УКРАЇНА

НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ БІОРЕСУРСІВ І ПРИРОДОКОРИСТУВАННЯ УКРАЇНИ

Кафедра вищої та прикладної математики

ПРАКТИКУМ З ВИБІРКОВОЇ ДИСЦИПЛІНИ

«МЕТОДИ МАТЕМАТИЧНОЇ СТАТИСТИКИ У НАУКОВИХ ДОСЛІДЖЕННЯХ»

Для магістрів І року навчання

Київ – 2024

УДК 378.022.51

Наведено необхідний теоретичний матеріал та навчальні завдання курсу “Методи математичної статистики в наукових дослідженнях». До кожної базової теми запропоновано типові завдання, що супроводжуються методичними рекомендаціями щодо їх розв’язання.

Рекомендовано вченою радою ННІ енергетики, автоматики і енергозбереження Національного університету біоресурсів і природокористування України,

Протокол № 6 від 21 червня 2024 р.

Укладач: доцент Л.А.Панталієнко

Рецензенти: доцент кафедри електротехніки, електромеханіки та електротехнологій ННІ ЕАіЕ Синявський О.Ю.; доцент кафедри вищої та прикладної математики ННІ ЕАіЕ Шостак С.В.

ПРАКТИКУМ З ВИБІРКОВОЇ ДИСЦИПЛІНИ «МЕТОДИ МАТЕМАТИЧНОЇ СТАТИСТИКИ У НАУКОВИХ ДОСЛІДЖЕННЯХ»

Для магістрів 1 року навчання

Укладач ПАНТАЛІЄНКО Людмила Анатоліївна

Програма вибіркової дисципліни

«Методи математичної статистики у наукових дослідженнях».

Змістовий модуль 1. Основи статистичного опису. − 8 год.

Тема лекційного заняття 1. Вибірка. Подія. Частота події. Принцип групування даних. − 2 год.

Основні задачі математичної статистики. Генеральна сукупність і вибірка. Вибірковий метод. Дискретні та неперервні випадкові величини. Статистичний розподіл вибірки. Варіаційний ряд. Полігон і гістограма. Емпірична функція розподілу.

Тема лекційного заняття 2. Закон стійкості частот. Частотне визначення ймовірностей. Завдання ймовірностей для дискретних та неперервних даних. − 2 год.

Відносна частота події. Статистичне означення ймовірності. Поняття про геометричні ймовірності. Передбачувальна властивість ймовірності. Завдання ймовірностей для дискретних та неперервних даних

Тема лекційного заняття 3. Основні дискретні та неперервні розподіли. − 2 год.

Основні закони розподілу дискретної випадкової величини (рівномірний, біноміальний, розподіл Пуассона, геометричний, гіпергеометричний), їх числові характеристики. Основні закони розподілу неперервних випадкових величин: рівномірний розподіл. нормальний та показниковий закони розподілу. Властивості та застосування.

Тема лекційного заняття 4. Теореми додавання та добутку подій. Формула повної ймовірності. Формули Байєса. − 2 год.

Алгебра подій: сума й добуток подій. Теореми додавання ймовірностей для сумісних та несумісних подій. Умовна ймовірність. Теореми множення ймовірностей для залежних та незалежних подій. Ймовірність появи хоча б однієї з кількох незалежних подій. Формула повної ймовірності. Формули Байєса.

Змістовий модуль 2. Статистичне оцінювання параметрів розподілу. Кореляційний та регресійний аналіз. − 8 год.

Тема лекційного заняття 5. Статистичні оцінки параметрів розподілу, їхні властивості. Точкове оцінювання параметрів основних розподілів. − 2 год.

Статистичні оцінки параметрів розподілу. Властивості оцінок. Точкові оцінки параметрів. Вибіркова середня. Вибіркова та виправлена дисперсія (середнє квадратичне відхилення).

Тема лекційного заняття 6. Інтервальні оцінки параметрів розподілу. Довірча ймовірність (надійність). Довірчий інтервал. − 2 год.

Інтервальні оцінки параметрів розподілу. Точність оцінки. Довірча ймовірність (надійність). Довірчий інтервал. Побудова довірчого інтервалу для математичного сподівання при відомому та невідомому σ.

Тема лекційного заняття 7. Перевірка статистичних гіпотез. Критерій Пірсона. Критерій згоди Колмогорова. − 2 год.

Постановка задачі. Статистичний критерій. Критична область. Перевірка гіпотез про закон розподілу (критерій згоди χ^2 Пірсона та критерій згоди λ Колмогорова). Статистична перевірка деяких параметричних гіпотез (про математичне сподівання, дисперсію нормально розподіленої величини, виключення грубих помилок при проведенні спостережень).

Тема лекційного заняття 8. Статистична (кореляційна) залежність між величинами. Вибірковий коефіцієнт кореляції. Лінійна регресія. − 2 год.

Статистична (кореляційна) залежність випадкових величин. Основні задачі кореляційного аналізу. Обчислення вибіркового коефіцієнта кореляції. Знаходження параметрів вибіркового лінійного рівняння регресії. Метод найменших квадратів в оцінюванні параметрів регресії..

5. Теми практичних занять

№ з/п	Назва теми	Кількість годин
1.	Статистичний розподіл вибірки. Варіаційний ряд. Полігон і гістограма.	2
2.	Ряд розподілу. Функція та щільність розподілу ймовірностей.	2
3.	Теореми додавання та добутку подій. Розрахунок надійності системи.	2
4.	Статистичні оцінки параметрів розподілу. Точкові оцінки параметрів нормального розподілу.	2
5.	Побудова довірчого інтервалу для математичного сподівання нормального розподілу.	2
6.	Критерій Пірсона χ^2. Критерій згоди Колмогорова.	2
7.	Обчислення вибіркового коефіцієнта кореляції. Побудова рівняння лінійної регресії.	2
Разом		14

Тема 1. Вибірка. Подія. Частота події. Принцип групування даних

1.1. Основні задачі математичної статистики

У науці вивчають ті явища, які підпорядковані умовам відтворюваності.

Масовими називають явища, умови спостереження яких можуть багаторазово відтворюватися у часі чи просторі.

Випадковими масовими явищами називають такі, умови спостереження яких неоднозначно пов'язані з результатами цих спостережень.

Закони теорії ймовірностей − це математичне вираження реальних закономірностей масових випадкових явищ, що вивчаються дослідним шляхом на підставі обробки спостережень над масовими випадковими явищами.

Математична статистика займається як статистичним описом результатів дослідів або спостережень, так і побудовою й перевіркою відповідних математичних моделей, що містять поняття ймовірності. Теоретичною базою математичної статистики слугує теорія ймовірностей.

Першою задачею математичної статистики є розробка методів збору та впорядкування статистичних даних (представлення їх в найбільш зручному для огляду й аналізу вигляді).

Другою задачею математичної статистики є визначення (оцінка) за статистичними даними характеристик випадкових величин.

Третьою задачею математичної статистики є задача перевірки правдоподібності статистичних гіпотез (про вигляд невідомого розподілу або про величину параметрів розподілу, вигляд якого невідомий).

У загальному випадку задача перевірки гіпотез формулюється так. Нехай маємо сукупність статистичних даних, які відносяться до однієї або кількох випадкових величин. Необхідно встановити чи суперечать ці дані тій чи іншій гіпотезі (наприклад, гіпотезі про те, що випадкова величина розподілена за певним законом зі щільністю $f(x)$ або про те, що дві випадкові величини некорельовані і т.і.)

Сучасна математична статистика розробляє способи визначення кількості необхідних дослідів до початку дослідження (планування експерименту), в процесі дослідження (послідовний аналіз) та розв'язує багато інших задач про прийняття рішень в умовах невизначеності.

1.2. Генеральна сукупність. Вибірка. Способи відбору

Нехай потрібно дослідити сукупність однорідних об'єктів щодо деякої якісної або ж кількісної ознаки X, що характеризує ці об'єкти. Наприклад, якщо є партія деталей, то кількісною ознакою може слугувати розмір деталі, що контролюється, а якісною − стандартність деталі. Вважаємо, що ознака X, яка вивчається, є випадковою величиною.

Генеральною сукупністю називають всю сукупність об'єктів, що підлягають вивченню. Сукупність n об'єктів, випадково відібраних з генеральної сукупності, називається ***вибірковою сукупністю*** або ***вибіркою.*** Число n − ***об'єм вибірки.***

Одним із основних методів статистичного дослідження є ***вибірковий метод***, суть якого полягає в тому, що на основі вивчення вибірки, що складається зі скінченого числа n об'єктів, робляться певні висновки для всієї генеральної сукупності. Природно, що ці висновки та оцінки відображують випадковий характер зібраних статистичних даних і тому повинні вважатись наближеними оцінками ймовірнісного характеру.

Для того, щоб за даними вибірки можна було одержати правильні висновки про ознаку X усієї генеральної сукупності, необхідно, щоб вибірка була ***репрезентативною***, тобто, щоб об'єкти вибірки правильно відображували генеральну сукупність, зберігаючи її пропорції.

В силу закону великих чисел вибірка буде репрезентативною, якщо її здійснювати випадково: кожний об'єкт вибірки відібраний випадково з генеральної сукупності, якщо всі об'єкти мають однакову ймовірність потрапити у вибірку.

Вибірка може проводитись за двома основними способами: відібраний об'єкт після обстеження може повертатися або не повертатися у генеральну сукупність. У відповідності з цим вибірки розділяють на ***повторні*** та ***безповторні.***

Повторною називають вибірку, за якою відібраний об'єкт (перед відбором наступного) повертається у генеральну сукупність.

Безповторною називають вибірку, за якою відібраний об'єкт у генеральну сукупність не повертається.

На практиці застосовують різні способи відбору. Принципово їх розділяють на два типи:

1. ***Відбір, що не вимагає поділу генеральної сукупності на частини (групи)***. Сюди відносять: а) ***простий випадковий безповторний відбір***; б) ***простий випадковий повторний відбір***.

2. ***Відбір, за яким генеральна сукупність розділяється на***

частини. Сюди відносять: а) ***типовий відбір***; б) ***механічний відбір***; в) ***серійний відбір***.

Простим випадковим називають такий відбір, за яким об'єкти беруть по одному з усієї генеральної сукупності. Здійснювати простий відбір можна різними способами. Наприклад, на картках записують номери від 1 до N (об'єм генеральної сукупності) та старанно перемішують. Далі навмання беруть одну картку і досліджують об'єкт з цим номером, потім картку повертають (або ж ні) і процес повторюють n разів (за об'ємом вибірки). При великому об'ємі генеральної сукупності цей спосіб відбору є досить громіздким.

Типовим називають відбір, за яким об'єкти беруть не з усієї генеральної сукупності, а з кожної її «типової» частини. Наприклад, якщо деталі виготовляють на декількох станках, відбір здійснюють не з усієї сукупності деталей, що виготовлені усіма станками, а з продукції кожного окремого станка. Типовий відбір застосовують тоді, коли значення досліджуваної ознаки значно коливаються у різних типових частинах. Наприклад, якщо продукція виготовляється на декількох станках, серед яких є більш і менш зношені.

Механічним називають відбір, за яким генеральну сукупність «механічно» розділяють на стільки груп, скільки об'єктів має війти у вибірку, і з кожної групи вибирають один об'єкт. Наприклад, якщо потрібно відібрати 20% виготовлених станком деталей, то відбирають кожну п'яту деталь; якщо потрібно відібрати 5% деталей, то відбирають кожну двадцяту деталь і т.д. Іноді механічний відбір може не забезпечувати вимоги репрезентативності для вибірки.

Серійним називають відбір, за яким об'єкти відбирають з генеральної сукупності не по одному, а «серіями», що підлягають суцільному дослідженню. Наприклад, якщо вироби виготовляють великою групою станків-автоматів, то суцільному дослідженню будуть підлягати тільки декілька станків. Серійний відбір застосовують тоді, коли значення досліджуваної ознаки коливаються незначно у різних серіях.

На практиці часто застосовують комбінований відбір, за яким комбінують вказані вище способи відбору. Наприклад, генеральну сукупність розбивають на серії однакового об'єму, потім простим випадковим відбором вибирають кілька серій і, нарешті, з кожної серії простим випадковим відбором вибирають окремі об'єкти.

1.3. Статистичний розподіл вибірки

Нехай для вивчення кількісної ознаки X з генеральної сукупності вибрано вибірку об'єму n, причому при n незалежних

випробуваннях ознака X прийняла значення x_1 – n_1 разів, x_2 – n_2 разів, ..., значення x_k – n_k разів; $n_1+n_2+...+n_k=n$ (*сума частот дорівнює об'єму вибірки*).

Значення x_i , $i=1,2,...,k$ ознаки X називають ***варіантами***, а послідовність варіант у зростаючому порядку – ***варіаційним рядом.***

Числа n_i (кількість спостережень) називаються ***частотами***, а числа $\omega_i=\frac{n_i}{n}$ – ***відносними частотами варіант*** x_i ($i=1,2,...,k$); причому $\sum_{i=1}^{k}\omega_i=1$ (*сума відносних частот дорівнює одиниці*).

Статистичним розподілом вибірки називається перелік варіант і відповідних їм частот або відносних частот. Статистичний розподіл можна задавати також у вигляді послідовності інтервалів зміни варіант і відповідних цим інтервалам частот або відносних частот (в якості частоти певного інтервалу приймають суму частот варіант, що потрапили в цей інтервал).

В результаті обробки статистичних даних одержується ***дискретний*** або ***інтервальний варіаційний ряд.***

Дискретний варіаційний ряд випадкової величини X подають у вигляді таблиці:

Таблиця 1.

X	x_1	x_2	…	x_k
n_i	n_1	n_2	…	n_k
ω_i	ω_1	ω_2	…	ω_k

Приклад 1.1. Заданий розподіл частот дискретної випадкової величини X (**ДВВ**) X :

X	2	6	12
n_i	3	10	7

Скласти розподіл відносних частот.

Розв'язання. За умовою задачі об'єм вибірки $n=20$ ($n=3+10+7=20$). Тепер знаходимо відносні частоти:

$$\omega_1=\frac{n_1}{n}=\frac{3}{20}=0{,}15\,;\ \omega_2=\frac{n_2}{n}=\frac{10}{20}=0{,}5\,;\ \omega_3=\frac{n_3}{n}=\frac{7}{20}=0{,}35\,.$$

Контроль: $0{,}15+0{,}5+0{,}35=1$. Результати обчислень зводимо в таблицю:

X	2	6	12
ω_i	0,15	0,5	0,35

Якщо вивчається неперервна випадкова величина X (**НВВ**) і число спостережень відносно велике ($n \geq 20$), будують інтервальний варіаційний ряд. Для цього проміжок, в якому містяться всі результати спостережень вибірки, розбивають на кілька (від 5 до 15) рівних або нерівних інтервалів, підраховують суми $n_1, n_2, ... n_r$ частот варіант, що потрапили в ці інтервали, відносні частоти $\omega_i = \frac{n_i}{n}$ та ***щільності відносних частот*** $p_i^* = \frac{\omega_i}{|\Delta_i|}$ ($|\Delta_i|$ – довжина i-го інтервалу, $i = 1,2,...,r$).

Довжина кожного інтервалу, зазвичай, береться сталою величиною $h = \frac{x_{\max} - x_{\min}}{M-1}$, $M = [1 + 3{,}32 \lg n]$ (n – об'єм вибірки). Межі послідовних інтервалів позначають $x_0, x_1, ... x_M$, кількість інтервалів – M, а самі інтервали записують у вигляді

$$[x_0, x_1[,\ [x_1, x_2[,\ [x_2, x_3[,\ ...,\ [x_{M-1}, x_M[, \text{ де } x_0 < x_1 < x_2 < ... < x_M .$$

Значення, що знаходяться на межі інтервалів, відносять до правої межі інтервалу.

Інтервальний варіаційний ряд випадкової величини X подають у вигляді таблиці:

Таблиця 2.

Інтервали	$[x_0, x_1[$	$[x_1, x_2[$	…	$[x_{M-1}, x_M[$
n_i	n_1	n_2	…	n_r
ω_i	ω_1	ω_2	…	ω_r
p_i^*	p_1^*	p_2^*	…	p_r^*

Межі часткових інтервалів визначають за формулами:

$$x_0 = x_{\min} - \frac{h}{2},\ x_1 = x_0 + h,\ x_2 = x_1 + h,\ \ldots, x_i = x_{i-1} + h, \ldots, x_M = x_{M-1} + h = x_{\max} + \frac{h}{2}.$$

Отже, розбиття на інтервали здійснюється так, щоб найменше значення вибірки потрапляло у середину першого інтервалу, а

найбільше значення вибірки – у середину останнього.

Зауваження. Для неперервної ознаки X закон розподілу можна будувати також у вигляді таблиці 1, вибираючи в якості x_i, $i = 1,2,...,k$ середнє значення з відповідного часткового інтервалу.

Приклад 1.2. У результаті 20-ти вимірювань деякої фізичної величини одним приладом (без систематичних похибок) одержано такі результати:

18,50; 15,97; 18,81; 18,93; 17,16; 19,04; 21,58; 18,50; 14,35; 13,21; 14,96; 19,25; 12,12; 14,55; 14,30; 14,39; 14,24; 18,48; 16,38; 16,45.

Скласти статистичний розподіл вибірки, розбивши увесь діапазон на $k = [1 + 3{,}32 \lg n]$ інтервалів (n − об'єм вибірки).

Розв'язання. 1) Запишемо послідовність варіант у зростаючому порядку, тобто варіаційний ряд

12,12; 13,21; 14,24; 14,30; 14,35; 14,39; 14,55; 14,96; 15,97; 16,38; 16,45; 17,16; 18,48; 18,50; 18,50; 18,81;18,93; 19,04; 19,25; 21,58.

Далі увесь діапазон розбивається на декілька часткових інтервалів, кількість яких обчислюється за формулою $k = [1 + 3{,}32 \lg n]$. У нашому випадку $n = 20$, а $k = [5{,}316] = 5.$ Отже, інтервалів буде п'ять:

12-14, 14-16, 16-18, 18-20, 20-22

однакової довжини $h = \frac{22-12}{5} = 2.$

Визначимо частоти варіант вибірки для кожного часткового інтервалу.

У перший інтервал 12-14 потрапляє 2 значення (12,12; 13,21). Тому $n_1 = 2$; у другий ,14-16, − 7 значень (14,24; 14,30; 14,35; 14,39; 14,55; 14,96; 15,97), тому $n_2 = 7$. Аналогічно: $n_3 = 3, n_4 = 7, n_5 = 1.$

Подамо інтервальний варіаційний ряд випадкової величини у вигляді таблиці :

Інтервали	[12-14[	[14-16[	[16-18[	[18-20[	[20-22[

n_i	2	7	3	7	1
ω_i	2/20=0,1	7/20=0,35	3/20=0,15	7/20=0,35	1/20=0,05
$p^*_i = \frac{\omega_i}{h}$	0,1/2=0,05	0,35/2=**0,175**	0,15/2=0,075	0,35/2=0,175	0,05/2=**0,025**

Контроль: 2+7+3+7+1=20 (сума всіх частот дорівнює об'єму вибірки).

0,1+0,35+0,15+0,35+0,05=1.

1.4. Полігон і гістограма

Графічним представленням для ряду розподілу у випадку дискретних числових даних є ***полігон частот***, а у випадку неперервних – ***гістограма***.

Полігоном частот називають ламану, відрізки якої з'єднують точки (x_1,n_1), (x_2,n_2), , (x_k,n_k). При цьому в прямокутній системі координат по осі абсцис слід відкласти значення x_1, x_2, ..., x_k варіант, а по осі ОУ – частот $n_1,n_2,...,n_k$. Аналогічно будують ***полігон відносних частот.***

Так, для ряду розподілу, що має числові результати

X	3	4	7	8	9
ω_i	0,1	0,2	0,3	0,1	0,3

полігон частот зображений на рис.1.1.

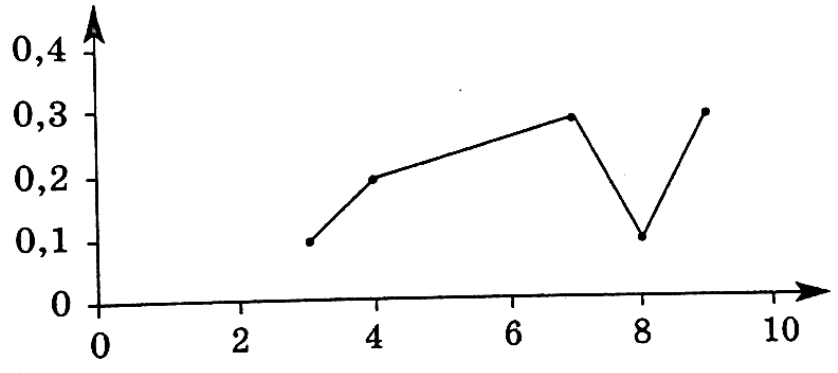


Рис. 1.1.

Функцію, визначену на всій числовій прямій, яка є кусково-сталою зі значенням $\frac{n_i}{h}$ ($\frac{\omega_i}{h}$) на інтервалах групування та нульовою – поза ними, називають ***гістограмою частот*** (***відносних частот***).

Для побудови гістограми по осі OX відкладають частинні інтервали, а над ними проводять відрізки, паралельні осі абсцис на відстані $p_i^* = \frac{n_i}{h}$ (або $\frac{\omega_i}{h}$), $i = 1,2,...,k$ відповідно (рис. 1.2). Площа i -того частинного прямокутника на гістограмі частот дорівнює $hn_i / h = n_i$ − сумі частот варіант i -го інтервалу, а площа гістограми частот дорівнює n − об'єму вибірки (площа гістограми відносних частот дорівнює одиниці).

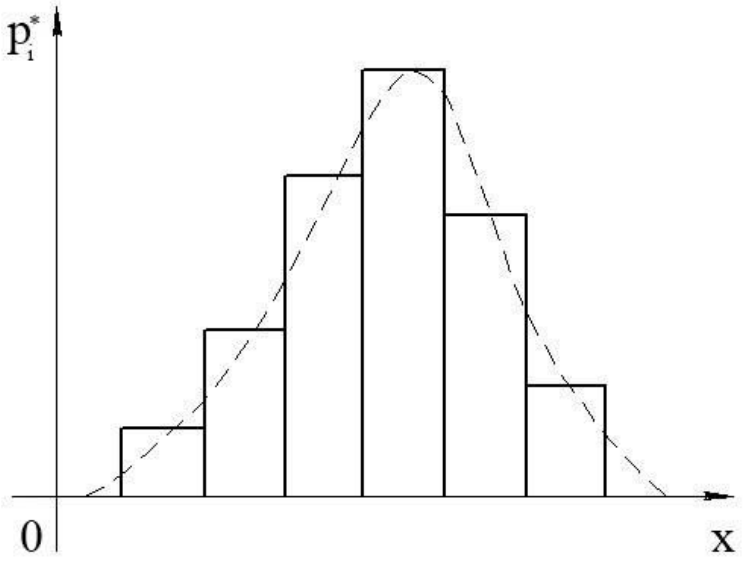


Рис.1.2.

Графічно гістограму зображують у вигляді графіка, що є сходинковою лінією, до якої належать верхні границі стовпчикової діаграми, коли стовпчики розташовані над відповідними інтервалами групування, а їх висотами є $\frac{n_i}{h}$ ($\frac{\omega_i}{h}$).

Якщо довжини частинних інтервалів малі, а об'єм вибірки n великий, то гістограма відносних частот слугує наближенням для щільності випадкової величини X (крива на рис. 1.2 близька до кривої розподілу величини X).

