частка деталей першого сорту всієї партії, якщо вибірка: а) повторна; б) безповторна.

○ За табл. 3 додатків знаходимо, що коренем рівняння $2\Phi(t) = 0{,}9542$ є $t = 2$. Частка вибіркова $w = 360/400 = 0{,}9$. Граничну помилку повторної вибірки знайдемо за формулою (2.33) при $t = 2$, $w = 0{,}9$ і $n = 400$:

$$\Delta = 2 \cdot \sqrt{\frac{0{,}9(1 - 0{,}9)}{400}} = 0{,}03.$$

Тоді для повторної вибірки довірчим є інтервал (0,9 – 0,03; 0,9 + 0,03) = (0,87; 0,93).

Для безповторної вибірки граничну помилку знайдемо за формулою (2.34) (при тих самих значеннях t, w та n):

$$\Delta = 2\sqrt{\frac{0{,}9 \cdot 0{,}1}{400}\left(1 - \frac{400}{9000}\right)} \approx 0{,}029.$$

В результаті отримаємо для безповторної вибірки такий довірчий інтервал (0,871; 0,929). ●

Задача 2.3. Знайти необхідні обсяги повторної і безповторної вибірок, щоб при визначенні середньої тривалості безперервної роботи блоків в партії із 6000 блоків з імовірністю 0,99 відхилення середньої генеральної від середньої вибіркової не перевищувало за абсолютною величиною 30 год. Середнє квадратичне відхилення генеральне вважати рівним 160 год.

○ Довірча імовірність 0,99 визначає $t = 2{,}58$ (за табл. 3 додатків з рівняння $\Phi(t) = 0{,}99/2 = 0{,}495$). За формулою (2.35) при $\Delta = 30$, $D_г = \sigma_г^2 = 160^2$ знайдемо необхідний обсяг повторної вибірки:

$$n = \frac{(2{,}58)^2 \cdot (160)^2}{(30)^2} \approx 189{,}34,$$

тобто вибірка повинна складатися з $n = 190$ електронних блоків.

Якщо вибірка безповторна, то обсяг її згідно формули (2.36) при $n = 190$, $N = 6000$ повинен складати

$$n' = \frac{190 \cdot 6000}{190 + 6000} \approx 185. \quad ●$$

Задача 2.4. Для визначення врожайності гречки зробили вибірку, до якої ввійшло вісім ділянок. Результати вибіркових спостережень за урожайністю (ц/га) наведені в таблиці:

Номер ділянки	1	2	3	4	5	6	7	8
Урожайність	18,2	15,1	16,9	17,8	19,1	15,4	20,5	16,3

Знайти довірчий інтервал, в якому з надійністю 0,95 перебуватиме середня врожайність $(\bar{x}_г)$ гречки всього поля.

○ Знайдемо точкові незміщені статистичні оцінки для $\bar{x}_г$ та $D_г$:

$$\bar{x}_{\text{в}} = \frac{\sum x_i}{n} = \frac{18{,}2 + 15{,}1 + 16{,}9 + 17{,}8 + 19{,}1 + 15{,}4 + 20{,}5 + 16{,}3}{8} = 17{,}4125;$$

$$\overline{x^2} = \sum x_i^2 / n = (331{,}24 + 228{,}01 + 285{,}61 + 316{,}84 + 364{,}81 + 237{,}16 + $$
$$+ 420{,}25 + 265{,}69)/8 = 306{,}20125;$$

$$D_{\text{в}} = \overline{x^2} - (\bar{x}_{\text{в}})^2 = 306{,}21125 - 303{,}19515 = 3{,}0614;$$

$$S^2 = \frac{n}{n-1} D_{\text{в}} = \frac{8}{7} \cdot 3{,}0061 = 3{,}4355; \quad S = 1{,}8535.$$

Будемо вважати, що врожайність усього поля розподілена за нормальним законом. Тоді за даною надійністю $\gamma = 0{,}95$ та числом ступенів вільності $k = n - 1 = 8 - 1 = 7$, користуючись табл. 4 додатків, знайдемо значення $t(\gamma, n-1) = 2{,}365$. Обчислимо межі довірчого інтервалу (2.39):

$$\bar{x}_{\text{в}} - t(\gamma, n-1)\, S/\sqrt{n-1} = 17{,}4125 - 2{,}265 \cdot 1{,}8535/7 =$$
$$= 17{,}4125 - 0{,}6263 = 16{,}7862;$$

$$\bar{x}_{\text{в}} + t(\gamma, n-1)\, S/\sqrt{n-1} = 17{,}4125 + 2{,}265 \cdot 1{,}8535/7 =$$
$$= 17{,}4125 + 0{,}6263 = 18{,}0388;$$

Отже, довірчий інтервал для середньої врожайності гречки всього поля має такий вид:

$$16{,}7862 < \bar{x}_{\text{г}} < 18{,}0388. \quad \bullet$$

§ 3. СТАТИСТИЧНА ПЕРЕВІРКА СТАТИСТИЧНИХ ГІПОТЕЗ

1. *Статистичні гіпотези та їх види.*
2. *Статистичний критерій перевірки основної гіпотези. Потужність критерію.*
3. *Параметричні статистичні гіпотези*
4. *Критерій узгодженості Пірсона* $\left(\chi^2\right)$.
5. *Перевірка гіпотези про нормальний розподіл генеральної сукупності.*

1. Статистичною називається гіпотеза про вид невідомого розподілу випадкової величини (кількісної ознаки об'єктів генеральної сукупності) або про параметри відомого розподілу.

Поряд із висунутою гіпотезою розглядають і гіпотезу, яка суперечить їй. Тому надалі будемо припускати, що у нас є дві гіпотези: H_0 та H_1, які не перетинаються. Гіпотезу H_0 будемо називати **основною** або **нульовою**, а гіпотезу H_1 — **конкуруючою** або **альтернативною**.

Гіпотези розрізняються за числом припущень. **Простою** називається гіпотеза, яка містить тільки одне припущення. **Складною** називається гіпотеза, яка складається із скінченного або нескінченного числа простих гіпотез.

Висунута статистична гіпотеза може бути правильною або хибною. Для перевірки її правильності використовують статистичні дані і статистичні методи, тому перевірку називають **статистичною**.

Для перевірки основної гіпотези потрібно мати критерій правильності цієї гіпотези. Як вже зазначалося, в розпорядженні дослідника є тільки вибірка $X_1, X_2, ..., X_n$, де X_i — значення кількісної ознаки i-ого об'єкта вибірки $(i = \overline{1,n})$, або значення вибіркової частки ω у випадку вивчення якісної ознаки об'єктів генеральної сукупності.

2. Статистичним критерієм (або просто **критерієм**, чи **статистикою**) називається випадкова великина K, яка використовується для перевірки основної гіпотези і закон розподілу якої (точний або наближений) відомий. Для кожного конкретного випадку величина K спеціально підбирається і може позначатися різними літерами: U або Z, якщо вона нормально розподілена, F або v^2 — по закону Фішера-Снедокора, T — по закону Ст'юдента, χ^2 — по закону "хі-квадрат", K — по закону Колмогорова і т. д.

Можливі значення випадкової величини (критерію) K розбиваються на дві непорожні множини Q та $\overline{Q}$ $\left(Q \cap \overline{Q} = \varnothing \right)$ такі, що Q складається із значень критерію, при яких H_0 приймається, а $\overline{Q}$ — із тих значень критерію, при яких H_0 відхиляється (а отже, приймається H_1). Множину $\overline{Q}$ називають **критичною областю**, а множину Q — **областю прийняття гіпотези**, або **областю допустимих значень**.

Для конкретної вибірки обчислюється значення критерію як функції варіант. Отримане значення позначається $K_{сп}$ і називається **спостереженим значеням критерію**.

Сформулюємо **основний принцип статистичної перевірки статистичної гіпотези**: якщо спостережене значення критерію належить критичній області $\left(K_{сп} \in \overline{Q} \right)$, тоді гіпотезу H_0 відхиляють; якщо спостережене значення критерію належить області допустимих значень $\left(K_{сп} \in Q \right)$, тоді гіпотезу H_0 приймають.

Зауваження. Припустимо, що для конкретних гіпотез H_0 і H_1 множини Q і $\overline{Q}$ **визначені**. В ряді посібників по математичній статистиці під статистичним критерієм розуміється правило, яке реалізує основний принцип статистичної перевірки статистичної гіпотези.

Оскільки висновок про правильність гіпотези робиться за результатами скінченної вибірки, а вона може бути "невдалою", то завжди існує ризик прийняти хибне рішення. При цьому можуть бути допущені помилки двох родів.

Якщо буде відхилена гіпотеза H_0 (і прийнята H_1), в той час як насправді правильною є H_0, тоді це є **помилка першого роду**; її імовірність позначають α:

$$\alpha = P\left(K_{сп} \in \overline{Q} / H_0 \right) = P\left(H_1 / H_0 \right),$$

де $P\left(H_1 / H_0 \right)$ — імовірність того, що буде прийнята гіпотеза H_1, якщо насправді для генеральної сукупності правильною є гіпотеза H_0. Число α називають **рівнем значущості**.

Якщо буде прийнята гіпотеза H_0, в той час як насправді правильною є H_1, тоді буде допущена **помилка другого роду**, її імовірність позначають β:

$$\beta = P\left(K_{сп} \in Q / H_1 \right) = P\left(H_0 / H_1 \right).$$

В цьому параграфі розглянемо один із методів перевірки статистичних гіпотез, який передбачає виконання таких кроків:

1) формулювання основної і конкуруючої гіпотез;

2) вибір відповідного рівня значущості α;

3) вибір статистичного критерія K перевірки гіпотези;

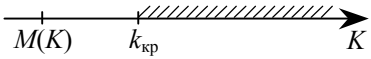
4) знаходження критичної області $\overline{Q}$ і області Q прийняття гіпотези;

5) формулювання правила перевірки гіпотези: гіпотеза H_0 приймається при заданому рівні значущості α, якщо $K_{сп} \in Q$; гіпотеза H_0 відкидається, якщо $K_{сп} \in \overline{Q}$.

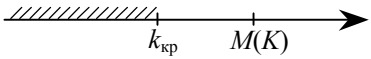
Розглянемо деякі особливості структури критичної області $\overline{Q}$. Якщо $\overline{Q}$ розташована зліва **і** справа від математичного сподівання випадкової величини K, то критична область називається **двосторонньою**, а критерій K — **двостороннім критерієм значущості**. Якщо ж $\overline{Q}$ розташована зліва **або** справа від математичного сподівання випадкової величини K, то критична область називається **односторонньою**, а критерій — **одностороннім**.

Реалізація основного принципу перевірки конкретної основної гіпотези H_0 передбачає знаходження множин Q та $\overline{Q}$. Оскільки ці множини числової осі не перетинаються, то існують точки, які їх розділяють. Такі точки називають **критичними точками** і позначаються $k_{кр}$.

Правосторонньою називається критична область, яка визначається нерівністю $K > k_{кр}$, де $k_{кр} > M(K)$:



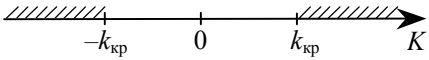
Лівосторонньою називається критична область, яка визначається нерівністю $K < k_{кр}$, де $k_{кр} < M(K)$:



Двосторонньою називається критична область, яка визначається сукупністю нерівностей

$$\left[\begin{array}{l} K < k_{кр}^{(1)}, \\ K > k_{кр}^{(2)}, \end{array}\right.$$

де $k_{кр}^{(1)} < M(K) < k_{кр}^{(2)}$. Зокрема, якщо критичні точки $k_{кр}^{(1)}$ та $k_{кр}^{(2)}$ відрізняються тільки знаком, тоді двостороння критична область визначається нерівністю $|K| > k_{кр}$:



Інтуїтивно зрозуміло, що множини Q та $\overline{Q}$ залежать від вибраного рівня значущості α, а тому слід очікувати залежності критичної точки $k_{\text{кр}}$ від α. Переконаємося в цьому, розглянувши підхід до використання основного принципу статистичної перевірки статистичної гіпотези у випадку правосторонньої критичної області. Аналіз решти випадків проводиться аналогічно. Для вибраного рівня значущості α $\overline{Q} = \{K > k_{\text{кр}}\}$. Критична точка $k_{\text{кр}}$ при умові правильності основної гіпотези H_0 задовольняє рівності

$$P(K > k_{\text{кр}}) = \alpha, \tag{3.1}$$

яка трактується таким чином: при правильності H_0 малоімовірно (з врахуванням малості α), що спостережене значення критерію K виявиться більшим від $k_{\text{кр}}$.

Для кожного практично важливого критерію складено таблиці, за допомогою яких знаходиться $k_{\text{кр}}$, що задовольняє рівність (3.1). Наступний крок — обчислення спостереженого значення $K_{сп}$ критерію за даною вибіркою. А далі в силу вступає **правило**: якщо $K_{сп} > k_{\text{кр}}$, то гіпотезу H_0 відхиляють (вважається, що вона є хибною), якщо ж $K_{сп} < k_{\text{кр}}$, тоді H_0 приймають (вважається, що H_0 — правильна).

Якщо основна гіпотеза H_0 містить твердження про параметри розподілу, вид якого відомий, тоді статистичний критерій перевірки цієї гіпотези називається **параметричним**. Якщо ж у гіпотезі H_0 мова ведеться про невідомий розподіл, тоді відповідний критерій називається **критерієм узгодженості**.

3. Параметричні статистичні гіпотези.

Розгляд статистичних гіпотез розпочнемо із параметричних гіпотез. Це зумовлено відносною простотою досліджень таких гіпотез в порівнянні із критеріями узгодженості.

Нижче будуть наведені тільки підсумки аналізу, який в повному обсязі можна знайти, зокрема, у посібнику [8].

Порівняння середньої вибіркової із гіпотетичною середньою генеральною нормальної сукупності

Нехай кількісна ознака X генеральної сукупності розподілена нормально з параметрами a та σ. При цьому середня генеральна a хоча і невідома, але є підстави припускати, що вона дорівнює гіпотетичному

значенню a_0. Для перевірки гіпотези $a = a_0$ потрібно знайти середню вибіркову $\bar{x}_в$ і з'ясувати, істотно чи ні $\bar{x}_в$ відхиляється від a_0.