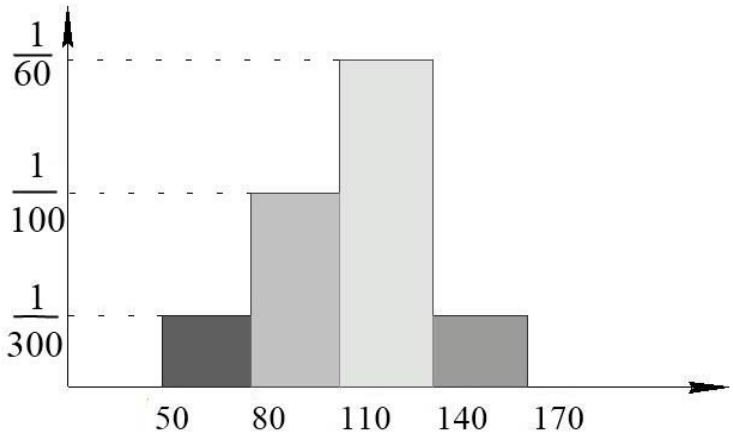
Приклад 1.3. Для групованої вибірки з довжиною інтервалу $h = 30$ побудувати: а) нормовану стовпчикову діаграму; б) гістограму відносних частот:

Інтервал	[50; 80)	[80; 110)	[110; 140)	[140; 170)
ω_i	0,1	0,3	0,5	0,1

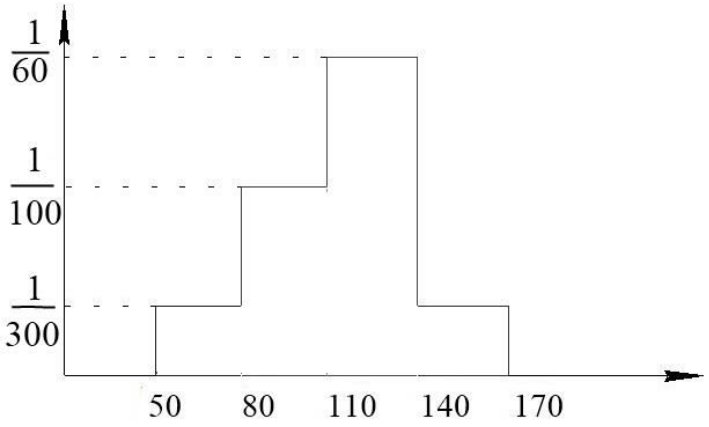
Розв’язання. За формулою $\frac{\omega_i}{h}$ $i=1,2,3,4$ обчислимо щільності відносних частот та доповнимо вихідну таблицю:

Інтервал	[50; 80)	[80; 110)	[110; 140)	[140; 170)
ω_i	0,1	0,3	0,5	0,1
$\frac{\omega_i}{h}$	$\frac{0,1}{30}$	$\frac{0,3}{30}$	$\frac{0,5}{30}$	$\frac{0,1}{30}$

Далі по осі OX відкладають частинні інтервали, а над ними проводять відрізки, паралельні осі абсцис на відстані $\frac{\omega_i}{h}$, $i=1,2,3,4$ відповідно.



За нормованою стовпчиковою діаграмою будують гістограму відносних частот:



1.5. Емпірична функція розподілу. Властивості

Нехай вивчається кількісна ознака X з невідомою функцією

розподілу $F(x)$. За даними вибірки x_1, x_2, .., x_n об'єму n побудований статистичний розподіл випадкової величини X.

Для будь-якого дійсного числа x позначимо через n_x число варіант, менших x, тобто частоту події $X < x$.

Означення. ***Емпіричною функцією розподілу (функцією розподілу вибірки)*** називається функція

$$F^*(x) = \frac{n_x}{n}, \qquad (1.1)$$

яка визначає для кожного значення x відносну частоту події $X < x$.

На відміну від емпіричної функції розподілу вибірки функцію розподілу $F(x)$ генеральної сукупності називають теоретичною функцію розподілу. Різниця між емпіричною і теоретичною функцію розподілу полягає в тому, що теоретична функція $F(x)$ визначає ймовірність появи події $X < x$, а емпірична функція − відносну частоту цієї події.

Емпірична функція $F^*(x)$ вибірки слугує для оцінки теоретичної функції розподілу $F(x)$ генеральної сукупності величини X.

Властивості емпіричної функції розподілу.

1. Значення емпіричної функції розподілу $F^*(x)$ належать відрізку $[0;1]$: $0 \le F^*(x) \le 1$.
2. Емпірична функція розподілу є неспадною функцією:
   при $x_1 < x_2$ $F^*(x_1) \le F^*(x_2)$.
3. Якщо x_1 − найменша варіанта, то $F^*(x) = 0$ при $x \le x_1$;
   якщо x_k − найбільша варіанта, то $F^*(x) = 1$ при $x > x_k$.

Приклад 1.4. За даним розподілом вибірки

X	1	4	6
n_i	10	15	25

побудувати емпіричну функцію розподілу $F^*(x)$.

Розв'язання. Знайдемо об'єм вибірки: $n = 10 + 15 + 25 = 50$. Найменша варіанта дорівнює 1, а найбільша − 6: $x_1 = 1$, $x_3 = 6$. Тому, за властивістю 3 $F^*(x) = 0$ при $x \le 1$ і $F^*(x) = 1$ при $x > 6$.

а) Значення $X < 4$, а саме $x_1 = 1$ спостерігалося 10 разів, тому

$$F^*(x) = \frac{10}{50} = 0{,}2 \text{ при } 1 < x \le 4.$$

б) Значення $X < 6$, а саме $x_1 = 1$ та $x_2 = 4$ спостерігалося

10+15=25 разів, звідси

$$F^*(x)=\frac{25}{50}=0{,}5 \text{ при } 4<x\le 6.$$

Отже, шукана емпірична функція розподілу має вигляд

$$F^*(x)=\begin{cases} 0 \ \textit{при} \ x\le 1; \\ 0{,}2 \ \textit{при} \ 1<x\le 4; \\ 0{,}5 \ \textit{при} \ 4<x\le 6; \\ 1 \ \textit{при} \ x>6, \end{cases}$$

а її графік зображено на рис. 1.3.

Емпірична функція є кусково-сталою функцією. Графік $F^*(x)$ являє сходинкову лінію, яка у напрямку числової осі змінює свої значення від нуля до одиниці.

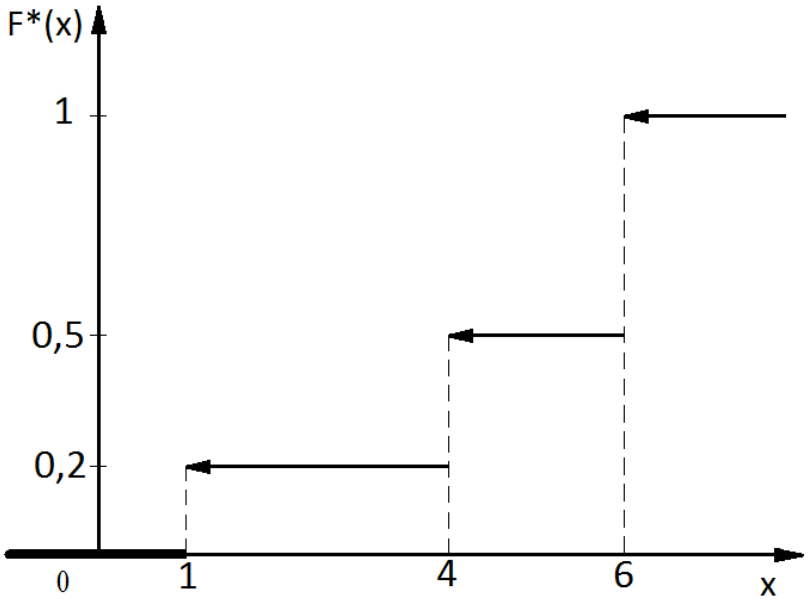


Рис. 1.3.

Зауваження. Розрахунки щодо емпіричної функції розподілу $F^*(x)$ зручно здійснювати за такою таблицею:

№	Варіанти X	Частота n_i	Накопичена частота $\sum n_i$	Відносна частота $W_i=\frac{n_i}{n}$	Накопичена відносна частота $\sum W_i$
1	1	10	10	0,2	**0,2**
2	4	15	10+15=25	0,3	0,2+0,3=**0,5**
3	6	25	10+15+25=50	0,5	0,2+0,3+0,5=**1**

1.6. П’ятиточкова характеристика вибірки

1.6.1. Інтегральні характеристики вибірки

До інтегральних характеристик виду «типовий представник» відносять ***моду***, яку визначають безпосередньо за вибіркою як значення, що найчастіше повторюється у вибірці, або за групованою вибіркою, коли вона є серединою того інтервалу групування, який спостерігають найчастіше (до якого потрапила найбільша кількість спостережувальних значень).

Модою **(**Mo**)** називається варіанта з найбільшою частотою.
Так, для дискретного статистичного розподілу,

X	1	4	**7**	9
n_i	5	1	**20**	6

згідно з означенням, мода дорівнює 7.

Для знаходження *моди* інтервального статистичного розподілу необхідно знайти модальний інтервал, тобто такий інтервал, що має найбільшу частоту. Тоді моду обчислюють за формулою

$$\text{Mo} = x_{i-1} + \frac{n_{\text{Mo}} - n_{\text{Mo}-1}}{2n_{\text{Mo}} - n_{\text{Mo}-1} - n_{\text{Mo}+1}} \cdot h, \qquad (1.2)$$

або

$$\text{Mo} = x_{i-1} + \frac{n_{\text{Mo}} - n_{\text{Mo}-1}}{(n_{\text{Mo}} - n_{\text{Mo}-1}) + (n_{M_0} - n_{\text{Mo}+1})} \cdot h, \qquad (1.3)$$

де x_{i-1} — початок модального інтервалу, h — довжина часткового інтервалу, n_{Mo} — частота модального інтервалу, $n_{\text{Mo}-1}$ — частота домодального інтервалу, $n_{\text{Mo}+1}$ — частота післямодального інтервалу.

Медіаною Me називається значення середнього елемента варіаційного ряду. Це варіанта, що ділить варіаційний ряд на дві частини, рівні за кількістю варіант.

1. Якщо обсяг вибірки $n = 2m + 1$ непарний, то медіаною буде значення елемента варіаційного ряду з номером $m + 1$:

$$\text{Me} = x_{m+1} \qquad (1.4)$$

2. Якщо обсяг вибірки $n = 2m$ парний, то медіаною буде середнє значення елементів варіаційного ряду з номерами m і $m + 1$(«середні за номерами»):

$$\text{Me} = \frac{x_m + x_{m+1}}{2}. \qquad (1.5)$$

Так, для ряду 2,3,5,6,7 ($n = 5$, непарне число) медіана дорівнює 5, а для ряду 2,3,**5,6,**7,9 ($n = 6$, парне число, $n = 2m$, $6 = 2m$, $m = 3$) медіана дорівнює (5+6)/2=11/2=5,5.

Для визначення медіани інтервального статистичного розподілу вибірки необхідно визначити медіанний інтервал.

Якщо на інтервалі $(x_{i-1}; x_i)$ значення функції $F^*(x_{i-1})<0,5$ і $F^*(x_i)>0,5$, то всередині інтервалу $(x_{i-1;}\ x_i)$ існує таке значення X=Me, що $F^*(\text{Me})=0,5$. Тоді медіану обчислюють за формулою

$$\text{Me} = x_{i-1} + \frac{0,5 - F^*(x_{i-1})}{F^*(x_i) - F^*(x_{i-1})} \cdot h\,. \qquad (1.6)$$

Медіанний інтервал можна визначати за графіком емпіричної функції $F^*(x)$.

До інтегральних характеристик вибірки відносять також квартилі (позначають Q_1, Q_2, Q_3) та характеристики, побудовані на їх основі.

Квартилями називають значення досліджуваної характеристики (випадкової величини X), які відтинають послідовно чверть, половину та три чверті менших за величиною значень: Q_1 − нижній (перший квартиль), Q_2 − медіана (другий квартиль), Q_3 − верхній (третий квартиль). Q_3 можна визначити також як значення досліджуваної характеристики, яка відтинає чверть більших за величиною значень.Значення досліджуваної характеристики, яка відтинає половину більших або менших за величиною значень, визначатиме медіану.

Інтерквартальний розмах R_l визначають як різницю між верхнім та нижнім квартилями:

$$R_l = Q_3 - Q_1\,. \qquad (1.7)$$

R_l визначає довжину інтервалу, який містить половину значень елементів вибірки, «середніх за величиною», крім чверті менших та чверті більших за величиною значень.

У **дискретному випадку** квартилі обчислюють за варіаційним рядом вибірки. За Q_1 беруть член варіаційного ряду з порядковим номером $\frac{n}{4}$, якщо n ділиться на 4, і $\left[\frac{n}{4}\right]+1$ − в іншому разі. Це буде елемент вибірки, який відділяє першу чверть елементів варіаційного ряду за номерами.

$$Q_1 = \begin{cases} a^*_{\left[\frac{n}{4}\right]},\ \textit{якщо}\ \left[\frac{n}{4}\right] - \textit{ціле}, \\ a^*_{\left[\frac{n}{4}\right]+1},\ \textit{якщо}\left[\frac{n}{4}\right] - \textit{не ціле}. \end{cases} \qquad (1.8)$$

Третій квартиль обчислюють за формулою:

$$Q_3 = \begin{cases} a^*_{\left[\frac{3n}{4}\right]}, \ якщо \left[\frac{3n}{4}\right] - ціле, \\ a^*_{\left[\frac{3n}{4}\right]+1}, \ якщо \left[\frac{3n}{4}\right] - не\ ціле. \end{cases} \quad (1.9)$$

Другий квартиль Q_2 (***медіану***) обчислюють в залежності від парності чи непарності обсягу вибірки.

1. Якщо обсяг вибірки $n = 2m+1$ непарний, то медіаною буде значення елемента варіаційного ряду з номером $m+1$:

$$\text{Me} = x_{m+1} \quad (1.10)$$

2. Якщо обсяг вибірки $n = 2m$ парний, то медіаною буде середнє значення елементів варіаційного ряду з номерами m і $m+1$(«середні за номерами»):

$$\text{Me} = \frac{x_m + x_{m+1}}{2}. \quad (1.11)$$

Приклад 1.5. Для вибірки зросту

170, 189, 182, 184, 182, 172, 181, 178, 171, 180, 183, 173, 181, 173, 189, 176, 179, 191, 186, 192
знайти квартилі.
Розв'язання. Запишемо варіаційний ряд:

170,171,172,173,**173**,176,178,179,180,**181,181**,182,182,183,**184**,186,189,189,191,192,

$n = 20 = 2\kappa$ – парне, $\kappa = 10$, тому: медіана $Q_2 = = \frac{a_{10} + a^*_{10+1}}{2} = \frac{181+181}{2} = 181$,

$\frac{n}{4} = \frac{20}{4} = 5$, $\left[\frac{n}{4}\right] = [5] = 5$, $Q_1 = a^*_5 = 173$, $\frac{3n}{4} = \frac{3\cdot 20}{4} = \frac{60}{4} = 15$, $\left[\frac{3n}{4}\right] = [15] = 15$, $Q_3 = a^*_{15} = 184$.

(Якщо, наприклад, об'єм вибірки n=15, $\frac{3n}{4} = \frac{3\cdot 15}{4} = \frac{45}{4} = 11{,}25$, $\left[\frac{3n}{4}\right] = [11{,}25] = 11$, то $Q_3 = a^*_{11+1} = a^*_{12} = 182$).

Квартилі для **неперервного випадку** обчислюють за групованою вибіркою. Інтервали, до яких потрапили квартилі називають ***квартильними.***
Квартилі обчислюють за формулами:

$$M_l = Q_2 = x_{M_l} + h_{M_l} \cdot \frac{\frac{n}{2} - m^*_{M_l}}{m_{M_l}},$$

де x_{M_l} – початок медіанного інтервалу, h_{M_l} – його довжина, $m^*_{M_l}$ – абсолютна частота, накопичена до початку медіанного інтервалу, m_{M_l} – абсолютна частота медіанного інтервалу.

Перший та третій квартилі обчислюють так:

$$Q_1 = x_{Q_1} + h_{Q_1} \cdot \frac{\frac{n}{4} - m^*_{Q_1}}{m_{Q_1}},\ Q_3 = x_{Q_3} + h_{Q_3} \cdot \frac{\frac{3n}{4} - m^*_{Q_3}}{m_{Q_3}},$$

де x_{Q_1} та x_{Q_3} – початки першого та третього квартильних інтервалів, h_{Q_1} та h_{Q_3} – довжини першого та третього інтервалів (у загальному випадку $h_{Q_1} = h_{M_l} = h_{Q_3}$, але якщо інтервал містить недостатню кількість елементів, його об'єднують з сусіднім, тому іноді довжина інтервалу може бути кратною h), $m^*_{Q_1}$ та $m^*_{Q_3}$ – абсолютні частоти, накопичені до початку першого та третього інтервалів, m_{Q_1} та m_{Q_3} – абсолютні частоти першого та третього інтервалів.

1.6.2. П'ятиточкова характеристика вибірки та її зображення

Складена інтегральна характеристика, яка містить п'ять чисел, є п'ятиточковою характеристикою вибірки: максимальне та мінімальне значення вибірки та три квартилі, записані у порядку зростання:

$$x_{\min},\ Q_1,\ Q_2,\ Q_3,\ x_{\max}.$$

П'ятиточкову характеристику зображають у вигляді рисунка, який називають «ящик з вусами» (від англ.терміну «box and whiskey plot»). Цей рисунок є прямокутником з перетином і сторонами, паралельними осям координат, та відрізками, паралельними осі Ox, що відходять від середин бічних сторін прямокутника. Прямокутник з відрізками розташований так, що бічні сторони спроектовані на точки осі Ox – квартилі, а кінці горизонтальних відрізків – на точки, що є мінімальними та максимальними значеннями вибірки.

Так, для п'ятиточкової характеристики прикладу 1.5

$$x_{\min} = 170,\ Q_1 = 173,\ Q_2 = 181,\ Q_3 = 184,\ x_{\max} = 192$$

«ящик з вусами» має вигляд

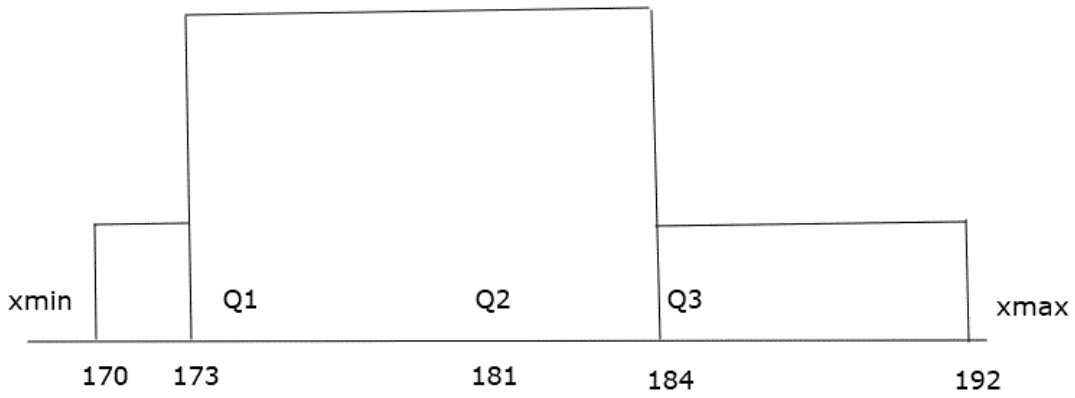


Рис. 1.4.

Контрольні запитання.

1. Як означається генеральна сукупність? Вибірка?
2. У чому полягає суть вибіркового методу?
3. Що називають варіантами? Варіаційним рядом? Розмахом вибірки?
4. Що називають частотою, відносною частотою варіант?
5. Яка вибірка називається згрупованою? Незгрупованою?
6. Яку вибірку називають повторною? Безповторною?
7. Який відбір називають простим випадковим? Механічним? Наведіть приклади.
8. Як означається статистичний розподіл вибірки? Як побудувати дискретний варіаційний ряд? Інтервальний варіаційний ряд?
9. Що являють собою полігон (відносних) частот, гістограма (відносних) частот вибірки?
10. Що називають діаграмою? Опишіть основні типи діаграм.
11. Як означається емпірична функція розподілу?
12. Навести основні властивості емпіричної функції розподілу.
13. Як зв’язані емпірична та теоретична функції розподілу?
14. Що називають модою? Модальним інтервалом?
15. Запишіть формулу знаходження моди для інтервального статистичного розподілу.
16. Як означають квартилі досліджуваної характеристики (випадкової величини)?
17. Що називають медіаною? медіанним інтервалом?
18. Як знайти медіану для дискретного статистичного розподілу? Опишіть алгоритм.
9. Запишіть формулу знаходження медіани для інтервального статистичного розподілу.
10. До якого квартилю (Q_1, Q_2, Q_3) відносять поняття медіани?

Тема 2. Закон стійкості частот. Частотне визначення ймовірностей. Завдання ймовірностей для дискретних та неперервних даних

2.1. Відносна частота події

Відносною частотою $W(A)$ ***події*** A називають відношення числа m елементарних результатів досліду, в яких подія A відбулась до загального числа n фактично проведених дослідів:

$$W(A)=\frac{m}{n}. \qquad (2.1)$$

На відміну від класичного означення ймовірності, означення відносної частоти припускає, що експеримент здійснено фактично. Іншими словами: ймовірність $P(A)$ обчислюють до проведення досліду, а відносну частоту $W(A)$ − після досліду.

Відносна частота події має ті ж самі найпростіші властивості, що і ймовірність $P(A)$: $0 \le W(A) \le 1$.

Емпіричним шляхом встановлено, що коли в однакових умовах проводити серії дослідів, збільшуючи їх кількість, відносна частота події дуже мало змінюється, коливаючись навколо деякого сталого числа − ймовірності цієї події. Така властивість відносної частоти називається її стійкістю.

Закон стійкості частот охоплює такі положення:

1. Відносна частота $W_n(A)$ будь-якої події A за нескінченного збільшення об'єму n вибірки наближується до граничного значення $P(A)$:

$$\lim_{n\to\infty} W_n(A)=P(A).$$

2. Граничне значення $P(A)$ не залежать від вибірки, за якою отримане, а характеризує явище загалом.

За законом стійкості частот визначають ймовірність: ймовірністю події A ($P(A)$) називають граничне значення відносної частоти за нескінченного збільшення об'єму вибірки.

Ймовірність, як і частота, є функцією подій, заданою на просторі елементарних подій U.

За законом стійкості, при достатньо великій кількості випробувань, відносну частоту події A можна прийняти за наближене значення ймовірності цієї події:

$$P(A)\approx W(A). \qquad (2.2)$$

Співвідношення (2.2) характеризує передбачувальну властивість ймовірності.

Передбачувальна властивість ймовірності полягає в тому, що значення ймовірності, отримане за однією вибіркою, уможливлює характеристику частоти іншої послідовності спостережень (вибірки) того самого масового явища.

На підставі рівності (2.2) разом з класичним використовують статистичне означення ймовірності, за яким в якості статистичної ймовірності події A приймають $W(A)$ − відносну частоту або число, близьке до неї.

Приклад 2.1. Відділ технічного контролю (ВТК) виявив 5 бракованих книг у партії з випадково відібраних 100 книг. Знайти відносну частоту появи бракованих книг.

Розв'язання. Подія A − «поява бракованої книги у партії з випадково відібраних 100 книг», $n=100$, $m=5$, $W(A)=\frac{5}{100}=0{,}05$.

2.2. Поняття про геометричні ймовірності

Геометричні ймовірності − це ймовірності попадання точки в область (відрізок, плоску область, просторову область (тіло) і т.і.).

Нехай, наприклад, на площині є деяка область g, що міститься всередині області G (рис.2.1). На фігуру G кинуто навмання точку. Тоді ймовірність попадання точки в область g визначається рівністю

$$P=\frac{площа\ g}{площа\ G}. \qquad (2.3)$$

При цьому: 1) кинута точка може опинитися в будь-якій точці області G; 2) ймовірність попадання точки на фігуру g пропорційна площі цієї фігури і не залежить від її форми та розташування.

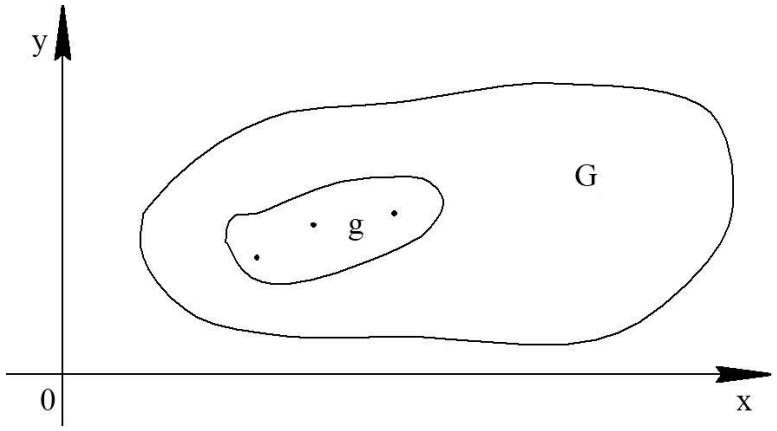


Рис.2.1.

Аналогічно означають ймовірності попадання точки на частину відрізка чи в деяку просторову область. Якщо позначити міру (довжину, площу, об'єм) області через mes, ймовірність попадання

точки в область g − частину області G визначається так:

$$P = \frac{mes\ g}{mes\ G}. \qquad (2.4)$$

Приклад 2.2. Під час бурі на ділянці між 10-м і 30-м кілометрами телефонної лінії стався обрив. Яка ймовірність того, що розрив лінії знаходиться між 20-м і 25-м кілометрами?

Розв'язання. Подія A − «розрив лінії знаходиться між 20-м і 25-м кілометрами». Ділянка між 20-м і 25-м кілометрами (область g) знаходиться всередині ділянки між 10-м і 30-м кілометрами (область G). Тому

$$P(A) = \frac{mes\ g}{mes\ G} = \frac{25-20}{30-10} = \frac{5}{20} = \frac{1}{4}.$$

Тут в якості міри областей g та G приймають їх довжини (рис.2.2).

Рис.2.2.

2.3. Завдання ймовірностей для дискретних та неперервних даних

Спосіб завдання ймовірностей залежить від дискретного чи неперервного типу даних. Так, у випадку дискретних даних (дискретного простору елементарних подій) ймовірність задають рядом розподілу ймовірностей, а у випадку неперервних даних, коли простором елементарних подій є відрізок числової осі, напівінтервал або вся числова вісь − функцією розподілу ймовірностей (граничним значенням емпіричної функції розподілу) або щільністю розподілу ((граничним значенням гістограми).

2.3.1. Ряд розподілу ймовірностей

Рядом розподілу ймовірностей **(*рядом розподілу)*** для дискретної випадкової величини (дискретних даних) називають відповідність між можливими значеннями та їх ймовірностями.

Якщо дискретна випадкова величина (ДВВ) X може приймати скінченне число можливих значень x_1, x_2,…, x_n відповідно з ймовірностями p_1, p_2,…, p_n то її закон розподілу можна подати таблицею

Таблиця 1.

X	x_1	x_2	x_3	…	x_n
p	p_1	p_2	p_3	…	p_n

Оскільки в кожному випробуванні випадкова величина X приймає одне і тільки одне можливе значення, події $X = x_1$, $X = x_2$,…, $X = x_n$ утворюють повну групу. Тому, сума ймовірностей цих подій дорівнює одиниці:

$$p_1 + p_2 + ,\ldots, + p_n = 1. \qquad (2.5)$$

Якщо множина можливих значень X є нескінченною, то числовий ряд $p_1 + p_2 + ,\ldots, + p_n + \ldots$ збігається до одиниці: $\sum_{i=1}^{\infty} p_i = 1$.

2.3.2. Функція розподілу ймовірностей

Розподіл ймовірностей у випадку неперервних даних задають функцією розподілу або щільністю розподілу ймовірностей.

Функцією розподілу (інтегральною функцією розподілу) ймовірностей випадкової величини X називається функція

$$F(x) = P(X < x) \qquad (2.6)$$

змінної x, яка визначає ймовірність того, що величина X в результаті випробування прийме значення менше, ніж число x.

Графік інтегральної функції розподілу називається інтегральною кривою розподілу. Функція розподілу є граничним значенням емпіричної функції розподілу:

$$F(x) = \lim_{n\to\infty} F_n^*(x),$$

де n – об'єм вибірки, і має всі її характеристичні та різницеві властивості.

Властивості функції розподілу.

1. Значення інтегральної функції розподілу належать відрізку $[0;1]$:
$0 \le F(x) \le 1$.

2. Інтегральна функція розподілу є неспадною функцією, тобто при $x_1 < x_2$ $F(x_1) \le F(x_2)$.

Наслідок 1. Ймовірність того, що випадкова величина X прийме значення з проміжку $[a,b)$ дорівнює приросту інтегральної функції розподілу на цьому проміжку

$$P(a \le X < b) = F(b) - F(a). \qquad (2.7)$$

Наслідок 2 (парадокс неперервності). Ймовірність того, що неперервна випадкова величина (НВВ) X прийме одне певне значення, дорівнює нулю:

$$P(X = x_1) = 0. \qquad (2.8)$$

Рівність (2.8) не означає, що подія $X = x_1$ неможлива. В результаті досліду величина X прийме одне з можливих значень, зокрема, цим значенням може бути і x_1, але ймовірність події $X = x_1$ нульова.

Отже, не має сенсу говорити про ймовірність того, що неперервна випадкова величина прийме одне певне значення, але має сенс розглядати ймовірність того, що НВВ потрапляє в інтервал, навіть скіль завгодно малий. Цей факт цілком відповідає вимогам прикладних задач. Наприклад, цікавляться ймовірністю того, що розміри деталей не виходять за встановлені межі, але не ставлять питання про ймовірність їх рівності з проектним розміром.

3. Якщо всі можливі значення випадкової величини належать інтервалу (a,b), то $F(x)=0$ при $x \le a$ та $F(x)=1$ при $x \ge b$.

Наслідок. Якщо можливі значення НВВ розташовані на всій осі x, то

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

На рис. 2.4 проілюстровано інтегральну криву розподілу у випадку, коли можливі значення НВВ належать $(-\infty,+\infty)$ та (a,b).

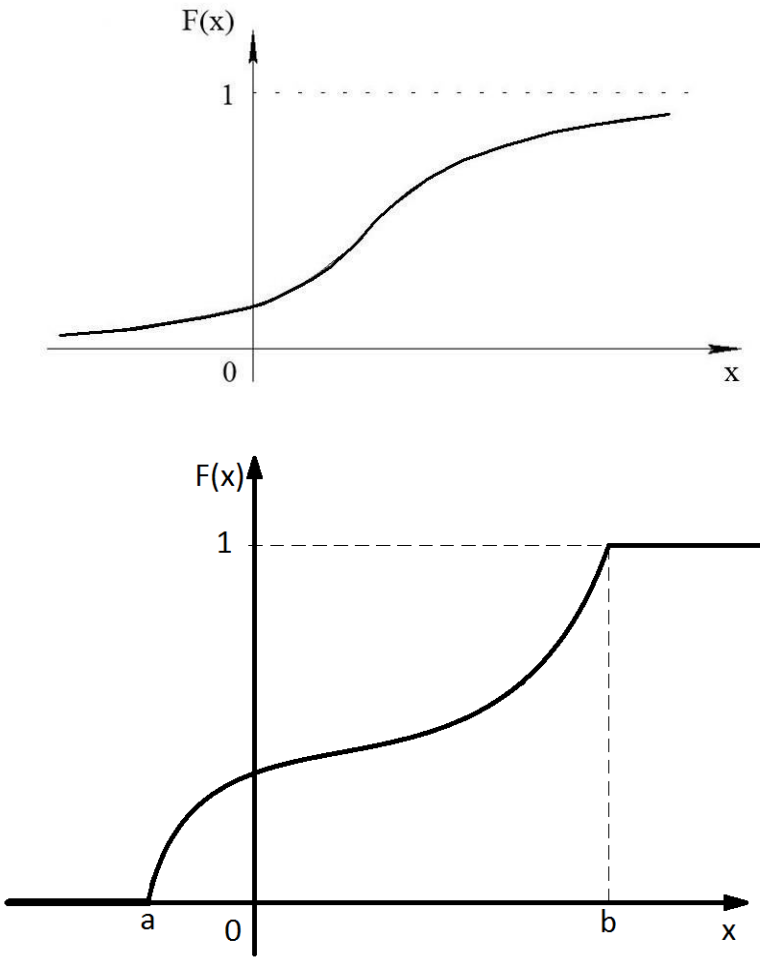


Рис.2.4

2.3.3. Щільність (диференціальна функція) розподілу

Щільність розподілу є граничним значенням гістограми. Згідно з цією властивістю щільністю розподілу ймовірностей називають невід'ємну функцію $f(x)$, за якою ймовірність будь якої події обчислюють як площу під графіком гістограми.

Нехай X − неперервна випадкова величина (НВВ), $F(x)$ − її інтегральна функція, диференційовна всюди, за винятком, можливо, скінченного числа точок.

Означення. Перша похідна від інтегральної функції розподілу

$$f(x) = F'(x) \tag{2.9}$$

називається ***диференціальною функцією розподілу*** (або ***щільністю розподілу ймовірностей***) неперервної випадкової величини X .

Графік диференціальної функції $y = f(x)$ називається ***кривою ймовірності*** (або кривою розподілу випадкової величини X):

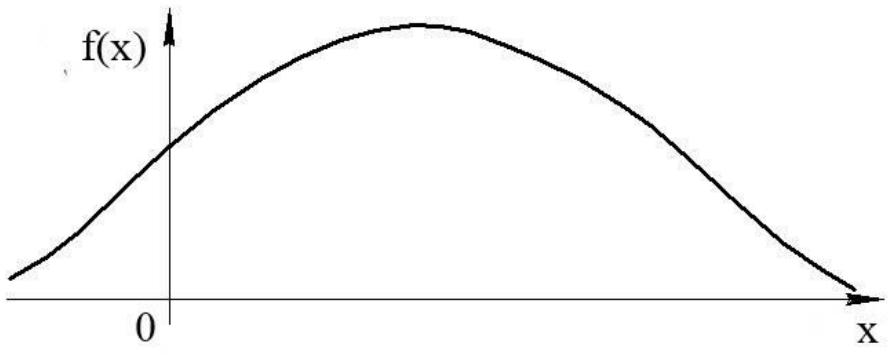


Рис. 2.5.

Властивості диференціальної функції розподілу

1. Диференціальна функція є невід'ємною в точках, де вона існує:

$$f(x) \geq 0 .$$

Дійсно, оскільки $f(x) = F'(x)$, а $F(x)$ − неспадна функція, то її похідна $f(x)$ невід'ємна.

2. Ймовірність $P(a < X < b)$ того, що НВВ X прийме значення з інтервалу $(a;b)$ дорівнює визначеному інтегралу від диференціальної функції, взятому в межах від a до b :

$$P(a < X < b) = \int_a^b f(x)dx . \tag{2.10}$$

Геометрично рівність (2.10) означає, що ймовірність $P(a < X < b)$ дорівнює площі криволінійної трапеції, яка прилягає до осі OX та

проектується у відрізок $(a;b)$ цієї осі (рис. 2.6).

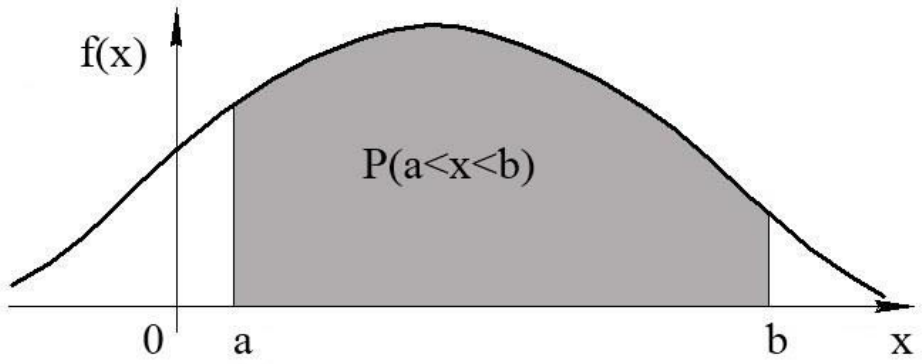


Рис.2.6.

3. Інтегральна функція розподілу $F(x)$ виражається через диференціальну за формулою

$$F(x)=\int_{-\infty}^{x} f(x)dx\,. \qquad (2.11)$$

Дійсно, за означенням $F(x)=P(X<x)$. Очевидно, нерівність $X<x$ рівносильна подвійній $-\infty<X<x$, тобто $F(x)=P(-\infty<X<x)$. Поклавши в формулі (2.10) $a=-\infty$, $b=x$, приходимо до (2.11).

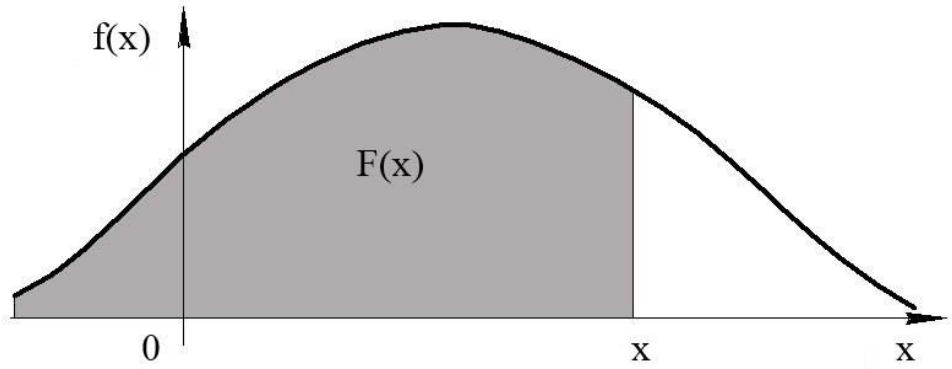


Рис.2.7.

4. Якщо $f(x)$ − диференціальна функція розподілу, то

$$\int_{-\infty}^{+\infty} f(x)dx=1\,. \qquad (2.12)$$

Невластивий інтеграл $\int_{-\infty}^{+\infty} f(x)dx$ виражає ймовірність того, що випадкова величина X прийме значення з інтервалу $(-\infty;+\infty)$. Очевидно, така подія достовірна, тому ймовірність її дорівнює

одиниці.

Наслідок. Зокрема, якщо всі можливі значення випадкової величини X належать інтервалу $(a;b)$, то

$$\int_a^b f(x)dx = 1 .$$

2.4. Інтегральні характеристики розподілу ймовірностей

Розглянемо спочатку інтегральні характеристики на базі середнього (поділ за засобами побудови). Такі характеристики є граничними значеннями частотних відповідників.

Як і для розподілу частот, розрізняють два основні типи інтегральних характеристик: 1) «типові представники» (або «міри центральної тенденції»); 2) «розсіювання» або «міри варіації».

2.4.1. Математичне сподівання.

Означення 1. Математичним сподіванням дискретної випадкової величини (***ДВВ***) називається сума добутків всіх її можливих значень на їх ймовірності.

Якщо випадкова величина X може приймати тільки значення x_1, x_2,…, x_n відповідно з ймовірностями p_1, p_2,…, p_n, то математичне сподівання величини X, що позначається символом $M(X)$, обчислюється за формулою

$$M(X) = x_1 p_1 + x_2 p_2 + ... + x_n p_n = \sum_{i=1}^{n} x_i p_i . \qquad (2.13)$$

Якщо ДВВ X приймає нескінченну множину можливих значень, то

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

причому ряд праворуч має бути абсолютно збіжним.

Як випливає із означення математичне сподівання ДВВ є величина невипадкова.

Ймовірнісний зміст $M(X)$***:*** математичне сподівання випадкової величини X наближено дорівнює (тим точніше, чим більше число випробувань) середньому арифметичному $\overline{X}$ всіх значень, які прийняла випадкова величина, тобто

$$M(X) \approx \overline{X} .$$

Нехай X − неперервна випадкова величина, що має

диференціальну функцію розподілу $f(x)$.

Математичним сподіванням НВВ X називається число

$$M(X)=\int_{-\infty}^{+\infty} xf(x)dx, \tag{2.14}$$

причому невластивий інтеграл (2.14) збігається абсолютно, тобто існує інтеграл $\int_{-\infty}^{+\infty}|x|f(x)dx$.

Зокрема, якщо всі можливі значення випадкової величини X належать проміжку $[a;b]$, то

$$M(X)=\int_{a}^{b} xf(x)dx. \tag{2.15}$$

Основні властивості $M(X)$**.**

1.Математичне сподівання сталої величини C дорівнює сталій:

$$M(C)=C,\ C=const.$$

2. Сталий множник C можна виносити за знак математичного сподівання

$$M(CX)=C\,M(X),\ C=const.$$

3. Математичне сподівання суми (різниці) двох випадкових величин X, Y дорівнює сумі (різниці) математичних сподівань цих величин

$$M(X\pm Y)=M(X)\pm M(Y).$$

4. Математичне сподівання добутку двох незалежних випадкових величин X, Y дорівнює добутку їх математичних сподівань

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

2.4.2. Дисперсія ДВВ

Означення 2. Дисперсією (розсіюванням) дискретної випадкової величини (***ДВВ***) називається математичне сподівання квадрата відхилення випадкової величини від її математичного сподівання. Дисперсія випадкової величини позначається через $D(X)$. Отже, згідно з означенням

$$D(X)=M[X-M(X)]^2. \tag{2.16}$$

Крім формули (2.16), для обчислення дисперсії застосовують і таку:

$$D(X)=M(X^2)-[M(X)]^2. \tag{2.17}$$

Основні властивості дисперсії.

1.Дисперсія сталої величини дорівнює нулю:

$$D(C)=0,\ C=const.$$

2. Сталий множник C можна виносити за знак дисперсії, піднісши його до квадрата

$$D(CX)=C^2\ D(X),\ C=const.$$

3. Дисперсія суми (різниці) двох незалежних випадкових величин X, Y дорівнює сумі дисперсій цих величин

$$D(X\pm Y)=D(X)+\ D(Y).$$

Дисперсією неперервної випадкової величини X, що має диференціальну функцію розподілу $f(x)$, називається число

$$D(X)=\int_{-\infty}^{+\infty}[x-M(X)]^2 f(x)dx.$$

Якщо всі можливі значення випадкової величини X належать проміжку $[a;b]$, то

$$D(X)=\int_{a}^{b}[x-M(X)]^2 f(x)dx. \qquad (2.18)$$

Дисперсію можна обчислювати також за формулою

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

або відповідно за формулою

$$D(X)=\int_{a}^{b}x^2 f(x)dx-[M(X)]^2. \qquad (2.19)$$

2.4.3. Середнє квадратичне відхилення

Означення 3. Середнім квадратичним відхиленням $\sigma(X)$ випадкової величини X називають число, що дорівнює квадратному кореню з її дисперсії

$$\sigma(X)=\sqrt{D(X)}. \qquad (2.20)$$

Приклад 2.3. Заданий закон розподілу дискретної випадкової величини X:

X	1	2	3
p	0,3	0,1	0,6

Знайти дисперсію та середнє квадратичне відхилення величини X.

Розв'язання. Знайдемо математичне сподівання ДВВ X:

$$M(X) = 1 \cdot 0{,}3 + 2 \cdot 0{,}1 + 3 \cdot 0{,}6 = 2{,}3 .$$

1 спосіб обчислення $D(X)$. Запишемо закони розподілу для відхилення $X - M(X)$ та квадрата відхилення випадкової величини від її математичного сподівання

$X - M(X)$	1-2,3=-1,3	2-2,3=-0,3	3-2,3=0,7
p	0,3	0,1	0,6

$[X - M(X)]^2$	$(-1{,}3)^2 = 1{,}69$	$(-0{,}3)^2 = 0{,}09$	$(0{,}7)^2 = 0{,}49$
p	0,3	0,1	0,6

Тоді, за формулою (2.16),

$$D(X) = 1{,}69 \cdot 0{,}3 + 0{,}09 \cdot 0{,}1 + 0{,}49 \cdot 0{,}6 = 0{,}81 .$$

2 спосіб обчислення $D(X)$. За законом розподілу ДВВ X запишемо закон розподілу величини X^2:

X^2	$1^2 = 1$	$2^2 = 4$	$3^2 = 9$
p	0,3	0,1	0,6

Тоді $M(X^2) = 1 \cdot 0{,}3 + 4 \cdot 0{,}1 + 9 \cdot 0{,}6 = 6{,}1$, $[M(X)]^2 = (2{,}3)^2 = 5{,}29$. За формулою (2.17) знаходимо дисперсію: $D(X) = 6{,}1 - 5{,}29 = 0{,}81$.

Згідно з (2.20) середнє квадратичне відхилення $\sigma(X) = \sqrt{0{,}81} = 0{,}9$.

2.5. Інтегральні характеристики не на базі середнього

До цього типу характеристик відносять квартилі (серед них медіана), інтерквартильний розмах та моду. Ці характеристики для випадкових величин називають так само, як для розподілів частот. Значення досліджуваної характеристики (випадкової величини *X*), які відтинають за ймовірністю відповідно чверть, половину та три чверті менших за величиною значень, називають ***квартилями*** розподілів ймовірностей (Q_1, Q_2, Q_3).

Другий квартиль, як і у частотному випадку, називають

медіаною. Медіану відносять до «типових представників» (або «міри центральної тенденції»). Q_3 може бути визначений як значення досліджуваної характеристики, яке відтинає одну чверть більших за ймовірністю значень.

Величину $R_j = Q_3 - Q_1$, де напівінтервал $[Q_1; Q_3)$ містить 50% середніх за ймовірністю значень досліджуваної характеристики, тобто за винятком j найбільших та j найменших, називають ***інтерквартильним розмахом***. Його відносять інтегральних характеристик «розсіювання» або «міри варіації».

Точку максимуму щільності розподілу у неперервному випадку називають ***модою розподілу ймовірностей***. У дискретному випадку мода − це значення елемента, для якого ймовірність ряду розподілу має найбільше значення.

Контрольні запитання

1. Що називають випадковою подією?
2. Які події називаються достовірними і неможливими? Які події називають несумісними, єдино можливими, рівно можливими?
3. У чому полягає класичне означення ймовірності і коли його застосовують?
4. Як означається відносна частота події? У чому полягає її зв'язок із класичним означенням ймовірності?
5. Сформулювати властивість стійкості.
6. Як означається геометрична ймовірність? Сформулювати постановку задачі та навести необхідні формули.
7. Яка випадкова величина називається дискретною? Неперервною?
8. Як задають розподіл дискретної випадкової величини?
9. Які дискретні випадкові величини називають незалежними?
10. Які основні числові характеристики дискретної випадкової величини? Дати їх означення.
11. У чому полягає ймовірнісний зміст математичного сподівання?
12. Які основні властивості математичного сподівання? Дисперсії?
13. Навести основні закони розподілу дискретної випадкової величини.
14. За якими формулами визначаються числові характеристики біномного закону? Закону розподілу Пуассона?
15. Як означається інтегральна функція розподілу? Для завдання яких величин вона застосовується?
16. Навести основні властивості інтегральної функції розподілу. Які з них притаманні лише неперервній випадковій величині?
17. Дати означення диференціальної функції розподілу неперервної випадкової величини. Сформулювати властивості.

Тема 3. Основні дискретні та неперервні розподіли

3.1. Дискретні розподіли

3.1.1. Біномний закон розподілу

Нехай здійснюється n незалежних випробувань, в кожному з яких подія A може з'явитись або ж ні. Ймовірність появи події A в кожному випробуванні стала і дорівнює p, $0 < p < 1$, $q = 1 - p$.

Розглянемо дискретну випадкову величину (ДВВ) X – «число появ події A в n випробуваннях», її можливі значення: 0, 1, 2, …, n. Ймовірності цих значень при наявності певних вимог для p, n ($p \geq 0{,}1$; $n < 25$) знайдемо за формулою Бернуллі

$$P(X = k) = P_n(k) = C_n^k \cdot p^k \cdot q^{n-k},\ k = 0,1,2,...,n. \qquad (3.1)$$

Формула (3.1) є аналітичним виразом закону розподілу величини X.