Розв'язування окресленої задачі суттєво залежить від того, чи відомий параметр σ.

А. Дисперсія генеральної сукупності відома

Нехай із генеральної сукупності, кількісна ознака якої нормально розподілена, організована вибірка обсягом n і знайдена середня вибіркова $\bar{x}_в$, при цьому $D_г = \sigma^2$ — відома. Потрібно при заданому рівні значущості α перевірити основну гіпотезу про рівність середньої генеральної a гіпотетичному значенню a_0, тобто $H_0 : a = a_0$.

Оскільки середня вибіркова $\overline{X}$ (вибірка поки не фіксована) є незміщеною оцінкою середньої генеральної, тобто $M(\overline{X}) = a$, то гіпотезу H_0 можна записати ще й таким чином: $M(\overline{X}) = a_0$.

Отже, потрібно перевірити, що математичне сподівання середньої вибіркової дорівнює гіпотетичній середній генеральній.

В якості критерія перевірки гіпотези H_0 використаємо випадкову величину:

$$U = (\overline{X} - a_0)/\sigma(\overline{X}) = (\overline{X} - a_0)\sqrt{n}/\sigma.$$

Ця величина розподілена нормально, причому якщо гіпотеза H_0 правильна, то $M(U) = 0,\ \sigma(U) = 1.$

Позначимо через $U_{\text{спост}}$ спостережене значення критерія для фіксованої вибірки:

$$U_{\text{спост}} = (\bar{x}_в - a_0)\sqrt{n}/\sigma. \tag{3.2}$$

Критичну область побудуємо в залежності від виду конкуруючої гіпотези.

Правило 1. Гіпотеза $H_0 : a = a_0$ при конкуруючій гіпотезі $H_1 : a \neq a_0$ для рівня значущості α приймається, якщо $|U_{\text{спост}}| < u_{\text{кр}}$, де $u_{\text{кр}}$ — корінь рівняння

$$\Phi(u_{\text{кр}}) = (1 - \alpha)/2\ . \tag{3.3}$$

Якщо $|U_{\text{спост}}| > u_{\text{кр}}$, то гіпотеза H_0 відкидається.

Правило 2. При конкуруючій гіпотезі $H_1 : a > a_0$ критичну точку правосторонньої критичної області для рівня значущості α знаходять за рівністю (табл. 3 додатків):

$$\Phi(u_{\text{кр}}) = (1 - 2\alpha)/2. \tag{3.4}$$

Якщо $U_{\text{спост}} < u_{\text{кр}}$, то основна гіпотеза H_0 приймається. Якщо ж $U_{\text{спост}} > u_{\text{кр}}$ — H_0 відкидається.

Правило 3. При конкуруючій гіпотезі $H_1 : a < a_0$ знаходять корінь $u_{\text{кр}}$ рівняння (3.4) (табл. 3 додатків).

Якщо $U_{\text{спост}} > -u_{\text{кр}}$, то нема підстав відкидати гіпотезу H_0. Якщо ж $U_{\text{спост}} < -u_{\text{кр}}$, то H_0 відкидається.

Б. Дисперсія генеральної сукупності невідома

В цьому випадку в якості критерія перевірки основної гіпотези береться випадкова величина

$$T = (\overline{X} - a_0)\sqrt{n}/S, \tag{3.5}$$

де $\overline{X}$ та S — середня та "виправлене" середнє квадратичне відхилення для нефіксованої вибірки обсягу n.

Величина T розподілена за законом Ст'юдента із $k = n - 1$ ступенями вільності.

Симетричність розподілу Ст'юдента дозволяє здійснювати побудову критичних областей (в залежності від виду конкуруючої гіпотези) аналогічно тому, як це робилося у випадку відомого параметра σ.

Правило 1*. Для того, щоб при заданому рівні значущості α перевірити основну гіпотезу $H_0 : a = a_0$ про рівність невідомої середньої генеральної a (нормально розподіленої генеральної сукупності з невідомою дисперсією) гіпотетичному значенню a_0 при конкуруючій гіпотезі $H_1 : a \neq a_0$, потрібно обчислити спостережене значення критерія

$$T_{\text{спост}} = (\overline{x}_в - a_0)\sqrt{n}/S$$

і за таблицею критичних точок розподілу Ст'юдента (табл. 5 додатків) по заданому рівню значущості α, розміщеному у верхньому рядку таблиці і числу ступенів вільності $k = n - 1$ знайти критичну точку $t_{\text{двост.кр}}(\alpha, k)$.

Якщо $|T_{\text{спост}}| < t_{\text{двост.кр}}$, то нема підстав відкидати основну гіпотезу H_0.

Якщо ж $|T_{\text{спост}}| > t_{\text{двост.кр}}$, то гіпотеза H_0 відкидається на користь альтернативної гіпотези H_1.

Зауваження. Критична точка $t_{\text{двост.кр}} = t_\alpha$ є коренем рівняння

$$\int_0^{t_\alpha} g_k(t)dt = (1-\alpha)/2,$$

де $g_k(t)$ — густина розподілу Ст'юдента, $k = n-1$ — число ступенів вільності. Це рівняння є аналогом рівняння (3.3).

Правило 2*. При конкуруючій гіпотезі $H_1 : a > a_0$ по рівню значущості α, розміщеному в нижньому рядку табл. 5 додатків, і числу ступенів вільності $k = n-1$ знаходимо критичну точку $t_{\text{правост.кр}}(\alpha, k)$ правосторонньої критичної області.

Якщо $T_{\text{спост}} < t_{\text{правост.кр}}$, то нема підстав відкидати основну гіпотезу H_0.

Зауваження. Вибір критичної точки в табл. 5 додатків по нижньому рядку значень рівня значущості зумовлений аналогом рівності (3.4) для випадку правосторонньої критичної області, тобто $t_{\text{правост.кр}} = t_{2\alpha}$ є коренем рівняння

$$\int_0^{t_{2\alpha}} g_k(t) = (1-2\alpha)/2.$$

Правило 3*. При конкуруючій гіпотезі $H_1 : a < a_0$ спочатку знаходять "допоміжну" критичну точку $t_{\text{правост.кр}}(\alpha; k)$, а потім покладають межу лівосторонньої критичної області:

$$t_{\text{правост.кр}} = t_{\text{лівост.кр}}.$$

Якщо $T_{\text{спост}} > -t_{\text{правост.кр}}$, то гіпотеза H_0 приймається.

Якщо $T_{\text{спост}} < -t_{\text{правост.кр}}$, то гіпотезу H_0 відкидають.

Порівняння виправленої дисперсії вибіркової з гіпотетичною дисперсією генеральною нормальної сукупності

Нехай кількісна ознака генеральної сукупності розподілена за нормальним законом, причому дисперсія генеральна хоча і невідома, проте є підстави припускати, що вона дорівнює гіпотетичному значенню σ_0^2.

На підставі виправленої дисперсії вибіркової S^2, знайденої для вибірки обсягом *n*, при заданому рівні значущості α потрібно перевірити основну гіпотезу H_0, яка полягає в тому, що дисперсія генеральна дорівнює гіпотетичному значенню σ_0^2. Враховуючи незміщеність S^2 як оцінки дисперсії генеральної, основну гіпотезу можна записати таким чином:

$$H_0 : M(S^2) = \sigma_0^2. \tag{3.6}$$

Іншими словами, потрібно з'ясувати, істотно чи неістотно відрізняються виправлена вибіркова і гіпотетична генеральна дисперсії.

В якості критерія перевірки гіпотези (3.6) візьмемо випадкову величину

$$\chi^2 = (n-1)S^2/\sigma_0^2, \tag{3.7}$$

де права частина є випадковою величиною, розподіленою за законом χ^2 з $k = n-1$ ступенями вільності.

Критична область будується в залежності від виду конкуруючої гіпотези.

Правило 1. Для того, щоб при заданому рівні значущості α перевірити основну гіпотезу $H_0 : \sigma^2 = \sigma_0^2$ при конкуруючій гіпотезі $H_1 : \sigma^2 > \sigma_0^2$, потрібно обчислити спостережене значення критерія $\chi^2_{\text{спост}}$ за формулою (3.7) і за табл. 6 додатків по заданому рівню значущості α і числу ступенів вільності $k = n-1$ знайти критичну точку $\chi^2_{\text{кр}}(\alpha; k)$.

Якщо $\chi^2_{\text{спост}} < \chi^2_{\text{кр}}$, то нема підстав відкидати основну гіпотезу.

Якщо $\chi^2_{\text{спост}} > \chi^2_{\text{кр}}$, то гіпотеза H_0 відкидається на користь гіпотези H_1.

Правило 2. Для того, щоб при заданому рівні значущості α перевірити гіпотезу $H_0 : \sigma^2 = \sigma_0^2$ при конкуруючій гіпотезі $H_1 : \sigma^2 \neq \sigma_0^2$, потрібно обчислити спостережене значення критерія $\chi^2_{\text{спост}}$ за формулою (3.7) і за табл. 6 додатків знайти ліву критичну точку $\chi^2_{\text{кр}}(1-\alpha/2; k)$ і праву критичну точку $\chi^2_{\text{кр}}(\alpha/2; k)$.

Якщо $\chi^2_{\text{лівост.кр}} < \chi^2_{\text{спост}} < \chi^2_{\text{правост.кр}}$, то нема підстав відкидати гіпотезу H_0.

Якщо $\chi^2_{\text{спост}} < \chi^2_{\text{лівост.кр}}$ або $\chi^2_{\text{спост}} > \chi^2_{\text{правост.кр}}$, то основна гіпотеза H_0 відкидається.

Правило 3. При конкуруючій гіпотезі $H_1 : \sigma^2 < \sigma_0^2$ знаходиться критична точка $\chi^2_{\text{кр}}(1-\alpha/2;k)$ лівосторонньої критичної області.

Якщо $\chi^2_{\text{спост}} > \chi^2_{\text{кр}}(1-\alpha/2;k),$ то нема підстав відкидати основну гіпотезу.

Якщо $\chi^2_{\text{спост}} < \chi^2_{\text{кр}}(1-\alpha/2;k),$ то гіпотеза H_0 відкидається.

4. Критерій узгодженості Пірсона (критерій χ^2).

Одна із найважливіших задач математичної статистики — знаходження невідомого закону розподілу випадкової величини X — кількісної ознаки об'єктів генеральної сукупності. В § 1 були висвітлені питання опрацювання вибірки з метою отримання інформації про вид емпіричного розподілу та його характеристик: середньої ознаки, величини розкиду, симетричності розподілу. Потім на основі цих даних підбирається той розподіл, який найкраще апроксимує дослідний розподіл випадкової величини.

Після вибору виду розподілу необхідно знайти (хоча б наближено) параметри того закону розподілу, який характеризує досліджувану випадкову величину.

Оскільки в більшості практично важливих випадків параметри теоретичного закону розподілу є або математичним сподіванням, або дисперсією випадкової великими, або виражаються через них [7, 5.3.4 Ч1], то отримані вище висновки дають можливість знайти ці параметри за дослідними даними. Наприклад, для нормального розподілу параметри a та σ визначається рівностями $a = \overline{x}_в$, $\sigma = \sigma_в$ (або $\sigma = S$). Таким чином, можна говорити про "відомість" теоретичного розподілу досліджуваної ознаки X. При цьому лапки вказують на гіпотетичність такого знання.

Критерій узгодженості Пірсона $\left(\chi^2\right)$ грунтується на виборі певної міри розбіжності між теоретичним і емпіричним (дослідним) розподілами. При цьому задачу перевірки узгодженості можна сформулювати таким чином: на основі вибірки спостережених значень деякої випадкової величини X потрібно визначити, що емпіричний розподіл належить певному розподілу (нормальному, показниковому, біноміальному і т. д.) із *визначеними* параметрами — гіпотеза H_0 проти альтернативної гіпотези H_1 — розподіл не належить вибраному розподілу.

Нехай у відповідності із гіпотезою H_0 **відома функція розподілу імовірностей *F(x)* досліджуваної ознаки *X***. Позначимо через $S_1, S_2, ..., S_m$ множини, на які розбивається вся область можливих значень випадкової величини *X*; ці множини — або інтервали для **неперервної** випадкової величини, або групи окремих значень **дискретної** випадкової величини, які не мають спільних точок. Тоді можна обчислити імовірності того, що при випробуванні випадкова величина *X* набере значення із множини S_i, тобто

$$p_i = P(X \in S_i), \quad i = \overline{1, m}. \qquad (3.8)$$

При цьому всі $p_i > 0, \ i = \overline{1, m}$, і

$$\sum_{i=1}^{m} p_i = 1.$$

Нехай $n_1, n_2, ..., n_m$ — відповідні емпіричні групові частоти, тобто суми частот тих варіант (значень випадкової величини *X* із вибірки), що потрапляють відповідно у множини $S_1, S_2, ..., S_m$.

Якщо гіпотеза H_0 правильна, то статистичний розподіл вибірки можна розглядати як емпіричний аналог для генерального розподілу, який визначається функцією $F(x)$. Це означає, що n_i є частотою (абсолютною) появи випадкової події $(X \in S_i), \ i = \overline{1, m}$, в послідовності із *n* спостережень. Отже, в першому розподілі кожній множині S_i ставляться у відповідність відносна частота n_i / n, а в другому — імовірність p_i. Тоді за методом найменших квадратів в якості міри розходження між статистичним розподілом відносних частот вибірки і теоретичним розподілом імовірностей отримується міра розбіжності виду

$$\chi^2 = \sum_{i=1}^{m} \frac{(n_i - np_i)^2}{np_i} \qquad (3.9)$$

така, що при збільшенні обсягу вибірки розподіл величини χ^2 наближається до граничного розподілу χ^2 із $k = m - r - 1$ ступенями вільності, де *m* — число інтервалів або груп, на які розбита вся множина спостережених даних, *r* — число параметрів гіпотетичного розподілу імовірностей, що оцінюється за даними вибірки.

Зауваження. Числа

$$n_i^0 = np_i, \quad i = \overline{1, m}, \qquad (3.10)$$

називаються **теоретичними частотами** відбуття випадкових подій $(X \in S_i)$. Враховуючи рівності (3.10), із (3.9) отримаємо таку міру розбіжності між теоретичними і емпіричними частотами:

$$\chi^2 = \sum_{i=1}^{m} \frac{(n_i - n_i^0)^2}{n_i^0}. \qquad (3.9^*)$$

Випадкова величина, що визначається рівностями (3.9) або (3.9*), називається **критерієм Пірсона**.

Для перевірки гіпотези H_0 задамо рівень значущості α і за табл. 6 додатків знайдемо критичну точку $\chi^2_{\text{кр}}(\alpha;k)$, де $p=\alpha$, $k=m-r-1$. Таким чином, правостороння критична область визначається нерівністю $\chi^2 > \chi^2_{\text{кр}}(\alpha;k)$, а область прийняття нульової гіпотези — нерівністю $\chi^2 < \chi^2_{\text{кр}}(\alpha;k)$. За результатами вибірки обчислюємо $\chi^2_{\text{сп}}$. Якщо $\chi^2_{\text{сп}} < \chi^2_{\text{кр}}$, то гіпотезу H_0 приймаємо, а у випадку $\chi^2_{\text{сп}} > \chi^2_{\text{кр}}$ H_0 відкидаємо.

5. Перевірка гіпотези про нормальний розподіл генеральної сукупності.

Використання критерія узгодженості Пірсона проілюструємо для випадку перевірки гіпотези H_0: кількісна ознака X об'єктів генеральної сукупності розподілена за нормальним законом.

Нехай статистичний розподіл має такий вид:

$$\begin{array}{c|cccc} [\,x_i;x_{i+1}) & [\,x_1;x_2) & [\,x_2;x_3) & \ldots & [x_m;x_{m+1}] \\ \hline n_i & n_1 & n_2 & \ldots & n_m \end{array}, \qquad (3.11)$$

де $x_{i+1}-x_i=h$, $i=\overline{1,m}$, $\sum_{i=1}^{k} n_i = n$, число частинних інтервалів визначається наближеною рівністю $m \approx \log_2 n$. Якщо ж статистичний розподіл є дискретним, тоді потрібно перейти до інтервального (див. розв'язування задачі 1.2). Знайдемо для розподілу (3.11) $\bar{x}_в$ та $D_в$. Тоді вважаємо, що густина розподілу досліджуваної ознаки X має такий вид:

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

Враховуючи те, що можливі значення нормально розподіленої випадкової величини заповнюють всю дійсну вісь, в якості множин $S_1, S_2, \ldots S_m$ візьмемо інтервали

$$(-\infty; x_2), (x_2; x_3), ...(x_m; \infty).$$

Для знаходження імовірностей $p_i = P(x_i < X < x_{i+1}), \quad i = \overline{1, m},$ де $x_1 = -\infty, \ x_{m+1} = \infty,$ використаємо формулу

$$P(\alpha < X < \beta) = \Phi\left(\frac{\beta - a}{\sigma}\right) - \Phi\left(\frac{\alpha - a}{\sigma}\right),$$

в якій *a* та σ — параметри нормального розподілу.
Отже, за табл. 3 додатків обчислимо

$$p_i = \Phi\left(\frac{x_{i+1} - \overline{x}_в}{\sigma_в}\right) - \Phi\left(\frac{x_i - \overline{x}_в}{\sigma_в}\right), \quad i = \overline{1, m},$$

поклавши $x_1 = -\infty, \ x_{m+1} = \infty.$

В результаті за формулою (3.9) можна знайти для даної вибірки (розподілу (3.11)) $\chi^2_{\text{сп}}$. Для заданого рівня значущості α і ступеня вільності $k = m - 3$ (число параметрів нормального розподілу $r = 2$) за табл. 6 додатків знаходимо $\chi^2_{\text{кр}}(\alpha; k)$. Тоді гіпотеза H_0 приймається, якщо $\chi^2_{\text{сп}} < \chi^2_{\text{кр}}(\alpha; k),$ і відхиляється у випадку $\chi^2_{\text{сп}} > \chi^2_{\text{кр}}(\alpha; k).$

Зауваження. Вказане вище число ступенів вільності $k = m - 3$ відноситься тільки до того випадку, коли обидва параметри нормального закону розподілу знаходяться за даними вибірки, тобто коли замість точних значень *a* і σ використовуються їх емпіричні значення $\overline{x}_в$ та $\sigma_в$. Якщо значення *a* точно відоме (наприклад, у випадку знаходження відхилень від еталону), то число ступенів вільності дорівнює $k = m - 2.$ Якщо ж відомі обидва параметри, то число ступенів вільності $k = m - 1.$ На практиці така ситуація зустрічається рідко, а тому для отримання числа ступенів вільності не менше п'яти потрібно, щоб число частинних інтервалів було не меншим восьми.

Критерій Пірсона можна використовувати як для випадку неперервної ознаки *X*, так і для дискретної. Проте недоліком його є те, що випадкова величина χ^2, як правило, залежить від групування варіаційного ряду вибірки. Тому іноді доцільним є використання інших критеріїв узгодженості.

РОЗВ'ЯЗУВАННЯ ТИПОВИХ ЗАДАЧ

Задача 3.1. Торгівельна фірма розглядає питання про відкриття в новому мікрорайоні міста філії. Відомо, що фірма буде працювати прибутково, якщо щомісячний середній дохід мешканців мікрорайону перевищує 500 грн. Відомо також, що дисперсія доходів $\sigma^2 = 400$ грн.

Знайти умови прийняття рішення, з допомогою якого на підставі вибірки обсягом $n = 100$ і рівня значущості $\alpha = 0{,}05$ можна встановити, що робота філії буде прибутковою.

○ Вважаємо, що середній місячний дохід навмання вибраного мешканця є нормально розподіленою випадковою величиною.

Фірма не відкриє філію, якщо середній місячний дохід мешканців не перевищить 500 грн. Тому будемо вважати, що $H_0 : a_0 = 500$, а $H_1 : a > 500$. Оскільки дисперсія відома, то згідно із правилом 2 гіпотеза H_1 приймається, якщо $U_{\text{спост}} > u_{\text{кр}}$, де $u_{\text{кр}}$ — корінь рівняння (3.4), тобто $\Phi(u_{\text{кр}}) = (1 - 2 \cdot 0{,}05)/2 = 0{,}45$. Звідки за табл. 3 додатків $u_{\text{кр}} = 1{,}65$. Із врахуванням (3.2) і умов задачі

$$U_{\text{спост}} = (\bar{x}_в - a_0)\sqrt{n}/\sigma = (\bar{x}_в - 500)\sqrt{100}/2.$$

Тому H_1 приймається і, отже, філію відкривають, якщо середній місячний дохід 100 мешканців

$$\bar{x}_в > 500 + 2 \cdot 1{,}65 = 503{,}3. \quad \bullet$$

Задача 3.2. Для уточнення норм виробки на підприємстві проведено 31 незалежне вимірювання продуктивності праці робітників, виконуючих однотипну операцію. Середня продуктивність праці склала $\bar{x}_в = 7{,}2$ (одиниць продукції за 1 год), а середнє квадратичне відхилення $\sigma_в = 0{,}5$ (одиниць продукції за год). Потрібно перевірити гіпотезу, що при масовому випуску цієї продукції середня продуктивність a_0 складе 7,5 одиниць продукції за 1 год при конкуруючій гіпотезі $a < 7{,}5$ одиниць продукції за 1 год при рівні значущості $\alpha = 0{,}01$.

○ Будемо вважати, що продуктивність праці навмання взятого робітника є нормально розподіленою випадковою величиною. Згідно із умовою задачі дисперсія цієї величини невідома, тому в якості статистичного критерія візьмемо випадкову величину (3.5), спостережене значення якої

$$T_{\text{спост}} = (\bar{x}_в - a_0)\sqrt{n}/S = (7{,}2 - 7{,}5)\sqrt{31}\Big/\left(0{,}5\sqrt{\frac{31}{31-1}}\right) = -3{,}286.$$

При обчисленні використовуємо такий зв'язок між $\sigma_в$ та S:

$$S = \sigma_в\sqrt{n/(n-1)}.$$

Оскільки $H_1 : a < a_0$, то потрібно побудувати лівосторонню критичну область. Згідно із правилом 3* знайдемо "допоміжну" критичну точку $t_{\text{правост.кр}}(0{,}01;30)$ по рівню значущості 0,01, розміщеному в нижньому рядку табл. 5 додатків: $t_{\text{правост.кр}}(0{,}01;30) = 2{,}457$.

Тоді $t_{\text{лівост.кр}} = -2{,}457$.

Так як $T_{\text{спост}} = -3{,}286 < -2{,}457 = t_{\text{лівост.кр}}$, то основну гіпотезу відкидаємо на користь альтернативної гіпотези $H_1 : a < 7{,}5$. ●

Задача 3.3. Кількісна ознака генеральної сукупності розподілена за нормальним законом. За вибіркою обсягом $n = 16$ знайдена дисперсія вибіркова $D_в = 10{,}6$. Для рівня значущості $\alpha = 0{,}02$ перевірити основну гіпотезу $H_0 : \sigma^2 = \sigma_0^2 = 12$, якщо конкуруюча гіпотеза $H_0 : \sigma^2 \neq 12$.

○ Знайдемо спочатку виправлену дисперсію

$$S^2 = \left[\, n/(n-1)\right] D_в = (16/15)10{,}6 = 11{,}307,$$

а потім спостережене значення критерію

$$\chi^2_{\text{спост}} = (n-1)\,S^2/\sigma_0^2 = 15 \cdot 11{,}307/12 = 14{,}133.$$

Оскільки $H_1 : \sigma^2 \neq 12$, то критична область є двосторонньою. Згідно із правилом 2 за табл. 6 додатків знаходимо критичні точки: ліву — $\chi^2_{\text{кр}}(1-\alpha/2\,;k) = \chi^2_{\text{кр}}(1-0{,}02/2\,;15) = \chi^2_{\text{кр}}(0{,}99;15) = 5{,}23$ і праву — $\chi^2_{\text{кр}}(\alpha/2\,;k) = \chi^2_{\text{кр}}(0{,}01;15) = 30{,}58$. Так як спостережене значення критерія належить області прийняття основної гіпотези ($5{,}23 < 14{,}133 < 30{,}58$), то нема підстав її відкидати. Іншими словами, виправлена дисперсія вибіркова (11,307) неістотно відрізняється від гіпотетичної дисперсії генеральної (12). ●

Задача 3.4. За даними табл. 1.2 про контрольні виміри товщини (в міліметрах) 200 вкладишів шатунних підшипників з допомогою критерію Пірсона перевірити гіпотезу H_0 про нормальний розподіл кількісної ознаки генеральної сукупності, якщо рівень значущості дорівнює 0,05.

○ Знайдемо число m частинних інтервалів однакової довжини, які отримуються при переході від дискретного ряду варіант з табл. 1.2 до інтервального розподілу частот, використовуючи формулу $m \approx \log_2 n$. Оскільки обсяг вибірки $n = 200$, то $m \approx \log_2 200 \approx 8$. Шуканий інтервальний розподіл наведений в табл. 1.3.

Для знаходження $\bar{x}_в$ та $\sigma_в$ від цього інтервального розподілу потрібно перейти до дискретного, "нові" варіанти якого є серединами частинних інтервалів. Для отриманого розподілу в задачі 1.2 знайдено: $\bar{x}_в = 1{,}7417$; $\sigma_в = 0{,}0113269$.

Результати допоміжних розрахунків в процесі знаходження теоретичних частот $n_i^0 = p_i n$ розташуємо в табл. 3.1, де $x_1 = -\infty$, $x_9 = \infty$. Відмітимо, що згідно із властивостями функції Лапласа $\Phi(-x) = -\Phi(x)$, $\Phi(-\infty) = -0{,}5$, $\Phi(\infty) = 0{,}5$. Крім цього, при знаходженні значень $\Phi(x)$ за допомогою табл. 3 додатків для аргументів, що мають третій знак після коми, здійснимо лінійну інтерполяцію. Наприклад, при знаходженні $\Phi(1{,}916)$ використаємо табличні значення: $\Phi(1{,}91) = 0{,}47193$, $\Phi(1{,}92) = 0{,}47257$. Тоді

$$\Phi(1{,}916) = \Phi(1{,}91) + [\Phi(1{,}92) - \Phi(1{,}91)] \cdot 0{,}6 =$$
$$= 0{,}47193 + (0{,}47257 - 0{,}47193) \cdot 0{,}6 = 0{,}472314 \approx 0{,}47231.$$

Для знаходження $\chi^2_{\text{сп}}$ складемо розрахункову табл. 3.2, з якої отримаємо

$$\chi^2_{\text{сп}} = 2{,}5393\,.$$

За таблицею критичних точок розподілу χ^2 (табл. 6 додатків) для рівня значущості $\alpha = 0{,}05$ і числа ступенів вільності $k = 8 - 3 = 5$ знайдемо $\chi^2_{\text{кр}}(0{,}05;5) = 11{,}07$.