Біномним називають закон розподілу ймовірностей, що визначається за формулою Бернуллі. Права частина рівності (3.1) є загальним членом розкладу бінома Ньютона і

$$C_n^n \cdot p^n + C_n^{n-1} \cdot p^{n-1} \cdot q^1 + ... + C_n^k \cdot p^k \cdot q^{n-k} + ... + C_n^0 \cdot p^0 \cdot q^n = (p+q)^n = 1.$$

Числові характеристики біномного закону розподілу визначаються за формулами

$$M(X) = np,\ D(X) = npq, \sigma(X) = \sqrt{npq}. \qquad (3.2)$$

3.1.2. Закон розподілу Пуасона

Якщо в схемі незалежних повторних випробувань n велике, а p або $q = 1 - p$ прямують до нуля ($p \leq 0{,}1$), то біномний закон апроксимується розподілом Пуассона

$$P(X = k) = P_n(k) = \frac{\lambda^k}{k!} \cdot e^{-\lambda},\ \lambda = np,\ k = 0,1,2,...,n. \qquad (3.3)$$

Дисперсія випадкової величини, розподіленої за законом Пуасона, дорівнює її математичному сподіванню (λ):

$$D(X) = M(X) = \lambda. \qquad (3.4)$$

Розподіл Пуасона описує також кількість незалежних подій, що відбуваються у випадкові моменти часу зі сталою інтенсивністю (***потік подій***). ***Інтенсивністю потоку*** λ_1 називається середнє число подій, що з'являються в одиницю часу. Якщо стала λ_1 відома, то ймовірність появи k подій потоку за час t обчислюється за формулою

Пуасона

$$P_t(k) = \frac{\lambda^k}{k!} \cdot e^{-\lambda},\ k = 0,1,2,...,n,\ \text{де}\ \lambda = \lambda_1 t.$$

3.1.3. Геометричний розподіл

Нехай незалежні випробування, в кожному з яких ймовірність появи події A стала і дорівнює p, $0 < p < 1$, проводять до першого настання події A. Отже, якщо подія A з'явилась в k-му випробуванні, то в попередніх $(k-1)$ її не було.

Позначимо через X ДВВ − число випробувань, які потрібно провести до першої появи події A. Можливими значеннями величини X є натуральні числа: 1, 2, …

Якщо в перших $(k-1)$ випробуваннях подія A не з'явилась, а в k-му випробуванні настала, то ймовірність такої «складеної події» за теоремою множення ймовірностей для незалежних подій

$$P(X = k) = q^{k-1} \cdot p,\ k = 1,2,... \qquad (3.5)$$

Поклавши в формулі (3.5) $k = 1,2,...$, n ,…, дістанемо геометричну прогресію

$$p,\ q \cdot p,\ q^2 \cdot p,\ \ldots,\ q^n \cdot p,\ldots$$

зі знаменником $q = 1 - p$ та одиничною сумою: $S = \frac{p}{1-q} = 1$. За цією причиною розподіл (3.5) називають ***геометричним***.

Числові характеристики геометричного розподілу визначаються за формулами

$$M(X) = \frac{1}{p},\ D(X) = \frac{1-p}{p^2}. \qquad (3.6)$$

Приклад 3.1. Стрільба з гармати здійснюється до першого улучання. Ймовірність улучання в ціль дорівнює 0,6. Знайти ймовірність того, що улучання відбудеться при третьому пострілі.

Розв'язання. Маємо схему повторних випробувань, X − число пострілів, які потрібно провести до першого улучання, $p = 0,6$; $q = 1 - 0,6 = 0,4$; $k = 3$. Тоді, ймовірність того, що випадкова величина X прийме значення 3 знаходимо за формулою (3.5):

$$P(X = 3) = 0,4^{3-1} \cdot 0,6 = 0,096.$$

3.1.4. Гіпергеометричний розподіл

Гіпергеометричний розподіл описує ймовірність настання m успішних результатів у n випробуваннях, якщо значення n мале порівняно з обсягом сукупності N:

$$P(X=m)=\frac{C_k^m \cdot C_{N-k}^{n-m}}{C_N^n},\ m=0,1,2,...,n;\ k\geq n. \qquad (3.7)$$

Наприклад, ймовірність того, що з n деталей, які випадково вибрано з партії обсягом N, m деталей виявляться бракованими, має гіпергеометричний закон розподілу (k − кількість бракованих деталей у партії).

Числові характеристики гіпергеометричного розподілу:

$$M(X)=\frac{kn}{N},\ D(X)=\frac{nk(N-k)(N-n)}{N^2(N-1)}.$$

Зі зменшенням відношення $\frac{n}{N}$ гіпергеометричний розподіл наближається до біномного з параметрами n і $p=\frac{k}{N}$. Дуже часто гіпергеометричний розподіл апроксимується розподілом Пуассона, якщо $\lambda=\frac{nk}{N}$.

3.2. Основні неперервні розподіли

3.2.1. Рівномірний закон розподілу

Рівномірним називають закон розподілу неперервної випадкової величини X, заданої на проміжку $[a;b]$ щільністю

$$f(x)=\begin{cases}0, & x<a;\\ \dfrac{1}{b-a}, & a\leq x\leq b;\\ 0, & x>b.\end{cases} \qquad (3.8)$$

Отже, при рівномірному розподілі на проміжку, що містить всі можливі значення НВВ, щільність розподілу зберігає стале значення.

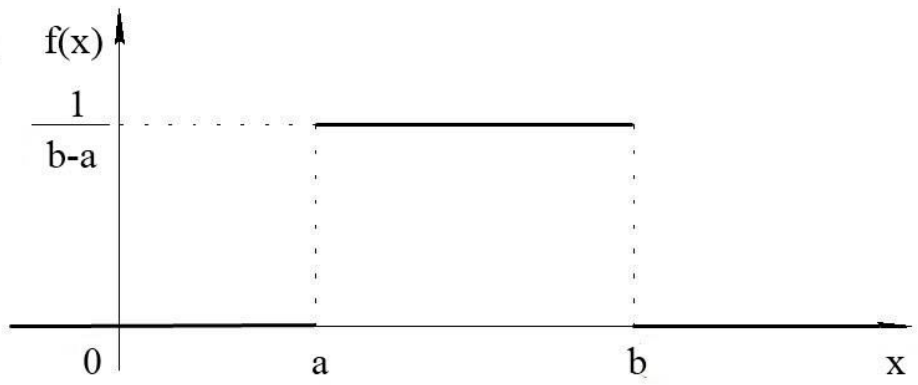


Рис.3.1.

Так, наприклад, похибку при округленні до найближчого цілої поділки на шкалі вимірювального приладу можна розглядати як випадкову величину X, яка може приймати зі сталою щільністю ймовірності будь-яке значення між двома сусідніми поділками. Тому величина X має рівномірний розподіл.

Рівномірно розподілена на відрізку $[a;b]$ випадкова величина X приймає значення тільки з цього проміжку. Її інтегральна функція визначається так:

$$F(x)=\begin{cases} 0, & x \le a; \\ \dfrac{x-a}{b-a}, & a < x < b; \\ 1, & x \ge b, \end{cases}$$

а числові характеристики − за формулами

$$M(X)=\frac{b+a}{2},\ D(X)=\frac{(b-a)^2}{12},\ \sigma(X)=\frac{b-a}{2\sqrt{3}}. \qquad (3.9)$$

Графіки функцій $f(x)$ і $F(x)$ представлено на рис.3.1, 3.2.

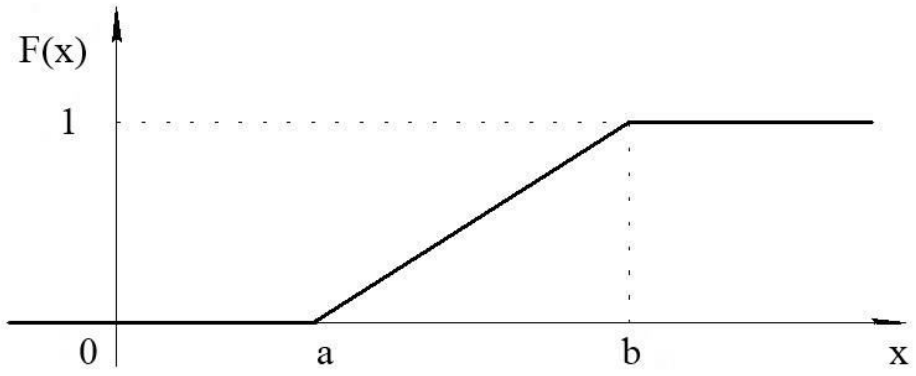


Рис.3.2.

Нехай проміжок $[\alpha;\beta]$ є частиною відрізка $[a;b]$: $[\alpha;\beta]\subset[a;b]$.Тоді ймовірність того, що рівномірно розподілена на $[a;b]$ випадкова величина X прийме значення з проміжку $[\alpha;\beta]$ визначається так:

$$P(\alpha \le X \le \beta)=F(\beta)-F(\alpha)=\frac{\beta-a}{b-a}-\frac{\alpha-a}{b-a}=\frac{\beta-\alpha}{b-a}. \qquad (3.10)$$

Приклад 3.2. Автобуси деякого маршруту рухаються строго за графіком з інтервалом руху 5хв. Знайти ймовірність того, що пасажир, який підійшов до зупинки, буде чекати чергового автобуса менше 3 хв.

Розв’язання. Неперервна випадкова величина X − «час до

чергового автобуса» розподілена рівномірно в інтервалі $(0;\ 5)$, оскільки автобуси рухаються строго за графіком. Шукану ймовірність $P(0 < X < 3)$ знайдемо за формулою (3.10)

$$P(0 < X < 3) = \frac{3-0}{5} - \frac{0-0}{5} = \frac{3}{5} = 0{,}6\,.$$

3.2.2. Нормальний закон розподілу

Нормальним називають розподіл ймовірностей неперервної випадкової величини X, щільність якого має вигляд

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x-a)^2}{2\sigma^2}}, \qquad (3.11)$$

де a − математичне сподівання, σ − середнє квадратичне відхилення X (***параметри розподілу***).

Інтегральна функція нормально розподіленої випадкової величини визначається через функцію Лапласа:

$$F(x) = \frac{1}{2} + \Phi\left(\frac{x-a}{\sigma}\right).$$

Графіки функцій $f(x)$ і $F(x)$ при різних значеннях σ зображено на рис.3.3, 3.4.

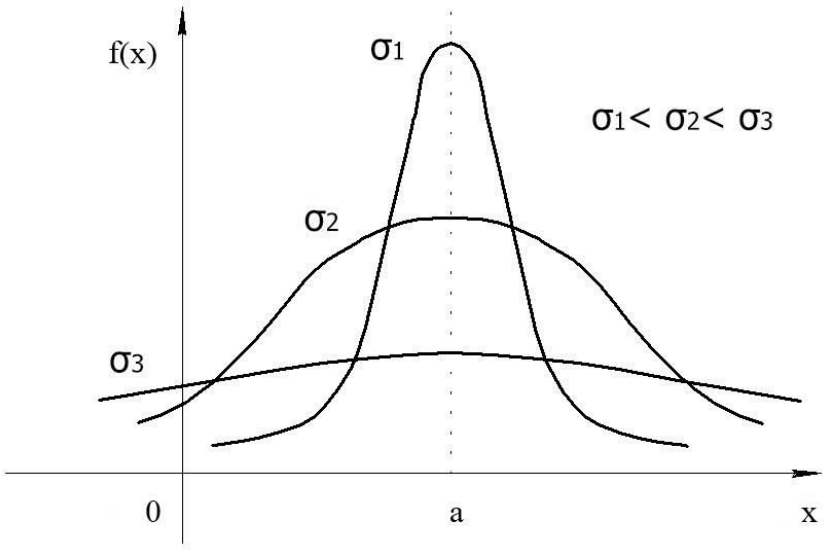


Рис. 3.3.

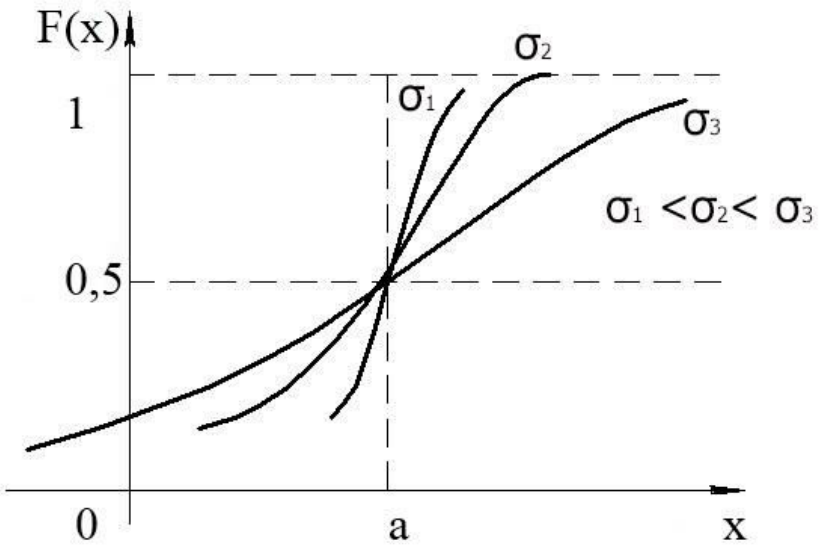


Рис. 3.4.

Ймовірність того, що нормально розподілена випадкова величина X прийме значення з інтервалу $(\alpha;\beta)$ обчислюється так:

$$P(\alpha < X < \beta) = \Phi\left(\frac{\beta - a}{\sigma}\right) - \Phi\left(\frac{\alpha - a}{\sigma}\right), \tag{3.12}$$

де $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{t^2}{2}} dt$ – ***функція Лапласа***.

Зокрема, ймовірність того, що модуль відхилення величини X від свого математичного сподівання менший заданого числа $\varepsilon > 0$, обчислюється за формулою

$$P(|X - a| < \varepsilon) = 2\Phi\left(\frac{\varepsilon}{\sigma}\right). \tag{3.13}$$

Приклад 3.4. Відомо, що відхилення довжини виготовлених деталей від стандарту є випадковою величиною, розподіленою за нормальним законом. Якщо стандартна довжина дорівнює $m = 40$ см, а середнє квадратичне відхилення $\sigma = 0{,}4\sqrt{2}$ см, то яку точність довжини виробу можна гарантувати з ймовірністю $0{,}8$?

Розв'язання. Потрібно знайти додатне число ε, для якого $P(|X - 40| < \varepsilon) = 0{,}8$. Оскільки

$$P(|X - 40| < \varepsilon) = 2\Phi\left(\frac{\varepsilon}{0{,}4\sqrt{2}}\right) = 2\Phi(1{,}77\varepsilon),$$

вихідна задача зводиться до розв'язання нерівності $2\Phi(1{,}77\varepsilon) > 0{,}8$, $\Phi(1{,}77\varepsilon) > 0{,}4$. За допомогою таблиці встановлюємо, що $1{,}77\varepsilon > 1{,}29$,

звідки $\varepsilon > 0{,}729$. Отже, найменше значення ε, що задовольняє останній нерівності, $\varepsilon = 0{,}72$.

Нормальні розподіли є основними в теорії помилок спостереження:

а) a − істинне значення величини, що вимірюють;

б) σ^2 − розсіювання точності вимірювання приладу, якщо цей прилад не має систематичних помилок (наприклад, зсув циферблату на 5 поділок);

в) результат вимірювання має $N(a,\sigma^2)$-розподіл.

Якщо прилад має систематичну помилку b, то результат має нормальний $N(a+b,\sigma^2)$- розподіл.

Нормальний розподіл з параметрами $a=0$ та $\sigma^2=1$ називають стандартним нормальним $(N(0,1))$. Випадкові величини зі стандартним нормальним розподілом позначають через Z.

3.2.3. Показниковий (експоненціальний) закон розподілу

Показниковим називають розподіл ймовірністей неперервної випадкової величини X, що описується диференціальною функцією

$$f(x)=\begin{cases} 0 \ \textit{при}\ x<0, \\ \lambda e^{-\lambda x} \ \textit{при}\ x\geq 0, \end{cases} \tag{3.14}$$

де λ − стала додатна величина (***параметр показникового розподілу***).

Графік щільності (3.14) представлено на рис.3.7.

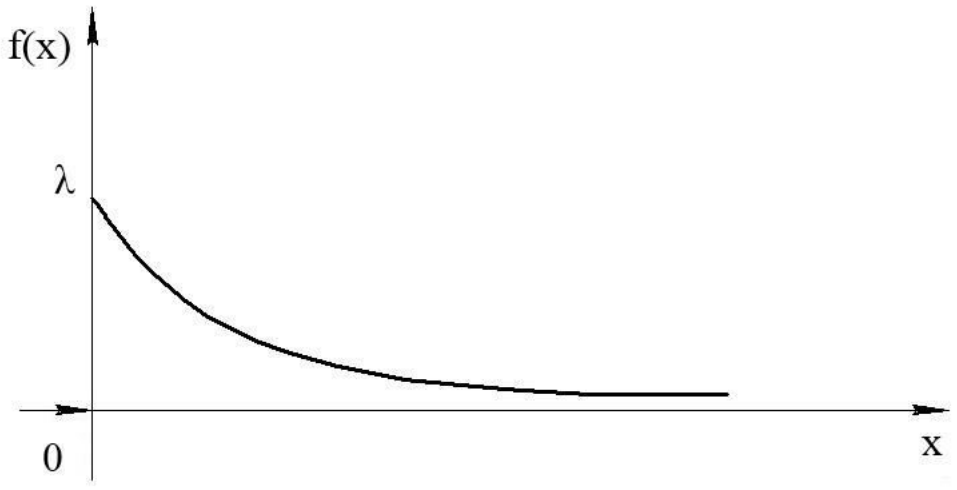


Рис. 3.5.

Функція розподілу показникового закону має вигляд (рис.3.6)

$$F(x)=\begin{cases} 0 \ при \ x<0, \\ 1-e^{-\lambda x} \ при \ x\geq 0. \end{cases} \tag{3.15}$$

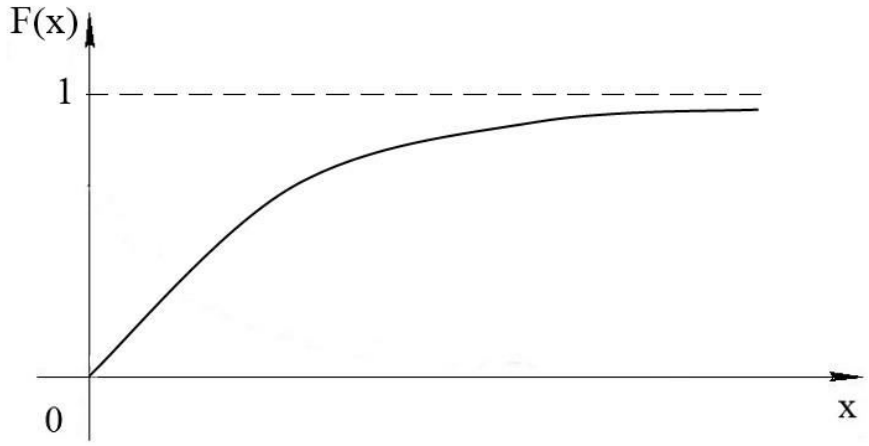


Рис. 3.6

Ймовірність потрапляння в інтервал $(a;b)$ неперервної випадкової величини X , розподіленої за показниковим законом, обчислюється так:

$$P(\alpha<X<\beta)=e^{-\lambda a}-e^{-\lambda\beta}\ . \tag{3.16}$$

Математичного сподівання, дисперсія та середнє квадратичне відхилення показникового розподілу відповідно дорівнюють

$$M(X)=\frac{1}{\lambda},\ D(X)=\frac{1}{\lambda^2},\ \sigma(X)=\frac{1}{\lambda}\ . \tag{3.17}$$

Отже, математичне сподівання та середнє квадратичне відхилення показникового розподілу рівні між собою.

Показниковий закон широко використовується в прикладаннях, зокрема, в теорії надійності, одним із основних понять якої є ***функція надійності.***

Нехай елемент (деякий пристрій) починає працювати в момент часу $t_0=0$, а через деякий час t настає відмова. Позначимо через T НВВ − час безвідмовної роботи елемента, а через λ − інтенсивність відмов (середнє число відмов в одиницю часу).

Часто тривалість безвідмовної роботи елемента має показниковий розподіл, інтегральна функція якого

$$F(t)=P(T<t)=1-e^{-\lambda t},\ t\geq 0\ ,\lambda>0 \tag{3.18}$$

визначає ймовірність відмови елемента за час t .

Функцією надійності $R(t)$ називають функцію, що визначає

ймовірність безвідмовної роботи елемента протягом часу t:

$$R(t) = P(T > t) = 1 - F(t) = e^{-\lambda t},\ t \geq 0,\ \lambda > 0. \qquad (3.19)$$

Приклад 3.5. Тривалість часу T безвідмовної роботи першого з двох незалежно працюючих елементів має показниковий розподіл $F_1(t) = 1 - e^{-0,02t}$, другого $F_2(t) = 1 - e^{-0,05t}$. Знайти ймовірність того, що за час $t = 6$ годин обидва елементи відмовлять.

Розв'язання. а) Ймовірність відмови першого елемента

$$p_1 = 1 - R_1(6) = 1 - e^{-0,02 \cdot 6} = 1 - 0,887 = 0,113.$$

Ймовірність відмови другого елемента

$$p_2 = 1 - R_2(6) = 1 - e^{-0,05 \cdot 6} = 1 - 0,741 = 0,259.$$

Ймовірність того, що обидва елементи відмовлять, за теоремою множення для незалежних подій, буде такою:

$$p_1 \cdot p_2 = 0,113 \cdot 0,259 \approx 0,03.$$

Контрольні запитання

1. Навести основні закони розподілу дискретної випадкової величини.
2. За якими формулами визначаються числові характеристики біномного закону? Закону розподілу Пуассона?
3. Як означається інтегральна функція розподілу? Для завдання яких величин вона застосовується?
4. Навести основні властивості інтегральної функції розподілу. Які з них притаманні лише неперервній випадковій величині?
5. Дати означення диференціальної функції розподілу неперервної випадкової величини. Чому її називають ще щільністю ймовірності (розподілу)?
6. Навести основні властивості диференціальної функції та їх геометричне тлумачення.
7. Навести основні розподіли неперервної випадкової величини. За якою характеристикою означається кожний з розподілів?
8. Якими є визначальні параметри нормального та показникового розподілу?
9. Як зв'язані математичне сподівання та середнє квадратичне відхилення показникового розподілу?
10. Як означається функція надійності?

Тема 4. Теореми додавання та добутку подій. Формула повної ймовірності. Формули Байєса

4.1. Характеризація подій

Введемо поняття достовірної (вірогідної) та неможливої події.

Означення 1. Подія називається ***достовірною***, якщо вона обов'язково відбудеться при умові виконання певної сукупності умов S.

Наприклад, при нормальному тиску та температурі 20^0C вода знаходиться у стані рідини.

Означення 2. Подія називається ***неможливою***, якщо вона ніколи свідомо не настане при умові виконання певного комплексу умов S.

Так, подія «вода при нормальному тиску та температурі 20^0C знаходиться у твердому стані» є неможливою подією.

Розглянемо основні типи випадкових подій.

Означення 3. Дві події називаються ***несумісними***, якщо поява однієї з них повністю виключає появу іншої в одному й тому ж самому випробуванні.

Наприклад, подія А − певний студент склав іспит з математики та подія В − цей студент не склав іспит з математики є несумісними подіями.

Означення 4. Декілька подій A_1 , A_2 ,…, A_n називають ***попарно-несумісними***, якщо поява однієї з них повністю виключає появу іншої в одному й тому ж самому випробуванні.

Так, події «випала одна певна цифра при однократному киданні грального кубика» – попарно-несумісні:

w_1 − випала цифра "1",

w_2 − випала цифра "2",

w_3 − випала цифра "3",

w_4 − випала цифра "4",

w_5 − -випала цифра "5",

w_6 − випала цифра "6".