Оскільки $\chi^2_{\text{сп}} < \chi^2_{\text{кр}}$, то нема підстав відкидати основну гіпотезу. Отже, дані спостережень узгоджуються із гіпотезою про нормальний розподіл кількісної ознаки генеральної сукупності. ●

Таблиця 3.1

Інтервал $(x_i; x_{i+1})$	n_i	$x_{i+1} - \bar{x}_в$	$x_i - \bar{x}_в$	$\frac{x_{i+1} - \bar{x}_в}{\sigma_в}$	$\frac{x_i - \bar{x}_в}{\sigma_в}$	$\Phi\left(\frac{x_{i+1} - \bar{x}_в}{\sigma_в}\right)$	$\Phi\left(\frac{x_i - \bar{x}_в}{\sigma_в}\right)$	$p_i = \Phi\left(\frac{x_{i+1} - \bar{x}_в}{\sigma_в}\right) - \Phi\left(\frac{x_i - \bar{x}_в}{\sigma_в}\right)$	$n_i^0 = p_i n = 200 \cdot p_i$
(–∞; 1,720)	3	–0,0217	–∞	–1,916	–∞	–0,47231	–0,5	0,0277	5,54
[1,720; 1,727)	13	–0,0147	–0,0217	–1,298	–1,916	–0,40285	–0,47231	0,06946	13,89
[1,727; 1,734)	32	–0,0077	–0,0147	–0,680	–1,298	–0,25175	–0,40285	0,1511	30,22
[1,734; 1,741)	49	–0,0007	–0,0077	–0,062	–0,680	–0,02472	–0,25175	0,2270	45,41
[1,741; 1,748)	42	0,0063	–0,0007	0,556	–0,062	0,21089	–0,02472	0,2356	47,12
[1,748; 1,755)	35	0,0133	0,0063	1,174	0,556	0,3798	0,21089	0,1689	33,78
[1,755; 1,762)	19	0,0203	0,0133	1,792	1,174	0,46336	0,3798	0,08356	16,71
(1,762; ∞)	7	∞	0,0203	∞	1,792	0,5	0,46336	0,03664	7,33
Усього	200	—	—	—	—	—	—	—	200

Таблиця 3.2

i	n_i	n_i^0	$n_i - n_i^0$	$(n_i - n_i^0)^2$	$(n_i - n_i^0)^2 / n_i^0$	n_i^2	n_i^2 / n_i^0
1	3	5,54	–2,54	6,4516	1,1646	9	1,6246
2	13	13,89	–0,89	0,7921	0,0570	169	12,1670
3	32	30,22	1,78	3,1684	0,1048	1024	33,8848
4	49	45,41	3,59	12,8881	0,2838	2401	52,8738
5	42	47,12	–5,12	26,2144	0,5563	1764	37,4363
6	35	33,78	1,22	1,4884	0,0441	1225	36,2641
7	19	16,71	2,29	5,2441	0,3138	361	21,6038
8	7	7,33	–0,33	0,1089	0,0149	49	6,6849
$\sum$	200	200			$\chi^2_{\text{сп}} = 2,5393$		202,5393

§ 4. КОРЕЛЯЦІЙНИЙ ТА РЕГРЕСІЙНИЙ АНАЛІЗ

1. *Поняття стохастичної та статистичної залежності, кореляції і регресії. Основні задачі кореляційного і регресійного аналізу.*
2. *Лінійні емпіричні рівняння парної кореляції.*
3. *Вибірковий коефіцієнт лінійної кореляції та його властивості.*
4. *Нелінійна парна кореляція.*

1. Поняття стохастичної і статистичної залежності, кореляції і регресії. Основні задачі кореляційного і регресійного аналізу

Дві випадкові величини можуть бути або незалежними, або пов'язані функціональною залежністю, або залежністю, яка називається стохастичною. Функціональна залежність, розглянута в [7, § 8], досить рідко зустрічається в економічних дослідженнях. Дуже часто доводиться вивчати такі випадкові величини, для яких зміна можливих значень (при проведенні випробувань) приводить до зміни закону (умовного) розподілу іншої (див. [7, § 7]). Такий зв'язок між випадковими величинами називається **стохастичним**; він виникає тоді, коли на обидві величини впливають випадкові фактори, серед яких є спільні.

Найбільш важливим випадком стохастичного зв'язку є так званий **кореляційний** зв'язок, який встановлює залежність між значеннями випадкової величини X і умовним математичним сподіванням $M(Y \mid X = x)$ [*)] випадкової величини Y:

$$M(Y \mid X = x) = g(x), \tag{4.1}$$

або між значенням випадкової величини Y і умовним математичним сподіванням $M(X \mid Y = y)$ [*)] випадкової величини X:

$$M(X \mid Y = y) = q(y). \tag{4.2}$$

Функції $g(x)$ і $q(y)$ називаються **функціями регресії** Y на X та X на Y відповідно, а їх графіки — лініями регресії Y на X та X на Y, рівняння (4.1) та (4.2) називаються **рівняннями регресії *Y* на *X* та *X* на *Y*** відповідно.

Використовуючи інформацію §§ 7, 8 [7] можна отримати важливу для наступного викладу основну властивість регресії величини Y на величину X: якщо $g(x)$ є функцією регресії величини Y на X, то мате-

[*)] Умовні математичні сподівання дискретних і неперервних випадкових величин визначаються в [7, п. 7.5].

матичне сподівання квадрата відхилення величини Y від функції $g(x)$ менше, ніж від будь-якої функції $h(X) \neq g(X)$, тобто

$$M\left[\, Y - g(X)\right]^{\,2} < M\left[\, Y - h(X)\right]^{\,2}. \qquad (4.3)$$

Аналогічною властивістю володіє і регресія величини X на величину Y.

Прикладами кореляційного зв'язку є стохастична взаємозалежність між: 1) окремими параметрами тіла людини або тварини; 2) обсягом виробництва підприємства та коефіцієнтом використання основних засобів.

На практиці сумісний закон розподілу випадкових величин X та Y (двовимірної випадкової величини $(X; Y)$ [7, §7]) невідомий. В розпорядженні дослідника є двовимірна вибірка

$$(x^{(1)}; y^{(1)}), (x^{(2)}; y^{(2)}), ..., (x^{(n)}; y^{(n)}), \qquad (4.4)$$

згрупувавши (для великих n) дані якої, можна отримати двовимірний статистичний розподіл, наведений в табл. 4.1.

Таблиця 4.1

X	Y						
	y_1	y_2	…	y_j	…	y_m	n_{x_i}
x_1	n_{11}	n_{12}	…	n_{1j}	…	n_{1m}	n_{x_1}
x_2	n_{21}	n_{22}	…	n_{2j}	…	n_{2m}	n_{x_2}
…	…	…	…	…	…	…	…
x_i	n_{i1}	n_{i2}	…	n_{ij}	…	n_{im}	n_{x_i}
…	…	…	…	…	…	…	…
x_k	n_{k1}	n_{k2}	…	n_{kj}	…	n_{km}	n_{x_k}
n_{y_j}	n_{y_1}	n_{y_2}	…	n_{y_j}	…	n_{y_m}	n

Тут n_{ij} — частота спільної появи ознак x_i, y_j (пари $(x_i; y_j)$); $\sum_{i=1}^{k} \sum_{j=1}^{m} n_{ij} = n$ — обсяг вибірки;

$$n_{x_i} = \sum_{j=1}^{m} n_{ij}, \ \ (i = \overline{1;k}); \quad n_{y_j} = \sum_{i=1}^{k} n_{ij}, \ \ (j = \overline{1;k}).$$

Кожному значенню X відповідає ряд значень Y, тобто зміна значень X приводить до зміни умовного статистичного розподілу $Y \mid x$. Аналогічно зміна значень Y веде до зміни умовного статистичного розподілу $X \mid y$.

Наприклад, при $X = x_1$ та $X = x_2$ відповідні умовні статистичні розподіли величини Y мають такі види:

$Y \mid x_1$	y_1	y_2	...	y_m
n_i	n_{11}	n_{12}	...	n_{1m}

;

$Y \mid x_2$	y_1	y_2	...	y_m
n_i	n_{21}	n_{22}	...	n_{2m}

.

Статистичною називається така залежність між величинами X та Y, для якої зміна спостережених значень однієї із величин зумовлює зміну умовного статистичного розподілу іншої.

Умовною середньою $\bar{y}_x$ називається середнє арифметичне спостережених значень Y, які відповідають значенню $X = x$. Наприклад, згідно із табл. 4.1

$$\bar{y}_{x_2} = \frac{y_1 n_{21} + y_2 n_{22} + ... + y_j n_{2j} + ... + y_m n_{2m}}{n_{x_2}}.$$

Умовною середньою $\bar{x}_y$ називається середнє арифметичне спостережених значень X, які відповідають значенню $Y = y$. Наприклад, у відповідності із табл. 4.1

$$\bar{x}_{y_1} = \frac{x_1 n_{11} + x_2 n_{21} + ... + x_k n_{k1}}{n_{y_1}}.$$

Можна довести, що умовні середні є незміщеними і спроможними точковими статистичними оцінками відповідних умовних математичних сподівань:

$$\bar{y}_x \approx M(Y \mid X = x); \quad \bar{x}_y \approx M(X \mid Y = y). \tag{4.5}$$

Вкажемо підхід до знаходження статистичних наближень (емпіричних функцій) $\hat{g}(x) = \hat{g}(x; a_0, a_1, ..., a_m)$ та $\hat{q}(y) = \hat{q}(y; b_0, b_1, ..., b_m)$ невідомих функцій регресій $g(x)$ та $q(y)$ відповідно, а отже, з врахуванням співвідношень (4.5) і емпіричних рівнянь Y на X

$$\bar{y}_x = \hat{g}(x, a_0, a_1, ..., a_m) \tag{4.6}$$

та X на Y

$$\bar{x}_y = \hat{q}(y, b_0, b_1, ..., b_m) \tag{4.7}$$

де $a_0, a_1, ..., a_m$ та $b_0, b_1, ..., b_m$ — невідомі параметри.

На першому кроці в прямокутній системі координат xOy будується послідовність пар чисел (4.4). Отримана сукупність точок називається **кореляційним полем** або **діаграмою розсіювання**. Конфігурація кореляційного поля дозволяє висунути гіпотезу про вид функції $\hat{g}(x)$. Наприклад, у випадку кореляційного поля, зображеного на рис. 4.1, $\hat{g}(x)$ доцільно шукати у вигляді лінійної функції $a_1 x + a_0$, а у випадку

рис. 4.2 у вигляді параболічної функції другого порядку $a_2x^2 + a_1x + a_0$.

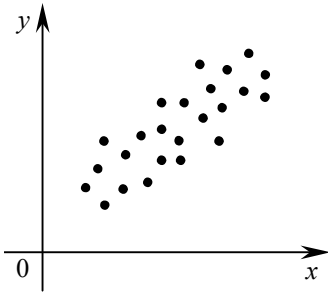


Рис. 4.1.

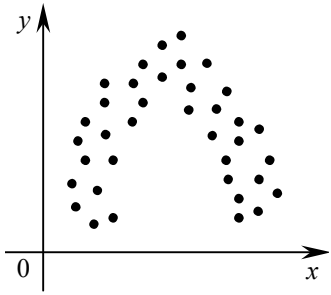


Рис. 4.2.

На другому кроці потрібно знайти невідомі параметри. Ця задача розв'язується з допомогою так званого **методу найменших квадратів** (МНК), суть якого полягає в наступному. Оскільки невідома функція регресії Y на X $g(X)$ згідно із (4.3) мінімізує величину $M\left[Y - h(X)\right]^2$, а оцінкою $M(Y \mid x)$ є статистична середня $\bar{y}_x$, що відповідає певним спостереженим значенням $X = x$, то емпірична лінія регресії (4.6) повинна задовольняти рівності

$$F(a_0, a_1, ..., a_m) \equiv \sum_{i=1}^{n}\left[\ y^{(i)} - \hat{g}(x^{(i)}; a_0, a_1, ..., a_m)\right]^2 \Rightarrow \min. \qquad (4.8)$$

Права частина цієї рівності є сума квадратів віддалей вздовж осі $0y$ від точок $(x^{(i)}; y^{(i)})$ послідовності (4.4) до відповідних точок лінії регресії із тією ж абсцисою.

Для випадку згрупованих даних із табл. 4.1 рівність (4.8) набирає такого виду

$$F(a_0, a_1, ..., a_m) \equiv \sum_{i=1}^{k}\ \sum_{j=1}^{m}\left[\ y_j - \hat{g}(x_i; a_0, a_1, ..., a_m)\right]^2 n_{ij} \Rightarrow \min.$$

(4.8^*)

Таким чином, емпірична функція $\hat{g}(x; a_0, a_1, ..., a_m)$ повинна усереднити (згладити) спостережені дані $(x^{(i)}; y^{(i)})$, $i = \overline{1, n}$, з послідовності (4.4) або (x_i, y_j), $i = \overline{1, k}$, $j = \overline{1, m}$, із табл. 4.1. При цьому невідомі параметри $a_0, a_1, ..., a_m$ повинні мінімізувати функцію $F(a_0, a_1, ..., a_m)$; необхідною умовою цього є виконання системи рівнянь

$$\begin{cases} \dfrac{\partial F(a_0, a_1, ..., a_m)}{\partial a_i} = 0, \quad i = \overline{0, m}, \end{cases} \tag{4.9}$$

яка називається **системою нормальних рівнянь**.

Метод найменших квадратів дозволяє **при встановленому виді емпіричної функції регресії** $\tilde{g}(x; a_0, a_1, ..., a_m)$ так знайти невідомі параметри $a_0, a_1, ..., a_m$, що вона буде найкращою оцінкою функції регресії $g(x)$ в тому розумінні, що сума квадратів відхилень спостережених значень випадкової величини Y від відповідних ординат емпіричної функції буде найменшою.

Аналогічно знаходиться емпірична функція $\tilde{q}(y; b_0, b_1, ..., b_m)$ регресії X на Y.

Зауваження. Враховуючи випадковий характер організації вибірки, можна зробити висновок, що для нефіксованої вибірки знайдені параметри є випадковими величинами.

Знаходження емпіричних рівнянь регресії — це тільки перший крок дослідження статистичних зв'язків між випадковими величинами. Наступним є встановлення сили або тісноти цих зв'язків. Відомо [7, п. 7.6], що мірою лінійного зв'язку (стохастичного) між двома випадковими величинами X та Y є коефіцієнт кореляції

$$r = r_{XY} = \frac{K_{XY}}{\sigma_X \sigma_Y} = \frac{M(XY) - M(X)M(Y)}{\sigma_X \sigma_Y}, \tag{4.10}$$

де K_{XY} — кореляційний момент або коефіцієнт коваріації (сумісної варіації), який ще часто позначають $\mathrm{cov}(X, Y)$, $\sigma_X = \sqrt{D(X)}$, $\sigma_Y = \sqrt{D(Y)}$. Зокрема, якщо величини X та Y незалежні, то коефіцієнт кореляції $r = 0$; якщо $r = \pm 1$, то X та Y пов'язані **лінійною** функціональною залежністю. Звідси випливає, що коефіцієнт кореляції r вимірює силу (тісноту) **лінійного** зв'язку між X та Y.

Використовуючи метод моментів, тобто замінивши числові характеристики їх статистичними оцінками, можна отримати точкову статистичну оцінку коефіцієнта кореляції — вибірковий (емпіричний) коефіцієнт кореляції.

$$r_в = r_{xy} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sigma_x \sigma_y}, \tag{4.11}$$

де для незгрупованих даних (4.4)

$$\overline{xy} = \sum_{i=1}^{n} x^{(i)} y^{(i)} / n, \tag{4.12}$$

а для згрупованих даних із табл. 4.1

$$\overline{xy} = \sum_{i=1}^{k} \sum_{j=1}^{m} n_{ij}\, x_i y_j / n = \sum n_{xy}\, xy / n, \tag{4.13}$$

σ_x та σ_y — середні квадратичні відхилення вибіркові для ознак X та Y відповідно.

Зауваження. Оскільки кореляційний момент рівносильно (4.10) визначається ще й таким чином: $K_{XY} = M\{ [\ X - M(X)]\,[Y - M(Y)]\ \}$, то формулу (4.11) можна подати і в такому вигляді

$$r_в = \frac{\sum_{i=1}^{n} (x^{(i)} - \overline{x})(y^{(i)} - \overline{y})}{n \sigma_x \sigma_y}. \tag{4.11*}$$

Оскільки $r_в$ для нефіксованої вибірки є випадковою величиною, то знайдене значення $r_в$ для конкретної вибірки може суттєво відрізнятись від r. В зв'язку із цим необхідно побудувати довірчі інтервали для оцінки r, а також перевірити статистичну гіпотезу про некорельованість X та Y ($r = o$).

У випадку нелінійності хоча б однієї із функцій регресії $g(x)$ та $q(y)$ коефіцієнт кореляції r вже не дає інформації про силу зв'язку між X та Y. Натомість ступінь концентрації розподілу поблизу лінії регресії показує **кореляційне відношення** Y на X:

$$\eta_{Y/X} = 1 - \frac{\sigma^2_{Y/X}}{\sigma^2_Y}, \tag{4.14}$$

де

$$\sigma^2_{Y/X} = D(Y \mid X = x) = M\left\{ [\ Y - g(x)\ ]^{\,2} \mid X = x \right\}. \tag{4.15}$$

Із означення випливає, що кореляційне відношення змінюється в межах від 0 до 1 включно; воно дорівнює нулю тоді і тільки тоді, коли $\sigma^2_{Y/X} = 0$, тобто весь розподіл зосереджений на лінії регресії (має місце функціональна залежність). Це відношення дорівнює нулю тоді і тільки тоді, коли лінія регресій Y на X являє собою горизонтальну пряму, що проходить через центр розподілу, тобто якщо Y та X некорельовані.

Аналогічно вводиться означення кореляційного відношення X на Y $\eta_{X/Y}$.

Можна довести, що у всіх випадках виконуються нерівності

$$r^2 \le \eta^2_{X/Y}, \quad r^2 \le \eta^2_{Y/X}. \tag{4.16}$$

За статистичними даними двовимірної вибірки можна знайти вибіркові кореляційні відношення $\hat{\eta}_{Y/X}$ та $\hat{\eta}_{X/Y}$, які є точковими статистичними оцінками $\eta_{Y/X}$ та $\eta_{X/Y}$ відповідно.

Сукупність методів оцінки кореляційних характеристик і перевірка статистичних гіпотез про них за даними вибірки називається **кореляційний аналізом**. В кореляційному аналізі використовуються такі основні методи:

1) побудова кореляційного поля і складання кореляційної таблиці;

2) знаходження вибіркового (емпіричного) коефіцієнта кореляції або кореляційного відношення;

3) перевірка статистичних гіпотез про значущість (істотність) зв'язку.

В практично важливих задачах одна із випадкових величин є "результуючою", а інша — від якої вона залежить (факторна величина). Наприклад, залежність урожайності сільськогосподарських культур Y від маси внесених добрив X. Це зразок односторонньої залежності між випадковими величинами, яка досліджується в регресійному аналізі.

Більш того, факторна величин може бути **детермінованою** (невипадковою), тобто значення якої не тільки відоме, але й ним можна керувати. При дослідженні лінійної залежності результуючої величини Y від факторної величини X загальний вираз емпіричного рівняння регресії має такий вид

$$\bar{y} = a_0 + a_1 x. \tag{4.17}$$

В загальному випадку емпіричне рівняння регресії визначається таким чином

$$\bar{y} = \bar{g}(x; a_0, a_1, ..., a_m), \tag{4.18}$$

де $a_0, a_1, ..., a_m$ — невідомі параметри.

Тепер, коли визначені об'єкти дослідження цього параграфа (рівняння регресії (4.6), (4.7), (4.17), (4.18)), можемо сформулювати основні задачі регресійного аналізу.

Регресійний аналіз — це дослідження односторонніх статистичних залежностей між випадковими величинами. При цьому деякі із факторних величин можуть бути невипадковими величинами.

Задачі регресійного аналізу:

1) визначення форми залежностей;

2) знаходження функції регресії;

3) побудова точкових та інтервальних оцінок параметрів функції регресії;

4) знаходження точкових та інтервальних оцінок умовних математичних сподівань, необхідних для визначення меж, в яких із заданою надійністю будуть міститися середні значення досліджуваної величини, якщо інші пов'язані з нею величини набирають певних значень;

5) перевірка узгодженості знайденої емпіричної функції регресії спостереженим даним.

Основна мета регресійного аналізу — **теоретично обгрунтований і статистично надійний точковий та інтервальний прогноз** значень залежної величини або умовного математичного сподівання цієї величини.

Зауваження. Якщо рівняння регресії описує об'єкт дослідження із економічної сфери і воно обгрунтоване в теоретично-економічному відношенні, то його називають **економетричним рівнянням**.

2. Лінійні емпіричні рівняння парної кореляції. Вибірковий коефіцієнт лінійної кореляції та його властивості.

Найбільш простим випадком є той, коли **обидві функції регресії** $g(x)$ і $q(y)$ в рівняннях регресії (4.1) та (4.2) є лінійними, тобто обидві лінії регресії є прямими лініями; вони називаються **прямими регресії**. В цьому випадку будемо говорити про лінійну кореляцію між випадковими величинами X та Y.

Можна довести, що коли сумісний розподіл імовірностей величин X і Y є **нормальним розподілом** на площині (див. [7, п. 7.8]), тоді **кореляційний зв'язок завжди є лінійним**. В зв'язку із цим слід очікувати лінійного кореляційного зв'язку між статистично залежними випадковими величинами X та Y, якщо кожну із них можна розглядати як суму великого числа незалежних або майже незалежних випадкових доданків.

Нехай конфігурація кореляційного поля, отриманого внаслідок зображення у вигляді точки в прямокутній декартовій системі координат $x0y$ кожної із пар чисел вибіркової послідовності (4.4), дозволяє висунути припущення про лінійну кореляційну залежність між X та Y. Тобто рівняння регресії (4.1) та (4.2) в цьому випадку набирають відповідно такого виду

$$M(Y \mid X = x) = \alpha_1 x + \alpha_0, \tag{4.19}$$

$$M(X \mid Y = y) = \beta_1 y + \beta_0, \tag{4.20}$$

де α_i, β_i $(i = 1,0)$ — сталі.

Тоді емпіричні рівняння регресії Y на X та X на Y будемо шукати у вигляді

$$\bar{y}_x = a_1 x + a_0, \tag{4.21}$$

$$\bar{x}_y = b_1 y + b_0, \tag{4.22}$$

де a_0, a_1 та b_0, b_1 — невідомі параметри, які є **точковими статистичними оцінками** відповідних **чисел** рівнянь (4.19) і (4.20).

Використавши метод найменших квадратів як для незгрупованих даних(4.4), так і для згрупованих даних з табл.4.1, можна отримати **систему нормальних рівнянь**

$$\begin{cases} \overline{x^2} a_1 + \bar{x} a_0 = \overline{xy}, \\ \bar{x} a_1 + a_0 = \bar{y}, \end{cases} \tag{4.23}$$

для знаходження параметрів рівняння регресії (Y на X) (4.21).

Система (4.23) має єдиний розв'язок

$$a_1 = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - (\bar{x})^2}, \quad a_0 = \bar{y} - \bar{x} a_1. \tag{4.24}$$

Коефіцієнт a_1 рівняння (4.21) емпіричної прямої регресії Y на X називається **вибірковим (емпіричним або статистичним) коефіцієнтом регресії *Y* на *X*** і позначається $a_{Y/X}$. Оскільки $\overline{x^2} - (\bar{x})^2 = \sigma_x^2$, а

$$\overline{xy} - \bar{x} \cdot \bar{y} = K^*_{XY} (= \overline{\mathrm{cov}}(X, Y)) \tag{4.25}$$

— **вибірковий (емпіричний або статистичний) кореляційний момент** або **вибіркова коваріація**, то першу формулу (4.24) можна записати в такому виді

$$a_{Y/X} = a_1 = \frac{K^*_{XY}}{\sigma_X^2} \left(\equiv \frac{\overline{\mathrm{cov}}(X, Y)}{\sigma_X^2} \right), \tag{4.26}$$

а шукане рівняння (4.21) із врахуванням другої рівності (4.25) набуде вигляду

$$\bar{y}_x - \bar{y} = \frac{K^*_{XY}}{\sigma_X^2} (x - \bar{x}). \tag{4.27}$$

Це рівняння показує, що емпірична пряма регресії Y на X проходить через точку з координатами $(\bar{x}, \bar{y})$, яка називається **середньою точкою кореляційного поля**.

Аналогічно можна отримати емпіричне рівняння прямої X на Y, якщо мінімізувати сумарні квадрати відхилень точок $(\bar{x}_i, y_i)$, $i = \overline{1, m}$, від шуканої прямої, тобто прямої

$$\bar{x}_y - \bar{x} = b_{X/Y}(y - \bar{y}), \tag{4.28}$$

де вибірковий (емпіричний або статистичний) коефіцієнт регресії X на Y визначається за формулою

$$b_{X/Y} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sigma_y^2} = \frac{K^*_{XY}}{\sigma_y^2} \left(\equiv \frac{\overline{\mathrm{cov}}(X,Y)}{\sigma_y^2} \right). \tag{4.29}$$

У формулах вибіркових коефіцієнтів регресії (4.26) і (4.29) чисельники співпадають, а знаменники завжди додатні, оскільки є дисперсіями вибірковими випадкових величин X та Y відповідно. Тому $a_{Y/X}$ і $b_{X/Y}$ мають однакові знаки. Відмітимо також, що вибірковий коефіцієнт регресії $a_{Y/X}(b_{X/Y})$ — це міра, яка на основі вибіркових даних в середньому вказує на вплив зміни змінної X (або Y) на змінну Y (або X).

Рівністю (4.11) **формально** був визначений вибірковий (емпіричний) коефіцієнт кореляції $r_в$ як точкова статистична оцінка коефіцієнта кореляції $r = r_{XY}$ — міри лінійного кореляційного зв'язку між випадковими величинами X та Y. Проте вид емпіричних коефіцієнтів регресії Y на X та X на Y вказує на природний зв'язок $r_в$ із $a_{Y/X}$ і $b_{X/Y}$ (за аналогією із зв'язком в теорії кореляції як розділу теорії імовірностей). А тому виникає необхідність детально вивчити властивості $r_в$ як **оцінки** сили лінійного кореляційного зв'язку між величинами X та Y. При цьому слід очікувати властивостей, аналогічних із властивостями r_{XY}.

3. **Вибірковим (емпіричним або статистичним) коефіцієнтом кореляції $r_в = r_{XY}$ випадкових величини X та Y, між якими припускається лінійний кореляційний зв'язок, називається відношення емпіричного кореляційного моменту (коефіцієнта коваріації) $K^*(X,Y)(= \overline{\mathrm{cov}}(X,Y))$ до добутку середніх квадратичних відхилень вибіркових σ_x та σ_y**:

$$r_в = r_{xy} = \frac{K^*_{XY}}{\sigma_x \sigma_y} \left(= \frac{\overline{\mathrm{cov}}(X,Y)}{\sigma_x \sigma_y} \right) = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sigma_x \sigma_y}, \tag{4.30}$$

де $\overline{xy}$ для незгрупованих даних (4.4) вибірки обчислюється за формулою (4.12), а для згрупованих даних із табл. 4.1 — за (4.13).

Зауваження. Для випадку, коли x_i та y_i $(i=\overline{1,k})$ є великими числами, а обсяг вибірки $n \geq 50$, то зручніше користуватися такою формулою

$$r_в = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]\left[\sum_{i=1}^{n}(y_i - \bar{y})^2\right]}}. \qquad (4.30^*)$$

Використовуючи формули (4.26) і (4.29), отримаємо вираз через емпіричні коефіцієнти регресій для вибіркового коефіцієнта кореляції:

$$r_в = \pm\sqrt{\frac{(K^*_{XY})^2}{\sigma_x^2\sigma_y^2}} = \pm\sqrt{\frac{K^*_{XY}}{\sigma_x^2}\cdot\frac{K^*_{XY}}{\sigma_y^2}} = \pm\sqrt{a_{Y/X}b_{X/Y}},$$

де знак перед коренем визначається знаком емпіричних коефіцієнтів регресій.