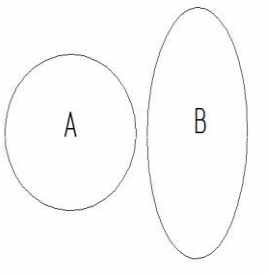
Рис. 4.1 − 4.3 ілюструють характеризацію подій. За несумісності подій чи попарної несумісності подій відповідні контури не перетинаються (рис.4.1, 4.2). Рис. 4.3 ілюструє ситуацію, коли події у сукупності є несумісними, тобто вони не мають жодного спільного елемента, але вони не є попарно-несумісними (кожна пара має спільні елементи).

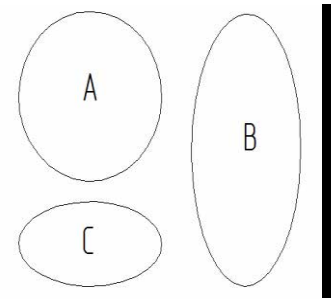
Означення 5. Події A , B , C ,.. … називають ***єдиноможливими***, якщо хоча б одна з них обов'язково відбувається.

Означення 6. Кажуть, що кілька подій утворюють ***повну групу***, якщо в результаті випробування з'явиться хоча б одна з них.

Інакше кажучи, поява хоча б однієї з подій повної групи є достовірна

подія.





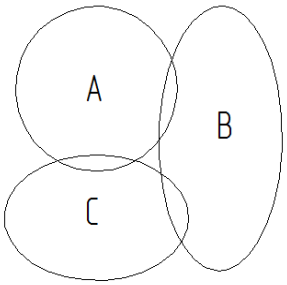


Рис.4.1. Рис.4.2. Рис.4.3.

Можна й іншим чином ввести поняття повної групи подій (через поняття єдиноможливих та несумісних подій).

Означення 7. Сукупність несумісних та єдиноможливих подій утворюють повну групу подій.

Означення 8. Події A, B, C,… називають ***рівноможливими***, якщо немає підстав вважати будь-яку одну з них більш або менш можливою, ніж якусь іншу.

Означення 9. Сукупність несумісних, єдиноможливих та рівноможливих подій називають ***сукупністю елементарних подій.***

Сукупність елементарних подій ще називають простором елементарних подій та позначають символом Ω. Так , події w_1, w_2, w_3, w_4, w_5, w_6 у прикладі з кубиком утворюють простір елементарних подій.

Означення 10. ***Протилежними*** називають дві єдиноможливі події, що утворюють повну групу.

Протилежні події прийнято позначати символами A і $\overline{A}$. Це, наприклад, улучання та промах при пострілі в мішень; випадіння “герба” чи “надпису” при однократному киданні монети тощо.

4.2. Операції над подіями

Означення 11. Сумою $A+B$ (або $A \cup B$) двох подій A і B називають подію C, що полягає або в появі події A, або події B, або ж в появі обох цих подій: $C = A + B$ (рис.4.4)

Сумою кількох подій $A_1, A_2, ..., A_n$ називають подію C, що полягає в появі хоча б однієї з цих подій, тобто $C = A_1 + A_2 + ... + A_n$.

Приклад 4.1. При однократному киданні кубика подія $\omega = w_2 + w_4 + w_6$ полягає в появі хоча б однієї з цифр 2, 4, 6 та означає «випала парна кількість очок».

Означення 12. Добутком $A \cdot B$ (або $A \cap B$) двох подій A і B називають подію C, що полягає в сумісній появі цих подій (рис.4.5)

Аналогічно, добутком кількох подій $A_1, A_2, ..., A_n$ називають подію

C, що полягає в сумісній появі всіх цих подій.

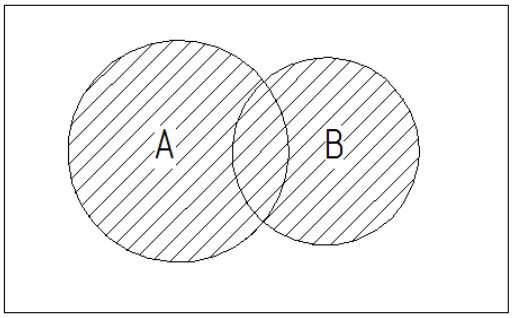


Рис.4.4.

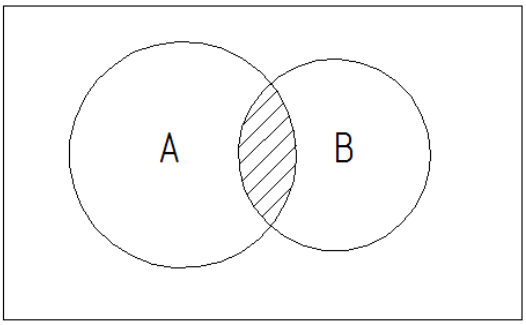


Рис.4.5.

4.3. Теорема додавання

Існують три основні варіанти теореми додавання:

1. Адитивність для попарно несумісних подій:

1.1) адитивність: якщо $A \cdot B = \emptyset$ (події несумісні), то

$$P(A+B) = P(A) + P(B). \qquad (4.1)$$

1.2) скінченна адитивність: якщо $A_1, A_2, ..., A_n$ такі, що $A_i \cdot A_j = \emptyset$, $i \neq j$ (події попарно несумісні), то

$$P(A_1 + A_2 + ... + A_n) = P(A_1) + P(A_2) + ... + P(A_n). \qquad (4.2)$$

Наслідок 1. Сума ймовірностей подій $A_1, A_2, ..., A_n$, що утворюють повну групу, дорівнює одиниці

$$P(A_1) + P(A_2) + ... + P(A_n) = 1. \qquad (4.3)$$

Доведення. Оскільки поява хоча б однієї з подій повної групи – достовірна подія, то $P(A_1 + A_2 + ... + A_n) = 1$. Тому, за формулою (4.2) маємо рівність (4.3).

Наслідок 2. Сума ймовірностей протилежних подій дорівнює одиниці

$$P(A) + P(\overline{A}) = 1. \qquad (4.4)$$

1.3) зліченна адитивність: якщо $A_1, A_2, ..., A_n, ...$ такі, що $A_i \cdot A_j = \emptyset$, $i \neq j$ (події попарно несумісні), то

$$P(A_1 + A_2 + ... + A_n + ...) = P(A_1) + P(A_2) + ... + P(A_n) + ...$$

2. *Загальний варіант для двох подій:*

$$P(A+B)=P(A)+P(B)-P(A\cdot B). \qquad (4.5)$$

3. ***Загальний варіант для незалежних у сукупності подій*** $A_1, A_2, ..., A_n$ **:**

$$P(A_1+A_2+...+A_n)=1-P(\overline{A}_1)\cdot P(\overline{A}_2)\cdot ...\cdot P(\overline{A}_n). \qquad (4.6)$$

Варіанти адитивності належать до характеристичних властивостей ймовірності.

Приклад 4.2. Нехай ймовірність захворіти на грип А дорівнює 0,6; на грип В − 0,5; на грип А та В одночасно − 0,2. Знайти ймовірність захворіти хоча б на один з типів грипу.

Розв'язання. Оскільки події А і В сумісні, ймовірність суми цих подій знаходимо за формулою (4.5):

$$P(A+B)=P(A)+P(B)-P(A\cdot B)=0{,}6+0{,}5-0{,}2=0{,}9.$$

4.4. Теорема множення

Означення 13. Умовною ймовірністю $P(A/B)$ (або $P_B(A)$) називають ймовірність події A, обчислену при умові, що подія B вже здійснилась.

Поняття незалежності подій означає нечутливість (стійкість) ймовірності однієї події до залучення додаткової умови у початкові умови спостереження.

Означення 14. Подію A називають ***незалежною*** від події B, $P(B)\neq 0$, якщо

$$P(A/B)=P(A)$$

Еквівалентне означення незалежності. Дві події A та B називають ***незалежними***, якщо ймовірність добутку цих подій дорівнює добутку їх ймовірностей:

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

(теорема множення для незалежних подій).

Означення 15. Кілька подій $A_1, A_2, ..., A_n$ називаються **попарно незалежними,** якщо кожні дві з них незалежні.

Означення 16. Кілька подій $A_1, A_2, ..., A_n$ називаються ***незалежними в сукупності*** (або просто ***незалежними***), якщо вони попарно незалежні та незалежні кожна подія і всі можливі добутки інших подій.

Для теореми множення існують такі основні варіанти:

1. Загальний варіант:

$$P(A\cdot B)=P(A)\cdot P(B/A)=P(B)\cdot P(A/B), \qquad (4.7)$$

причому $P(B)\neq 0$, $P(B)\neq 0$ для відповідної частини запису.

2. Для незалежних подій:

$$P(A \cdot B) = P(A) \cdot P(B). \qquad (4.8)$$

3. **Для незалежних в сукупності подій**:

$$P(A_1 \cdot A_2 \cdot \ldots \cdot A_n) = P(A_1) \cdot P(A_2) \cdot \ldots \cdot P(A_n). \qquad (4.9)$$

Доведення першого варіанту теореми випливає із зв'язку умовної та безумовної ймовірності.

Щодо варіантів цієї теореми для незалежних подій, то відповідні рівності збігаються з визначенням незалежності чи безпосередньо випливають із них.

Приклад 4.3. Нехай ймовірність потрапити до групи ризику дорівнює 0,1; а ймовірність захворіти на грип у цій групі − 0,8. Знайти ймовірність одночасного виконання двох умов (належності групі ризику та захворювання на грип).

Розв'язання. Введемо такі позначення: подія А − «потрапити до групи ризику», подія В − «захворіти на грип у цій групі». Шукану ймовірність обчислюємо за формулою (4.8):

$$P(A \cdot B) = P(A) \cdot P(B/A) = 0{,}1 \cdot 0{,}8 = 0{,}08.$$

Незалежність певних об'єктів в умові задачі однозначно тлумачать як незалежність подій, що цим об'єктам відповідають, а незалежність ознак означає незалежність подій.

Зауваження. Існує клас задач, для яких потрібно спочатку провести спостереження за фіксованих умов A, а потім з одержаної вибірки додатково відібрати ті спостереження, які задовольняють додаткову вимогу (чи вимоги) B. Частоту, побудовану за сформованою таким чином вибіркою називають умовною і позначають символом $W_n(A/B)$, n − об'єм вибірки. Умовна частота містить усі характеристичні властивості частоти і є звичайною частотою.

Граничними значеннями умовних частот є умовні ймовірності:

$$\lim_{n\to\infty} W_n(A/B) = P(A/B).$$

4.5. Формули повної ймовірності та Байєса

Суть формули повної ймовірності полягає у можливості обчислення події A, коли її можна реалізувати одним із кількох способів, що взаємно виключають один одного. Зміст виразу «спосіб реалізації» пояснюють через поняття «повна група подій».

Повною групою подій називають набір попарно несумісних подій $B_1, B_2, ..., B_n$, для яких справджується одна з двох умов:

1) хоча б одна з подій $B_1, B_2, ..., B_n$ реалізується:

$$B_1 + B_2 + ... + B_n = \mathrm{U}; \tag{4.10}$$

2) сума ймовірностей цих подій дорівнює одиниці.

Теорема. Нехай подія A може наступити при умові появи однієї з несумісних подій $B_1, B_2, ..., B_n$, що утворюють повну групу:

$$P(B_1) + P(B_2) + ... + P(B_n) = 1.$$

Тоді ймовірність події A обчислюється за формулою

$$P(A) = P(B_1) \cdot P(A/B_1) + P(B_2) \cdot P(A/B_2) + ... + P(B_n) \cdot P(A/B_n), \tag{4.11}$$

що називається ***формулою повної ймовірності.***

Доведення. За умовою подія A може наступити при умові появи однієї з несумісних подій $B_1, B_2, ..., B_n$. Це означає появу однієї, не має значення якої, з несумісних подій $B_1 \cdot A, B_2 \cdot A, ..., B_n \cdot A$, тобто

$$A = B_1 \cdot A + B_2 \cdot A + ... + B_n \cdot A.$$

За теоремами додавання для несумісних подій та множення для залежних подій дістанемо

$$P(A) = P(B_1 \cdot A) + P(B_2 \cdot A) + ... + P(B_n \cdot A) =$$
$$= P(B_1) \cdot P(A/B_1) + P(B_2) \cdot P(A/B_2) + ... + P(B_n) \cdot P(A/B_n).$$

Оскільки наперед невідомо, яка з подій $B_1, B_2, ..., B_n$ настане, їх називають ***гіпотезами.***

Припустимо тепер, що подія A вже здійснилась в результаті випробування. Як змінилися у зв'язку з цим ймовірності гіпотез, тобто $P(B_i/A)$, $i = 1,2,...,n$? Формули, що дозволяють переоцінити ймовірності гіпотез, називають ***формулами Байєса*** та мають вигляд

$$P(B_i/A) = \frac{P(B_i) \cdot P(A/B_i)}{\sum_{i=1}^{n} P(B_i) \cdot P(A/B_i)}, \; i = 1,2,...,n. \tag{4.12}$$

Доведемо, наприклад формулу (4.12) для $i = 1$. За теоремою множення для залежних подій маємо

$$P(A \cdot B_1) = P(A) \cdot P(B_1/A) = P(B_1) \cdot P(A/B_1),$$

звідки

$$P(B_1 / A) = \frac{P(B_1) \cdot P(A / B_1)}{P(A)}.$$

Підставивши замість $P(A)$ праву частину формули (4.11), приходимо до формули Байєса при $i = 1$.

Приклад 4.4. Виріб перевіряється на стандартність одним з двох контролерів. Ймовірність того, що виріб потрапить до І-го контролера дорівнює 0,55; до ІІ-го контролера − 0,45. Ймовірність того, що виріб буде визнаний стандартним І-м контролером дорівнює 0,9; до ІІ-м − 0,98. Виріб при перевірці було визнано стандартним. Знайти ймовірність того, що цей виріб перевіряв ІІ-й контролер.

Розв'язання. Позначимо через A подію: «виріб при перевірці було визнано стандартним». Тут можна зробити два припущення (гіпотези): B_1 − «виріб потрапив до І-го контролера», B_2 − «виріб потрапив до ІІ-го контролера». Події B_1, B_2 є несумісними та утворюють повну групу.

За умовою задачі $P(B_1) = 0{,}55$, $P(B_2) = 0{,}45$; $P(A / B_1) = 0{,}9$, $P(A / B_2) = 0{,}98$. Ймовірність події A знаходимо за формулою (4.11) при $n = 2$:

$$P(A) = 0{,}55 \cdot 0{,}9 + 0{,}45 \cdot 0{,}98 = 0{,}495 + 0{,}441 = 0{,}936.$$

Застосовуючи формули Байєса (4.12) при $i = 2$, $n = 2$

$P(B_2 / A) = \dfrac{P(B_2) \cdot P(A / B_2)}{\sum_{i=1}^{n} P(B_i) \cdot P(A / B_i)}$, дістанемо $P(B_2 / A) = \dfrac{0{,}45 \cdot 0{,}98}{0{,}936} = 0{,}471$.

Контрольні запитання

1. Як означається сума двох (скінченної кількості) подій?.
2. Як означається добуток двох (скінченної кількості) подій?.
3. У чому полягає теорема додавання ймовірностей? Сформулювати для випадку несумісних та сумісних подій.
4. Чому дорівнює сума ймовірностей подій, що утворюють повну групу? Протилежних подій?
5. Які події називають незалежними, залежними?
6. Як означається умовна ймовірність?
7. У чому полягає теорема множення ймовірностей? Сформулювати для випадку незалежних, залежних подій.
8. Чому дорівнює ймовірність появи хоча б однієї з подій, незалежних в сукупності?
9. Сформулювати постановку задачі формули повної ймовірності, формул Байєса.
10. Які події називають гіпотезами? У якому варіанті постановки задачі формулу повної ймовірності?

Тема 5. Статистичні оцінки параметрів розподілу, їхні властивості. Точкове оцінювання параметрів основних розподілів

5.1. Постановка задачі статистичного оцінювання

Основною метою статистичного оцінювання є визначення дійсних параметрів генеральної сукупності на основі вивчення вибіркових показників. Параметри звичайно характеризують певну властивість теоретичного розподілу величини X. Так, якщо відомо, що розподіл в генеральній сукупності нормальний, то необхідно оцінити (наближено знайти) $M(X)=a$ і $\sigma(X)=\sigma$, якщо є підстави вважати, що X має показниковий розподіл − оцінити параметр λ.

Статистичні методи дозволяють враховувати дані досліду (вибіркові значення) для уточнення ймовірнісної моделі, наприклад для оцінки щільності ймовірностей або функції розподілу випадкової величини X. Останнє надає можливість спрогнозувати подальші події, що є важливим для прийняття рішень.

Нехай потрібно вивчити кількісну ознаку X генеральної сукупності. Припустимо, що з теоретичних міркувань встановлено, який саме розподіл має ознака X. Природно виникає питання оцінювання параметрів, якими визначається цей розподіл.

Нехай з генеральної сукупності з функцією розподілу $F(x,\theta)$, де θ − невідомий параметр, здійснено вибірку об'єму n:

$$x_1,\ x_2,\ \ldots,\ x_n. \tag{5.1}$$

Це значення величини X, що одержані в результаті n незалежних спостережень. Для знаходження наближених значень (оцінки) невідомого параметра θ будемо розглядати функції вигляду

$$\theta^* = f\ (x_1, x_2, ..., x_n), \tag{5.2}$$

які називають ***вибірковими функціями*** або ***статистиками***.

Задача оцінки невідомого параметра θ зводиться до знаходження таких вибіркових функцій (5.2), які можна використовувати як оцінку параметра θ.

Припустимо, що (5.2) − статистична оцінка невідомого параметра θ теоретичного розподілу та нехай за допомогою вибірки об'єму n (5.1) знайдено оцінку θ_1^*. Повторимо дослід, тобто візьмемо з генеральної сукупності іншу вибірку того ж об'єму n і на її основі знайдемо оцінку θ_2^*. Повторюючи дослід багатократно, дістанемо числа θ_1^*, θ_2^*, …, θ_k^*, що будуть у загальному випадку різними. Оскільки кожна вибірка випадкова, то оцінку θ^* можна розглядати як випадкову величину, а числа θ_1^*, θ_2^*, …, θ_k^* − як її можливі значення.

5.2. Властивості оцінок

Для того, щоб оцінки давали «добрі» наближення параметрів, вони мають задовольняти певним вимогам, а саме: бути ***незміщеними, ефективними*** та ***спроможними***.

Означення 1. Оцінку θ^* невідомого параметра θ називають ***незміщеною***, якщо її математичне сподівання дорівнює оцінюваному параметру:

$$M(\theta^*) = \theta, \tag{5.3}$$

де $M(\theta^*)$ − математичне сподівання оцінки θ^*.

Якщо умова (5.3) не виконується, оцінка θ^* параметра θ називається зміщеною. Вимога незміщеності гарантує відсутність систематичних (одного знаку) помилок при оцінці параметрів.

Незміщена оцінка θ^* дає хороший результат, якщо її можливі значення мало розсіяні навколо свого середнього значення, тобто дисперсія $D(\theta^*)$ мала.

Означення 2. Статистична оцінка θ^* називається ***ефективною***, якщо вона має найменшу дисперсію серед усіх можливих оцінок параметра θ, обчислених за вибірками одного й того ж об'єму.

При розгляді вибірок великого об'єму статистичні оцінки мають задовольняти умову спроможності.

Означення 3. Статистичну оцінку θ^* параметра θ називають ***спроможною***, якщо при збільшенні кількості незалежних випробувань n вона прямує за ймовірністю до значення параметра θ:

$$\lim_{n\to\infty} P(|\theta^* - \theta| < \varepsilon) = 1. \tag{5.4}$$

для довільного $\varepsilon > 0$.

Зокрема, якщо дисперсія незміщеної оцінки при $n \to \infty$ прямує до нуля, то така оцінка виявляється і спроможною.

Розрізняють точкові та інтервальні оцінки параметрів.

Точковою називають оцінку невідомого параметра θ, що визначається одним числом. При вибірках малого об'єму точкові оцінки можуть приводити до грубих помилок. Тоді застосовують так звані інтервальні оцінки.

Інтервальною називають оцінку, яка визначається двома числами − кінцями інтервалу, відносно якого з наперед заданою ймовірністю можна стверджувати, що оцінюваний параметр знаходиться всередині цього інтервалу.

Зазначимо, що для малих об'ємів вибірки точкові оцінки можуть слугувати в якості першого наближення невідомого параметра.

5.3. Точкові оцінки

Нехай вивчається випадкова величина (ознака) X генеральної сукупності за результатами вибірки x_1, x_2, …, x_n. об'єму n.

Для побудови точкових оцінок математичного сподівання $M(X)=a$ і дисперсії $D(X)=\sigma^2$ ознаки X (їх називають також ***генеральною середньою*** та ***генеральною дисперсією***) використовують відповідні ***вибіркові характеристики*** (***середні показники***).

5.3.1. Вибіркова середня

Вибірковою середньою $\bar{x}_B$ називають середнє арифметичне значень вибірки x_1, x_2, …, x_n:

$$\bar{x}_B = \frac{1}{n}\sum_{i=1}^{n} x_i \ . \qquad (5.5)$$

Якщо значення ознаки X : x_1 , x_2 , …, x_k мають відповідно частоти $n_1, n_2, ..., n_k$; причому $n_1 + n_2 + ... + n_k = n$, n − об'єм вибірки, то обчислення вибіркової середньої здійснюють за формулою

$$\bar{x}_B = \frac{1}{n}\sum_{i=1}^{k} x_i n_i \ . \qquad (5.6)$$

В якості оцінки генеральної середньої приймають вибіркову середню. Така оцінка є ***незміщеною*** та ***спроможною***.

Ефективність оцінки залежить від закону розподілу ознаки X. Так, доведено, що коли випадкова величина X розподілена за нормальним законом, то оцінка $\bar{x}_B$ буде ефективною.

5.3.2. Вибіркова та виправлена дисперсія

Щоб охарактеризувати розсіювання вибіркових значень навколо свого середнього значення $\bar{x}_B$ вводять показник «вибіркова дисперсія».

Вибірковою дисперсією D_B називається середнє арифметичне квадратів відхилень вибіркових значень X від вибіркової середньої $\bar{x}_B$.

а) Якщо всі варіанти x_1, x_2, …, x_n різні (не згруповані дані), то

$$D_B = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x}_B)^2 \ . \qquad (5.7)$$

б) Для згрупованих даних (значення ознаки x_1, x_2, …, x_k мають відповідно частоти $n_1, n_2, ..., n_k$; причому $n_1 + n_2 + ... + n_k = n$) вибіркова дисперсія D_B обчислюється так:

$$D_B = \frac{1}{n}\sum_{i=1}^{k} n_i (x_i - \bar{x}_B)^2 . \qquad (5.8)$$

Обчислення вибіркової дисперсії D_B можна значно спростити, якщо скористатися формулою

$$D_B = \overline{x^2} - [\bar{x}]^2 \qquad (5.9)$$

(дисперсія дорівнює середньому квадратів значень ознаки X мінус квадрат загальної середньої).

Вибіркова дисперсія D_B є спроможною, але зміщеною оцінкою генеральної дисперсії D_Γ:

$$M[D_B] = \frac{n-1}{n} D_\Gamma . \qquad (5.10)$$

Користуючись оцінкою D_B ми будемо допускати деяку систематичну похибку у бік менших значень: $M[D_B] < D_\Gamma$. За формулою (5.10) легко «підправити» вибіркову дисперсію так, щоб її математичне сподівання дорівнювало D_Γ . Для цього достатньо помножити D_B на дріб $\frac{n}{n-1}$. У зв'язку з цим вводять так звану ***виправлену дисперсію*** S^2:

$$S^2 = \frac{n}{n-1} D_B . \qquad (5.11)$$

Виправлена дисперсія S^2 вже є незміщеною оцінкою генеральної дисперсії D_Γ: $M[S^2] = D_\Gamma$.

Для оцінки середнього квадратичного відхилення генеральної сукупності використовують „виправлене" середнє квадратичне відхилення

$$S = \sqrt{S^2} . \qquad (5.12)$$

Така оцінка є зміщеною.

Зауваження. При достатньо великому об'ємі вибірки ($n \geq 30$) вибіркова D_B та "виправлена" дисперсія S^2 відрізняються мало (вже при $n = 30$ на 3%). Тому при $n \geq 30$ вибіркова дисперсія D_B може слугувати оцінкою генеральної дисперсії D_Γ . На практиці виправленою дисперсію S^2 користуються при $n < 30$.