З другого боку, і емпіричні коефіцієнти регресій можна виразити через вибірковий коефіцієнт кореляції:

$$a_{Y/X} = \frac{K^*_{XY}}{\sigma_x^2} = \frac{K^*_{XY}}{\sigma_x\sigma_y}\cdot\frac{\sigma_y}{\sigma_x} = r_в\frac{\sigma_y}{\sigma_x},$$

$$b_{X/Y} = \frac{K^*_{XY}}{\sigma_y^2} = \frac{K^*_{XY}}{\sigma_x\sigma_y}\cdot\frac{\sigma_x}{\sigma_y} = r_в\frac{\sigma_x}{\sigma_y}.$$

Тоді емпіричне рівняння прямих регресій із врахуванням цих формул можна записати в такому вигляді:

$$\bar{y}_x - \bar{y} = r_в\frac{\sigma_y}{\sigma_x}(x - \bar{x}), \qquad (4.31)$$

$$\bar{x}_y - \bar{x} = r_в\frac{\sigma_x}{\sigma_y}(y - \bar{y}). \qquad (4.32)$$

Наведемо еквівалентні (4.30) вирази вибіркового коефіцієнта кореляції.

Нехай $\{(x_i, y_i), i=\overline{1,n}\}$ — незгруповані дані двовимірної вибірки обсягом *n*. Величину

$$D_{yx} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{n}\sum_{i=1}^{n}\left[y_i - \bar{y} - r_в\frac{\sigma_y}{\sigma_x}(x_i - \bar{x})\right]^2 \qquad (4.33)$$

природно називати дисперсією спостережених значень y_i навколо емпіричної прямої регресії (4.31) *Y* на *X*, врахувавши при цьому, що гео-

метрично різниця $y_i - \hat{y}_i$ означає відхилення по ординаті точки (x_i, y_i) від точки $(x_i, \hat{y}_i)$ прямої (4.31).

Можна довести, що

$$D_{yx} = \sigma_y^2(1 - r_в^2), \qquad r_в = \pm\sqrt{1 - \frac{D_{yx}}{\sigma_y^2}}, \tag{4.30**}$$

де знак перед коренем вибирається у відповідності із знаком коефіцієнта регресії.

Аналогічно ввівши дисперсію спостережених значень x_i навколо емпіричної прямої регресії (4.32) X на Y за формулою

$$D_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x}_i)^2 = \frac{1}{n}\sum_{i=1}^{n}\left[x_i - \bar{x} - r_в\frac{\sigma_x}{\sigma_y}(y_i - \bar{y})\right]^2,$$

можна отримати рівність

$$r_в = \pm\sqrt{1 - \frac{D_{xy}}{\sigma_x^2}}. \tag{4.30***}$$

Властивості вибіркового коефіцієнта кореляції

Властивість 1. Величина $r_в$ є безрозмірною, тобто вона не залежить від вибору одиниць виміру випадкових величин X та Y.

Властивість 2. Вибірковий коефіцієнт кореляції $r_в$ не перевищує за абсолютною величиною одиницю, тобто $|r_в| \le 1$.

Властивість 3. Вибірковий коефіцієнт кореляції $r_в = \pm 1$ тоді і тільки тоді, коли між випадковими величинами X та Y існує лінійний функціональний зв'язок.

Властивість 4. Якщо між випадковими величинами X і Y відсутній хоча б один із кореляційних зв'язків, то вибірковий коефіцієнт кореляції $r_в$ дорівнює нулю.

Властивість 5. Рівність $r_в = \pm 1$ є необхідною і достатньою умовою співпадань регресій Y на X і X на Y.

Із розглянутих властивостей $r_в$ можна зробити висновок про те, що вибірковий коефіцієнт кореляції є мірою тісноти (сили) лінійного кореляційного зв'язку між випадковими величинами X і Y. Справді, якщо $|r_в| = 1$, то між X і Y існує лінійний функціональний зв'язок, а якщо $r_в = 0$, то лінійний кореляційний зв'язок відсутній. А із формул

(4.30**) і (4.30***) випливає, що у випадку збільшення $|r_в|$ до одиниці сила кореляційного зв'язку зростає, оскільки сума квадратів відхилень спостережених значень від прямих регресій прямує до нуля. Якщо ж $|r_в| \to 0,$ то сила кореляційного зв'язку зменшується, бо сума квадратів відхилень зростає.

4. Нелінійна парна кореляція. Вибіркове кореляційне відношення та його властивості.

Розглянемо тепер випадок, коли хоча б одна із двох послідовностей точок

$$(x_1, \overline{y}_{x_1}), (x_2, \overline{y}_{x_2}), ..., (x_k, \overline{y}_{x_k}), \tag{4.34}$$

$$(y_1, \overline{x}_{y_1}), (y_2, \overline{x}_{y_2}), ..., (y_m, \overline{x}_{y_m}) \tag{4.35}$$

дає підстави зробити висновок про існування нелінійного кореляційного зв'язку між випадковими величинами X та Y. На основі цих статистичних даних потрібно оцінити параметри нелінійної регресії та силу кореляційної залежності.

Нехай, наприклад, конфігурація точок (4.34) дозволяє зробити припущення про наявність параболічної кореляції другого порядку. В цьому випадку вибіркове рівняння регресії Y на X слід шукати в такому вигляді

$$\overline{y}_x = a_2 x^2 + a_1 x + a_0, \tag{4.36}$$

де $a_i (i = 0,1,2)$ — невідомі параметри.

Користуючись методом найменших квадратів, можна отримати систему нормальних рівнянь:

$$\begin{cases} \overline{x^4} a_2 + \overline{x^3} a_1 + \overline{x^2} a_0 = \overline{\overline{y}_x x^2}, \\ \overline{x^3} a_2 + \overline{x^2} a_1 + \overline{x} a_0 = \overline{\overline{y}_x x}, \\ \overline{x^2} a_2 + \overline{x} a_1 + a_0 = \overline{\overline{y}_x}, \end{cases} \tag{4.37}$$

де

$$\overline{x^l} = \left(\sum_{i=1}^{k} x_i^l n_{x_i} \right) \Big/ n, \quad l = \overline{1,4}, \tag{4.38}$$

$$\overline{\overline{y}_x x^l} = \left(\sum_{i=1}^{k} \overline{y}_{x_i} x_i^l n_{x_i} \right) \Big/ n, \quad l = \overline{0,2}.$$

Оскільки $\overline{\overline{y}_x} = \overline{y},$ то розв'язавши третє рівняння системи (4.37) відносно a_0 і підставивши в (4.36), після простих перетворень отримаємо

$$\overline{y}_x = \overline{y} + a_1 (x - \overline{x}) + a_2 (x^2 - \overline{x^2}). \tag{4.36*}$$

Знайдені із системи рівнянь (4.37) параметри a_1 та a_2 підставимо в (4.36*); в підсумку отримуємо шукане рівняння регресії Y на X.

У випадку параболічної регресії (другого порядку) X на Y

$$\bar{x}_y = b_2 y^2 + b_1 y + b_0$$

невідомі параметри b_2, b_1, b_0 знаходяться як розв'язок такої системи нормальних рівнянь

$$\begin{cases} \overline{y^4} b_2 + \overline{y^3} b_1 + \overline{y^2} b_0 = \overline{\bar{x}_y y^2}, \\ \overline{y^3} b_2 + \overline{y^2} b_1 + \bar{y} b_0 = \overline{\bar{x}_y y}, \\ \overline{y^2} b_2 + \bar{y} b_1 + b_0 = \bar{x}. \end{cases}$$

РОЗВ'ЯЗУВАННЯ ТИПОВИХ ЗАДАЧ

Задача 4.1. Отримані статистичні дані десяти однотипних підприємств стосовно коефіцієнта використання основних засобів X і добового обсягу виробництва Y(тис. грн.):

x_i	0,4	0,45	0,5	0,55	0,6	0,65	0,7	0,75	0,8	0,9
y_i	2,3	2,4	3,2	3,8	4,5	4,7	5,1	5,6	7,2	8,4

1) скласти систему нормальних рівнянь і знайти коефіцієнти рівняння $\bar{y}_x = \alpha_1 x + a_0$ прямої регресії Y на X;

2) обчислити вибірковий коефіцієнт кореляції $r_в$.

○ 1) Для знаходження коефіцієнтів системи нормальних рівнянь (4.23)

$$\begin{cases} \overline{x^2} a_1 + \bar{x} a_0 = \overline{xy}, \\ \bar{x} a_1 + a_0 = \bar{y}, \end{cases}$$

складемо розрахункову табл. 4.2,

Таблиця 4.2

i	x_i	y_i	x_i^2	$x_i y_i$	y_i^2
1	0,4	2,3	0,16	0,92	5,29
2	0,45	2,4	0,2025	1,08	5,76
3	0,5	3,2	0,25	1,6	10,24
4	0,55	3,8	0,3025	2,09	14,44
5	0,6	4,5	0,36	2,7	20,25
6	0,65	4,7	0,4225	3,055	22,09
7	0,7	5,1	0,49	3,57	26,01
8	0,75	5,6	0,5625	4,2	31,36

9	0,8	7,2	0,64	5,76	54,76
10	0,9	8,4	0,81	7,56	70,56
Σ	6,3	47,2	4,2	32,535	260,76

останній стовпець якої потрібний для обчислення σ_y. Використовуючи нижній рядок табл. 4.2, отримаємо (обсяг вибірки n=10):

$$\bar{x} = \sum_{i=1}^{10} x_i / n = 6{,}3/10 = 0{,}63\ ;\ \bar{y} = \sum_{i=1}^{10} y_i / n = 47{,}2/10 = 4{,}72\ ;$$

$$\overline{x^2} = \sum_{i=1}^{10} x_i^2 / n = 4{,}2/10 = 0{,}42\ ;\ \overline{xy} = \sum_{i=1}^{10} x_i y_i / n = 32{,}535/10 = 3{,}2535\ ;$$

$$\overline{y^2} = \sum_{i=1}^{10} y_i^2 / n = 260{,}76/10 = 26{,}076;$$

$$\begin{cases} 0{,}42a_1 + 0{,}63a_0 = 3{,}2535, \\ 0{,}63a_1 + a_0 = 4{,}72. \end{cases}$$

Знайдемо єдиний розв'язок системи нормальних рівнянь згідно із формулами (4.24):

$$a_1 = \frac{\overline{xy} - \bar{x}\cdot\bar{y}}{\overline{x^2} - (\bar{x})^2} = \frac{3{,}2535 - 0{,}63\cdot 4{,}72}{0{,}42 - (0{,}63)^2} = \frac{0{,}2799}{0{,}0231} \approx 12{,}117\ ,$$

$$a_0 = \bar{y} - \bar{x}a_1 = 4{,}72 - 0{,}63\cdot 12{,}117 = -2{,}914\ .$$

Отже, емпіричне рівняння прямої регресії Y на X має такий вид:

$$\bar{y}_x = 12{,}117x - 2{,}914\ .$$

2) Вибірковий коефіцієнт кореляції $r_в$ знайдемо за формулою (4.30):

$$r_в = \frac{\overline{xy} - \bar{x}\cdot\bar{y}}{\sigma_x \sigma_y} = \frac{\overline{xy} - \bar{x}\cdot\bar{y}}{\sqrt{\overline{x^2} - (\bar{x})^2}\sqrt{\overline{y^2} - (\bar{y})^2}} =$$

$$= \frac{3{,}2535 - 0{,}63\cdot 4{,}72}{\sqrt{0{,}42 - (0{,}63)^2}\sqrt{26{,}076 - (4{,}72)^2}} \approx 0{,}945.$$

При цьому вибірковий коефіцієнт кореляції $r_в \approx 0{,}945$ свідчить про щільний лінійний зв'язок між X та Y, до того ж, оскільки $r_в > 0$,

то має місце додатна кореляція, тобто при зростанні X зростає відповідне значення результативної ознаки Y.

Задача 4.2 Дано результати статистичних досліджень факторіальної ознаки x та результативної ознаки y:

x	2	3	4	2	3	4	2	3	3	4	2	3	4	5	5	5	5	5	7	7
y	4	5	1	3	4	3	4	4	3	3	4	4	3	5	4	5	4	5	6	6

Побудувати рівняння кореляційної залежності.

○ Оскільки деякі пари чисел (x,y) зустрічаються декілька разів (наприклад, пара (2,4) — 3 рази, а пара (5,4) — 2 рази), тому для зручності згрупуємо дані, заповнивши табл. 4.3 (на перетині значень x та y вкажемо частоту появи даної пари)

Таблиця 4.3

X \ Y	1	3	4	5	6	n_x
2	—	1	3	—	—	4
3	—	1	3	1	—	5
4	1	—	3	—	—	4
5	—	—	2	3	—	5
7	—	—	—	—	—	2
n_y	1	2	11	4	2	$n = 20$

Визначимо характер зв'язку між X та Y.

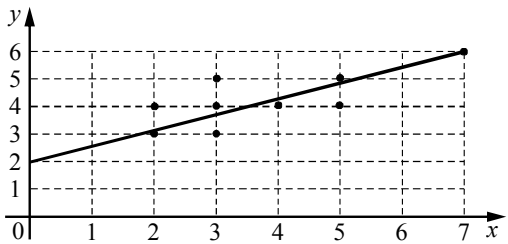


Рис. 4.3

Вважатимемо, що між x та y існує прямолінійна залежність, рівняння якої $\bar{y}_x = \alpha_1 x + a_0$ Щоб знайти параметри a_i та a_0, запишемо систему нормальних рівнянь: $\begin{cases} \overline{x^2} a_1 + \bar{x} a_0 = \overline{xy}, \\ \bar{x} a_1 + a_0 = \bar{y}, \end{cases}$

Проте рівняння прямолінійної регресії у вигляді $\bar{y}_x - \bar{y} = r_в \dfrac{\sigma_y}{\sigma_x}(x - \bar{x})$ зручніше, оскільки за величиною $r_в$ можна зробити висновок про тісноту зв'язку між X та Y. Нагадаємо, що

$$\bar{y} = \frac{\sum_{j=1}^{m} y_j n_{y_j}}{n}; \quad \bar{x} = \frac{\sum_{i=1}^{k} x_i n_{x_i}}{n}; \quad \sigma_y = \sqrt{D_y}; \quad D_y = \sigma_y^2 = \frac{\sum_{j=1}^{m} (y_j - y)^2 n_{y_j}}{n},$$

або за розрахунковою формулою

$$\sigma_y^2 = \overline{y^2} - (\bar{y})^2 = \frac{\sum_{j=1}^{m} y_j^2 n_{y_j}}{n} - (\bar{y})^2, \; \sigma_x^2 = \overline{x^2} - (\bar{x})^2 = \frac{\sum_{i=1}^{k} x_i^2 n_{y_i}}{n} - (\bar{x})^2,$$

$$r_b = \frac{\overline{xy} - \bar{x}\,\bar{y}}{\sigma_x \sigma_y}, \text{ де } \overline{xy} = \frac{\sum_{i=1}^{k} \sum_{j=1}^{m} x_i y_j n_{ij}}{n}.$$

Отже, $\bar{x} = \dfrac{2 \cdot 4 + 3 \cdot 5 + 4 \cdot 4 + 5 \cdot 5 + 7 \cdot 2}{20} = \dfrac{78}{20} = 3{,}90;$

$$\bar{y} = \frac{1 \cdot 1 + 3 \cdot 2 + 4 \cdot 11 + 5 \cdot 4 + 6 \cdot 2}{20} = \frac{83}{20} = 4{,}15;$$

$$\sigma_x^2 = \frac{2^2 \cdot 4 + 3^2 \cdot 5 + 4^2 \cdot 4 + 5^2 \cdot 5 + 7^2 \cdot 2}{20} - \left(\frac{78}{20}\right)^2 = 2{,}19; \quad \sigma_x = \sqrt{2{,}19} = 1{,}48;$$

$$\sigma_y^2 = \frac{1^2 \cdot 1 + 3^2 \cdot 2 + 4^2 \cdot 11 + 5^2 \cdot 4 + 6^2 \cdot 2}{20} - \left(\frac{83}{20}\right)^2 = 1{,}13; \quad \sigma_y = \sqrt{1{,}13} = 1{,}06.$$

Обчислимо спочатку $\overline{xy} - \bar{x}\,\bar{y}$:

$$\overline{xy} - \bar{x}\,\bar{y} = \frac{2\cdot 3\cdot 1+2\cdot 4\cdot 3+3\cdot 3\cdot 1+3\cdot 4\cdot 3+3\cdot 5\cdot 3+4\cdot 1\cdot 1+4\cdot 4\cdot 3+5\cdot 4\cdot 2+5\cdot 5\cdot 3+7\cdot 6\cdot 2}{20} -$$

$$-\frac{78}{20}\cdot\frac{83}{20} = 0,87.$$

Тоді вибірковий коефіцієнт кореляції дорівнюватиме:

$$r_{\text{в}} = \frac{0,87}{1,48\cdot 1,06} \approx 0,55.$$

Таким чином, рівняння прямолінійної регресії має вигляд:

$$\bar{y}_x - 4,15 = 0,55\cdot\frac{1,06}{1,48}(x-3,90), \text{ або } \bar{y}_x = 0,39x + 2,61.$$

Отже, між величинами X та Y існує прямолінійна кореляція, рівняння регресії якої $\bar{y}_x = 0,39x + 2,61$. При цьому вибірковий коефіцієнт кореляції $r_{\text{в}} \approx 0,55$ свідчить про середній лінійний зв'язок між X та Y, до того ж, оскільки $r_{\text{в}} > 0$, то має місце додатна кореляція, тобто при зростанні X зростає відповідне значення результативної ознаки Y.

ДОДАТКИ

Значення функції Гаусса $\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$

Таблиця 1

x	Соті долі x									
	0	1	2	3	4	5	6	7	8	9
0,0	0,39894	39892	39886	39876	39862	39844	39822	39797	39767	39733
0,1	39695	39654	39608	39559	39505	39448	39387	39322	39253	39181
0,2	39104	39024	38940	38853	38762	38667	38568	38466	38361	38251
0,3	38139	38023	37903	37780	37654	37524	37391	37255	37115	36973
0,4	36827	36678	36526	36371	36213	36053	35889	35723	35553	35381
0,5	35207	35029	34849	34667	34482	34294	34105	33912	33718	33521
0,6	33322	33121	32918	32713	32506	32297	32086	31874	31659	31443
0,7	31225	31006	30785	30563	30339	30114	29887	29659	29431	29200
0,8	28969	28737	28504	28269	28034	27798	27562	27324	27086	26848
0,9	26609	26369	26129	25888	25647	25406	25164	24923	24681	24439
1,0	24197	23955	23713	23471	23230	22988	22747	22506	22265	22025
1,1	21785	21546	21307	21069	20831	20594	20357	20121	19886	19652
1,2	19419	19186	18954	18724	18494	18265	18037	17810	17585	17360
1,3	17137	16915	16694	16474	16256	16038	15822	15608	15395	15183
1,4	14973	14764	14556	14350	14146	13943	13742	13542	13344	13147
1,5	12952	12758	12566	12376	12188	12001	11816	11632	11450	11270
1,6	11092	10915	10741	10567	10396	10226	10059	09893	09728	09566
1,7	09405	09246	09089	08933	08780	08628	08478	08329	08183	08038
1,8	07895	07754	07614	07477	07341	07206	07074	06943	06814	06687
1,9	06562	06438	06316	06195	06077	05959	05844	05730	05618	05508
2,0	05399	05292	05186	05082	04980	04879	04780	04682	04586	04491
2,1	04398	04307	04217	04128	04041	03955	03871	03788	03706	03626
2,2	03547	03470	03394	03319	03246	03174	03103	03034	02965	02898
2,3	02833	02768	02705	02643	02582	02522	02463	02406	02349	02294
2,4	02239	02186	02134	02083	02033	01984	01936	01888	01842	01797
2,5	01753	01709	01667	01625	01585	01545	01506	01468	01431	01394
2,6	01358	01323	01289	01256	01223	01191	01160	01130	01100	01071
2,7	01042	01014	00987	00961	00935	00909	00885	00861	00837	00814
2,8	00792	00770	00748	00727	00707	00687	00668	00649	00631	00613
2,9	00595	00578	00562	00545	00530	00514	00499	00485	00470	00457
3,0	00443	00430	00417	00405	00393	00381	00370	00358	00348	00337
3,1	00327	00317	00307	00298	00288	00279	00271	00262	00254	00246
3,2	00238	00231	00224	00216	00210	00203	00196	00190	00184	00178
3,3	00172	00167	00161	00156	00151	00146	00141	00136	00132	00127

Продовження табл. 1

x	Соті долі x									
	0	1	2	3	4	5	6	7	8	9
3,4	00123	00119	00115	00111	00107	00104	00100	00097	00094	00090
3,5	00087	00084	00081	00079	00076	00073	00071	00068	00066	00063
3,6	00061	00059	00057	00055	00053	00051	00049	00047	00046	00044
3,7	00042	00041	00039	00038	0037	00035	00034	00033	00031	00030
3,8	00029	00028	00027	00026	00025	00024	00023	00022	00021	00021
3,9	00020	00019	00018	00018	00017	00016	00016	00015	00014	00014
x	Десяті долі x									
	0		2		4		6		8	
4,	0,0001338		0000589		0000249		0000101		0000040	
5,	0000015									

Значення функції $P(m;\lambda) = \dfrac{\lambda^m e^{-\lambda}}{m!}$

Таблиця 2

m	λ									
	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1,0
0	0,90484	81873	74082	67032	60653	54881	49659	44933	40657	36788
1	09048	16375	22225	26813	30327	32929	34761	35946	36591	36788
2	00452	01637	03334	05363	07582	09879	12166	14379	16466	18394
3	00015	00109	00333	00715	01264	01976	02839	03834	04940	06131
4		00005	00025	00072	00158	00296	00497	00767	01111	01533
5			00002	00006	00016	00036	00070	00123	00200	00307
6					00001	00004	00008	00016	00030	00051
7							00001	00002	00004	00007
8										00001

Продовження табл. 2

m	λ									
	1,5	2,0	2,5	3,0	3,5	4,0	4,5	5,0	5,5	6,0
0	0,22313	13534	08208	04979	03020	01832	01111	00674	00409	00248
1	33470	27067	20521	14936	10569	07326	04999	03369	02248	01487
2	25102	27067	25652	22404	18496	14653	11248	08422	06181	04462
3	12551	18045	21376	22404	21579	19537	16872	14037	11332	08924
4	04707	09022	13360	16803	18881	19537	18981	17547	15582	13385
5	01412	03609	06680	10082	13217	15629	17083	17547	17140	16062
6	00353	01203	02783	05041	07710	10420	12812	14622	15712	16062
7	00076	00344	00994	02160	03855	05954	08236	10444	12345	13768
8	00014	00086	00311	00810	01687	02977	04633	06528	08487	10326
9	00002	00019	00086	00270	00656	01323	02316	03627	05187	06884
10		00004	00022	00081	00230	00529	01042	01813	02853	04130
11		00001	00005	00022	00073	00192	00426	00824	01426	02253
12			00001	00006	00021	00064	00160	00343	00654	01126
13				00001	00006	00020	00055	00132	00277	00520
14					00001	00006	00018	00047	00109	00223
15						00002	00005	00016	00040	00089
16							00002	00005	00014	00033
17								00001	00004	00012
18									00001	00004
19										00001

Значення функції Лапласа $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{\frac{t^2}{2}} dt$

Таблиця 3

x	Соті долі x									
	0	1	2	3	4	5	6	7	8	9
0,0	0,00000	00399	00798	01197	01595	01994	02392	02790	03188	03586
0,1	03983	04380	04776	05172	05567	05962	06356	06749	07142	07535
0,2	07926	08317	08706	09095	09483	09871	10257	10642	11026	11409
0,3	11791	12172	12552	12930	13307	13683	14058	14431	14803	15173
0,4	15542	15910	16276	16640	17003	17364	17724	18082	18439	18793
0,5	19146	19497	19847	20194	20540	20884	21226	21566	21904	22240
0,6	22575	22907	23237	23565	23891	24215	24537	24857	25175	25490
0,7	25804	26115	26424	26730	27035	27337	27637	27935	28230	28524
0,8	28814	29103	29389	29673	29955	30234	30511	30785	31057	31327
0,9	31594	31859	32121	32381	32639	32894	33147	33398	33646	33891

Продовження табл. 3

x	Соті долі x									
	0	1	2	3	4	5	6	7	8	9
1,0	34134	34375	34614	34850	35083	35314	35543	35769	35993	36214
1,1	36433	36650	36864	37076	37286	37493	37698	37900	38100	38298
1,2	38493	38686	38877	39065	39251	39435	39617	39796	39973	40147
1,3	40320	40490	40658	40824	40988	41149	41308	41466	41621	41774
1,4	41924	42073	42220	42364	42507	42647	42786	42922	43056	43189
1,5	43319	43448	43574	43699	43822	43943	44062	44179	44295	44408
1,6	44520	44630	44738	44845	44950	45053	45154	45254	45352	45449
1,7	45543	45637	45728	45818	45907	45994	46080	46164	46246	46327
1,8	46407	46485	46562	46638	46712	46784	46856	46926	46995	47062
1,9	47128	47193	47257	47320	47381	47441	47500	47558	47615	47670
2,0	47725	47778	47831	47882	47932	47982	48030	48077	48124	48169
2,1	48214	48257	48300	48341	48382	48422	48461	48500	48537	48574
2,2	48610	48645	48679	48713	48745	48778	48809	48840	48870	48899
2,3	48928	48956	48983	49010	49036	49061	49086	49111	49134	49158
2,4	49180	49202	49224	49245	49266	49286	49305	49324	49343	49361
2,5	49379	49396	49413	49430	49446	49461	49477	49492	49506	49520
2,6	49534	49547	49560	49573	49585	49598	49609	49621	49632	49643
2,7	49653	49664	49674	49683	49693	49702	49711	49720	49728	49736
2,8	49744	49752	49760	49767	49774	49781	49788	49795	49801	49807
2,9	49813	49819	49825	49831	49836	49841	49846	49851	49856	49861
3,0	49865	49869	49874	49878	49882	49886	49889	49893	49897	49900
3,1	49903	49906	49910	49913	49916	49918	49921	49924	49926	49929
3,2	49931	49934	49936	49938	49940	49942	49944	49946	49948	49950
3,3	49952	49953	49955	49957	49958	49960	49961	49962	49964	49965
3,4	49966	49968	49969	49970	49971	49972	49973	49974	49975	49976
3,5	49977	49978	49978	49979	49980	49981	49981	49982	49983	49983
3,6	49984	49985	49985	49986	49986	49987	49987	49988	49988	49989
3,7	49989	49990	49990	49990	49991	49991	49992	49992	49992	49992
3,8	49993	49993	49993	49994	49994	49994	49994	49995	49995	49995
3,9	49995	49995	49996	49996	49996	49996	49996	49996	49997	49997

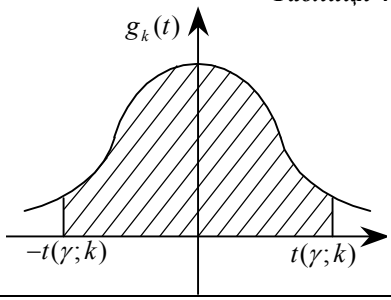
x	Десяті долі x				
	0	2	4	6	8
4,	0,4999683	4999867	4999946	4999979	4999992
5,	4999997				

Таблиця 4

Значення $t = t(\gamma; k)$,
що задовільняють рівнянню

$$P\left(|T| < t\right) = 2\int_0^t g_k(t)dt = \gamma,$$

де $g_k(t)$ — густина **розподілу Ст'юдента (*t*-розподілу)**, $k = n-1$ — число ступенів вільності



$k = n-1$	γ				
	0,9	0,95	0,98	0,99	0,999
1	6,314	12,706	31,821	63,657	636,619
2	2,920	4,303	6,965	9,925	31,599
3	2,353	3,182	4,541	5,841	12,924
4	2, 132	2,776	3,747	4,604	8,610
5	2,015	2,571	3,365	4,032	6,869
6	1,943	2,447	3,143	3,707	5,969
7	1,895	2,365	2,998	3,499	5,408
8	1,860	2,306	2,896	3,355	5,041
9	1,833	2,262	2,821	3,250	4,781
10	1,812	2,228	2,764	3,169	4,587
11	1,796	2,201	2,718	3,106	4,437
12	1,782	2,179	2,681	3,055	4,318
13	1,771	2,160	2,650	3,012	4,221
14	1,761	2,145	2,624	2,977	4,140
15	1,753	2,131	2,602	2,947	4,073
16	1,746	2,120	2,583	2,921	4,015
17	1,740	2,110	2,567	2,898	3,965
18	1,734	2,101	2,552	2,878	3,922
19	1,729	2,093	2,539	2,861	3,883
20	1,725	2,086	2,528	2,845	3,850
25	1,708	2,060	2,485	2,785	3,725
30	1,697	2,042	2,457	2,750	3,646
40	1,684	2,021	2,423	2,704	3,551
50	1,676	2,009	2,403	2,678	3,496
60	1,671	2,000	2,390	2,660	3,460
70	1,667	1,994	2,381	2,648	3,435

Продовження табл. 4

80	1,664	1,990	2,374	2,639	3,416
90	1,662	1,987	2,368	2,632	3,402
100	1,660	1,984	2,364	2,626	3,390
120	1,658	1,980	2,358	2,617	3,373
∞	1,645	1,960	2,326	2,576	3,291

Таблиця 5

Критичні точки розподілу Ст'юдента (*t*-розподілу)

Для двосторонньої критичної області критична точка $t_{\text{двост.кр}}(\alpha;k)=t_\alpha$ є коренем рівняння $\int\limits_0^{t_\alpha} g_k(t)dt=(1-\alpha)/2$; для односторонньої (правосторонньої) критичної області точка $t_{\text{правост.кр}}(\alpha;k)=t_{2\alpha}$ є коренем рівняння $\int\limits_{t_{2\alpha}}^{\infty} g_k(t)dt=(1-2\alpha)/2$, де $g_k(t)$ — густина розподілу Ст'юдента, $k=n-1$ — число ступенів вільності. Для лівосторонньої критичної області $t_{\text{лівост.кр}}(\alpha;k)=-t_{2\alpha}$.