Приклад 5.1. За рядом розподілу відносних частот

X	-1	1	2	3
$\omega_i = \frac{n_i}{n}$	0,2	0,3	0,1	0,4

обчислити вибіркову середню та вибіркову дисперсію.

Розв'язання. Оскільки дані згруповані, вибіркову середню $\bar{x}_B$ обчислюємо за формулою (5.6) згідно з таким перетворенням:

$$\bar{x}_B = \sum_{i=1}^{k} x_i \frac{n_i}{n} = \sum_{i=1}^{k} x_i \omega_i .$$

Отже, $\bar{x}_B = -1 \cdot 0{,}2 + 1 \cdot 0{,}3 + 2 \cdot 0{,}1 + 3 \cdot 0{,}4 = 1{,}5$. Так само перетворимо формулу (5.8) для розрахунку вибіркової дисперсії D_B:

$$D_B = \sum_{i=1}^{k} \frac{n_i}{n} \left(x_i - \bar{x}_B\right)^2 = \sum_{i=1}^{k} \omega_i \left(x_i - \bar{x}_B\right)^2 .$$

Таким чином, остаточно дістанемо

$$D_B = (-1-1{,}5)^2 \cdot 0{,}2 + (1-1{,}5)^2 \cdot 0{,}3 + (2-1{,}5)^2 \cdot 0{,}1 + (3-1{,}5)^2 \cdot 0{,}4 = 2{,}25 .$$

Приклад 5.2. За вибіркою об'єму $n = 7$ знайдено зміщену оцінку генеральної дисперсії $D_в = 4$. Знайти незміщену оцінку генеральної дисперсії.

Розв'язання. За формулою (5.11) знаходимо виправлену дисперсію S^2

$$S^2 = \frac{n}{n-1} D_B = \frac{7}{6} \cdot 4 \approx 4{,}67 ,$$

що слугує незміщеною оцінкою генеральної дисперсії.

Контрольні запитання

1. Що розуміють під статистичною оцінкою параметра розподілу? Сформулювати постановку задачі статистичного оцінювання.
2. Які функції називають вибірковими функціями або статистиками?
3. Яка статистична оцінка називається незміщеною? Ефективною? Спроможною?
4. Яка статистична оцінка називається точковою? Інтервальною?
5. Як означають генеральну середню та генеральну дисперсію?
6. Яку оцінку приймають в якості незміщеної оцінки генеральної середньої?
7. Яка характеристика досліджуваної ознаки є визначальною для ефективності оцінки генеральної середньої?
8. Навести формули для обчислення точкових оцінок математичного сподівання, дисперсії, середнього квадратичного відхилення.
9. Як «підправити» вибіркову дисперсію, щоб оцінка стала незміщеною?
10. При яких об'ємах вибірки в якості оцінки генеральної дисперсії приймають "виправлену" дисперсію? Вибіркову дисперсію?

Тема 6. Інтервальні оцінки параметрів розподілу. Довірча ймовірність (надійність). Довірчий інтервал.

6.1. Постановка задачі інтервального оцінювання

При малому числі спостережень точкові оцінки параметрів, значною мірою випадкові, змінюються від вибірки до вибірки. Тому важливо знати, які помилки може спричинити заміна параметру його оцінкою, і в яких можливих межах лежить дійсне значення оцінюваного параметру. Щоб мати уявлення про точність і надійність оцінки в математичній статистиці користуються ***довірчими інтервалами*** та ***довірчою ймовірністю.***

Нехай за даними вибірки об'єму n знайдено статистичну оцінку $\theta^*\ (x_1, x_2, ..., x_n)$, що є точковою характеристикою параметра θ. Вочевидь, що оцінка θ^* тим точніше визначає параметр θ, чим меншою буде абсолютна величина різниці $|\theta - \theta^*|$. Інакше кажучи, якщо $\delta > 0$ і

$$|\theta - \theta^*| < \delta, \qquad (6.1)$$

то число δ характеризує ***точність оцінки.*** Оскільки θ^* випадкова, а θ − невипадкова величина, то категорично стверджувати, що знайдена оцінка задовольняє нерівності (6.1) неможливо. Можна казати лише про ймовірність γ, з якою ця нерівність виконується.

Довірчою ймовірністю (надійністю) оцінки θ^* невідомого параметра θ називається ймовірність γ, з якою здійснюється нерівність (6.1), $\delta > 0$:

$$P(|\theta - \theta^*| < \delta) = \gamma. \qquad (6.2)$$

Звичайно надійність оцінки γ задається наперед, причому в якості γ беруть число, близьке до одиниці: 0,95; 099; 0,999.

Перетворивши (6.1) за властивостями модуля:

$$|\theta - \theta^*| < \delta,\ -\delta < \theta - \theta^* < \delta,$$

або,

$$\theta^* - \delta < \theta < \theta^* + \delta,$$

подамо довірчу ймовірність (6.2) у вигляді:

$$P(\theta^* - \delta < \theta < \theta^* + \delta) = \gamma. \qquad (6.3)$$

Формулу (6.3) слід розуміти так: ймовірність того, що інтервал

$(\theta^* - \delta, \theta^* + \delta)$ містить в собі (покриває) невідомий параметр θ, дорівнює γ. Інтервал $(\theta^* - \delta, \theta^* + \delta)$ називається ***довірчим інтервалом***. Кінці довірчого інтервалу називають *довірчими межами*.

6.2. Довірчий інтервал для оцінки генеральної середньої при відомому σ.

Нехай кількісна ознака X генеральній сукупності розподілена нормально з параметрами $M(X) = a$ і $\sigma(X) = \sigma$. Необхідно побудувати довірчий інтервал для оцінки невідомого математичного сподівання a за даними вибірки об'єму n.

Якщо випадкова величина X розподілена нормально із заданим середнім квадратичним відхиленням σ та невідомим математичним сподіванням a, формула (6.3) набуває вигляду

$$P\left(\bar{x}_B - \frac{t\sigma}{\sqrt{n}} < a < \bar{x}_B + \frac{t\sigma}{\sqrt{n}}\right) = \gamma. \qquad (6.4)$$

Тут $\delta = \frac{t\sigma}{\sqrt{n}}$ – точність оцінки, n – об'єм вибірки, t – значення аргументу функції Лапласа $\Phi(t)$ (див. додаток 2), при якому $\Phi(t) = \frac{\gamma}{2}$.

Доведення. Щоб одержати розрахункову формулу (6.4), потрібно за вибіркою об'єму n x_1, x_2, …, x_n знайти оцінку параметра a (вибіркову середню):

$$\bar{x}_B = \frac{1}{n}\sum_{i=1}^{n} x_i .$$

Розглядаючи вибіркову середню як випадкову величину $\bar{X}_B = \frac{1}{n}\sum_{i=1}^{n} \bar{X}_i$, можна довести, що при вказаних умовах вона також буде розподілена нормально з параметрами $M(\bar{X}_в) = a$, $\sigma(\bar{X}_в) = \frac{\sigma}{\sqrt{n}}$.

Будемо вимагати, щоб виконувалось співвідношення

$$P\left(\left|\bar{X}_в - a\right| < \delta\right) = \gamma. \qquad (6.5)$$

З іншого боку, враховуючи нормальний розподіл $\bar{X}_B$ та його параметри, маємо

$$P\left(\left|\bar{X}_в - a\right| < \delta\right) = 2\Phi\left(\frac{\delta\sqrt{n}}{\sigma}\right). \qquad (6.6)$$

Тут скористалися відомою формулою відхилення

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

нормальної величини X з параметрами $M(X)=a$ та $\sigma(X)=\sigma$).

Нехай ймовірність γ відома, тоді з рівностей (6.5), (6.6) дістанемо

$$2\Phi\left(\frac{\delta\sqrt{n}}{\sigma}\right)=\gamma\,.$$

Покладемо $t=\frac{\delta\sqrt{n}}{\sigma}$, тоді точність оцінки $\delta=\frac{t\sigma}{\sqrt{n}}$, а формула (6.6) набуває вигляду

$$P\left(|\bar{x}_в-a|<\frac{t\sigma}{\sqrt{n}}\right)=2\Phi(t)=\gamma$$

або

$$P\left(\bar{x}_B-\frac{t\sigma}{\sqrt{n}}<a<\bar{x}_B+\frac{t\sigma}{\sqrt{n}}\right)=2\Phi(t)=\gamma\,.$$

Приклад 6.1. Проведено 5 рівно точних вимірювань одним приладом відстані від гармати до цілі, $\sigma=40$ м (середнє квадратичне відхилення). Знайти довірчий інтервал для оцінки «істинної» відстані a до цілі з надійністю 0,95, якщо середнє арифметичне результатів вимірювань $\bar{x}_B=2000$ м (припускається, що результати розподілені нормально).

Розв'язання. Довірчий інтервал для оцінки математичного сподівання a величини X − відстані від гармати до цілі знайдемо згідно з формулою (6.4):

$$\bar{x}_B-t\frac{\sigma}{\sqrt{n}}<a<\bar{x}_B+t\frac{\sigma}{\sqrt{n}}\,, \qquad (6.7)$$

оскільки σ відоме.

Для нашої задачі $\bar{x}_B=2000;\quad n=5;\quad \sigma=40.$

Отже, усі величини, крім t, відомі. Знайдемо t за співвідношенням $\Phi(t)=\frac{\gamma}{2}$, де $\Phi(t)$ − функція Лапласа (див. додаток 2). Дістанемо: $\Phi(t)=\frac{0{,}95}{2}=0{,}475,$ а значення аргументу функції $\Phi(t)$ знаходимо за додатком 2: $t=1{,}96.$

Тепер обчислимо $\delta=\frac{t\sigma}{\sqrt{n}}$ − точність оцінки: $\delta=\frac{1{,}96\cdot 40}{\sqrt{5}}\approx 35{,}06$.

Підставивши усі дані в формулу (6.7), дістанемо шуканий

довірчий інтервал: 1964,94< a <2035,06.

Знайдений довірчий інтервал з надійністю $\gamma = 0{,}95$ покриває невідомий параметр $M(X) = a$ (тобто у 95% параметр a знаходиться всередині цього інтервалу).

Приклад 6.2. Знайти мінімальний об'єм вибірки, при якому з надійністю $\gamma = 0{,}975$ точність оцінки математичного сподівання a генеральної сукупності за вибірковою середньою дорівнює $\delta = 0{,}3$, якщо відоме середнє квадратичне відхилення $\sigma = 1{,}2$ нормально розподіленої генеральної сукупності.

Розв'язання. За формулою $n = \dfrac{t^2\sigma^2}{\delta^2}$ знайдемо об'єм вибірки. Згідно з умовою $\gamma = 0{,}975$, тому $\Phi(t) = \dfrac{\gamma}{2} = \dfrac{0{,}975}{2} = 0{,}4875$. Звідси, за таблицею значень функції Лапласа (див. додаток 2), дістанемо значення $t = 2{,}24$.

Знаходимо $n = \dfrac{2{,}24^2 \cdot 1{,}2^2}{0{,}3^2} \approx 80{,}28$. Оскільки об'єм вибірки ціле додатне число, потрібно взяти число $n = 81$.

Приклад 6.3. Одержано 36 значень нормально розподіленої величини X, причому середнє квадратичне відхилення відоме: $\sigma = 1{,}6$. Задано точність $\delta = 0{,}5$, з якою потрібно оцінити математичне сподівання a за допомогою довірчого інтервалу. Встановити надійність γ такої інтервальної оцінки.

Розв'язання. Визначимо число t з умови $\delta = \dfrac{t\sigma}{\sqrt{n}}$, де δ, σ, n – задані числа. Маємо

$$t = \frac{\sqrt{n}\delta}{\sigma} = \frac{6 \cdot 0{,}5}{1{,}6} = 1{,}88.$$

За допомогою таблиці значень функції Лапласа (див. додаток 2) знайдемо $\Phi(t) = \Phi(1{,}88) = 0{,}47 = \dfrac{\gamma}{2}$. Звідси надійність інтервальної оцінки $\gamma = 2 \cdot 0{,}47 = 0{,}94$, а довірчий інтервал довжини $2\delta = 1$ з надійністю $\gamma = 0{,}94$ покриває математичне сподівання a величини X.

6.3 Побудова довірчого інтервалу для математичного сподівання нормального розподілу при невідомому σ.

Якщо середнє квадратичне відхилення σ невідоме, то довірчий інтервал для оцінки невідомого математичного сподівання a

визначатиметься так:

$$\left(\bar{x}_B - \frac{t_\gamma S}{\sqrt{n}}; \bar{x}_B + \frac{t_\gamma S}{\sqrt{n}}\right), \qquad (6.8)$$

де S – „виправлене” середнє квадратичне відхилення, а величина $t_\gamma = t(\gamma, n)$ знаходиться за відомими n, γ за допомогою таблиці значень розподілу Стьюдента (див. додаток 3).

Приклад 6.4. За вибіркою об’єму $n = 10$:

X	-2	1	2	3	4	5
n_i	2	1	2	2	2	1

оцінити з надійністю 0,95 математичне сподівання a нормальної ознаки X генеральної сукупності за вибірковою середньою, побудувавши довірчий інтервал.

Розв’язання. Вибіркову середню та «виправлене» середнє квадратичне відхилення знайдемо відповідно за формулами:

$$\bar{x}_B = \frac{1}{n}\sum_{i=1}^{k} x_i n_i \,,\ S^2 = \frac{1}{n-1}\sum_{i=1}^{k} n_i (x_i - \bar{x}_B)^2 \,,\ S = \sqrt{S^2} \,.$$

Підставивши у ці формули дані задачі, дістанемо

$$\bar{x}_B = \frac{-2\cdot 2 + 1\cdot 1 + 2\cdot 2 + 3\cdot 2 + 4\cdot 2 + 5\cdot 1}{10} = \frac{20}{10} = 2\,,$$

$$S^2 = \frac{2\cdot(-2-2)^2 + 1\cdot(1-2)^2 + 2\cdot(2-2)^2 + 2\cdot(3-2)^2 + 2\cdot(4-2)^2 + 1\cdot(5-2)^2}{9} =$$

$$= \frac{2\cdot 16 + 1\cdot 1 + 2\cdot 1 + 2\cdot 4 + 1\cdot 9}{9} = \frac{52}{9} \approx 5{,}78\,,$$

$$S = \sqrt{5{,}78} \approx 2{,}4\,.$$

За таблицею розподілу Стьюдента (див. додаток 3) знаходимо величину t_γ за відомими $\gamma = 0{,}95$ і $n = 10$: $t_\gamma = 2{,}26$. Довірчий інтервал для математичного сподівання a будуємо за формулою (6.8):

$$\left(2-\frac{2{,}26\cdot 2{,}4}{\sqrt{10}};\ 2+\frac{2{,}26\cdot 2{,}4}{\sqrt{10}}\right),\ \text{або}\ (0{,}3;\ 3{,}7).$$

Шуканий довірчий інтервал покриває математичне сподівання a з надійністю $\gamma = 0{,}95$.

Зауваження. Із зростанням об'єму вибірки розподіл Стьюдента наближується до нормального. Тому вже при $n \geq 30$ при побудові довірчих інтервалів для оцінки математичного сподівання a можна використовувати формулу (6.7), в якій покласти $\sigma = S$.

Контрольні запитання

1. Яка статистична оцінка називається інтервальною? Точковою?
2. Сформулювати постановку задачі інтервального оцінювання.
3. Для яких об'ємів вибірки застосовується інтервального оцінювання? Точкове оцінювання?
4. Як означається довірча ймовірність (надійність) статистичної оцінки?
5. Як визначити довірчий інтервал для математичного сподівання нормального розподілу при відомому середньому квадратичному відхиленні σ?
6. Як можна підвищити точність статистичної оцінки?
7. За якою формулою знаходиться мінімальний об'єм вибірки для оцінки математичного сподівання з наперед заданою точністю δ і надійністю γ?
8. Як знайти довірчий інтервал для математичного сподівання нормального розподілу при невідомому середньому квадратичному відхиленні σ?
9. Як побудувати довірчий інтервал для середнього квадратичного відхилення?
10. Як застосовують точкові оцінки в інтервальному оцінюванні?

Тема 7. Перевірка статистичних гіпотез. Критерій Пірсона. Критерій згоди Колмогорова

7.1. Постановка задачі. Статистичний критерій. Критична область

Статистичною гіпотезою називають гіпотезу про вигляд невідомого розподілу генеральної сукупності (***непараметрична гіпотеза***) або про параметри відомих розподілів (***параметрична гіпотеза***).

Так, наприклад, гіпотеза про те, що час безвідмовної роботи певного елементу приладу має показниковий розподіл, є непараметричною гіпотезою (гіпотезою про закон розподілу). Гіпотеза про те, що середні розміри деталей, які виготовляються на однотипних паралельно працюючих станках, не відрізняються між собою, є параметричною гіпотезою (гіпотезою про параметри розподілу).

Найбільш точні висновки про справедливість статистичної гіпотези можна зробити на підставі дослідження всієї генеральної сукупності. Проте на практиці, як правило, таке дослідження неможливе, і тому висновки про істинність (хибність) статистичних гіпотез приймають на основі вибірки об'єму n.

Процес використання вибірки для перевірки істинності (хибності) статистичної гіпотези називають ***статистичною перевіркою*** цієї гіпотези.

Гіпотезу, що підлягає перевірці, називають ***основною*** або ***нульовою гіпотезою*** і позначають H_0. Разом з основною H_0 розглядають гіпотезу, що суперечить основній. Її називають ***конкуруючою (альтернативною) гіпотезою*** і позначають H_1.

Наприклад, якщо нульова гіпотеза полягає в тому, що математичне сподівання a нормального розподілу дорівнює числу a_0, то альтернативна гіпотеза може стверджувати, зокрема, що $a \neq a_0$. Записують це так:

$$H_0: a = a_0 \quad H_1: a \neq a_0.$$

Висунута нульова гіпотеза H_0 може бути вірною або невірною. В результаті її статистичної перевірки можна допустити помилку одного з двох типів.

Відхилення основної гіпотези, коли вона істинна (прийняття альтернативної), називають ***помилкою першого роду***; прийняття основної гіпотези, коли істинною є альтернативна (відхилення альтернативної), називають ***помилкою другого роду***.

Для перевірки основної гіпотези H_0 використовують спеціально

підібрану випадкову величину, яка є функцією вибіркових даних і розподіл якої відомий. Таку величину називають ***статистичним критерієм*** або просто ***критерієм*** і позначають через K.

За вибірковими даними обчислюють значення всіх величин, які входять у критерій і отримують окреме значення критерію $K_{спост}$. Множину всіх можливих значень критерію розбивають на дві підмножини: критичну область і область прийняття гіпотези.

Сукупність значень критерію, при яких основну гіпотезу H_0 відхиляють, називають ***критичною областю***. Сукупність значень критерію, при яких гіпотеза H_0 вважається правдоподібною, називають ***областю прийняття гіпотези.***

Оскільки критерій K − одновимірна випадкова величина, всі її можливі значення належать деякому інтервалу. Тому критична область і область прийняття гіпотези також є інтервалами, а отже існують точки, які їх розділяють. Ці ***точки*** називають ***критичними***.

Розрізняють правосторонню, лівосторонню і двосторонню критичні області.

1). Нехай, наприклад, статистичний критерій K має ***правосторонню критичну область***: $K > K_{крит}$, де $K_{крит}$ − критична точка (рис.7.1). Щоб побудувати цю область (знайти $K_{крит}$), задають досить малу імовірність α (***рівень значущості***) і потім за допомогою таблиці критичних точок критерію K знаходять $K_{крит}$, виходячи з умови

$$P\left(K > K_{крит}\right) = \alpha . \qquad (7.1)$$

Після того, як критичну точку знайдено, за вибірковими даними обчислюють значення критерію $K_{спост}$ і якщо $K_{спост} > K_{крит}$, нульову гіпотезу H_0 відкидають. Якщо ж $K_{спост} < K_{крит}$, то гіпотезу H_0 вважають такою, що узгоджується з дослідними даними.

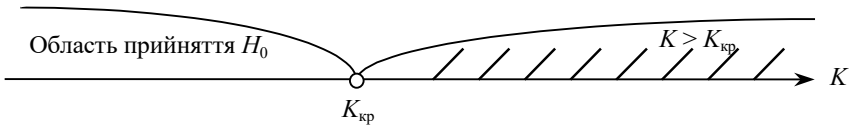


Рис. 7.1.

Зауваження. 1. Довірча ймовірність γ та рівень значущості α зв'язані між собою таким співвідношенням: $\alpha = 1 - \gamma$.

2. У процесі перевірки $K_{спост}$ може виявитись більшим за $K_{крит}$ не тому, що гіпотеза H_0 невірна, а з інших причин (недостатній об'єм

вибірки, недоліки у проведенні експерименту і т.і.). У цьому випадку, відхиливши правильну основну гіпотезу, ми допустимо помилку першого роду. Ймовірність такої помилки дорівнює рівню значущості α.

На практиці для більшої впевненості прийняття гіпотези її перевіряють за допомогою інших критеріїв або повторюють експеримент, збільшивши об'єм вибірки.

Відхиляють гіпотезу з більшою категоричністю, ніж її приймають.

2). Якщо при $K < K_{кр}$ нульова гіпотеза відхиляється, то в цьому разі ми маємо ***лівосторонню критичну область***. (рис.7.2).

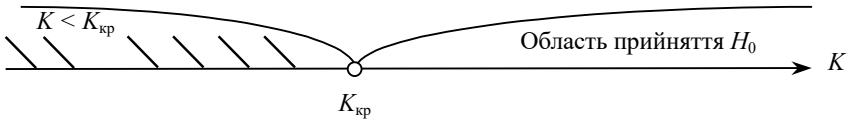


Рис. 7.2.

3). Якщо ж при $K < K'_{кр}$ і при $K > K''_{кр}$ нульова гіпотеза відхиляється, то маємо двосторонню критичну область (рис. 7.3).

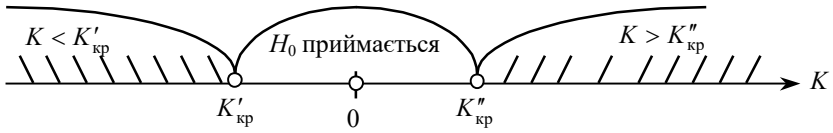


Рис. 7.3.

7.2. Перевірка гіпотез про закон розподілу. Критерій Пірсона

7.2.1. Постановка задачі.

На практиці часто доводиться мати справу з генеральною сукупністю, закон розподілу якої невідомий. Результати попереднього статистичного аналізу вибіркових даних (вигляд гістограми, вибіркові числові характеристики і т.і.) дозволяють висунути гіпотезу про вид закону розподілу, яка потребує статистичної перевірки.

Нехай потрібно перевірити гіпотезу H_0 про те, що ознака X генеральної сукупності має функцію розподілу $F(x)$. За вибіркою можна побудувати емпіричну функцію розподілу $F^*(x)$ випадкової величини X. Порівняння емпіричного $F^*(x)$ та теоретичного $F(x)$ розподілів здійснюється за допомогою спеціально підібраної випадкової величини − ***критерію згоди***.

До найбільш поширених критеріїв відносять критерій згоди χ^2 Пірсона (для ДВВ та НВВ) та критерій λ Колмогорова (для НВВ).

7.2. 2. Критерій згоди Пірсона

І. Випадок інтервального варіаційного ряду.

Нехай з генеральної сукупності здійснено вибірку об'єму n. Весь діапазон результатів розбиваємо на l інтервалів Δ_1, Δ_2,…, Δ_l та будуємо інтервальний варіаційний ряд

Інтервали Δ_i	Δ_1	Δ_2	…	Δ_i	…	Δ_l
Частоти n_i	n_1	n_2	…	n_i	…	n_l

де $\sum_{i=1}^{l} n_i = n$.

Для перевірки гіпотези про те, що випадкова величина X має функцію розподілу $F(x)$, додержуються такого ***алгоритму.***

1. За допомогою гіпотетичної функції розподілу $F(x)$ обчислюють ймовірності p_i потрапляння випадкової величини X в інтервали Δ_i. Якщо, наприклад, $\Delta_i = [x_{i-1}, x_i[$, то

$$p_i = P(x_{i-1} \le X < x_i) = F(x_i) - F(x_{i-1}) \ \ (i = \overline{1,l}).$$

2. Перемножуючи отримані ймовірності p_i та об'єм вибірки n, знаходять теоретичні частоти $n \cdot p_i$ для інтервалів Δ_i.

3. Обчислюють вибіркову статистику (критерій) χ^2:

$$\chi^2_{спост} = \sum_{i=1}^{l} \frac{(n_i - np_i)^2}{np_i}. \qquad (7.2)$$

4. Визначають число k ступенів свободи за формулою $k = l - r - 1$, де l – число частинних інтервалів, r – число параметрів, які характеризують гіпотетичний розподіл $F(x)$.

5. За заданим рівнем значущості α та числом ступенів свободи $k = l - r - 1$ знаходять з таблиці додатку 4 критичну точку $\chi^2_{кр}(\alpha; k)$.

6. Якщо $\chi^2_{спост} > \chi^2_{кр}(\alpha; k)$, висунуту гіпотезу відхиляють, тобто вважається, що гіпотетична функція розподілу $F(x)$ не узгоджується з дослідними даними. Якщо ж $\chi^2_{спост} < \chi^2_{кр}(\alpha; k)$, то вважається, що

гіпотетична функція розподілу $F(x)$ узгоджується з дослідними даними.

Зауваження. 1. Доведено, що при $n \to \infty$ закон розподілу випадкової величини (3.2) незалежно від того, як розподілена генеральна сукупність, наближається до закону розподілу χ^2 зі ступенями свободи $k = l - r - 1$.

2. Критерій χ^2 побудовано так, що чим ближче до нуля значення $\chi^2_{спост}$, обчислене за формулою (7.2), тим більш ймовірно, що нульова гіпотеза є вірною. Тому для перевірки нульової гіпотези потрібно за заданим рівнем значущості α і числом ступенів свободи $k = l - r - 1$ знаходять за таблицею додатку 4 критичну точку $\chi^2_{кр}(\alpha; k)$, яка задовольняє умову

$$P(\chi^2 > \chi^2_{кр}(\alpha; k)) = \alpha \qquad (7.3)$$

(правостороння критична область).

3. Інтервали, до яких потрапило менше трьох спостережень, об'єднують.

Аналогічно використовують критерій Пірсона для перевірки гіпотези про показниковий ($r = 1$) чи рівномірний розподіл ($r = 2$) ознаки генеральної сукупності.

II. Випадок дискретного варіаційного ряду (для рівновіддалених варіант).

Нехай статистичний розподіл вибірки задано у вигляді послідовності ***рівновіддалених*** варіант і відповідних їм частот:

x_i	x_1	x_2	...	x_k
n_i	n_1	n_2	...	n_k

Для того, щоб при рівні значущості α перевірити гіпотезу про нормальний розподіл ознаки генеральної сукупності, необхідно:

1) обчислити вибіркове середнє $\bar{x}$ і вибіркове середнє квадратичне відхилення σ_B;

2) визначити теоретичні частоти

$$n'_i = \frac{nh}{\sigma}\varphi(u_i), \qquad (7.4)$$

де n — обсяг вибірки; h — крок (різниця між двома сусідніми варіантами); $\varphi(u)$ — функція Гауса;

$$u_i = \frac{x_i - \bar{x}}{\sigma}, \tag{7.5}$$

3) порівняти емпіричні та теоретичні частоти, обчисливши спостережуване значення критерію Пірсона

$$\chi^2_{спост} = \sum_{i=1}^{k} \frac{(n_i - n'_i)^2}{n'_i}, \tag{7.6}$$

4) за таблицею критичних точок розподілу χ^2 (додаток 5) при рівні значущості α і числі ступенів свободи $k = s - (r+1)$ (s - кількість варіант вибірки, r – кількість параметрів розподілу) знайти критичну точку $\chi^2_{\text{кр}}(\alpha; k)$ правосторонньої критичної області;

5) якщо $\chi^2_{спост} < \chi^2_{кр}$, то немає підстав відхиляти гіпотезу про нормальний розподіл генеральної сукупності. Іншими словами, емпіричні та теоретичні частоти різняться несуттєво (випадково).

Якщо $\chi^2_{спост} \geq \chi^2_{кр}$, то гіпотезу відхиляють, оскільки емпіричні та теоретичні частоти різняться суттєво.

Зауваження. Критерій Пірсона також застосовують для перевірки гіпотез про значення ймовірностей групованої вибірки. Гіпотеза стосується ймовірності належності кожного із можливих спостережень цим групам. Належність відповідним групам визначають інтервалом можливих значень досліджуваної характеристики для неперервних даних або окремими значеннями для дискретних даних. Отже гіпотеза критерію Пірсона має вигляд:

$$H_0: \; p_1 = p_1^{(0)}, \; p_2 = p_2^{(0)}, \; \ldots, \; p_M = p_M^{(0)}.$$

Тут p_i, $i = 1,2,...M$ − позначення для істинних, але невідомих значень ймовірностей відповідним групам, а $p_i^{(0)}$, $i = 1,2,...M$ − припущення про значення відповідних ймовірностей (одночасно про всі). Вочевидь, сума гіпотетичних ймовірностей $p_i^{(0)}$, $i = 1,2,...M$ дорівнює одиниці, оскільки групи разом утворюють повну групу подій.

7.3. Критерій згоди Колмогорова

При застосуванні критерію Колмогорова порівнюється емпірична $F^*(x)$ і гіпотетична $F(x)$ функції розподілу.

Постановка задачі. Нехай висунута гіпотеза, що випадкова величина X має неперервну функцію розподілу $F(x)$. З генеральної сукупності здійснено вибірку об’єму n ($n \geq 50$).

Перевірку нульової гіпотези здійснюють за таким ***алгоритмом***:

1. За результатами вибірки будують емпіричну функцію розподілу $F^*(x)$.

2. За допомогою гіпотетичної функції розподілу обчислюють значення теоретичної функції розподілу, що відповідають вибірковим значенням ознаки X.

3. Знаходять міру D відхилення теоретичної функції розподілу від емпіричної:

$$D = \max_x \left| F^*(x) - F(x) \right| .$$

4. Обчислюють значення вибіркової статистики

$$\lambda = D\sqrt{n} = \max_x \left| F^*(x) - F(x) \right| \cdot \sqrt{n} . \qquad (7.7)$$

Академік А.М. Колмогоров довів, що коли нульова гіпотеза вірна, то розподіл вибіркової статистики $\lambda = D\sqrt{n}$ при $n \to \infty$ наближається до розподілу Колмогорова

$$K(\lambda) = P(D\sqrt{n} < \lambda) = \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2\lambda^2} .$$

Якщо задати рівень значущості α, то із співвідношення

$$P(\lambda > \lambda_\alpha) = P(D\sqrt{n} > \lambda_\alpha) = 1 - \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2\lambda_\alpha^2} = \alpha$$

за таблицею критичних точок розподілу Колмогорова (див. додаток 6) знаходять відповідну критичну точку λ_α.

5. Порівнюємо обчислене за вибіркою $\lambda_{спост} = D\sqrt{n}$ з критичною точкою λ_α.

Якщо $\lambda_{спост} > \lambda_\alpha$, то нульова гіпотеза відхиляється. Якщо ж $\lambda_{спост} < \lambda_\alpha$, то вважається, що гіпотетична $F(x)$ функція розподілу узгоджується з дослідними даними.

Зауваження. Критерій Колмогорова доцільно застосовувати тоді, коли відомий не тільки вигляд гіпотетичної функції розподілу $F(x)$, а й параметри, що її визначають. Якщо ж параметри оцінюються за вибіркою, критерій не дає високої степені точності.

Приклад 7.1. З генеральної сукупності X здійснено вибірку об'єму n=100, яка задана у вигляді послідовності інтервалів і відповідних їм частот:

Інтервали	4 - 6	6 - 8	8 - 10	10 - 12	12 - 14	14 - 16
Частоти n_i	11	16	21	23	19	10

Вважаючи відомими $a=M(X)=10$, $\sigma=\sigma(X)=4$, перевірити за допомогою критерію λ Колмогорова при рівні значущості 0,05 гіпотезу про те, що вибірка здійснена з нормально розподіленої генеральної сукупності.

Розв'язання. Функція розподілу випадкової величини X, розподіленої за нормальним законом з параметрами $a = 10$, $\sigma = 4$, має вигляд $F(x)=\frac{1}{2}+\Phi\left(\frac{x-a}{\sigma}\right)$, тобто

$$F(x)=\frac{1}{2}+\Phi\left(\frac{x-10}{4}\right), \qquad (7.8)$$

де $\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_0^x e^{-\frac{t^2}{2}}dt$ – функція Лапласа (додаток 2); $\Phi(-x)=-\Phi(x)$ **(**функція Лапласа непарна). Для $x>5$ приймають $\Phi(x)=0{,}5$.

За вибірковими даними (n = 100) обчислюємо значення емпіричної функції розподілу $F^*(x)=\frac{n_x}{n}$, теоретичної функції розподілу $F(x)$ та абсолютні величини різниць $F^*(x_i) - F(x_i)$. Результати обчислень зводимо в таблицю

Таблиця 7.1**.**

x_i	$F^*(x_i)$	$F(x_i)$	$\lvert F^*(x_i) - F(x_i)\rvert$
4	0	0,0668	**0,0668**
6	0,11	0,1587	0,0487
8	0,27	0,3085	0,0385
10	0,48	0,5	0,02
12	0,71	0,6915	0,0185
14	0,9	0,8413	0,0587
16	1	0,9332	**0,0668**

Розрахунки $F(x_i)$, третій стовпець таблиці 7.1.

$F(4)=\frac{1}{2}+\Phi\left(\frac{4-10}{4}\right)=\frac{1}{2}+\Phi\left(\frac{-6}{4}\right)=0{,}5+\Phi(-1{,}5)=0{,}5-\Phi(+1{,}5)=$ 0,5-0,4332=0,0668;

$$F(6)=\frac{1}{2}+\Phi\left(\frac{6-10}{4}\right)=\frac{1}{2}+\Phi\left(\frac{-4}{4}\right)=0{,}5+\Phi(-1)=0{,}5-\Phi(+1)=$$

=0,5-0,3413=0,1587; і т.д.

$$F(16)=\frac{1}{2}+\Phi\left(\frac{16-10}{4}\right)=\frac{1}{2}+\Phi\left(\frac{6}{4}\right)=0{,}5+\Phi(1{,}5)=0{,}5+\Phi(1{,}5)=0{,}5+0{,}4332=0{,}9332.$$

З останнього стовпця таблиці 7.1 знаходимо

$$D=\max_x\left|F^*(x_i)-F(x_i)\right|=0{,}0668\,.$$

Обчислюємо значення вибіркової статистики λ

$$\lambda_{спост}=D\sqrt{n}=0{,}0668\cdot\sqrt{100}=0{,}668$$

і за таблицею критичних точок розподілу Колмогорова (Додаток 5) при рівні значущості $\alpha = 0{,}05$ знаходимо критичну точку $\lambda_\alpha = 1{,}358$.

Оскільки $\lambda_{спост} < \lambda_\alpha$, то немає підстав для відхилення гіпотези про нормальний розподіл генеральної сукупності.

Контрольні запитання

1. Що називають непараметричною (параметричною) гіпотезою?
2. Що розуміють під процесом статистичної перевірки гіпотези?
3. Яку гіпотезу називають основною (нульовою) гіпотезою?
4. Яку гіпотезу називають конкуруючою (альтернативною)?
5. Як означається помилка першого роду? Помилка другого роду?
6. Як перевіряються гіпотези про закон розподілу?
7. Що називають статистичним критерієм?
8. Як означається критична область? Критична точка?
9. Які існують типи критичних областей?
10. Що таке рівень значущості? Як рівень значущості зв'язаний з довірчою ймовірністю (надійністю)?
11. Як перевіряється гіпотеза про закон розподілу за допомогою критерію згоди χ^2 (критерію Пірсона)? Опишіть алгоритм.
12. Як перевіряється гіпотеза про закон розподілу за допомогою критерію згоди λ Колмогорова? Опишіть алгоритм.

Тема 8. Статистична (кореляційна) залежність між величинами. Вибірковий коефіцієнт кореляції. Лінійна регресія.

8.1. Залежність між випадковими величинами. Рівняння регресії

У багатьох наукових дослідженнях виникає необхідність проводити одночасно спостереження над кількома випадковими величинами, щоб встановити та оцінити їх взаємозв'язок.

Дві випадкові величини X та Y можуть бути зв'язані або функціональною, або статистичною залежністю, або ж бути взагалі незалежними.

Якщо кожному можливому значенню x величини X відповідає певне значення y іншої величини Y, то кажуть, що випадкові величини X та Y зв'язані функціональною залежністю.

Строга функціональна залежність реалізується дуже рідко, оскільки випадкові величини X та Y (або ж одна з них) зазнають впливу багатьох випадкових факторів, серед яких можуть бути і спільні.

У цьому випадку між двома величинами X та Y виникає ***статистична залежність***.

Залежність між випадковими величинами X та Y, за якою кожному значенню однієї величини відповідає розподіл іншої, називається ***статистичною***.

Зокрема, якщо кожному можливому значенню однієї величини ставиться у відповідність середнє значення іншої, то така статистична залежність називається ***кореляційною***.

Для випадку кореляційної залежності, якщо величина X прийняла значення x, то математичне сподівання величини Y є при цьому функцією від x:

$$M_x Y = f(x). \qquad (8.1)$$

Рівняння (8.1) називається ***рівнянням регресії*** Y ***на*** X**.**

Оскільки математичне сподівання є істинним (справжнім) значенням величини Y, що спостерігається, то рівняння регресії (8.1) дає справжню залежність між величинами X та Y. Тому кінцевою метою багатьох досліджень є знаходження вибіркового рівняння регресії (8.1), яке прийнято записувати у вигляді

$$\bar{y}_x = f(x). \qquad (8.2)$$

Тут $\bar{y}_x$ – ***умовне середнє*** (це середнє арифметичне значень випадкової величини Y, що відповідають значенню $X = x$); f – функція регресії Y на X.

Графік (8.2) називають ***вибірковою лінією регресії*** Y ***на*** X**.**
Аналогічно рівняння

$$\bar{x}_y = \varphi(y) \qquad (8.3)$$

називається ***вибірковим рівнянням регресії*** X ***на*** Y**.** При цьому рівнянням (8.2) та (8.3) у загальному випадку відповідають дві різні лінії на площині XOY.

Нехай результати вибірки подано у вигляді так званої ***кореляційної таблиці***.

Кореляційна таблиця.

X/Y	y_1	y_2	…	y_k	n_{x_i}	$\bar{y}_{x_i}$
x_1	n_{11}	n_{12}	…	n_{1k}	n_{x_1}	$\bar{y}_{x_1}$
x_2	n_{21}	n_{22}	…	n_{2k}	n_{x_2}	$\bar{y}_{x_2}$
…	…	…	…	…	…	…
x_m	n_{m1}	n_{m2}	…	n_{mk}	n_{x_m}	$\bar{y}_{x_m}$
n_{y_j}	n_{y_1}	n_{y_2}	…	n_{y_k}	n	–

Якщо розглядати таблицю за рядками, то кожному значенню x_i відповідає деякий розподіл випадкової величини Y. Обчислимо для

ціх розподілів умовні середні значення

$$\bar{y}_{x_i} = \frac{\sum_{j=1}^{k} y_j n_{ij}}{n_{x_i}},\ i = 1,2,...,m\ . \qquad (8.4)$$

Отже, маємо залежність (8.2). Аналогічно, розглядаючи таблицю за стовпцями, визначаємо умовні середні величини

$$\bar{x}_{y_j} = \frac{\sum_{i=1}^{m} x_i n_{ij}}{n_{y_j}},\ j = 1,2,...,k\ . \qquad (8.5)$$

Приходимо до залежності вигляду (8.3).

У кореляційному аналізі при дослідженні залежності кількісних ознак X та Y розглядають дві основні задачі:

1) знайти наближену функцію регресії, що характеризує основну тенденцію залежності Y від X (або X від Y) та належить одному з відомих типів функцій (лінійна, квадратична, показникова, логарифмічна і т.і.);

2) оцінити силу, тісноту цієї залежності, тобто визначити ступінь розсіювання можливих значень однієї випадкової величини відносно лінії регресії, якщо одна із величин набуває певних значень.

8.2. Лінійна регресія. Метод найменших квадратів

Нехай вивчається вибірка об'єму n з двох кількісних ознак X та Y: $(x_1; y_1),(x_2; y_2),...,(x_n; y_n)$, причому значення $x_i; y_i, i = \overline{1,n}$ зустрічаються по одному разу.

У цьому випадку немає необхідності групувати дані та використовувати поняття умовної середньої. Тому шукане рівняння регресії можна записати так;

$$y = f(x) \qquad (8.6)$$

або

$$x = \varphi(y). \tag{8.7}$$

Наближений вигляд згладжувальної функції f (або φ) можна визначити виходячи з теоретичних міркувань або ж за характером розташування на координатній площині експериментальних точок $(x_1; y_1), (x_2; y_2), \ldots, (x_n; y_n)$. Це так зване ***поле розсіювання.***

При вибраному вигляді згладжувальної функції $y = f(x, a, b)$ параметри a та b потрібно підібрати так, щоб сума квадратів відхилень y_i від $f(x_i, a, b)$ була найменшою:

$$S = S(a,b) = \sum_{i=1}^{n} [y_i - f(x_i, a, b)]^2 \to \min_{a,b} \tag{8.8}$$

У цьому розумінні функція f за методом найменших квадратів "найкращим чином" описує відповідний процес.

Задача (8.8) – це задача на безумовний екстремум функції двох змінних $S = S(a,b)$. На підставі необхідної умови екстремуму невідомі параметри a, b знаходимо за умовою

$$\begin{cases} \dfrac{\partial S}{\partial a} = 0\ , \\ \dfrac{\partial S}{\partial b} = 0\ . \end{cases} \tag{8.9}$$

Припустимо, що точки $(x_i; y_i)$, $i = 1,2,...,n$ групуються навколо прямої лінії, як на рис.8.1.

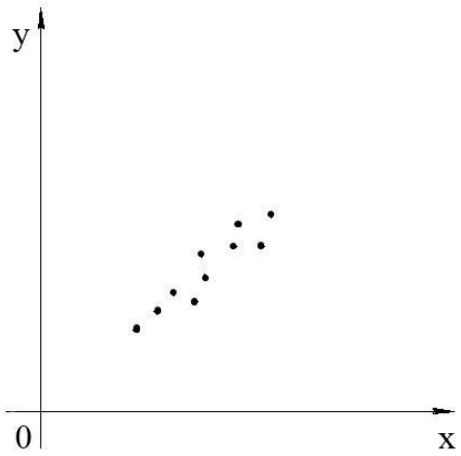


Рис. 8.1.

Тоді емпіричну функцію регресії шукають у вигляді прямої

$$y = ax + b, \tag{8.10}$$

де a, b – невідомі параметри.

Для випадку лінійної згладжувальної функції маємо

$$S(a,b) = \sum_{i=1}^{n} \left[y_i - (ax_i + b) \right]^2,$$

а система (8.9) набуває вигляду

$$\begin{cases} \sum_{i=1}^{n} (y_i - ax_i - b) x_i = 0, \\ \sum_{i=1}^{n} (y_i - ax_i - b) = 0 \end{cases}$$

або, остаточно

$$\begin{cases} a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i, \\ a\sum_{i=1}^{n} x_i + b \cdot n = \sum_{i=1}^{n} y_i. \end{cases} \tag{8.11}$$

Система (8.11) називається ***нормальною системою методу найменших квадратів*** для відшукання параметрів лінійної залежності. Це неоднорідна система двох лінійних рівнянь відносно невідомих a, b. Знайшовши розв'язок системи (8.11)

$$a = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2}, \; b = \frac{\sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} x_i}{n}, \tag{8.12}$$

дістанемо лінійну функцію регресії Y на X.

Аналогічно знаходиться емпірична лінійна функція регресії X на Y: $x = cy + d$. Для цього випадку параметри c, d будуть розв'язками системи

$$\begin{cases} c\sum_{i=1}^{n} y_i^2 + d\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} x_i y_i \ , \\ c\sum_{i=1}^{n} y_i + d \cdot n = \sum_{i=1}^{n} x_i \ . \end{cases} \tag{8.13}$$

та знаходяться за формулами

$$c = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}, \ d = \frac{\sum_{i=1}^{n} x_i - c\sum_{i=1}^{n} y_i}{n}. \tag{8.14}$$

Приклад 8.1. За даними спостережень величин X та Y

x_i	1	2	3	4
y_i	3	4	6	8

знайти функцію регресії Y на X.

Розв'язання. Побудуємо точки $M_i(x_i, y_i)$, $i = 1,2,3,4$ в прямокутній системі координат на площині XOY (рис. 8.2).

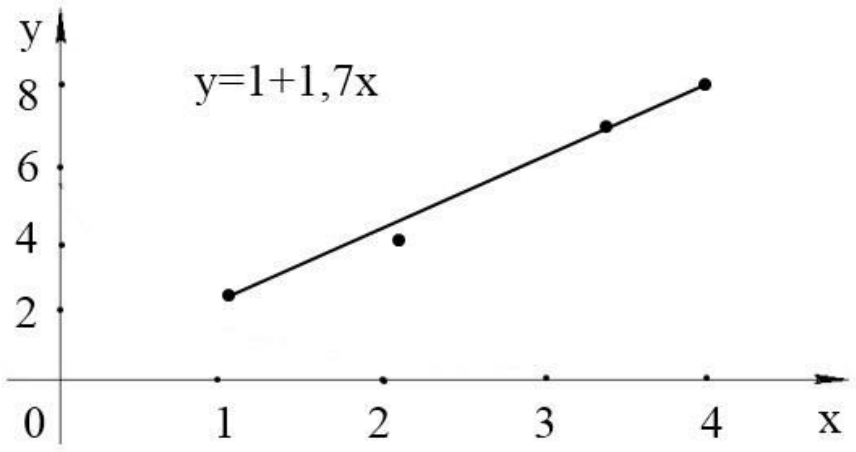


Рис. 8.2.

Ці точки групуються навколо прямої лінії, тому будемо підбирати лінійну функцію регресії. Обчислимо суми, що входять у формули (8.12).

Для зручності результати обчислень заносимо в таблицю:

	x_i	y_i	$x_i\ y_i$	x_i^2
	1	3	3	1
	2	4	8	4
	3	6	18	9
	4	8	32	16
$\sum$	10	21	61	30

Знаходимо розв'язок нормальної системи за готовими формулами (8.12):

$$a = \frac{4 \cdot 61 - 10 \cdot 21}{4 \cdot 30 - 10^2} = \frac{244 - 210}{120 - 100} = \frac{34}{20} = 1{,}7\,,\ b = \frac{21 - 1{,}7 \cdot 10}{4} = \frac{4}{4} = 1\,.$$

Отже, емпірична регресія (оцінка істинної регресії) має вигляд $y = 1{,}7x + 1$. Графічно емпіричну регресію зображують у вигляді прямої, що групує спостережувані дані (рис. 8.2).

Якщо число спостережень n велике ($n \geq 50$), експериментальні дані прийнято групувати та записувати у вигляді кореляційної таблиці (див. п.8.2). У цій таблиці враховується частота n_x повторення значення x випадкової величини X у вибірці; частота n_y повторення значення y випадкової величини Y і частота n_{xy} повторення пари (x, y) значень випадкових величин (X, Y) у вибірці. При цьому $\sum n_x = \sum n_y = \sum n_{xy} = n$ (для спрощення поточні індекси в сумах опускаємо).