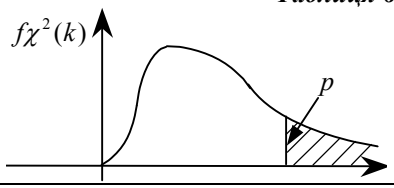
Число ступенів вільності $k=n-1$	Рівень значущості α (**двостороння** критична область)					
	0,10	0,05	0,02	0,01	0,002	0,001
1	6,314	12,71	31,82	63,66	318,3	637,0
2	2,920	4,303	6,965	9,925	22,33	31,60
3	2,353	3,182	4,541	5,841	10,22	12,94
4	2,132	2,776	3,747	4,604	7,173	8,610
5	2,015	2,571	3,365	4,032	5,893	6,859
6	1,943	2,447	3,143	3,707	5,208	5,959
7	1,895	2,365	2,998	3,499	4,785	5,405
8	1,860	2,306	2,896	3,355	4,501	5,041
9	1,833	2,262	2,821	3,250	4,297	4,781
10	1,812	2,228	2,764	3,169	4,144	4,587
11	1,796	2,201	2,718	3,106	4,025	4,437
12	1,782	2,179	2,681	3,055	3,930	4,318
13	1,771	2,160	2,650	3,012	3,852	4,221
14	1,761	2,145	2,624	2,977	3,787	4,140

Продовження табл.5

15	1,753	2,131	2,602	2,947	3,733	4,073
16	1,746	2,120	2,583	2,921	3,686	4,015
17	1,740	2,110	2,567	2,898	3,646	3,965
18	1,734	2,101	2,552	2,878	3,611	3,922
19	1,729	2,093	2,539	2,861	3,579	3,883
20	1,725	2,086	2,528	2,845	3,562	3,850
21	1,721	2,080	2,518	2,831	3,527	3,819
22	1,717	2,074	2,508	2,819	3,505	3,792
23	1,714	2,069	2,500	2,807	3,485	3,767
24	1,711	2,064	2,492	2,797	3,467	3,745
25	1,708	2,060	2,485	2,787	3,450	3,725
26	1,706	2,056	2,479	2,779	3,435	3,707
27	1,703	2,052	2,473	2,771	3,421	3,690
28	1,701	2,048	2,467	2,763	3,408	3,674
29	1,699	2,045	2,462	2,756	3,396	3,659
30	1,697	2,042	2,457	2,750	3,385	3,646
40	1,684	2,021	2,423	2,704	3,307	3,551
60	1,671	2,000	2,390	2,660	3,232	3,460
120	1,658	1,981	2,362	2,624	3,172	3,374
∞	1,645	1,960	2,326	2,576	3,090	3,291
Число ступенів вільності $k = n - 1$	0,05	0,025	0,01	0,005	0,001	0,0005
	Рівень значущості α (**одностороння** критична область)					

Таблиця 6

Значення
$P\left(\chi^2(k) > \chi^2(p;k) \right) = p,$
де k — число ступенів вільності



k	p							
	0,999	0,99	0,95	0,90	0,10	0,05	0,01	0,001
1	$0{,}157 \cdot 10^{-5}$	0,0002	0,004	0,02	2,71	3,84	6,63	10,83
2	0,002	0,02	0,10	0,21	4,61	5,99	9,21	13,82

Продовження табл. 6

3	0,02	0,12	0,35	0,58	6,25	7,82	11,34	16,27
4	0,09	0,30	0,71	1,06	7,78	9,49	13,28	18,47
5	0,21	0,55	1,15	1,61	9,24	11,07	15,08	20,51
6	0,38	0,87	1,64	2,20	10,65	12,59	16,81	22,46
7	0,60	1,24	2,17	2,83	12,02	14,06	18,48	24,32
8	0,86	1,65	2,73	3,49	13,36	15,51	20,09	26,12
9	1,15	2,09	3,33	4,17	14,68	16,92	21,67	27,88
10	1,48	2,56	3,94	4,87	15,99	18,31	23,21	29,59
11	1,83	3,05	4,58	5,58	17,28	19,68	24,72	31,26
12	2,21	3,57	5,23	6,30	18,55	21,03	26,22	32,91
13	2,62	4,11	5,89	7,04	19,81	22,36	27,69	34,53
14	3,04	4,66	6,57	7,79	21,06	23,69	29,14	36,12
15	3,48	5,23	7,26	8,55	22,31	25,00	30,58	37,70
16	3,94	5,81	7,96	9,31	23,54	26,30	32,00	39,25
17	4,42	6,41	8,67	10,09	24,77	27,59	33,41	40,79
18	4,90	7,02	9,39	10,86	25,99	28,87	34,81	42,31
19	5,41	7,63	10,12	11,65	27,20	30,14	36,19	43,82
20	5,92	8,26	10,85	12,44	28,41	31,41	37,57	45,31
21	6,45	8,90	11,59	13,24	29,62	32,67	38,93	46,80
22	6,98	9,54	12,34	14,04	30,81	33,92	40,29	48,27
23	7,53	10,20	13,20	14,85	32,01	35,17	41,64	49,73
24	8,08	10,86	13,85	15,66	33,19	36,42	43,98	51,18
25	8,65	11,52	14,61	16,47	34,38	37,65	44,31	52,62
26	9,22	12,20	15,37	17,29	35,56	38,89	45,64	54,05
27	9,80	12,88	16,15	18,11	36,74	40,11	46,96	55,48
28	10,39	13,56	16,93	18,94	37,92	41,34	48,28	56,89
29	10,99	14,26	17,71	19,77	39,09	42,56	49,59	58,30
30	11,59	14,95	18,49	20,60	40,26	43,77	50,89	59,70
40	17,92	22,16	26,51	29,05	51,81	55,76	63,69	73,40
50	24,67	29,71	34,76	37,69	63,17	67,51	76,15	86,66
100	61,92	70,07	77,93	82,36	118,50	124,34	135,81	149,45

Таблиця 7

Критичні точки $F_{кр}(\alpha;k_1,k_2)$ розподілу Фішера-Снедокора,

що задовільняють рівнянню $P\left[F > F_{кр}(\alpha;k_1,k_2)\right] = \alpha$ при $\alpha = 0{,}05$

k_2	k_1									
	1	2	3	4	5	6	8	12	24	∞
1	161,45	199,50	215,71	224,58	230,16	233,99	238,88	243,91	249,05	254,32
2	18,51	19,00	19,16	19,25	19,30	19,33	19,37	19,41	19,45	19,50
3	10,13	9,55	9,28	9,12	9,01	8,94	8,84	8,74	8,64	8,53
4	7,71	6,94	6,59	6,39	6,26	6,16	6,04	5,91	5,77	5,63
5	6,61	5,79	5,41	5,19	5,05	4,95	4,82	4,68	4,53	4,36
6	5,99	5,14	4,76	4,53	4,39	4,28	4,15	4,00	3,84	3,67
7	5,59	4,74	4,35	4,12	3,97	3,87	3,73	3,57	3,41	3,23
8	5,32	4,46	4,07	3,84	3,69	3,58	3,44	3,28	3,12	2,93
9	5,12	4,26	3,86	3,63	3,48	3,37	3,23	3,07	2,90	2,71
10	4,96	4,10	3,71	3,48	3,33	3,22	3,07	2,91	2,74	2,54
11	4,84	3,98	3,59	3,36	3,20	3,09	2,95	2,79	2,61	2,40
12	4,75	3,88	3,49	3,26	3,11	3,00	2,85	2,69	2,50	2,30
13	4,67	3,80	3,41	3,18	3,02	2,92	2,77	2,60	2,42	2,21
14	4,60	3,74	3,34	3,11	2,96	2,85	2,70	2,53	2,35	2,13
15	4,54	3,68	3,29	3,06	2,90	2,79	2,64	2,48	2,29	2,07
16	4,49	3,63	3,24	3,01	2,85	2,74	2,59	2,42	2,24	2,01
17	4,45	3,59	3,20	2,96	2,81	2,70	2,55	2,38	2,19	1,96
18	4,41	3,55	3,16	2,93	2,77	2,66	2,51	2,34	2,15	1,92
19	4,38	3.52	3,13	2,90	2,74	2,63	2,48	2,31	2,11	1,88
20	4,35	3,49	3,10	2,87	2,71	2,60	2,45	2,28	2,08	1,84
25	4,24	3,38	2,99	2,76	2,60	2.49	2,34	2.16	1,96	1,71
30	4,17	3,32	2,92	2,69	2,53	2,42	2,27	2,09	1,89	1,62
40	4,08	3,23	2,84	2,61	2,45	2,34	2,18	2,00	1,79	1,51
60	4,00	3,15	2,76	2,52	2,37	2,25	2,10	1,92	1,70	1,39
120	3,92	3,07	2,68	2,45	2,29	2,17	2,02	1,83	1,61	1,25
∞	3,84	2,99	2,60	2,37	2,21	2,09	1,94	1,75	1,52	1,00

ЛІТЕРАТУРА

1. *Бочаров П. П., Печенкин А. В.* **Теория вероятностей. Математическая статистика**. — М.: Гардарика, 1998. — 328 с.
2. *Бугір М. К.* **Практикум з теорії імовірностей та математичної статистики.** — Тернопіль: ЦМДС, 1998. — 171 с.
3. *Булдик Г. М.* **Теория вероятностей и математичская статистика.** — Минск: Вышэйшая школа, 1989. — 285 с.
4. *Гмурман В. Е.* **Руководство к решению задач по теории вероятностей и математической статистике.** — М.: Высш. шк., 1979. — 400 с.
5. *Гмурман В. Е.* **Теория вероятностей и математическая статистика.** — М.: Высш. шк., 2004. — 479 с.
6. *Грубер Й.* **Эконометрия.** — Том 1. **Введение в эконометрию.** — К.: Астарта, 1996. — 397 с.
7. *Єрьоменко В. О., Шинкарик М. І.* **Теорія імовірностей.** — Тернопіль: Економічна думка, 2000. — 176 с.
8. *Єрьоменко В. О., Шинкарик М. І.* **Математична статистика.** — Тернопіль: Економічна думка, 2002. — 247 с.
9. *Жлуктенко В. І., Наконечний С. І.* **Теорія ймовірностей і математична статистика:** Навч. метод. посібник. У 2ч. — ч.І. Теорія ймовірностей. — К.: КНЕУ, 2000. — 304 с.
10. *Карасев А. И.* **Теория вероятностей и математическая статистика.** — М.: Статистика, 1977. — 279 с.
11. *Кибзун А.И., Горяинов Е.Р. Наумов А.В., Сиротин А.Н.* **Теория вероятностей и математическая статистика. Базовый курс с примерами и задачами** / Учебн. пособие. М.: ФИЗМАТЛИТ, 2002. —224с.
12. *Коваленко И. Н., Гнеденко Б. В.* **Теория вероятностей.** — К.: Вища школа, 1990. — 328 с.
13. *Колемаев В. А., Калинина В. Н.* **Теория вероятностей и математическая статистика:** Учебник / Под ред. В. А. Колемаева. — М.: ИНФРА-М, 2000. — 302 с.
14. *Колемаев В. А., Староверов О. В., Турундаевский В. Б.* **Теория вероятностей и математическая статистика.** — М.: Высш. шк., 1991. — 400 с.

15. *Кремер М.Ш.* **Теория вероятностей и математическая статистика.** — М.: ЮНИТИ-ДАНА, 2002. — 543 с.

16. *Мармоза А.Т.* **Практикум з математичної статистики:** Навч. посібник. —К.: Кондор, 2004. —264с.

17. *Мюллер П., Найман П., Шторм Р.* **Таблицы по математической статистике.** — М.: Финансы и статистика, 1982. — 278 с.

18. *Румшиский Л. З.* **Математическая обработка результатов эксперимента.** — М.: Наука, 1971. — 192 с.

19. **Статистика підприємництва:** Навч. посібник. — Вашків П. Г., Пастер П. І., Сторожук В. П., Ткач Є. І. — К.: Слобожанщина, 1999. — 600 с.

20. *Черняк І.О., Обушна О.М., Ставицький А.В.* **Теорія ймовірностей та математична статистика:** Збірник задач: Навч. посібник. — К.: Т-во "Знання", КОО, 2001. — 199 с. — (Вища освіта XXI століття).

ЗМІСТ