У випадку згрупованих даних система для знаходження коефіцієнтів a, b лінійної емпіричної функції регресії Y на X

$$\bar{y}_x = ax + b \tag{8.15}$$

набуває вигляду

$$\begin{cases} a\sum n_x x^2 + b\sum n_x x = \sum n_{xy} xy, \\ a\sum n_x x + nb = \sum n_y y, \end{cases} \tag{8.16}$$

а її розв'язок:

$$a = \frac{n\sum n_{xy} xy - \sum n_x x \sum n_y y}{n\sum n_x x^2 - \left(\sum n_x x\right)^2};\ b = \frac{1}{n}\left(\sum n_y y - a\sum n_x x\right). \tag{8.17}$$

Аналогічно трансформується система (8.13) для знаходження коефіцієнтів c, d лінійної регресії X на Y $\bar{x}_y = cy + d$ та формули (8.14) для обчислення цих коефіцієнтів.

Слід зазначити, що $\bar{y}_x = ax + b$ та $\bar{x}_y = cy + d$ – різні прямі (рис.8.3).

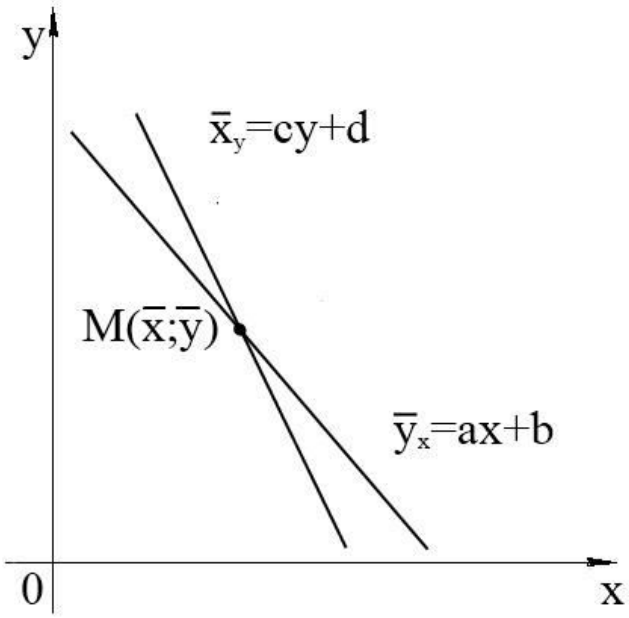


Рис. 8.3.

Перша пряма одержується в результаті розв'язання задачі про мінімізацію суми квадратів відхилень по вертикалі, а друга − по горизонталі. Прямі лінійної регресії перетинаються в точці $M(\bar{x},\bar{y})$, що називається ***центром кореляції***.

8.3. Вибірковий коефіцієнт кореляції

До основних характеристик, що описують силу зв'язку між випадковими величинами X та Y, відносять ***кореляційний момент***

$$\mu_{xy} = M[(X - M(X))(Y - M(Y))] \tag{8.18}$$

та ***коефіцієнт кореляції***

$$r_{xy} = \frac{\mu_{xy}}{\sigma(X)\sigma(Y)}, \tag{8.19}$$

для обчислення яких потрібно знати закони розподілу величин X та Y.

При обробці експериментальних даних, як правило, закони розподілу невідомі. Тому для оцінки сили зв'язку між величинами X та Y застосовують точкові оцінки μ_{xy} і r_{xy} − ***вибірковий кореляційний момент***

$$K_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \tag{8.20}$$

та ***вибірковий коефіцієнт кореляції***

$$r_B = \frac{K_{xy}}{\sigma_x \sigma_y} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - y)^2}} \tag{8.21}$$

або

$$r_B = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}\sqrt{\sum_{i=1}^{n}\left(y_i - y\right)^2}}, \qquad (8.22)$$

де $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ − відповідні середні значення (вибіркові середні).

Вибірковий коефіцієнт кореляції, як оцінка коефіцієнта кореляції генеральної сукупності, слугує для вимірювання лінійного зв'язку між величинами − кількісними ознаками X та Y.

Властивості вибіркового коефіцієнта кореляції:

1. $-1 \le r_B \le 1$ (або $|r_B| \le 1$).

2. Чим більшою є величина r_B , тим тісніший зв'язок між досліджуваними ознаками X та Y.

3. Якщо $|r_B| = 1$, то кореляційна залежність між X та Y стає лінійною функціональною.

4. Якщо $r_B = 0$, то між досліджуваними ознаками X та Y немає лінійної кореляційної залежності, але умова $r_B = 0$ не виключає існування будь-якої іншої кореляційної залежності (параболічної, показникової і т.і.).

Якщо з деяких теоретичних міркувань заздалегідь відомо, що величини X та Y мають нормальний розподіл, то рівність $r_B = 0$ свідчить про відсутність будь-якої залежності між ознаками X та Y (тобто величини X , Y − незалежні).

Щоб одержати ще одну формулу для розрахунку r_B , скористаємося відомими формулами

$$D(X) = M\left(X^2\right) - \left[M(X)\right]^2,\ \sigma(X) = \sqrt{D(X)},$$
$$D(Y) = M\left(Y^2\right) - \left[M(Y)\right]^2,\ \sigma(Y) = \sqrt{D(Y)}.$$

та перетворимо (8.18) до вигляду

$$\mu_{xy} = M(X \cdot Y) - M(X) \cdot M(Y).$$

Тоді

$$\sigma_x^2 = \overline{x^2} - (\bar{x})^2,\ \sigma_y^2 = \overline{y^2} - (\bar{y})^2,\ K_{xy} = = \overline{x \cdot y} - \bar{x} \cdot \bar{y},$$

а формула для обчислення вибіркового коефіцієнта кореляції буде такою:

$$r_B = \frac{\overline{x \cdot y} - \bar{x} \cdot \bar{y}}{\sqrt{\overline{x^2} - (\bar{x})^2} \cdot \sqrt{\overline{y^2} - (\bar{y})^2}} \tag{8.23}$$

або

$$r_B = \frac{\overline{x \cdot y} - \bar{x} \cdot \bar{y}}{\sigma_x \cdot \sigma_y}, \tag{8.24}$$

де $\sigma_x = \sqrt{\overline{x^2} - (\bar{x})^2}$, $\sigma_y = \sqrt{\overline{y^2} - (\bar{y})^2}$ – середні квадратичні відхилення величин X та Y відповідно.

Щоб встановити зв'язок r_B з вибірковим лінійним рівнянням регресії $\bar{y}_x = ax + b$, запишемо нормальну систему для визначення коефіцієнтів a, b у вигляді

$$\begin{cases} a\overline{x^2} + b\bar{x} = \overline{x \cdot y}, \\ a\bar{x} + b = \bar{y}. \end{cases} \tag{8.25}$$

Далі знайдемо з другого рівняння системи $b = \bar{y} - a\bar{x}$ та підставимо в рівняння регресії:

$$\bar{y}_x = ax + \bar{y} - a\bar{x}$$

або

$$\bar{y}_x - \bar{y} = a(x - \bar{x}). \tag{8.26}$$

Для визначення a помножимо друге рівняння системи (8.25) на $\bar{x}$ та віднімемо його від першого рівняння:

$$a\overline{x^2} - a(\bar{x})^2 = \overline{x \cdot y} - \bar{x} \cdot \bar{y}.$$

Звідси

$$a = \frac{\overline{x \cdot y} - \bar{x} \cdot \bar{y}}{\overline{x^2} - (\bar{x})^2}$$

або

$$a = \frac{\overline{x \cdot y} - \bar{x} \cdot \bar{y}}{\sigma_x^2}. \qquad (8.27)$$

Порівнюючи (8.27) з виразом для коефіцієнта кореляції (8.24), приходимо до такої формули зв'язку:

$$a = \frac{\sigma_y}{\sigma_x} r_{xy}. \qquad (8.28)$$

Тепер вираз для a (8.28) підставимо в лінійне рівняння регресії Y на X:

$$\bar{y}_x - \bar{y} = r_{xy} \frac{\sigma_y}{\sigma_x} (x - \bar{x}). \qquad (8.29)$$

Аналогічно можна знайти вибіркове рівняння прямої лінії регресії X на Y:

$$\bar{x}_y - \bar{x} = r_{xy} \frac{\sigma_x}{\sigma_y} (y - \bar{y}). \qquad (8.30)$$

Контрольні запитання

1. Яку залежність між випадковими величинами називають функціональною? статистичною? кореляційною?
2. Як означається рівняння регресії Y на X (X наY)?
3. Що називають умовною середньою?
4. Як записують вибіркове рівняння регресії?
5. Як називаютьграфік вибіркового рівняння регресії?
6. Опишіть структуру кореляційної таблиці. Чи можливо скласти кореляційну таблицю для неперервних випадкових величин?
7. Які дві основні задачі розглядають у кореляційному аналізі?
8. Як знайти наближений вигляд згладжувальної функції? Що таке поле розсіювання?
9. У чому полягає суть методу найменших квадратів?
10. Який вигляд має нормальна система методу найменших квадратів для випадку лінійної згладжувальної функції?
11. Як означається вибірковий коефіцієнт кореляції? Оцінкою чого слугує вибірковий коефіцієнт кореляції?
12. Сформулювати властивості вибіркового коефіцієнта кореляції.

Додатки.

Додаток 1. Таблиця значень функції Лапласа $\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_0^x e^{-\frac{z^2}{2}}dz$.

x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$
0,00	0,0000	0,32	0,1255	0,64	0,2389	0,96	0,3315
0,01	0,0040	I0,33	0,1293	0,65	0,2422	0,97	0,3340
0,02	0,0080	0,34	0,1331	0,66	0,2454	0,98	0,3365
0,03	0,0120	0,35	0,1368	0,67	0,2486	0,99	0,3389
0,04	0,0160	0,36	0,1406	0,68	0,2517	1,00	0,3413
0,05	0,0199	0,37	0,1443	0,69	0,2549	1,01	0,3438
0,06	0,0239	0,38	0,1480	0,70	0,2580	1,02	0,3461
0,07	0,0279	0,39	0,1517	0,71	0,2611	1,03	0,3485
0,08	0,0319	0,40	0,1554	0,72	0,2642	1,04	0,3508
0,09	0,0359	0,41	0,1591	0,73	0,2673	1,05	0,3531
0,10	0,0398	0,42	0,1628	0,74	0,2703	1,06	0,3554
0,11	0,0438	0,43	0,1664	0,75	0,2734	1,07	0,3577
0,12	0,0478	0,44	0,1700	0,76	0,2764	1,08	0,3599
0,13	0,0517	0,45	0,1736	0,77	0,2794	1,09	0,3621
0,14	0,0557	0,46	0,1772	0,78	0,2823	1,10	0,3634
0,15	0,0596	0,47	0,1808	0,79	0,2852	1,11	0,3665
0,16	0,0636	0,48	0,1844	0,80	0,2881	1,12	0,3686
0,17	0,0675	0,49	0,1879	0,81	0,2910	1,13	0,3708
0,18	0,0714	0,50	0,1915	0,82	0,2939	1,14	0,3729
0,19	0,0753	0,51	0,1950	0,83	0,2967	1,15	0,3749
0,20	0,0793	0,52	0,1985	0,84	0,2995	1,16	0,3770
0,21	0,0832	0,53	0,2019	0,85	0,3023	1,17	0,3790
0,22	0,0871	0,54	0,2054	0,86	0,3051	1,18	0,3810
0,23	0,0910	0,55	0,2088	0,87	0,3078	1,19	0,3830
0,24	0,0948	0,56	0,2123	0,88	0,3106	1,20	0,3849
0,25	0,0987	0,57	0,2157	0,89	0,3133	1,21	0,3869
0,26	0,1026	0,58	0,2190	0,90	0,3159	1,22	0,3883
0,27	0,1064	0,59	0,2224	0,91	0,3186	1,23	0,3907
0,28	0,1103	0,60	0,2257	0,92	0,3212	1,24	0,3925
0,29	0,1141	0,61	0,2291	0,93	0,3238	1,25	0,3944
0,30	0,1179	0,62	0,2324	0,94	0,3264		
0,31	0,1217	0,63	0,2357	0,95	0,3289		

Продовження додатку 1.

x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$
1,26	0,3962	1,59	0,4441	1,92	0,4726	2,50	0,4938
1,27	0,3980	1,60	0,4452	1,93	0,4732	2,52	0,4941
1,28	0,3997	1,61	0,4463	1,94	0,4738	2,54	0,4945
1,29	0,4015	1,62	0,4474	1,95	0,4744	2,56	0,4948
1,30	0,4032	1,63	0,4484	1,96	0,4750	2,58	0,4951
1,31	0,4049	1,64	0,4495	1,97	0,4756	2,60	0,4953
1,32	0,4066	1,65	0,4505	1,98	0,4761	2,62	0,4956
1,33	0,4082	1,66	0,4515	1,99	0,4767	2,64	0,4959
1,34	0,4099	1,67	0,4525	2,00	0,4772	2,66	0,4961
1,35	0,4115	1,68	0,4535	2,02	0,4783	2,68	0,4963
1,36	0,4131	1,69	0,4545	2,04	0,4793	2,70	0,4965
1,37	0,4147	1,70	0,4554	2,06	0,4803	2,72	0.4967
1,38	0,4162	1,71	0,4564	2,08	0,4812	2,74	0,4969
1,39	0,4177	1,72	0,4573	2,10	0,4821	2,76	0,4971
1,40	0,4192	1,73	0,4582	2,12	0,4830	2,78	0,4973
1,41	0,4207	1,74	0,4591	2,14	0,4838	2,80	0,4974
1,42	0,4222	1,75	0,4599	2,16	0,4846	2,82	0,4976
1,43	0,4256	1,76	0,4608	2,18	0,4854	2,84	0,4977
1,44	0,4251	1,77	0,4616	2,20	0,4861	2,86	0,4979
1,45	0,4265	1,78	0,4625	2,22	0,4868	2,88	0,4980
1,46	0,4279	1,79	0,4633	2,24	0,4875	2,90	0,4981
1,47	0,4292	1,80	0,4641	2,26	0,4881	2,92	0,4982
1,48	0,4306	1,81	0,4649	2,28	0,4887	2,94	0,4984
1,49	0,4319	1,82	0,4656	2,30	0,4893	2,96	0,4985
1,50	0,4332	1,83	0,4664	2,32	0,4898	2,98	0,4986
1,51	0,4345	1,84	0,4671	2,34	0,4904	3,00	0,49865
1,52	0,4357	1,85	0,4678	2,36	0,4909	3,20	0,49951
1,53	0,4370	1,86	0,4686	2,38	0,4913	3,40	0,49966
1,54	0,4382	1,87	0,4693	2,40	0,4918	3,60	0,499841
1,55	0,4394	1,88	0,4699	2,42	0,4922	3,80	0,499928
1,56	0,4406	1,89	0,4706	2,44	0,4927	4,00	0,499968
1,57	0,4418	1,90	0,4713	2,46	0,4931	4,50	0,499997
1,58	0,4429	1,91	0,4719	2,48	0,4934	5,00	0,499997

Додаток 2. Таблиця значень функції Гаусса $\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$.

x	0	1	2	3	4	5	6	7	8	9
0,0	0,3989	3989	3989	3988	3986	3984	3982	3980	3977	3973
0,1	3970	3965	3961	3956	3951	3945	3939	3932	3925	3918
0,2	3910	3902	3894	3885	3876	3867	3857	3847	3836	3825
0,3	3814	3802	3790	3778	3765	3752	3739	3726	3712	3697
0,4	3683	3668	3653	3637	3621	3605	3589	3572	3555	3538
0,5	3521	3503	3485	3467	3448	3429	3410	3391	3372	3352
0,6	3332	3312	3292	3271	3251	3230	3209	3187	3166	3144
0,7	3123	3101	3079	3056	3034	3011	2989	2966	2943	2920
0,8	2897	2874	2850	2827	2803	2780	2755	2732	2709	2685
0,9	2661	2637	2613	2589	2565	2541	2516	2492	2468	2444
1,0	0,2420	2396	2371	2347	2323	2299	2275	2251	2227	2203
1,1	2179	2155	2131	2107	2083	2059	2036	2012	1989	1965
1,2	1942	1919	1895	1872	1849	1826	1804	1781	1758	1736
1,3	1714	1691	1669	1647	1626	1604	1582	1561	1539	1518
1,4	1497	1476	1456	1435	1415	1394	1374	1354	1334	1315
1,5	1295	1276	1257	1238	1219	1200	1182	1163	1145	1127
1,6	1109	1092	1074	1057	1040	1023	1006	0989	0973	0957
1,7	0940	0925	0909	0893	0878	0863	0848	0833	0818	0804
1,8	0790	0775	0761	0748	0734	0721	0707	0694	0681	0669
1,9	0656	0644	0632	0620	0608	0596	0584	0573	0562	0551
2,0	0,0540	0529	0519	0508	0498	0488	0478	0468	0459	0449
2,1	0440	0431	0422	0413	0404	0396	0387	0379	0371	0363
2,2	0355	0347	0339	0332	0325	0317	0310	0303	0297	0290
2,3	0283	0277	0270	0264	0258	0252	0246	0241	0235	0229
2,4	0224	0219	0213	0208	0203	0198	0194	0189	0184	0180
2,5	0175	0171	0167	0163	0158	0154	0151	0147	0143	0139
2,6	0136	0132	0129	0126	0122	0119	0116	0113	0110	0107
2,7	0104	0101	0099	0096	0093	0091	0088	0086	0084	0081
2,8	0079	0077	0075	0073	0071	0069	0067	0065	0063	0061
2,9	0060	0058	0056	0055	0053	0051	0050	0048	0047	0046
3,0	0,0044	0043	0042	0040	0039	0038	0037	0036	0035	0034
3,1	0033	0032	0031	0030	0029	0028	0027	0026	0025	0025
3,2	0024	0023	0022	0022	0021	0020	0020	0019	0018	0018
3,3	0017	0017	0016	0016	0015	0015	0014	0014	0013	0013
3,4	0012	0012	0012	0011	0011	0010	0010	0010	0009	0009
3,5	0009	0008	0008	0008	0008	0007	0007	0007	0007	0006
3,6	0006	0006	0006	0005	0005	0005	0005	0005	0005	0004
3,7	0004	0004	0004	0004	0004	0004	0003	0003	0003	0003
3,8	0003	0003	0003	0003	0003	0002	0002	0002	0002	0002
3,9	0002	0002	0002	0002	0002	0002	0002	0002	0001	0001

Додаток 3. **Таблиця значень** $t_\gamma = t(\gamma, n)$.

γ / n	**0,95**	**0,99**	**0,999**	γ / n	**0,95**	**0,99**	**0,999**
5	2,78	4,60	8,61	20	2,093	2,861	3,883
6	2,57	4,03	6,86	25	2,064	2,797	3,745
7	2,45	3,71	5,96	30	2,045	2,756	3,659
8	2,37	3,50	5,41	35	2,032	2,720	3,600
9	2,31	2,36	5,04	40	2,023	2,708	3,558
10	2,26	3,25	4,78	45	2,016	2,692	3,527
11	2,23	3,17	4,59	50	2,009	2,679	3,502 '
12	2,20	3,11	4,44	60	2,001	2,662	3,464
13	2,18	3,06	4,32	70	1,996	2,649	3,439
14	2,16	3,01	4,22	80	1,991	2,640	3,418
15	2,15	2,98	4,14	90	1,987	2,633	3,403
16	2,13	2,95	4,07	100	1,984	2,627	3,392
17	2,12	2,92	4,02	120	1,980	2,617	3,374
18	2,11	2,90	3,97	∞	1,960	2,576	3,291
19	2,10	2,88	3,92				

Додаток 4. Критичні точки розподілу χ^2

Число ступенів свободи k	Рівень значущості α					
	0,01	0,025	0,05	0,95	0,975	0,99
1	6,6	5,0	3,8	0,0039	0,00098	0,00016
2	9,2	7,4	6,0	0,103	0,051	0,020
3	11,3	9,4	7,8	0,352	0,216	0,115
4	13,3	11,1	9,5	0,711	0,484	0,297
5	15,1	12,8	11,1	1,15	0,831	0,554
6	16,8	14,4	12,6	1,64	1,24	0,872
7	18,5	16,0	14,1	2,17	1,69	1,24
8	20,1	17,5	15,5	2,73	2,18	1,65
9	21,7	19,0	16,9	3,33	2,70	2,09
10	23,2	20,5	18,3	3,94	3,25	2,56
11	24,7	21,9	19,7	4,57	3,82	3,05
12	26,2	23,3	21,0	5,23	4,40	3,57
13	27,7	24,7	22,4	5,89	5,01	4,11
14	29,1	26,1	23,7	6,57	5,63	4,66
15	30,6	27,5	25,0	7,26	6,26	5,23
16	32,0	28,8	26,3	7,96	6,91	5,81
17	33,4	30,2	27,6	8,67	7,56	6,41
18	34,8	31,5	28,9	9,39	8,23	7,01
19	36,2	32,9	30,1	10,1	8,91	7,63
20	37,6	34,2	31,4	10,9	9,59	8,26
21	38,9	35,5	32,7	11,6	10,3	8,90
22	40,3	36,8	33,9	12,3	11,0	9,54
23	41,6	38,1	35,2	13,1	11,7	10,2
24	43,0	39,4	36,4	13,8	12,4	10,9
25	44,3	40,6	37,7	14,6	13,1	11,5
26	45,6	41,9	38,9	15,4	13,8	12,2
27	47,0	43,2	40,1	16,2	14,6	12,9
28	48,3	44,5	41,3	16,9	15,3	13,6
29	49,6	45,7	42,6	17,7	16,0	14,3
30	50,9	47,0	43,8	18,5	16,8	15,0

Додаток 5. Критичні точки розподілу Колмогорова

Рівень значущості α	0,20	0,10	0,05	0,02	0,01	0,001
λ_α	1,073	1,224	1,358	1,520	1,627	1,950

Список рекомендованої літератури.

1. Барковський В.В., Барковська Н.В., Лопатін О.К. Теорія ймовірностей та математична статистика: Навч. посіб. К.: ЦНЛ, 2019. – 424 с.
2. Свердан П.Л. Вища математика. Математичний аналіз і теорія ймовірностей. Підручник. К.: Знання, 2008. – 450 с.
3. Герич М.С., Синявська О.О. Математична статистика: Навч. посіб. Ужгород: ДВНЗ «УжНУ, 2021. – 146 с.
4. Процеров Ю. С. Математична статистика: навч.-метод. посіб. для студентів ф-туматематики, фізики та інформ. технологій спец. 113 Прикладна математика / Ю. С. Процеров. Одеса: Одес. нац. ун-т ім. І. І. Мечникова, 2023. – 132 с.
5. Руденко В.М. Математична статистика: Навч. посіб. К.: ЦУЛ, 2019. – 304 с.
6. Гончаров О.А., Князь І.О., Хоменко О.В. Теорія ймовірностей та математична статистика: Навч. посіб. Суми: Сумський державний університет, 2022. – 174 с.
7. Панталієнко Л.А. Методичні рекомендації до виконання індивідуальних завдань з вибіркової дисципліни «Методи математичної статистики у наукових дослідженнях» для студентів магістратури І року навчання. − ЦП «КОМПРИНТ» К., 2021. – 90с.
8. Клепко В.Ю., Голець В.Л. Вища математика в прикладах і задачах: Навч. посібник. 2-ге видання. К.: Центр навч. літератури, 2019. – 594 с.
9. Руська Р. В. Теорія імовірності та математична статистика в психології : навч. посіб. Тернопіль. – 2020. – 112 с.
10. Стрелковська І.В., Паскаленко В.М. Математична статистика Навчальний посібник для фахівців у галузі зв'язку. Одеса: Одеська національна академія зв'язку ім. О.С. Попова, 2019 – 110 с.
11. Ямненко Р.Є. Математична статистика. КНУ імені Тараса Шевченка (механіко-математичний факультет). II семестр 2020. https://probability.knu.ua/userfiles/yamnenko/ms_lecture-1.pdf

12. Горбачук, В. М. Теорія ймовірностей та математична статистика [Електронний ресурс]: підручник для здобувачів ступеня бакалавра за технічними та економічними спеціальностями / В. М. Горбачук, О. І. Кушлик-Дивульська ; КПІ ім. Ігоря Сікорського. – Електронні текстові дані (1 файл: 7,93 Мбайт). Київ : КПІ ім. Ігоря Сікорського, 2023. – 351 с. https://ela.kpi.ua/handle/123456789/52357

З М І С Т